



UNIVERSIDAD  
DE MÁLAGA

PROGRAMA DE DOCTORADO EN MATEMÁTICAS  
ESCUELA DE INGENIERÍAS INDUSTRIALES  
DEPARTAMENTO DE MATEMÁTICA APLICADA

PHD THESIS

---

THEORY AND APPLICATIONS OF  
DISTRIBUTIONALLY ROBUST OPTIMIZATION  
WITH SIDE DATA

---

*Author:*

*Adrián Esteban Pérez*

*Advisor:*


Prof. Dr. *Juan Miguel Morales González*

Universidad de Málaga, 2022



UNIVERSIDAD  
DE MÁLAGA

AUTOR: Adrián Esteban Pérez

 <https://orcid.org/0000-0003-0124-8772>

EDITA: Publicaciones y Divulgación Científica. Universidad de Málaga



Esta obra está bajo una licencia de Creative Commons Reconocimiento-NoComercial-SinObraDerivada 4.0 Internacional:

<http://creativecommons.org/licenses/by-nc-nd/4.0/legalcode>

Cualquier parte de esta obra se puede reproducir sin autorización pero con el reconocimiento y atribución de los autores.

No se puede hacer uso comercial de la obra y no se puede alterar, transformar o hacer obras derivadas.

Esta Tesis Doctoral está depositada en el Repositorio Institucional de la Universidad de Málaga (RIUMA): [riuma.uma.es](http://riuma.uma.es)



## DECLARACIÓN DE AUTORÍA Y ORIGINALIDAD DE LA TESIS PRESENTADA PARA OBTENER EL TÍTULO DE DOCTOR

D./Dña ADRIÁN ESTEBAN PÉREZ

Estudiante del programa de doctorado MATEMÁTICAS de la Universidad de Málaga, autor/a de la tesis, presentada para la obtención del título de doctor por la Universidad de Málaga, titulada:  
THEORY AND APPLICATIONS OF DISTRIBUTIONALLY ROBUST OPTIMIZATION WITH SIDE DATA

Realizada bajo la tutorización de JUAN MIGUEL MORALES GONZÁLEZ y dirección de JUAN MIGUEL MORALES GONZÁLEZ (si tuviera varios directores deberá hacer constar el nombre de todos)

DECLARO QUE:

La tesis presentada es una obra original que no infringe los derechos de propiedad intelectual ni los derechos de propiedad industrial u otros, conforme al ordenamiento jurídico vigente (Real Decreto Legislativo 1/1996, de 12 de abril, por el que se aprueba el texto refundido de la Ley de Propiedad Intelectual, regularizando, aclarando y armonizando las disposiciones legales vigentes sobre la materia), modificado por la Ley 2/2019, de 1 de marzo.

Igualmente asumo, ante a la Universidad de Málaga y ante cualquier otra instancia, la responsabilidad que pudiera derivarse en caso de plagio de contenidos en la tesis presentada, conforme al ordenamiento jurídico vigente.

En Málaga, a 25 de JULIO de 2022

|  |   |
|--|---|
| Fdo.: ADRIÁN ESTEBAN PÉREZ<br>Doctorando/a                 | Fdo.: JUAN MIGUEL MORALES GONZÁLEZ<br>Tutor/a |
| Fdo.: JUAN MIGUEL MORALES GONZÁLEZ<br>Director/es de tesis |   |



D. Juan Miguel Morales, profesor titular del departamento de Matemática Aplicada de la Universidad de Málaga, en calidad de tutor y director de la tesis realizada por el doctorando D. Adrián Esteban Pérez, dentro del programa de doctorado en Matemáticas, certifica que:

- D. Adrián Esteban Pérez ha realizado en dicho departamento bajo mi dirección, el trabajo de investigación correspondiente a su tesis doctoral, titulado:

**Theory and applications of Distributionally Robust Optimization with side data**

- Autorizo su presentación para la lectura y defensa de la tesis doctoral ante el tribunal que ha de juzgarlo en la Universidad de Málaga, para que así conste a efectos de lo establecido en el artículo octavo del Real Decreto 99/2011.

Fdo. Juan Miguel Morales González

Director y tutor de la tesis

Málaga, a 25 de Julio de 2022

*A mi familia*



# Acknowledgments

This PhD thesis concludes an amazing four-year period as a Ph.D. student and proud member of the Optimization and Analytics for Sustainable energyY Systems (OASYS) group at the University of Málaga.

First of all, the first acknowledgment goes to my PhD supervisor, Juan M. Morales. From him, I have learnt a lot about what is to be a researcher. Many thanks for the guidance, patience and support during my time as a PhD student and for suggesting the topic of this beautiful area in data-driven decision making uncertainty called *distributionally robust optimization*. Also, many thanks to the European Research Council (ERC) under the European Union’s Horizon 2020 research and innovation programme (Grant agreement No. 755705) for funding the research conducted in this thesis.

As an OASYS member, I have had the opportunity to meet a lot of amazing people. Salva, many thanks for all the support and advice. I especially learnt from you how to make an excellent presentation. Miguel Ángel, thanks for all the support, help and advice with the coding, wind power data, and computer stuff!!

Ricardo, thanks for the support and for giving me the chance to meet Laura, and Júcar! Álvaro, I met you some years ago, and I am so grateful to have you as a friend. Many thanks for all the support and for being the bridge to meeting the lovely Paloma. Paloma, many thanks for your friendship during my time in Málaga! I have so many incredible adventures to remember! Thanks.

Also, I had the opportunity to meet Mirna. Many thanks for your friendship during your stay, visiting the OASYS group. I will never forget our adventure visiting *El Torcal de Antequera*. Thanks to the other OASYS members who I’ve met along the way: Lisa (thanks for all your English language and administrative support), José and his beautiful garden at the office, Antonio, Conchi, Jesús and Asun. Also, I would like to thank Rafael for the support with the Picasso supercomputer.

Thanks to the ERC grant, I had the opportunity to visit the Risk Analytics and Optimization (RAO) lab at EPFL under the supervision of Prof. Daniel Kuhn in Lausanne (Switzerland). Daniel, many thanks for hosting me at EPFL and being an inspiration and icon in distributionally robust optimization. I learnt so much about it from you!



Many thanks to all the RAO members I met during my stay in Lausanne: Yves, Wouter, Bahar, Soorosh, Mengmeng, Tobias, Dirk, Roland, Tianshu, Cagil and Trevor. Also, many thanks to Amandine for all the administrative support in Lausanne. I will never forget the conversations and funny moments! Being an RAO member allowed me to discover Switzerland's beauty: hiking, chocolate and cheese. I very much enjoyed the school in Zinal and the hiking. I will never forget the lunch at *le Banana* and the cappuccino coffee with the amazing views of the *Lac Léman*. Barbara, I am so grateful for meeting you. Thanks for the support during my stay, especially, for all the moments and the support when I went hiking for the first time. Tobias, many thanks for the fruitful conversations about distributionally robust optimization and entropy!

Mengmeng, many thanks for being iconic, and I will never forget our adventure in Zürich and of course, the *maître chocolatier* experience! You are an amazing friend and researcher!

Special thanks to the *party office members*: Yves, Wouter and Bahar. Thanks to them, the hours spent in the office were so much fun!

Finally, many thanks to Yves for being an amazing collaborator and friend. Many thanks for your support during (and after) my PhD stay in Lausanne and the fruitful conversations about distributionally robust optimization and life. I hope I will get the chance to work with you and Daniel on distributionally robust optimization again. Many thanks Yves, for also introducing me to Seongmin. Seongmin, you are amazing, and I am very lucky to have you as a friend.

Also, I would like to thank Juan, Jesús and Augusto for teaching me the scientific field called Operations Research during my bachelor and masters.

Last but not least, I would like to thank three special people: my mum, Raquel, my dad, José María, and my sister, Andrea. I could never have done this thesis without their constant encouragement and support.



# Resumen

## Motivación

El tema de esta tesis se inspira en el hecho de que en el mundo actual se han de tomar decisiones basadas en datos, tal y como ocurre en muchas aplicaciones, incluyendo la operación de sistemas de energía eléctrica, la logística, las finanzas, etc. Estos problemas de toma de decisiones basadas en datos pueden formularse como programas matemáticos de optimización que están sujetos a importantes incertidumbres, ya que los modelos de optimización se construyen a partir de datos ruidosos e inciertos.

Hoy en día se genera una gran cantidad y variedad de datos que, puestos a disposición del decisor, constituyen un valioso recurso en los problemas de optimización. Estos datos, sin embargo, no están exentos de incertidumbre sobre el contexto físico, económico o social, sistema o proceso del que proceden; incertidumbre que, por otra parte, el decisor debe tener en cuenta en su proceso de toma de decisiones. El objetivo de esta tesis es desarrollar fundamentos teóricos e investigar métodos para resolver problemas de optimización en los que existe una gran diversidad de datos sobre fenómenos aleatorios. De forma muy general, el objetivo de la toma de decisiones bajo incertidumbre es encontrar soluciones óptimas correspondientes a problemas de optimización en los que existe (o se supone) un fenómeno aleatorio. Frecuentemente, en la toma de decisiones bajo incertidumbre, el decisor tiene disponible datos históricos sobre el fenómeno aleatorio subyacente al problema y, posteriormente, realiza un proceso de estimación/predicción. El esquema de este enfoque puede representarse como:

$$\text{datos} \longrightarrow \text{estimación/predicciones} \longrightarrow \text{decisiones}$$

De hecho, en cierto modo, el decisor está tratando de resolver un problema de inferencia, ya que quiere inferir una solución óptima basada en la distribución de probabilidad de los datos, que es desconocida e incierta, y, en consecuencia, también lo es el valor objetivo que se pretende optimizar.

Sin embargo, se plantea la siguiente cuestión: *¿Un buen estimador de la verdadera distribución de la que proceden los datos conduce necesariamente a un buen estimador de la decisión óptima?* Aunque a primera vista pueda parecer lógico que así sea, la realidad

es que este enfoque natural puede ser superado por métodos en los que las decisiones son *directamente* inferidas de los datos disponibles, según el siguiente esquema simplificado:

$$\text{datos} \longrightarrow \text{decisiones}$$

El paso intermedio en el que se infiere una distribución de probabilidad a través de los datos supone añadir una fuente de error innecesaria, ya sea por el error de calibración de la distribución de probabilidad elegida como generadora de los datos o por el error intrínseco al propio proceso de inferencia. Este hecho se conoce en la literatura como *la maldición del optimizador* y es una reminiscencia del fenómeno del *sobreajuste* en el aprendizaje estadístico.

Los problemas de toma de decisiones bajo incertidumbre han sido abordados por diferentes comunidades científicas desde varios paradigmas distintos, que difieren en el tratamiento y la representación de los fenómenos aleatorios. A continuación, discutimos los más destacados y relevantes para esta disertación.

## Programación Estocástica

En la Programación Estocástica se suele suponer que la incertidumbre sigue una distribución de probabilidad conocida [125]. Típicamente, un problema de Programación Estocástica se puede formular como sigue:

$$\inf_{\mathbf{x} \in X} \mathbb{E}_{\mathbb{Q}}[f(\mathbf{x}, \boldsymbol{\xi})] = \inf_{\mathbf{x} \in X} \int_{\Xi} f(\mathbf{x}, \boldsymbol{\xi}) \mathbb{Q}(d\boldsymbol{\xi}) \quad (1)$$

donde el objetivo es calcular la decisión óptima  $\mathbf{x}^*$  dentro del conjunto factible  $X$  que minimiza el valor esperado de una cierta función (real) de coste objetivo  $f$ ,  $\mathbb{E}_{\mathbb{Q}}[f(\mathbf{x}, \boldsymbol{\xi})]$ , con respecto a la distribución de probabilidad  $\mathbb{Q}$  de la variable aleatoria  $\boldsymbol{\xi}$  soportada en  $\Xi$  [125]. Aunque en esta sección, en la ecuación (1) se minimiza el valor esperado,  $\mathbb{E}$ , se puede considerar alternativamente una medida del riesgo, como puede ser, por ejemplo, el *valor en riesgo condicional*.

Las características clave de este paradigma son:

- La Programación Estocástica normalmente lleva a tener que resolver programas de optimización a gran escala.
- Si la distribución de la incertidumbre es desconocida, como suele ser el caso, entonces el problema de optimización bajo incertidumbre no puede resolverse. Sin embargo, aunque a veces la distribución de probabilidad sea desconocida para el decisor, ésta podría estimarse o sustituirse por una *distribución nominal* predeterminada utilizando información previa conocida por el decisor. Una elección común para esta distribución nominal sería la distribución empírica estimada a partir de algunas muestras de datos de entrenamiento o una estimación paramétrica de

la verdadera distribución generadora de los datos consistente con la información disponible para el decisor. No obstante, aunque un programa estocástico fuera calibrado a un conjunto de datos fijo con parámetros de distribución insesgados, el resultado de la optimización podría estar sesgado de forma optimista (de nuevo, esto se conoce como la *maldición del optimizador*).

- Incluso en el caso de que la distribución de la incertidumbre sea totalmente conocida, la evaluación del coste de una decisión factible fija requiere el cálculo de una integral multidimensional, lo que se sabe que es computacionalmente difícil [71].

## Optimización Robusta

La Optimización Robusta surge como un paradigma alternativo a la Programación Estocástica que se basa en resolver un problema *minimax* donde el decisor se protege frente a la realización del peor caso/escenario de la incertidumbre dentro de un conjunto  $\Xi$  conocido como *conjunto de incertidumbre* [12], esto es,

$$\inf_{\mathbf{x} \in X} \sup_{\xi \in \Xi} f(\mathbf{x}, \xi) \quad (2)$$

Las características clave de este paradigma son:

- El problema minimax definido en (2) debe admitir una reformulación tratable, lo que depende de las hipótesis sobre la función de coste  $f$  y la estructura del conjunto de incertidumbre  $\Xi$ .
- El conjunto de incertidumbre  $\Xi$  debe definirse adecuadamente para cubrir todas las posibles realizaciones de la incertidumbre. Además, el “tamaño” del conjunto de incertidumbre debe reflejar el grado de conservadurismo adoptado por el decisor, mientras que la “forma” del mismo determina el conocimiento previo sobre la distribución de probabilidad de la incertidumbre.
- Mientras que la Programación Estocástica adopta un punto de vista *optimista* sobre la toma de decisiones, al considerar que se conoce perfectamente la distribución de probabilidad de la incertidumbre, la Optimización Robusta adopta un enfoque *pesimista*, al trabajar con la peor realización posible de la incertidumbre, lo que puede conducir a tomar decisiones demasiado conservadoras.

## Optimización Distribucionalmente Robusta

Aunque el paradigma de la Optimización Distribucionalmente Robusta no es novedoso, sino que surgió en los años 50 en el entorno de la optimización estocástica de inventarios, es en esta década cuando ha crecido sustancialmente como área activa de investigación

(véase [85, 117, 124]). La Optimización Distribucionalmente Robusta puede considerarse como una metodología a medio camino entre la optimización estocástica y la optimización robusta. En este caso, se supone que existe un fenómeno aleatorio que sigue una distribución de probabilidad prescrita/fija. Sin embargo, el decisor desconoce de qué distribución se trata (es decir, es ambigua), ya que no dispone de información completa, por lo que trata de protegerse contra la peor distribución dentro un determinado conjunto de distribuciones de probabilidad (el llamado *conjunto de ambigüedad*).

Un problema distribucionalmente robusto genérico con conjunto de ambigüedad  $\mathcal{U}$  puede formularse como:

$$\inf_{\mathbf{x} \in X} \sup_{Q \in \mathcal{U}} \mathbb{E}_Q [f(\mathbf{x}, \boldsymbol{\xi})] \quad (3)$$

La Optimización Distribucionalmente Robusta ofrece un marco unificado. En particular, si el conjunto de ambigüedad  $\mathcal{U}$  es unipuntual (esto es, está formado por una única distribución de probabilidad), entonces el problema (3) se reduce al problema de Programación Estocástica definido en (1). Por el contrario, si  $\mathcal{U}$  es el conjunto de todas las distribuciones soportadas en  $\Xi$ , entonces el problema (3) se reduce al problema de Optimización Robusta definido en (2).

Un punto clave es que el conjunto de ambigüedad determina el problema. Si bien es habitual en la literatura encontrar diversas formas de especificar el conjunto de ambigüedad, bien sea basándose en momentos o bien por medio de hipótesis paramétricas, en esta tesis nos centramos en conjuntos de ambigüedad definidos por métricas o distancias de probabilidad. En este sentido, la *distancia de Wasserstein*, íntimamente relacionada con el *problema del transporte óptimo*, ha resultado tener bastante éxito debido a sus buenas propiedades estadísticas [99]:

**Definición 1 (Distancia de Wasserstein de orden  $p$ ).** Sea  $p \in [1, \infty)$ . Dadas dos medidas de probabilidad  $P, Q$  con  $p$ -ésimo momento finito soportadas en  $\Xi$ , esto es,  $P, Q \in \mathcal{P}_p(\Xi)$ , la distancia de Wasserstein de orden  $p$  entre  $P$  y  $Q$ ,  $\mathcal{W}_p(P, Q)$ , se define como el valor

$$\left( \inf_{\pi \in \Pi(P, Q)} \left\{ \int_{\Xi^2} \|\boldsymbol{\xi}_1 - \boldsymbol{\xi}_2\|^p \pi(d\boldsymbol{\xi}_1, d\boldsymbol{\xi}_2) \right\} \right)^{1/p}$$

donde  $\Pi(P, Q)$  denota el conjunto de todas las distribuciones conjuntas de  $\boldsymbol{\xi}_1$  y  $\boldsymbol{\xi}_2$  con marginales  $P$  y  $Q$ , respectivamente.

La popularidad y el atractivo de los modelos de optimización distribucionalmente robustos en los últimos años es debido a las siguientes razones (véase [30, 87, 99]):

- **Fidelidad:** Los modelos de optimización distribucionalmente robustos reconocen la existencia de incertidumbre distribucional (esto es, ambigüedad en la distribución generadora de las observaciones) y, con ello, son capaces de proteger al decisor frente a la magnitud y el tipo de error de medición o estimación.

- **Tratabilidad computacional:** Los modelos de optimización distribucionalmente robustos suelen poder reformularse (o ser aproximados) como programas convexos finitos computacionalmente tratables.
- **Garantías de rendimiento:** Para conjuntos de ambigüedad adecuadamente calibrados, el coste óptimo esperado en el peor de los casos que proporciona el problema de optimización distribucionalmente robusto ofrece una cota superior del coste real de las decisiones calculadas mediante el modelo distribucionalmente robusto bajo un cierto nivel de confianza (*garantía de muestra finita o cota de generalización*). La otra garantía se conoce como *consistencia asintótica*: Cuando el número de muestras de entrenamiento tiende a infinito y el conocimiento sobre la distribución de probabilidad es revelado, las decisiones óptimas del problema de optimización distribucionalmente robusto convergen a una decisión óptima del problema estocástico con información perfecta.
- **Regularización y robustez distribucional:** Se ha demostrado la conexión entre los modelos de aprendizaje con regularización y algunos modelos de optimización distribucionalmente robustos (ver [87]). Esto es otra prueba más de la analogía que se puede establecer entre la *maldición del optimizador* en Optimización Estocástica y el fenómeno del *sobreajuste* en el Aprendizaje Estadístico.
- **Anticipación a los “cisnes negros”:** Los modelos de optimización distribucionalmente robustos pueden prevenir futuras realizaciones de la incertidumbre que pueden tener consecuencias devastadoras.
- **Optimalidad:** Se ha demostrado que, en cierto sentido, los modelos de optimización distribucionalmente robustos son “óptimos” cuando el decisor trata de obtener buenas decisiones directamente de los datos (véanse [130, 131]).

## Optimización con restricciones probabilísticas

Si consideramos un problema de optimización en el que algunas restricciones contienen parámetros aleatorios, una forma conservadora de garantizar factibilidad ante tal incertidumbre, sería plantear un número infinito de restricciones (lo que se conoce como *Programación Semi-infinita*), calculando la solución óptima que satisface cualquier realización de la incertidumbre.

No obstante, es posible que algunas realizaciones aumenten significativamente el coste de una solución factible. Si se descartaran tales realizaciones “perjudiciales” o “dañinas”, el coste de la solución podría reducirse considerablemente, manteniendo una alta garantía de factibilidad. La optimización con restricciones probabilísticas considera problemas de optimización en los que algunas restricciones implican parámetros aleato-

rios que deben satisfacerse con un umbral de probabilidad preestablecido (denominado *fiabilidad*).

La optimización con restricciones probabilísticas presenta algunos desafíos:

- Comprobar la factibilidad de una solución requiere un procedimiento de integración multidimensional. Además, es poco frecuente contar con información perfecta acerca de la distribución de la incertidumbre, lo que nos lleva a un problema de optimización con restricciones distribucionalmente robustas.
- En general, el conjunto factible determinado por las restricciones probabilísticas es no convexo e incluso desconexo y, por lo tanto, los modelos con restricciones probabilísticas son computacionalmente difíciles de resolver. Por ello, en la literatura técnica se han propuesto aproximaciones convexas [81, 105], donde la más popular es la basada en el *valor en riesgo condicional*:

**| Definición 2 (Valor en riesgo condicional, CVaR [118]).** *El valor en riesgo condicional a nivel  $\epsilon \in (0, 1)$  de una variable aleatoria real bajo la medida de probabilidad  $Q$ ,  $Q\text{-CVaR}_\epsilon(\omega)$ , se define como el valor  $\inf_{\tau \in \mathbb{R}} \left\{ \tau + \frac{1}{\epsilon} \mathbb{E}_Q[(\omega - \tau)^+] \right\}$  y cuando el ínfimo se alcanza,  $\tau$  representa el valor en riesgo con nivel de confianza  $1 - \epsilon$ .*

## Objetivos

*En esta tesis pretendemos desarrollar, en el ámbito de la Optimización Distribucionalmente Robusta basada en el problema de transporte óptimo, una metodología puramente basada en datos que explote cierta información extra/previa sobre el fenómeno aleatorio. Esta información extra cristaliza en dos ejes o focos sobre la naturaleza del fenómeno aleatorio: en primer lugar, alguna información previa sobre, por ejemplo, la forma y/o estructura de la distribución de probabilidad; en segundo lugar, alguna información previa condicional como la dada por algunas covariables que ayudan a explicar el fenómeno aleatorio subyacente al problema de optimización sin recurrir a técnicas de regresión previa. Los desafíos a la hora de abordar la inclusión de información extra o previa en problemas de optimización distribucionalmente robustos por medio de un conjunto de ambigüedad son fundamentalmente el modelado, la derivación de “buenas” propiedades teóricas y la disponibilidad de reformulaciones que sean tratables computacionalmente. Si bien en la literatura técnica existe una gran cantidad de trabajos (véanse, por ejemplo, [70, 89] y referencias posteriores) donde se han propuesto conjuntos de ambigüedad que incluyen información extra acerca, por ejemplo, de la forma de la distribución de probabilidad que gobierna la incertidumbre, esto generalmente conduce a una complejidad computacional mayor que la del problema distribucionalmente robusto original*



debido al modelado. En esta tesis, se pretende aplicar el uso de conos de orden, ampliamente utilizados en inferencia estadística con restricciones o en regresión isotónica [126], como herramienta de modelado sencilla para incorporar información previa sobre la distribución de la incertidumbre en problemas de Optimización Distribucionalmente Robusta.

Sin embargo, la literatura técnica es escasa en cuanto a problemas de Optimización Distribucionalmente Robusta que incluyan cierta información condicional sobre la incertidumbre, ya que básicamente se han tratado problemas puramente estocásticos o robustos en los que la información condicional está determinada por ciertas covariables (véanse, por ejemplo, [17, 19, 9]).

Por otra parte, el uso de métricas probabilísticas para la toma de decisiones bajo incertidumbre ha crecido notablemente con el auge de la Optimización Distribucionalmente Robusta, y en particular, gracias al trabajo seminal [99] donde se muestra la potencia y beneficios del uso de la métrica de Wasserstein, ya que goza de buenas propiedades estadísticas [109]. La extensión del enfoque propuesto en [99] a problemas de toma de decisiones bajo incertidumbre condicionales no es para nada trivial. De hecho, en nuestro trabajo publicado [53] se prueba que, en general, la aplicación directa de la metodología propuesta en [99] conduce a un problema *mal planteado*. La formulación usando la métrica de Wasserstein requiere considerar la introducción de recortes de probabilidades, íntimamente relacionados con problemas de transporte óptimo *parciales*. Éste carácter *parcial* permite una formulación rigurosa de un problema de optimización distribucionalmente robusto con información condicional con una métrica probabilística bajo condiciones muy generales. De hecho, dadas sus bondades y su aplicabilidad práctica, en el trabajo [53] nos centramos en el uso de la métrica de Wasserstein. Aunque el uso de los recortes de probabilidades ha sido aplicado con asiduidad en el campo de la Estadística (véase [2] y referencias posteriores a este artículo), salvo mayor conocimiento, no ha sido usado en el campo de la toma de decisiones bajo incertidumbre. Un objetivo de esta tesis es mostrar la potencia y versatilidad del uso de recortes de probabilidad en el campo de la toma de decisiones bajo incertidumbre (en nuestro trabajo [53] se muestra, por ejemplo, una aplicación del marco propuesto a problemas de optimización bajo incertidumbre con muestras contaminadas). Otro objetivo es la formalización rigurosa y la extensión de garantías teóricas propuestas en [99] al caso condicional, lo que ha requerido un profundo análisis y desarrollo teórico. Asimismo, hemos querido dotar de robustez distribucional a la metodología propuesta en un trabajo previo sobre Programación Estocástica Prescriptiva o con información contextual [17].

Por último, se han aplicado las técnicas desarrolladas a varios problemas de interés eminentemente práctico procedentes de los ámbitos de las finanzas, la operación de los sistemas eléctricos o la gestión de inventarios. Hemos considerado el problema del vendedor de periódicos y el problema de selección de carteras, que son comúnmente

utilizados en optimización bajo incertidumbre como problemas test o de referencia con los que comparar diferentes modelos o alternativas existentes en la literatura. Más allá de estos problemas de referencia clásicos, también hemos aplicado nuestra metodología al campo de la operación de los sistemas eléctricos, donde la creciente integración de energías renovables como la eólica introduce una alta componente aleatoria. En particular, se ha desarrollado una metodología capaz de explotar cierta información previa o extra para el operador del sistema eléctrico. El objetivo de este último es obtener decisiones de operación más rentables satisfaciendo la fiabilidad del sistema protegiéndose frente a eventos imprevistos derivados de la ambigüedad de la distribución de la incertidumbre.

## Contribuciones

Las principales contribuciones de esta tesis son:

1. La revisión de los principales retos teóricos y de modelización de los problemas de toma de decisiones basados en datos que están sujetos a aleatoriedad y ruido.
2. La ilustración de los beneficios de usar información previa/extra para enriquecer el conjunto de ambigüedad de un modelo de optimización distribucionalmente robusto utilizando resultados empíricos y simulados.
3. La formulación de un enfoque distribucionalmente robusto para modelar cierta información estructural sobre la distribución de probabilidad de la incertidumbre. Esta se modela a través de un enfoque basado en una partición del soporte explotando el problema del transporte óptimo y las restricciones de conos de orden. Además, se proporcionan reformulaciones tratables, que son el resultado de la capacidad de modelar información de forma (como puede ser la multimodalidad) sin incrementar la complejidad del problema de optimización distribucionalmente robusto por medio de la inclusión de restricciones lineales.
4. La formulación de versiones distribucionalmente robustas de programas estocásticos condicionales explotando el concepto de conjunto de recortes de probabilidad, el problema del transporte óptimo parcial y una métrica de probabilidad. Además, la metodología propuesta puede verse como una extensión natural de la Optimización Distribucionalmente Robusta estándar basada en la métrica de Wasserstein al caso con información condicional.
5. El desarrollo de un marco distribucionalmente robusto utilizando recortes de probabilidad para abordar problemas de toma de decisiones con muestras contaminadas.

6. La obtención de versiones distribucionalmente robustas de algunos métodos locales de predicción no paramétrica, como la regresión kernel de Nadaraya-Watson y el método de los  $K$  vecinos más cercanos, que se utilizan frecuentemente en la optimización estocástica contextual o Programación Estocástica Prescriptiva.
7. Las garantías de rendimiento teórico de los marcos distribucionalmente robustos que proponemos se exponen y discuten formalmente. En particular, para los dos marcos distribucionalmente robustos desarrollados en esta tesis, se proporcionan garantías de muestra finita y consistencia asintótica. Además, la introducción de los recortes de probabilidad en Optimización Distribucionalmente Robusta abre la puerta a su aplicación en otros ámbitos del campo de la toma de decisiones bajo incertidumbre.
8. El desarrollo de un modelo del problema del Flujo Óptimo de Cargas con restricciones probabilísticas y robusto desde el punto de vista de la distribución que es capaz de explotar la información contextual. Además, a través de un conjunto de ambigüedad basado en recortes de probabilidad se proporciona una reformulación tratable utilizando la conocida aproximación basada en el *valor en riesgo condicional*.
9. El análisis de los modelos y metodologías de optimización distribucionalmente robustos propuestos en esta tesis mediante ejemplos ilustrativos y estudios de casos realistas en finanzas, gestión de inventarios y operación de sistemas de energía eléctrica.

## Estructura de la tesis

Los capítulos de esta tesis se esquematizan de la siguiente manera:

- El Capítulo 2 contiene una visión general de los principales fundamentos de la toma de decisiones bajo incertidumbre. En concreto, introducimos la Programación Estocástica, Robusta, Distribucionalmente Robusta y con restricciones probabilísticas. Asimismo, se introducen en este capítulo definiciones y conceptos que son usados a lo largo de esta tesis, tales como el valor en riesgo condicional y la conocida métrica o distancia de Wasserstein. Se incluye aquí también una revisión de la literatura sobre Optimización Distribucionalmente Robusta.
- Los Capítulos 3 y 4 proporcionan el núcleo principal de esta tesis. En concreto, se desarrollan dos marcos distribucionalmente robustos para abordar problemas de toma de decisiones basados en datos bajo ambigüedad sobre su distribución de probabilidad:

- El Capítulo 3 discute el primer marco, que se construye a partir de alguna información previa o extra sobre la distribución de probabilidad de la incertidumbre. Más concretamente, la información extra o previa considerada es cierto conocimiento estructural de la distribución, por ejemplo, sobre su forma. Por medio de un enfoque que combina un procedimiento basado en el problema del transporte óptimo usando particiones del soporte y restricciones de conos de orden proponemos una metodología de modelado de la forma o tendencia de la subyacente distribución de probabilidad generadora de los datos. Además, probamos que la metodología propuesta conduce a la resolución de problemas tratables desde un punto de vista computacional y proporcionamos resultados teóricos de garantía de rendimiento. Finalmente, se incluyen unos experimentos numéricos para ilustrar la metodología propuesta a través de dos aplicaciones. En primer lugar, nos centramos en el problema del vendedor de periódicos, donde asumimos que se posee cierta información previa acerca de la distribución de probabilidad de los datos. Se incluye una discusión y comparativa con otros enfoques alternativos disponibles en la literatura técnica. En segundo lugar, se presenta el problema de una empresa estratégica que compite *à la Cournot* en un mercado donde cierta información estructural sobre la distribución de la incertidumbre está disponible para el decisor.
- El Capítulo 4 elabora el segundo marco, que se se construye a partir de alguna información condicional disponible para el decisor. Esta información condicional viene dada por medio de un evento medible genérico. Este evento proporciona a su vez cierta información contextual que podría venir dada por medio de algunas covariables (también conocidas como variables exógenas, características o atributos). Estas covariables pueden ser modeladas por medio de un vector aleatorio que se presupone que presenta un cierto poder predictivo acerca de la incertidumbre y que afecta al valor de la decisión del problema de optimización. Un caso particularmente relevante es aquel en que la información condicional se proporciona en términos de un conjunto de confianza para las covariables; o más concretamente, cuando se reduce a un valor predicho de las mismas. Así pues, en este último caso, la decisión óptima estaría parametrizada en función de este valor predicho y proporcionaría la mejor respuesta en términos de coste (esperado) habiendo predicho u observado dichas covariables. En primer lugar, establecemos el problema estocástico condicional y hacemos una revisión y comparativa de las metodologías existentes en la literatura técnica. Seguidamente, introducimos el concepto de recorte de una distribución de probabilidad y haciendo uso de la conexión existente entre los recortes y el problema del transporte

parcial proporcionamos teóricamente la justificación del uso de los recortes en problemas estocásticos condicionales. Además, se proporcionan resultados teóricos que muestran que la metodología propuesta satisface buenas garantías de rendimiento y convergencia asintótica. También se proporciona una reformulación tratable del problema con la misma complejidad que el problema distribucionalmente robusto basado en la métrica de Wasserstein que ignora la información condicional. Asimismo, se demuestra que la metodología propuesta puede aplicarse en problemas estocásticos donde hay presencia de datos contaminados. Finalmente, se incluyen tres aplicaciones que ilustran la metodología propuesta a través de una serie de experimentos numéricos: en primer lugar, se aborda el problema del vendedor de periódicos en el caso en que se tiene disponible cierta información contextual. En segundo lugar, se aborda el problema de selección de carteras utilizando la metodología distribucionalmente robusta basada en recortes que hemos desarrollado. Por último, se discute en profundidad una aplicación del marco propuesto al problema del Flujo Óptimo de Cargas donde se asume que el operador del sistema eléctrico tiene disponible cierta información contextual sobre la incertidumbre. En concreto, formulamos el problema como un problema distribucionalmente robusto con restricciones probabilísticas y explotamos como información contextual la producción predicha para el conjunto de centrales eólicas del sistema. Se proporciona una reformulación tratable del problema de optimización resultante y se ilustra todo ello en un caso de estudio realista bajo dos escenarios de penetración eólica en el sistema.

- El Capítulo 5 concluye esta tesis y proporciona sugerencias sobre posibles trabajos futuros de investigación.
- El Apéndice A proporciona pruebas teóricas y resultados complementarios al marco desarrollado en el Capítulo 3.
- El Apéndice B contiene material teórico adicional a la metodología desarrollada en el Capítulo 4. En particular, la Sección B.1 de dicho apéndice incluye nociones y resultados teóricos de probabilidad y topología relacionados con la métrica de Wasserstein; una reformulación tratable, bajo hipótesis débiles, del problema de optimización considerado como un programa finito convexo; y un procedimiento constructivo de obtención de una distribución dentro del conjunto de ambigüedad que maximiza el supremo sobre el que se formula el problema de optimización distribucionalmente robusta que se postula. La Sección B.2 contiene las principales pruebas teóricas correspondientes al Capítulo 4. Finalmente, en la Sección B.3 se prueba la consistencia asintótica de la metodología propuesta usando un enfoque basado en técnicas de vecinos más cercanos.

- Por último, el Apéndice C enumera la notación y proporciona otros conceptos, demostraciones teóricas y experimentos adicionales sobre el problema del Flujo Óptimo de Cargas que se discute en la Sección 4.4 del Capítulo 4.

## Conclusiones

Muchos problemas de toma de decisiones en el mundo real se construyen a partir de parámetros que se corresponden con datos que son aleatorios y ruidosos. Es habitual formular estos problemas como programas matemáticos de optimización bajo incertidumbre, en los que dichos parámetros se tratan como variables aleatorias. No tener en cuenta esta incertidumbre puede conducir a decisiones infactibles/subóptimas. En la actualidad, los decisores no sólo recogen observaciones de las incertidumbres que afectan directamente a sus procesos de decisión, sino que también reúnen cierta información previa sobre la distribución generadora de los datos de la incertidumbre. Esta información previa es utilizada por el decisor para prescribir un conjunto más preciso de posibles distribuciones de probabilidad, el llamado *conjunto de ambigüedad* en la Optimización Distribucionalmente Robusta. La información previa estudiada en esta tesis puede ser:

- *Información estructural*, que puede venir dada por algún conocimiento experto del problema de optimización a resolver. Esta información estructural puede ser información de forma como la multimodalidad o unimodalidad de la distribución generadora de los datos. En particular, la introducción de restricciones de cono de orden permite el modelado de esta información estructural de un modo sencillo sin añadir complejidad.
- *Información condicional* dada en términos de un evento genérico medible. Este evento puede transmitir alguna información contextual ligada, por ejemplo, a *co-variables* (también conocidas como variables exógenas, características o atributos).

En esta tesis, se desarrollan varios modelos de optimización distribucionalmente robustos haciendo uso de herramientas de análisis convexo, teoría de la probabilidad, estadística y optimización bajo incertidumbre. Los contenidos de esta tesis se recogen en los artículos publicados [55], [53] y el preprint [51]:

- En nuestro artículo [55], se presenta un nuevo marco para la Optimización Distribucionalmente Robusta basada en la teoría del transporte óptimo en combinación con restricciones de conos de orden para aprovechar información previa o extra sobre la verdadera distribución generadora de los datos. Motivados por el exceso de conservadurismo del enfoque tradicional distribucionalmente robusto basado en la métrica de Wasserstein, se formula un conjunto de ambigüedad capaz de incorporar información sobre el orden entre las probabilidades que la verdadera

distribución de los parámetros inciertos del problema asigna a algunas subregiones del soporte. Nuestro enfoque es capaz de modelar una amplia gama de información sobre la forma (como la relacionada con la monotonicidad o la multimodalidad) de forma práctica e intuitiva. Además, bajo hipótesis débiles, el problema resultante de optimización distribucionalmente robusto puede, de hecho, reformularse como un problema convexo finito en el que la información extra (expresada a través de las restricciones del cono de orden) se presenta como restricciones lineales, a diferencia de las formulaciones con mayor complejidad computacional que existen en la literatura. Además, nuestro enfoque está respaldado por garantías teóricas de rendimiento y es capaz de convertir la información proporcionada en soluciones con mayor fiabilidad y mejor rendimiento, como ilustran los experimentos numéricos efectuados sobre el conocido problema del vendedor de periódicos y el problema de una empresa estratégica que compite *à la Cournot* en un mercado de producto homogéneo.

- En nuestro trabajo [53], se explota la conexión entre los recortes de probabilidad y el transporte parcial de masa para proporcionar una forma fácil, pero potente y novedosa, de extender la Optimización Distribucionalmente Robusta estándar basada en la métrica de Wasserstein al caso de los programas estocásticos condicionales. Nuestro enfoque produce decisiones que son distribucionalmente robustas frente a la incertidumbre en el proceso de inferir la medida de probabilidad condicional de los parámetros aleatorios a partir de una muestra finita procedente de la verdadera distribución conjunta generadora de los datos. A través de una serie de experimentos numéricos contruidos sobre el problema del vendedor de periódicos de un solo artículo y un problema de selección de carteras, se demuestra que nuestro método alcanza un rendimiento notablemente mejor fuera de la muestra que algunas alternativas existentes. Hemos apoyado estos resultados empíricos con un análisis teórico, mostrando que nuestro enfoque goza de atractivas garantías de rendimiento.
- En nuestro preprint [51], se desarrolla un modelo de Flujo Óptimo de Cargas con restricciones probabilísticas, robusto desde el punto de vista de la distribución, que es capaz de explotar la información contextual a través de un conjunto de ambigüedad basado en recortes de probabilidad. Hemos proporcionado una reformulación de este modelo como un programa lineal continuo utilizando la conocida aproximación basada en el valor en riesgo condicional. Mediante una serie de experimentos numéricos realizados en una red eléctrica modificada de 118 nodos con incertidumbre eólica, se demuestra que, explotando la dependencia estadística entre la predicción de la producción eólica y su error de predicción asociado, nuestro enfoque puede identificar soluciones de despacho que, satisfaciendo la fi-

abilidad requerida del sistema, conducen a un ahorro de costes de hasta varios puntos porcentuales con respecto a las soluciones del problema del Flujo Óptimo de Cargas proporcionadas por un método distribucionalmente robusto alternativo que ignora dicha dependencia estadística.

## Trabajo futuro

A continuación, se enumeran las posibles futuras líneas de investigación resultantes del trabajo realizado en esta tesis:

1. El desarrollo de métodos de descomposición para resolver problemas a gran escala de optimización distribucionalmente robustos basados en la métrica de Wasserstein. Esto es un problema fundamentalmente derivado de la dependencia del número de restricciones con respecto al tamaño de la muestra de entrenamiento.
2. Se requiere desarrollo teórico en profundidad para intentar eliminar la dependencia de la dimensión de la incertidumbre en las garantías de muestra finita en el ámbito de la optimización estocástica condicional. El uso de otras métricas probabilísticas es otra potencial futura línea de investigación.
3. Es necesario el desarrollo de procedimientos de calibración (basados, por ejemplo, en validación cruzada o remuestreo) para elegir adecuadamente los parámetros de robustez de los modelos de optimización distribucionalmente robustos presentes en esta tesis.
4. La aplicación de los recortes de probabilidades en los problemas de Programación Estocástica en dos etapas requiere un estudio adicional.
5. Es necesario estudiar procedimientos basados en datos para calibrar adecuadamente los parámetros de robustez en nuestro modelo del Flujo Óptimo de Cargas distribucionalmente robusto con restricciones probabilísticas de acuerdo con las preferencias de riesgo del operador del sistema (por ejemplo, recurriendo a la validación cruzada o al remuestreo). También, es muy interesante la extensión de este modelo para tener en cuenta restricciones intertemporales, lo que, entre otras cosas, implicaría la adaptación de nuestro conjunto de ambigüedad basado en los recortes de probabilidades para tratar con procesos estocásticos y datos provenientes de series temporales.

## Lista de publicaciones

- Esteban-Pérez, A., Morales, J.M. Partition-based distributionally robust optimization via optimal transport with order cone constraints. *4OR-Q J Oper Res*



(2021).

- Esteban-Pérez, A., Morales, J.M. Distributionally robust stochastic programs with side information based on trimmings. *Math. Program.* (2021).
- Esteban-Pérez, A., Morales, J. M. Distributionally Robust Optimal Power Flow with Contextual Information. *arXiv preprint* arXiv:2109.07896. (2021).

# Abstract

Nowadays, a large amount of varied data is being generated which, when made available to the decision maker, constitutes a valuable resource in optimization problems. These data, however, are not free from uncertainty about the physical, economic or social context, system or process from which they originate; uncertainty that, on the other hand, the decision maker must take into account in his/her decision making process. The objective of this PhD dissertation is to develop theoretical foundations and investigate methods for solving optimization problems where there is a great diversity of data on uncertain phenomena. Today's decision makers not only collect observations from the uncertainties directly affecting their decision-making processes, but also gather some prior information about the data-generating distribution of the uncertainty. This information is used by the decision maker to prescribe a more accurate set of potential probability distributions, the so-called *ambiguity set* in distributionally robust optimization. Our intention, therefore, is to develop a purely data-driven methodology, within the scope of distributionally robust optimization based on the optimal transportation problem, which exploits some extra/prior information about the random phenomenon. This extra information crystallizes in two axes on the nature of the random phenomenon: first, some prior information about, for example, the shape/structure of the probability distribution; second, some conditional information such as that given by various covariates, which help explain the random phenomenon underlying the optimization problem without resorting to prior regression techniques.

We propose a formulation of a distributionally robust approach to model certain structural information about the probability distribution of the uncertainty. This is given in terms of a partition-based approach, exploiting the optimal transport problem and order cone constraints. In addition, tractable reformulations are provided, and by the same token, the power of modeling shape information (such as multimodality), without jeopardizing the complexity of the distributionally robust optimization problem by adding linear constraints.

Moreover, by leveraging *probability trimmings* and their connection with the partial optimal transport problem, we formulate a distributionally robust version of conditional stochastic programs. The theoretical performance guarantees of the distributionally

robust frameworks we propose are also formally stated and discussed. In addition, we show that the proposed methodology based on probability trimmings can be applied to decision-making problems under uncertainty with contaminated samples.

Furthermore, we develop a distributionally robust chance-constrained Optimal Power Flow model that is able to exploit contextual/side information through an ambiguity set based on probability trimmings, providing a tractable reformulation using the well-known conditional value-at-risk approximation.

Finally, we test, analyze, and discuss the proposed optimization models and methodologies developed in this PhD dissertation through illustrative examples and realistic case studies in finance, inventory management and power systems operation.



# Contents

|          |   |           |
|----------|---|-----------|
| <b>1</b> | <b>Introduction</b>   | <b>2</b>  |
| 1.1      | Background and Motivation . . . . .   | 2         |
| 1.2      | Contributions . . . . .   | 4         |
| 1.3      | Thesis Outline . . . . .  | 5         |
| 1.4      | List of Publications . . . . .  | 6         |
| <b>2</b> | <b>Essentials of decision-making under uncertainty</b>  | <b>7</b>  |
| 2.1      | Stochastic programming . . . . .  | 8         |
| 2.2      | Robust optimization . . . . .   | 9         |
| 2.3      | Distributionally robust optimization . . . . .  | 9         |
| 2.3.1    | Improving the specification of the ambiguity set with prior information . . . . .                   | 12        |
| 2.4      | Chance-constrained programming . . . . .  | 15        |
| 2.5      | Notation . . . . .  | 17        |
| 2.6      | Summary . . . . .   | 17        |
| <b>3</b> | <b>Leveraging structural information via optimal transport and order cone constraints</b>           | <b>20</b> |
| 3.1      | Methodology and theoretical foundations . . . . .   | 21        |
| 3.1.1    | Introduction . . . . .  | 21        |
| 3.1.2    | Tractable reformulations . . . . .  | 24        |
| 3.1.3    | Separable objective function . . . . .  | 25        |
| 3.1.4    | Order cone constraints . . . . .  | 26        |
| 3.1.5    | On convergence and out-of-sample performance guarantees . . . . .                                   | 28        |
| 3.2      | Numerical experiments . . . . .   | 30        |
| 3.2.1    | Application I. Newsvendor problems . . . . .  | 34        |
|          | The single-item newsvendor problem . . . . .  | 34        |
|          | The multi-item newsvendor problem . . . . .   | 37        |
| 3.2.2    | Application II. The problem of a strategic firm competing <i>à la Cournot</i> in a market . . . . . | 38        |

|          |   |           |
|----------|---|-----------|
| 3.3      | Summary . . . . .   | 41        |
| <b>4</b> | <b>Conditional stochastic programs: A distributionally robust solution approach based on probability trimmings</b>                                  | <b>42</b> |
| 4.1      | Methodology and theoretical foundations . . . . .   | 43        |
| 4.1.1    | Preliminaries and motivation . . . . .  | 43        |
| 4.1.2    | The Partial Mass Transportation Problem and Trimmings . . . . .   | 45        |
| 4.1.3    | Tractable reformulation of the partial mass transportation problem  | 50        |
| 4.1.4    | Finite sample guarantee and asymptotic consistency . . . . .  | 51        |
|          | Case $\mathbb{Q}(\tilde{\Xi}) = \alpha > 0$ . Applications in data-driven decision making under contaminated samples . . . . .                      | 52        |
|          | The case of unknown $\mathbb{Q}(\tilde{\Xi}) = \alpha > 0$ . . . . .  | 55        |
|          | The case $\mathbb{Q} \ll \lambda^d$ and $\mathbb{Q}(\tilde{\Xi}) = \alpha = 0$ . . . . .  | 56        |
| 4.2      | Application I. Newsvendor problem . . . . .   | 60        |
| 4.3      | Application II. Portfolio allocation problem . . . . .  | 65        |
| 4.3.1    | Case $\mathbb{Q}(\tilde{\Xi}) = \alpha = 0$ . . . . .   | 65        |
| 4.3.2    | Case $\mathbb{Q}(\tilde{\Xi}) = \alpha > 0$ . . . . .   | 70        |
| 4.4      | Application III. Optimal Power Flow problem . . . . .   | 74        |
| 4.4.1    | Introduction . . . . .  | 74        |
| 4.4.2    | DC-OPF under uncertainty: Mathematical Formulation . . . . .  | 76        |
|          | Variables and constraints . . . . .   | 77        |
|          | Dealing with uncertainty in the DC-OPF problem: Joint chance constraints, Distributionally Robust Optimization and contextual information . . . . . | 78        |
| 4.4.3    | A tractable and conservative <b>CVaR</b> -based approximation of the distributionally robust joint chance constraints . . . . .                     | 80        |
| 4.4.4    | An exact tractable reformulation of the worst-case expected cost  | 81        |
| 4.4.5    | Numerical results . . . . .   | 83        |
| 4.4.6    | Evaluation of the out-of-sample performance via re-optimization   | 84        |
| 4.4.7    | A 118-bus case study . . . . .  | 85        |
|          | Medium wind penetration case . . . . .  | 86        |
|          | High wind penetration case . . . . .  | 89        |
| 4.5      | Summary . . . . .   | 93        |
| <b>5</b> | <b>Conclusions and future work</b>  | <b>94</b> |
| 5.1      | Summary and conclusions . . . . .   | 94        |
| 5.2      | Directions for future research . . . . .  | 96        |

|          |   |            |
|----------|---|------------|
| <b>A</b> | <b>Proofs of Chapter 3</b>  | <b>97</b>  |
| A.1      | Proof of Theorem 3.1 . . . . .                                    | 97         |
| A.2      | Proof of Corollary 3.1 . . . . .                                  | 100        |
| A.3      | Proof of Theorem 3.2 . . . . .                                    | 101        |
| A.4      | Proof of Theorem 3.3 . . . . .                                    | 101        |
| A.5      | Proof of Theorem 3.4 . . . . .                                    | 101        |
| A.6      | Proof of Theorem 3.5 . . . . .                                    | 102        |
| <b>B</b> | <b>Additional material to Chapter 4</b>                           | <b>103</b> |
| B.1      | Complementary definitions and technical results . . . . .         | 104        |
| B.1.1    | Concepts from measure theory and the Wasserstein metric . . . .   | 104        |
| B.1.2    | Concepts and technical results from probability trimmings . . . . | 105        |
| B.1.3    | Topological properties of the ambiguity set . . . . .             | 106        |
| B.1.4    | Tractable reformulation and maximizer of problem (SP2) . . . .    | 107        |
| B.2      | Proofs of Chapter 4 . . . . .                                     | 108        |
| B.2.1    | Proof of Lemma 4.1 . . . . .                                      | 108        |
| B.2.2    | Proof of Proposition 4.1 . . . . .                                | 109        |
| B.2.3    | Proof of Theorem 4.1 . . . . .                                    | 110        |
| B.2.4    | Proof of Proposition 4.2 . . . . .                                | 112        |
| B.2.5    | Proof of Theorem 4.2 . . . . .                                    | 112        |
| B.2.6    | Proof of Lemma 4.4 . . . . .                                      | 113        |
| B.2.7    | Proof of Theorem 4.3 . . . . .                                    | 114        |
| B.2.8    | Proof of Proposition 4.3 . . . . .                                | 114        |
| B.2.9    | Proof of Theorem 4.4 . . . . .                                    | 114        |
| B.2.10   | Proof of Lemma 4.5 . . . . .                                      | 115        |
| B.3      | Asymptotic consistency under a nearest neighbors lens . . . . .   | 115        |
| <b>C</b> | <b>Additional material to Section 4.4</b>                         | <b>129</b> |
| C.1      | Notation used in Section 4.4 . . . . .                            | 129        |
| C.1.1    | Sets, numbers and indices . . . . .                               | 129        |
| C.1.2    | Parameters and functions . . . . .                                | 129        |
| C.1.3    | Random variables and uncertain parameters . . . . .               | 130        |
| C.1.4    | Variables . . . . .   | 131        |
| C.1.5    | Other symbols . . . . .   | 131        |
| C.2      | Proof of Proposition 4.4 . . . . .                                | 132        |
| C.3      | Proof of Proposition 4.5 . . . . .                                | 132        |
| C.4      | Real-time re-dispatch problem . . . . .                           | 133        |
| C.5      | Illustrative example (3-bus system) . . . . .                     | 134        |
| C.6      | Data for the illustrative example (3-bus system) . . . . .        | 141        |





# List of Figures

|      |  |    |
|------|--|----|
| 3.1  | Single-item newsvendor problem: (Approximate) true data-generating distribution, order quantity and performance metrics . . . . .  | 36 |
| 3.2  | Multi-item newsvendor problem: Performance metrics . . . . .   | 38 |
| 3.3  | Strategic firm problem: (Approximate) true data-generating distribution, optimal solution and performance metrics . . . . .  | 40 |
| 4.1  | Probability simplex (in blue) corresponding to the trimming set $\mathcal{R}_{0.5}(\hat{\mathbb{Q}}_N)$  | 47 |
| 4.2  | Newsvendor problem with features: True distributions, quantile estimate and performance metrics . . . . .  | 62 |
| 4.3  | Portfolio problem with features: Performance metrics . . . . .   | 67 |
| 4.4  | Impact of the robustness parameter with 200 training samples, $K_N = \lfloor N/(\log(N+1)) \rfloor$ and $\delta = 0.5$ , $\lambda = 0.1$ . . . . .   | 68 |
| 4.5  | Portfolio problem with features: Varying context under an optimal selection of the robustness parameters, $K_N = \lfloor N/(\log(N+1)) \rfloor$ and $\delta = 0.5$ , $\lambda = 0.1$ . . . . . | 70 |
| 4.6  | Portfolio problem with features: Performance metrics. Case $\alpha > 0$ and $\delta = 0.5$ , $\lambda = 0.1$ . . . . .   | 72 |
| 4.7  | Case $\alpha > 0$ , impact of the robustness parameter with 200 training samples and $\delta = 0.5$ , $\lambda = 0.1$ . . . . .  | 73 |
| 4.8  | Portfolio problem with features: Performance metrics under an optimal selection of the robustness parameters. Case $\alpha > 0$ and $\delta = 0.5$ , $\lambda = 0.1$                           | 74 |
| 4.9  | Medium level of wind penetration, $N = 100$ and $\epsilon = 0.1$ : Total downward and upward reserve capacity and performance metrics . . . . .  | 87 |
| 4.10 | Medium level of wind penetration, $N = 300$ and $\epsilon = 0.1$ : Total downward and upward reserve capacity and performance metrics . . . . .  | 88 |
| 4.11 | High level of wind penetration, $N = 100$ and $\epsilon = 0.1$ : Total downward and upward reserve capacity and performance metrics . . . . .  | 91 |
| 4.12 | High level of wind penetration, $N = 300$ and $\epsilon = 0.1$ : Total downward and upward reserve capacity and performance metrics . . . . .  | 92 |

|     |  |     |
|-----|--|-----|
| C.1 | Heat map of the true joint distribution and kernel estimate of the true conditional density given $z^* = 30$ MW . . . . .            | 135 |
| C.2 | Three-bus system, sample size $N = 30$ and $\epsilon = 0.1$ : Total downward and upward reserves and performance metrics . . . . .   | 136 |
| C.3 | Three-bus system, sample size $N = 30$ and $\epsilon = 0.1$ : Generators' dispatch and participation factors . . . . .               | 137 |
| C.4 | Three-bus system, sample size $N = 2000$ and $\epsilon = 0.1$ : Total downward and upward reserves and performance metrics . . . . . | 138 |
| C.5 | Three-bus system, sample size $N = 2000$ and $\epsilon = 0.1$ : Generators' dispatch and participation factors . . . . .             | 139 |

# List of Tables

|     |  |     |
|-----|--|-----|
| 3.1 | Single-item newsvendor problem: Values for parameters $\varepsilon, \rho$ in DROC and $\rho$ in DROW. . . . .  | 35  |
| 3.2 | Multi-item newsvendor problem: Values for parameters $\varepsilon, \rho$ in DROC and $\rho$ in DROW . . . . .  | 37  |
| 4.1 | Medium level of wind penetration, summary data for total expected cost [\$] under the optimal value of the robustness parameter for methods DROW and DROTRIMM. . . . . | 90  |
| 4.2 | High level of wind penetration, summary data for total expected cost [\$] under the optimal value of the robustness parameter for DROW and DROTRIMM. . . . .           | 91  |
| C.1 | Generators' location, power output limits and reserve capacity costs . .   | 141 |
| C.2 | Slopes ( $m_s$ ) and intercepts ( $n_s$ ) of the generators' piecewise linear cost functions . . . . .   | 141 |
| C.3 | Transmission line parameters . . . . .   | 141 |



# Chapter 1

## Introduction

The subject matter of this thesis is inspired by the fact that today's world is guided by data-driven decisions, given that they appear in many applications, such as energy systems operation, logistics, and finance. These decisions have to be made to accommodate the world's uncertain nature efficiently. These data-driven decision-making problems can be formulated as mathematical optimization programs containing noisy and/or uncertain data. This first chapter discusses the motivation behind the methods developed and formally states the main contributions of this thesis. The outline of the thesis is also provided, and the published papers are listed.

### 1.1 Background and Motivation

Nowadays, a large amount of varied data is being generated which, when made available to the decision maker, constitutes a valuable resource in optimization problems. These data, however, are not free of uncertainty about the physical, economic or social context, system or process from which they originate; uncertainty that, on the other hand, the decision maker must take into account in his/her decision making process. The objective of this thesis is to develop theoretical foundations and investigate methods for solving optimization problems where there is a great diversity of data on uncertain phenomena. In a very general way, the objective of *decision making under uncertainty* is to find optimal solutions corresponding to optimization problems where a random phenomenon exists (or it is assumed). To date, decision-making problems under uncertainty have been addressed by different scientific communities from several, distinct paradigms, which differ in the treatment and representation of the uncertain phenomena.

Traditionally, in decision making under uncertainty, the decision maker collects historical data on the random phenomena underlying the problem and subsequently performs an estimation/prediction process. The scheme of this approach can be represented as:

$$data \longrightarrow estimation/predictions \longrightarrow decisions$$

In fact, in a way, the decision maker is trying to solve an inference problem since she wants to infer an optimal decision using the limited information on the probability distribution of the uncertain phenomena that is conveyed by the input data and, consequently, given the inherited uncertainty in the decision's value.

However, the following question arises: *Does a good estimator of the true data-generating distribution necessarily lead to a good estimator of our optimal decision variable?* Although at first glance, it might seem logical that it does, the reality is that this natural approach can be superseded by methods in which decisions are directly inferred from data according to the following simplified scheme:

$$data \longrightarrow decisions$$

The intermediate step in which a probability distribution is inferred through the data involves adding an unnecessary source of error, either due to the calibration error of the probability distribution chosen as the data generator or due to the error intrinsic to the inference process itself. This is known in the literature as *the optimizer's curse* and is reminiscent of the phenomenon of *overfitting* in statistical learning.

When making a decision under uncertainty, it is usually assumed that the uncertainty follows a known probability distribution in *stochastic programming*. In contrast, *robust optimization* has emerged as an alternative paradigm to stochastic programming relying, as it does, on hedging against the worst-case realization of the uncertainty. Although the *distributionally robust optimization* (DRO) paradigm is not novel (it emerged in the 1950s in the environment of stochastic inventory optimization), it is in this decade that it has grown substantially as an active area of research. Distributionally robust optimization can be seen as a methodology which bridges the gap between stochastic programming and robust optimization. In this case, it is assumed that there is a random phenomenon that follows a prescribed probability distribution. However, the decision maker does not know which distribution that is, as s/he does not have complete information and therefore seeks to protect against the worst distribution in a certain set of probability distributions (the so-called *ambiguity set*). The key point is that the ambiguity set determines the problem. While it is common in the literature to find various ways of specifying the ambiguity set, such as moment-based sets or parametric assumptions, in this thesis, we focus on ambiguity sets defined by probability metrics or distances. In this sense, the *Wasserstein metric*, intimately related to the *optimal transport problem*, has proved to be quite successful due to its good statistical properties. In this particular case, as in all extreme cases, it is clear that if the ambiguity set is reduced to a single element, we would end up working with stochastic programming and, to the contrary, if the ambiguity set contains all probability distributions with a given support, then we would find ourselves with the case of robust optimization.

In this thesis, we intend to develop, within the scope of distributionally robust optimization based on the optimal transportation problem, a purely data-driven methodology that exploits some extra/prior information about the random phenomenon. This extra information crystallizes in two axes on the nature of the random phenomenon: first, some prior information about, for example, the shape/structure of its probability distribution; second, some conditional information such as that given by some covariates, which help to explain the random phenomenon underlying the optimization problem without resorting to prior regression techniques.

This thesis intends to delve deeper into this line by applying the methodology to be developed to several problems of eminently practical interest taken from the realms of finance, power systems or inventory management.

## 1.2 Contributions

The main contributions of this thesis are:

1. The review of the major theoretical and modeling challenges of data-driven decision making problems originating from a large amount of data subject to randomness and noise.
2. The illustration of the prior information used to improve the specification of the ambiguity set of a distributionally robust optimization model using both empirical and simulated results.
3. The formulation of a distributionally robust model to factor in some structural information about the probability distribution of the uncertainty. This is given in terms of a partition-based approach, exploiting the optimal transport problem and order cone constraints. In addition, we provide tractable reformulations and investigate the power of modeling shape information (such as multimodality) without jeopardizing the complexity of the DRO problem by adding linear constraints.
4. The formulation of distributionally robust versions of *conditional* stochastic programs exploiting *probability trimmings*, the *partial* optimal transport problem and probability metrics.
5. The development of a distributionally robust framework using probability trimmings to address data-driven decision-making problems under contaminated samples.
6. The use of probability trimmings to produce distributionally robust versions of some local nonparametric predictive methods, such as Nadaraya-Watson kernel

regression and  $K$ -nearest neighbors, which are often used in contextual stochastic optimization/prescriptive stochastic programming.

7. Theoretical performance guarantees of the distributionally robust frameworks we propose are formally stated and discussed.
8. The development of a distributionally robust chance-constrained Optimal Power Flow (OPF) model that is able to exploit contextual/side information through an ambiguity set based on probability trimmings, providing a tractable reformulation using the well-known Conditional Value-at-Risk approximation.
9. The analysis and testing of the proposed optimization models and methodologies through illustrative examples and realistic case studies in finance, inventory management and power systems operation.

### 1.3 Thesis Outline

The chapters of this thesis are outlined as follows:

- Chapter 2 provides an overview of the main essentials of decision making under uncertainty.
- Chapter 3 provides theoretical foundations and applications of a distributionally robust framework that exploits some prior information about the characteristics of the probability distribution of the uncertainty via optimal transport and order cone constraints.
- Chapter 4 lays out the theoretical foundations and discusses applications of a distributionally robust framework that exploits some conditional information available to the decision maker through partial optimal transport and probability trimmings.
- Chapter 5 concludes this thesis and provides suggestions for future work.
- Appendix A provides theoretical proofs and complementary results to the framework developed in Chapter 3.
- Appendix B compiles supporting theoretical material to the framework developed in Chapter 4. More specifically, Section B.1 in this appendix provides notions and theoretical results in probability and topology related to the Wasserstein metric; a tractable reformulation, under weak assumptions, of the distributionally robust optimization approach we propose as a finite convex program; and a constructive procedure for obtaining a distribution within the ambiguity set that maximizes



the supremum that leads to our DRO problem. Section B.2 contains the main theoretical proofs to the framework developed in Chapter 4. Finally, Section B.3 provides a proof of the asymptotic consistency of the proposed methodology using a nearest neighbors lens.

- Lastly, Appendix C lists the notation and provides other concepts and additional experiments about the Optimal Power Flow problem studied in Chapter 4.

## 1.4 List of Publications

- Esteban-Pérez, A., Morales, J.M. Partition-based distributionally robust optimization via optimal transport with order cone constraints. *4OR-Q J Oper Res* (2021).
- Esteban-Pérez, A., Morales, J.M. Distributionally robust stochastic programs with side information based on trimmings. *Math. Program.* (2021).
- Esteban-Pérez, A., Morales, J. M. Distributionally Robust Optimal Power Flow with Contextual Information. *arXiv preprint* arXiv:2109.07896. (2021).

## Chapter 2

# Essentials of decision-making under uncertainty

### Contents

|            |   |           |
|------------|---|-----------|
| <b>2.1</b> | <b>Stochastic programming . . . . .</b>   | <b>8</b>  |
| <b>2.2</b> | <b>Robust optimization . . . . .</b>  | <b>9</b>  |
| <b>2.3</b> | <b>Distributionally robust optimization . . . . .</b>                             | <b>9</b>  |
| 2.3.1      | Improving the specification of the ambiguity set with prior information . . . . . | 12        |
| <b>2.4</b> | <b>Chance-constrained programming . . . . .</b>                                   | <b>15</b> |
| <b>2.5</b> | <b>Notation . . . . .</b>   | <b>17</b> |
| <b>2.6</b> | <b>Summary . . . . .</b>  | <b>17</b> |

Many real-world decision-making problems involve input data that are random and noisy. It is customary to formulate these problems as mathematical optimization programs under uncertainty, whose parameters are treated as random variables. Disregarding this uncertainty may lead to infeasible/suboptimal decisions. To date, decision-making problems under uncertainty have been addressed by different scientific communities from several, distinct paradigms, which differ in the treatment and representation of the uncertain phenomena. This chapter introduces some of these paradigms, in particular, those that are the most relevant to the primary purpose of the research in this thesis, and briefly describes the main types of prior information that can be used to improve the specification of the ambiguity set in distributionally robust optimization. The chapter concludes with the main notation employed throughout the dissertation.

## 2.1 Stochastic programming

In the realm of stochastic programming, a full and accurate knowledge about the probability distribution of the random phenomena  $\xi$  (defined over a given probability space) in the optimization problem ([125]) is assumed. Essentially, a typical stochastic program can be formulated as follows:

$$\inf_{\mathbf{x} \in X} \mathbb{E}_{\mathbb{Q}}[f(\mathbf{x}, \xi)] = \inf_{\mathbf{x} \in X} \int_{\Xi} f(\mathbf{x}, \xi) \mathbb{Q}(d\xi) \quad (2.1)$$

where the goal is to compute an optimal decision  $\mathbf{x}^*$  within a feasible set  $X$  which minimizes the expected value of a given objective cost function (real-valued)  $f$ ,  $\mathbb{E}_{\mathbb{Q}}[f(\mathbf{x}, \xi)]$ , with respect to the probability distribution  $\mathbb{Q}$  of the random variable  $\xi$  with support set  $\Xi$ . Although, in this section, in Eq. (2.1) we minimize the expectation,  $\mathbb{E}$ , it can be replaced with an alternative risk measure like, for example, the Conditional Value-at-Risk, **CVaR**:

**| Definition 2.1 (Conditional Value-at-Risk, CVaR [118]).** *The CVaR at level  $\epsilon \in (0, 1)$  of a univariate random variable  $\omega$  under the probability measure  $Q$ ,  $Q - \text{CVaR}_{\epsilon}(\omega)$ , is defined as the value  $\inf_{\tau \in \mathbb{R}} \left\{ \tau + \frac{1}{\epsilon} \mathbb{E}_Q[(\omega - \tau)^+] \right\}$  and when the infimum is attained,  $\tau$  represents the Value-at-Risk with confidence level  $1 - \epsilon$ .*

However, stochastic programming has some drawbacks (see [99]):

- It typically requires having to solve large-scale optimization programs.
- If the distribution  $\mathbb{Q}$  is unknown, as is often the case, then problem (2.1) cannot be solved. However, sometimes the unknown probability distribution  $\mathbb{Q}$  could be estimated/replaced with a predetermined *nominal distribution* using some prior information known by the decision-maker. The common choice for this nominal distribution is the *empirical distribution* estimated from some training data samples or a parametric estimate of  $\mathbb{Q}$  consistent with the aforementioned information. Nevertheless, although a stochastic program were calibrated to a fixed dataset with unbiased distributional parameters, the optimization output could possibly be optimistically biased. This phenomenon is known as *optimizer's curse* or *optimization bias* and is reminiscent of the well known *overfitting* problem in Statistical Learning.
- Even in the case that the probability  $\mathbb{Q}$  is fully known, evaluating the cost of a fixed feasible decision requires a multidimensional integration process that is known to be computationally difficult [71].

## 2.2 Robust optimization

Robust optimization, and its adoption of the worst-case perspective, emerges as an alternative paradigm to stochastic programming. This paradigm relies on hedging against the worst-case realization of the uncertainty supported in a given set called the *uncertainty set* [12]. This framework leads to the following *minimax* program:

$$\inf_{\mathbf{x} \in X} \sup_{\boldsymbol{\xi} \in \Xi} f(\mathbf{x}, \boldsymbol{\xi}) \quad (2.2)$$

The key features of this paradigm are:

- The minimax program (2.2) should admit a tractable reformulation. A computationally tractable reformulation depends on assumptions about the cost function  $f$  and the structure of the uncertainty set  $\Xi$ .
- The uncertainty set  $\Xi$  should be defined suitably to cover all the possible realizations of the uncertainty  $\boldsymbol{\xi}$ . In addition, the *size* of  $\Xi$  should reflect the degree of conservatism adopted by the decision-maker, while the *shape* of  $\Xi$  determines the prior knowledge about the probability distribution of  $\boldsymbol{\xi}$ .
- While stochastic programming adopts the *optimistic* angle of knowing with perfect information the probability distribution of  $\boldsymbol{\xi}$ , robust optimization takes a *pessimistic* view on the uncertainty  $\boldsymbol{\xi}$ , which can lead to over-conservative decisions.

## 2.3 Distributionally robust optimization

Distributionally robust optimization (DRO) emerges as an alternative paradigm to bridge the gap between the explicitness and optimism of stochastic programming and the conservatism and pessimism of robust optimization. It seeks to compute the optimal decision which minimizes the worst-case expectation over any probability distribution within the so-called *ambiguity set*, that is, a set of potential probability distributions consistent with the given prior knowledge about the uncertainty. We refer to [85, 117, 124] for recent surveys on DRO and optimization under uncertainty.

A generic DRO model with ambiguity set  $\mathcal{U}$  can be stated as follows:

$$\inf_{\mathbf{x} \in X} \sup_{Q \in \mathcal{U}} \mathbb{E}_Q [f(\mathbf{x}, \boldsymbol{\xi})] \quad (2.3)$$

DRO offers a unified framework as, in the case that the ambiguity set  $\mathcal{U}$  reduces to a singleton, then problem (2.3) reduces to the stochastic program (2.1). In contrast, if

$\mathcal{U}$  collects all the distributions supported in  $\Xi$ , then problem (2.3) collapses into the robust optimization problem (2.2).

The ambiguity set  $\mathcal{U}$  is a key ingredient of any distributionally robust optimization model and can be constructed in a data-driven manner based on  $N$  training samples, in which case we explicitly denote the dependence of  $\mathcal{U}$  on  $N$  by  $\mathcal{U}_N$ . A good ambiguity set should be rich enough to contain the true data-generating distribution  $\mathbb{Q}$  of  $\xi$  with a high confidence level, but at the same time not too large, so that unrealistic/pathological distributions can be excluded, thereby avoiding over-conservative decisions. Moreover, ideally, the ambiguity set should be easily parametrized from data and computationally tractable reformulated, so that it can be solved by off-the-shelf optimization software. It is no wonder, therefore, that much effort has been expended on this issue, resulting in several ways to specify and characterize the ambiguity set, namely:

1. *Moment-based approach:* The ambiguity set is defined as the set of all probability distributions whose moments satisfy certain constraints; see [42, 64, 93, 92, 98, 102, 139, 145], to name just a few.
2. *Dissimilarity-based approach:* The ambiguity set is defined as the set of all probability distributions whose *dissimilarity* to a prescribed distribution (often referred to as the *nominal distribution*) is lower than or equal to a given value. Within this category, the choice of the *dissimilarity* function leads to a wealth of distinct variants:
  - (a) *Optimal-transport-based (OTP) approach:* Here, we include the work in [29, 28, 63, 99, 122], among many others, all of which use, as the dissimilarity function, the well known *Wasserstein* distance:

**Definition 2.2 ( $p$ -Wasserstein distance).** For any  $p \in [1, \infty)$ , given two probability measures  $P, Q$  with finite  $p$ -th moment supported on  $\Xi$ , that is,  $P, Q \in \mathcal{P}_p(\Xi)$ , the Wasserstein metric of order  $p$  between  $P$  and  $Q$ ,  $\mathcal{W}_p(P, Q)$ , is given by the value

$$\left( \inf_{\pi \in \Pi(P, Q)} \left\{ \int_{\Xi^2} \|\xi_1 - \xi_2\|^p \pi(d\xi_1, d\xi_2) \right\} \right)^{1/p}$$

where  $\Pi(P, Q)$  denotes the set of all joint distributions of  $\xi_1$  and  $\xi_2$  with marginals  $P$  and  $Q$ , respectively.

The Wasserstein metric exhibits some nice statistical convergence properties (see [109] and references therein).

- (b)  *$\phi$ -divergences-based approach:* This class comprises all those approaches which use  $\phi$ -divergences (such as the Kullback-Leibler divergence) as a quantifier

of dissimilarity, for instance, [11, 13, 103]. We also include in this group the likelihood-based approaches, proposed by [47] and [135].

(c) *Other measures of dissimilarity*: This category includes all other dissimilarity-based procedures for constructing ambiguity sets than those already mentioned, such as the ones that utilize the family of  $\zeta$ -structure probability metrics (for example, the total variation metric, the Bounded Lipschitz metric ...), see, for instance, the work in [116] and [142], and the Prokhorov metric [49].

3. *Hypothesis-test-based approach*: The ambiguity set is made up of all those probability distributions which, given a data sample, pass a certain hypothesis test with a prescribed confidence level; see, for example, the work in [15, 16, 36].

The popularity and attractiveness of distributionally robust optimization models in recent years is due to the following reasons (see [30, 87, 99]):

- **Fidelity**: DRO models agree and benefit from the existence of distributional uncertainty and the magnitude and type of measurement/estimation error, respectively.
- **Computational tractability**: DRO models can usually be reformulated (or approximated by) as computationally tractable finite convex programs.
- **Performance guarantees**: For suitably calibrated data-driven ambiguity sets, the worst-case optimal expected cost delivered by problem (2.3) offers an upper confidence bound on the out-of-sample expected cost attained by the optimizers of (2.1) (*finite sample guarantee or generalization bound*). To be more precise:

**Definition 2.3.** A data-driven solution for problem (2.1) is a feasible solution  $\hat{\mathbf{x}}_N \in X$  which is constructed from the sample data, and its out-of-sample performance is defined as  $\mathbb{E}_{\mathbb{Q}}[f(\hat{\mathbf{x}}_N, \boldsymbol{\xi})]$ .

**Definition 2.4 (Finite sample guarantee [99]).** Given a data-driven solution  $\hat{\mathbf{x}}_N$ , a finite sample guarantee is a relation in the form

$$\mathbb{Q}^N \left[ \mathbb{E}_{\mathbb{Q}}[f(\hat{\mathbf{x}}_N, \boldsymbol{\xi})] \leq \hat{J}_N \right] \geq 1 - \beta \quad (2.4)$$

where  $\hat{J}_N$  is a certificate for the out-of-sample performance of  $\hat{\mathbf{x}}_N$  (i.e., an upper bound that is generally contingent on the training dataset),  $\beta \in (0, 1)$  is a significance parameter with respect to the  $N$ -fold product of distribution  $\mathbb{Q}$ , i.e.,  $\mathbb{Q}^N$ , on which both  $\hat{\mathbf{x}}_N$  and  $\hat{J}_N$  depend. Moreover, we refer to the probability on the left-hand side of (2.4) as the reliability of  $(\hat{\mathbf{x}}_N, \hat{J}_N)$  and can be understood

as a confidence level. Similarly, we say that a data-driven method built to address problem (2.1) enjoys a finite sample guarantee, if it produces pairs  $(\hat{\mathbf{x}}_N, \hat{J}_N)$  satisfying a relation in the form (2.4).

The other guarantee is known as *asymptotic consistency*: When the number of training samples grows to infinity and the knowledge about  $\mathbb{Q}$  is revealed, the optimizers of (2.3) converge to an optimizer of (2.1).

- **Regularization and distributional robustness:** The connection between learning models with regularization and some DRO models can be proved (see [87]). This further emphasizes the connection between the optimizer’s curse in optimization and overfitting in Statistical Learning.
- **Foreseeing black swans:** DRO models can prevent future realizations of the uncertainty that can have cataclysmal consequences.
- **Optimality:** It has been shown that, in some sense, DRO models are “optimal” when the decision-maker tries to retrieve good decisions directly from data (see [130, 131]).

### 2.3.1 Improving the specification of the ambiguity set with prior information

Today’s decision makers not only collect observations of the uncertainties directly affecting their decision-making processes, but may also have available some prior information about the probability law generating those observations. That information can be exploited for data-driven decision making using distributionally robust optimization (DRO).

The ambiguity set  $\mathcal{U}$  in (2.3) is as a set of distributions consistent with the information about  $\mathbb{Q}$  and serves as a confidence set. Ideally, one would like to have the smallest ambiguity set that contains the true data-generating distribution  $\mathbb{Q}$  in problem (2.3). In this vein, if we have available some prior information about  $\mathbb{Q}$ , we should use it to discard all those other distributions that do not conform with that information from the ambiguity set. That information used to strengthen the specification of the ambiguity set can be, for example:

- **Dependence information:** Some dependence structure about  $\xi$  can be exploited using *copula* theory. This may occur when there is available some prior knowledge of the marginal distributions of  $\mathbb{Q}$ . This has been applied in conjunction with a Wasserstein-based DRO model in references [6, 64].
- **Shape/structural information:** Structural information can result from expert knowledge of the problem. This information about  $\mathbb{Q}$  may be, e.g., symmetry,

unimodality or multimodality, and has been considered in [55, 72, 90, 91], among others. In our work [55] we propose an optimal transport-based DRO model exploiting some prior information on the *order* among the probabilities that  $\mathbb{Q}$  assigns to some regions of its support set. This type of order is enforced by means of order cone constraints and can encode a wide range of information on the shape of the probability distribution of the uncertain parameters such as information related to monotonicity or multi-modality. The authors in [90, 91] propose a moment-based ambiguity set where unimodality is considered as prior information for solving chance-constrained DRO problems. The advantage of our proposed methodology in [55] over other alternative approaches is that the inclusion of such prior information does not jeopardize the computational tractability of the underlying mathematical program (as it translates into adding linear constraints). Chapter 3 contains the main core of our work [55].

- **Conditional information:** Some information is available to the decision maker in terms of a (measurable) event  $\xi \in \tilde{\Xi}$ . Hence, problem (2.1) turns into the following *conditional* stochastic program:

$$J^* := \inf_{\mathbf{x} \in X} \mathbb{E}_{\mathbb{Q}} [f(\mathbf{x}, \xi) \mid \xi \in \tilde{\Xi}] = \inf_{\mathbf{x} \in X} \mathbb{E}_{\mathbb{Q}_{\tilde{\Xi}}} [f(\mathbf{x}, \xi)] \quad (2.5)$$

where  $\mathbb{Q}_{\tilde{\Xi}}$  is the  $\mathbb{Q}$ -conditional distribution of  $\xi$  given  $\xi \in \tilde{\Xi}$ . It is notable that since  $\mathbb{Q}$  is rarely known, its conditional version,  $\mathbb{Q}_{\tilde{\Xi}}$ , is even less known.

The generic event  $\xi \in \tilde{\Xi}$  can represent contextual/side information delivered by some *covariates* (also known as exogenous variables, features or attributes). In this setting, we have  $\xi := (\mathbf{z}, \mathbf{y})$  with  $\mathbf{z}$  being a random vector modeling features that may have some predictive power over the uncertainty  $\mathbf{y}$  that affects the value of the decision  $\mathbf{x}$ . A marked case is when the conditional information is given in terms of a set  $\mathcal{Z}$  (which could be a confidence set for  $\mathbf{z}$ ); or more specifically, when it reduces to a singleton, i.e.  $\mathcal{Z} := \{\mathbf{z}^*\}$ . Thus, in the latter case, the optimal decision delivered by problem (2.5) is parametrized on  $\mathbf{z}^*$  and provides the best-response in terms of expected cost having predicted/observed  $\mathbf{z} = \mathbf{z}^*$ . This side information acts by changing the probability measure of the uncertainties. In fact, if the joint distribution of the features and the uncertainties  $\mathbb{Q}$  were known, this measure change would correspond to conditioning that distribution on the side information given. Unfortunately, in practice, the decision maker only has an incomplete picture of such a joint distribution in the form of a finite data sample.

The development of optimization methods capable of exploiting the side information to make improved decisions, in a context of limited knowledge of its explanatory power on the uncertainties, defines the ultimate purpose of the so-



called *Prescriptive Stochastic Programming* or *Conditional Stochastic Optimization* paradigm. This paradigm has recently become very popular in the technical literature, see, for instance, [9, 17, 112] and references therein. More specifically, a data-driven approach to address the newsvendor problem, whereby the decision is explicitly modeled as a parametric function of the features, is proposed in [9]. This approach thus seeks to optimize said function. In contrast, the authors in [17] formulate and formalize the problem of minimizing the conditional expectation cost given the side information, and develop various schemes based on machine learning methods (typically used for regression and prediction) to get data-driven solutions. Their approach is *non-parametric* in the sense that the optimal decision is not constrained to be a member of a certain family of the features' functions. The inspiring work in [17] has been subject to further study and improvement in two principal directions, namely, the design of efficient algorithms to trim down the computational burden of the optimization [45] and the development of strategies to reduce the variance and bias of the decision obtained and its associated cost (the pairing of both interpreted as a statistical estimator). In the latter case, we can cite the work in [22], where they leverage ideas from bootstrapping and machine learning to confer robustness on the decision and acquire asymptotic performance guarantees. Similarly, the authors in [18] and [112] propose regularization procedures to reduce the variance of the data-driven solution to the conditional expectation cost minimization problem which is formalized and studied in [17]. A scheme to robustify the data-driven methods introduced in this work is also proposed in [19] for dynamic decision-making.

A different, but related thrust of research focuses on developing methods to construct predictions specifically tailored to the optimization problem that is to be solved and where those predictions are then used as input information. Essentially, the predictions are intended to yield decisions with a low disappointment or regret. This framework is known in the literature as (smart) *Predict-then-Optimize*, see, e.g., [8, 46, 48, 101], and references therein.

The methodology we propose, which is described in [53] and developed in Chapter 4 is built, in contrast, upon Distributionally Robust Optimization. Accordingly, we address problem (2.3) by way of the following DRO formulation *conditional* on the event  $\xi \in \tilde{\Xi}$ :

$$\inf_{\mathbf{x} \in X} \sup_{Q_{\tilde{\Xi}} \in \mathcal{U}_{\tilde{\Xi}}} \mathbb{E}_Q[f(\mathbf{x}, \xi)] \quad (2.6)$$

where  $\mathcal{U}_{\tilde{\Xi}}$  is an ambiguity set for  $\mathbb{Q}_{\tilde{\Xi}}$ .

Nevertheless, the technical literature on the use of DRO to address Prescriptive

or Conditional Stochastic Programming problems is still relatively scarce. We highlight papers [19, 38, 69, 82, 84, 106, 107], with [107] being a generalization of [106]. In [38], they resort to a scenario-dependent ambiguity set to exploit feature information in a DRO framework. However, their objective is to minimize a joint expectation and consequently, their approach cannot directly handle the Conditional Stochastic Optimization setting we consider here. In [69], the authors deal with a stochastic control problem with time-dependent data. They extend the idea of [74] to a fully dynamic setting and robustify the control policy against the worst-case weight vector that is within a certain  $\chi^2$ -distance from the one originally given by the Nadaraya-Watson estimator. In the case of [19], the authors propose using the conditional empirical distribution given by a local predictive method as the center of the Wasserstein ball that characterizes the DRO approach in [99]. This proposal, nonetheless, fails to explicitly account for the inference error associated with the local estimation. In [82, 84], the authors develop a two-step procedure whereby a regression model between the uncertainty and the features is first estimated and then a distributionally robust decision-making problem is formulated, considering a Wasserstein ball around the empirical distribution of the residuals. Finally, the authors in [107] also consider a Wasserstein-ball ambiguity set as in [19, 82, 84], but centered at the empirical distribution of the joint data sample of the uncertainty and the features. In addition, they further constrain the ambiguity set by imposing that the worst-case distribution assigns some probability mass to the support of the uncertainty conditional on the values taken on by the features.

Unlike the modeling approaches discussed above, ours constitutes a general framework to handle conditional stochastic programs within the DRO paradigm. In particular, our DRO framework is based on a new class of ambiguity sets that exploits the close and convenient connection between probability trimmings and the partial mass problem to immunize the decision against the error incurred in the process of inferring *conditional* information from *joint* (limited) data. We refer the reader to Chapter 4 for a full description and analysis of our proposal.

## 2.4 Chance-constrained programming

If we consider an optimization problem where some constraints involve random parameters, e.g.,  $g(\mathbf{x}, \xi) \leq 0$ , one conservative way to deal with randomness would be to state the following *semi-infinite* constraint:

$$g(\mathbf{x}, \xi) \leq 0, \forall \xi \in \Xi \quad (2.7)$$

which is equivalent to the *robust* constraint  $\sup_{\xi \in \Xi} g(\mathbf{x}, \xi) \leq 0$ . Nevertheless, a small number of samples in  $\Xi$  can significantly increase the cost of a feasible solution  $\mathbf{x}$ . Disregarding such “harmful” samples can result in a considerably reduced solution cost, maintaining high reliability. Chance-constrained programming considers optimization problems where some constraints involve random parameters which need to be satisfied with a pre-fixed probability threshold  $1 - \epsilon$  (*reliability*). This may be formulated by the following *single* (or individual) chance/probabilistic constraint:

$$\mathbb{Q}(g(\mathbf{x}, \xi) \leq 0) \geq 1 - \epsilon \quad (2.8)$$

**Remark 2.1.** If  $g(\mathbf{x}, \xi) \leq 0$  in (2.8) is replaced by  $K$  constraints,  $g_k(\mathbf{x}, \xi) \leq 0, \forall k = 1, \dots, K$ , then the respective chance-constraint would be the following joint chance constraint:

$$\mathbb{Q}(g_k(\mathbf{x}, \xi) \leq 0, \forall k = 1, \dots, K) \geq 1 - \epsilon \quad (2.9)$$

which can be recast equivalently as the following single chance constraint:

$$\mathbb{Q}\left(\max_{k=1, \dots, K} g_k(\mathbf{x}, \xi) \leq 0\right) \geq 1 - \epsilon$$

Chance-constrained programming inherits some difficult challenges:

- Checking the feasibility of a solution  $\mathbf{x}$  in (2.8) requires a multidimensional integration procedure. Furthermore, having a full knowledge of  $\mathbb{Q}$  is rare, and therefore, the decision-maker is rather interested in solving the following distributionally robust version of (2.8):

$$\mathbb{Q}(g(\mathbf{x}, \xi) \leq 0) \geq 1 - \epsilon, \forall Q \in \mathcal{U} \iff \inf_{Q \in \mathcal{U}} \mathbb{Q}(g(\mathbf{x}, \xi) \leq 0) \geq 1 - \epsilon \quad (2.10)$$

where  $\mathcal{U}$  is an ambiguity set for  $\mathbb{Q}$ .

- Usually, the feasible set given by (2.8) is non-convex and even disconnected, and hence, chance-constrained models are hard to solve and convex approximations have been proposed in the technical literature ([81, 105]). The most popular is the one based on the **CVaR**. In particular, by definition of **CVaR**, the following holds:

$$\mathbb{Q}\left(g(\mathbf{x}, \xi) \leq \mathbb{Q} - \mathbf{CVaR}_\epsilon(g(\mathbf{x}, \xi))\right) \geq 1 - \epsilon \quad (2.11)$$

Therefore,

$$\mathbb{Q} - \mathbf{CVaR}_\epsilon(g(\mathbf{x}, \xi)) \leq 0 \implies \mathbb{Q}(g(\mathbf{x}, \xi) \leq 0) \geq 1 - \epsilon \quad (2.12)$$

which implies that

$$\sup_{Q \in \mathcal{U}} Q - \text{CVaR}_\epsilon(g(\mathbf{x}, \boldsymbol{\xi})) \leq 0 \implies \inf_{Q \in \mathcal{U}} Q(g(\mathbf{x}, \boldsymbol{\xi}) \leq 0) \geq 1 - \epsilon \quad (2.13)$$

being  $\mathcal{U}$  an ambiguity set for  $\mathbb{Q}$ .

## 2.5 Notation

Next, we introduce the main notation used in this dissertation. Other notation is defined as required throughout the main text.

We use  $\overline{\mathbb{R}}$  to denote the extended real line, and adopt the conventions of its associated arithmetic. Furthermore,  $\mathbb{R}_+$  denotes the set of non-negative real numbers. We employ lower-case bold face letters to represent vectors and bold face capital letters for matrices. We use  $\text{diag}(a_1, \dots, a_m)$  for a diagonal matrix of size  $m \times m$  whose diagonal elements are equal to  $a_1, \dots, a_m$ . Moreover, given a matrix  $\mathbf{M}$ , its transpose matrix will be written as  $\mathbf{M}^\top$ . We define  $\mathbf{e}$  as the array with all its components equal to 1. The inner product of two vectors  $\mathbf{u}, \mathbf{v}$  (in a certain space) is denoted  $\langle \mathbf{u}, \mathbf{v} \rangle = \mathbf{u}^\top \mathbf{v}$ . Given any norm  $\|\cdot\|$  in the Euclidean space (of a given dimension  $d$ ), the dual norm is defined as  $\|\mathbf{u}\|_* = \sup_{\|\mathbf{v}\| \leq 1} \langle \mathbf{u}, \mathbf{v} \rangle$ . Given a function  $f : \mathbb{R}^d \rightarrow \overline{\mathbb{R}}$ , we will say that  $f$  is a proper function if  $f(\mathbf{x}) < +\infty$  for at least one  $\mathbf{x}$  and  $f(\mathbf{x}) > -\infty$  for all  $\mathbf{x} \in \mathbb{R}^d$ . Additionally, the convex conjugate function of  $f$ ,  $f^*$ , is defined as  $f^*(\mathbf{y}) := \sup_{\mathbf{x} \in \mathbb{R}^d} \langle \mathbf{y}, \mathbf{x} \rangle - f(\mathbf{x})$ . It is well known that if  $f$  is a proper function, then  $f^*$  is also a proper function. Given a set  $A \subseteq \mathbb{R}^d$ , we denote its interior (resp. relative interior) as  $\text{int}(A)$  (resp.  $\text{relint}(A)$ ), and its indicator function  $\mathbb{I}_A(\mathbf{a})$  is defined through  $\mathbb{I}_A(\mathbf{a}) = 1$  if  $\mathbf{a} \in A$ ;  $= 0$ . The support function of set  $A$ ,  $S_A$ , is defined as  $S_A(\mathbf{b}) := \sup_{\mathbf{a} \in A} \langle \mathbf{b}, \mathbf{a} \rangle$ . The dual cone  $\mathcal{C}^*$  of a cone  $\mathcal{C}$  is given by  $\mathcal{C}^* := \{\mathbf{y} / \langle \mathbf{y}, \mathbf{x} \rangle \geq 0, \forall \mathbf{x} \in \mathcal{C}\}$ . We use the symbol  $\delta_{\boldsymbol{\xi}}$  to represent the Dirac distribution supported on  $\boldsymbol{\xi}$ . In addition, we reserve the symbol “ $\hat{\cdot}$ ” for objects which are dependent on the sample data. The Lebesgue measure in  $\mathbb{R}^d$  is denoted as  $\lambda^d$ . The  $K$ -fold product of a distribution  $\mathbb{Q}$  will be denoted as  $\mathbb{Q}^K$ . The symbols  $\mathbb{E}$  and  $\mathbb{P}$  denote, respectively, “expectation” and “probability.” Finally, for the rest of the dissertation we assume that we always have measurability for those objects whose expected values we consider.

## 2.6 Summary

This chapter has introduced the essentials of the different paradigms to hedge against the uncertainty in decision-making problems. Moreover, we have presented the general formulations of a classical stochastic program, a robust optimization model, a distributionally robust optimization model and the chance-constrained setting. Additionally, we have described the principal types of prior information that can be used to improve

the specification of the ambiguity set in distributionally robust optimization. Finally, we have presented the main notation used throughout this dissertation and introduced several concepts relevant to this thesis such as the Wasserstein metric and the **CVaR**.



## Chapter 3

# Leveraging structural information via optimal transport and order cone constraints

### Contents

|            |   |           |
|------------|---|-----------|
| <b>3.1</b> | <b>Methodology and theoretical foundations . . . . .</b>  | <b>21</b> |
| 3.1.1      | Introduction . . . . .  | 21        |
| 3.1.2      | Tractable reformulations . . . . .  | 24        |
| 3.1.3      | Separable objective function . . . . .  | 25        |
| 3.1.4      | Order cone constraints . . . . .  | 26        |
| 3.1.5      | On convergence and out-of-sample performance guarantees . . . . .                                   | 28        |
| <b>3.2</b> | <b>Numerical experiments . . . . .</b>  | <b>30</b> |
| 3.2.1      | Application I. Newsvendor problems . . . . .  | 34        |
|            | The single-item newsvendor problem . . . . .  | 34        |
|            | The multi-item newsvendor problem . . . . .   | 37        |
| 3.2.2      | Application II. The problem of a strategic firm competing <i>à la Cournot</i> in a market . . . . . | 38        |
| <b>3.3</b> | <b>Summary . . . . .</b>  | <b>41</b> |

In this chapter, we propose a methodology to strengthen the specification of the ambiguity set in an optimal transport-based DRO problem by means of order cone constraints. Essentially, these constraints allows us to factor in some information on the probability masses that the true data-generating distribution assigns to events within a partition of its support set. Therefore, this chapter deals with our work [55]. We also discuss computational aspects and applications of the proposed framework. For ease of reading, all the proofs of the theoretical results that are presented next have been moved to the appendices.

### 3.1 Methodology and theoretical foundations

This section elaborates on the theoretical core of our work [55]. We begin by introducing the mathematical formulation of the distributionally robust optimization problem we propose to solve.

#### 3.1.1 Introduction

In an attempt to avoid overly conservative solutions of the program (2.3), we seek to sharpen the specification of Wasserstein ambiguity sets with prior information on the true probability distribution of the problem's uncertain parameters. We represent this information in the form of order cone constraints on the probability masses associated with a partition of the sample space. Problem (POC) below formulates the data-driven distributionally robust optimization (DDRO) framework we propose.

$$(\text{POC}) \quad \inf_{\mathbf{x} \in X} \sup_{Q \in \mathcal{Q}} \mathbb{E}_Q[f(\mathbf{x}, \boldsymbol{\xi})] \quad (3.1a)$$

$$\text{s.t. } \mathbb{P}_Q[\boldsymbol{\xi} \in \Xi_i] = p_i, \forall i \in \mathcal{I} \quad (3.1b)$$

$$\tilde{c}(\mathbf{p} - \hat{\mathbf{p}}) \leq \rho \quad (3.1c)$$

$$\sum_{i \in \mathcal{I}} p_i C(Q_i, \hat{Q}_i) \leq \varepsilon \quad (3.1d)$$

$$Q_i \in \mathcal{Q}_i, \forall i \in \mathcal{I} \quad (3.1e)$$

$$\mathbf{p} \in \Theta \quad (3.1f)$$

where  $X \subseteq \mathbb{R}^n$  is the set of feasible decisions,  $\boldsymbol{\xi} : \Omega \rightarrow \Xi \subseteq \mathbb{R}^d$  is a random vector defined on the measurable space  $(\Omega, \mathcal{F})$  with  $\sigma$ -algebra  $\mathcal{F}$ , and  $\mathcal{Q}$  is the set of all probability distributions over the measurable space  $(\Omega, \mathcal{F})$ . Moreover, for each  $i \in \mathcal{I}$ ,  $Q_i$  is the conditional distribution of  $Q$  given  $\boldsymbol{\xi} \in \Xi_i$ , that is  $Q_i = Q(\boldsymbol{\xi} \mid \boldsymbol{\xi} \in \Xi_i) \in \mathcal{Q}_i$ , with  $\mathcal{Q}_i$  being the set of all conditional probability distributions of  $Q$  given  $\boldsymbol{\xi} \in \Xi_i$ . In this setting,  $\mathcal{I}$  is the set of regions  $\Xi_i$  with pairwise disjoint interiors into which the support set  $\Xi$  is partitioned, that is,  $\bigcup_{i \in \mathcal{I}} \Xi_i = \Xi$  and  $\text{int}(\Xi_i) \cap \text{int}(\Xi_j) = \emptyset, \forall i, j \in \mathcal{I}, i \neq j$ . Furthermore, we assume that  $\mathbb{Q}(\Xi_i \cap \Xi_j) = 0, \forall i, j \in \mathcal{I}, i \neq j$ , where  $\mathbb{Q}$  is the true data-generating distribution. This is equivalent to stating that  $\{\Xi_i\}_{i \in \mathcal{I}}$  constitutes a  $\mathbb{Q}$ -packing (see a formal definition of this concept in page 50 of [66]) and will allow us to unequivocally assign samples from  $\mathbb{Q}$  to the partitions  $\Xi_i, i \in \mathcal{I}$ . Finally, constraint (3.1c) defines the set of all probability vectors  $\mathbf{p}$  that differ from the nominal empirical probability vector  $\hat{\mathbf{p}}$  in at most  $\rho$  according to the cost function  $\tilde{c}$ . This is a function that quantifies how dissimilar two probability vectors  $\mathbf{p}$  and  $\mathbf{q}$  are. For this purpose, we require that  $\tilde{c}$  be a non-negative jointly convex lower semicontinuous function such that if  $\mathbf{p} = \mathbf{q}$ , then  $\tilde{c}(\mathbf{p}, \mathbf{q}) = 0$ . As mentioned further on, function  $\tilde{c}$



could, for example, take the form of a norm or a  $\phi$ -divergence. To ease the notation and the formulation, we use  $\boldsymbol{\xi}$  to represent either the random vector  $\boldsymbol{\xi}(\omega)$ , with  $\omega \in \Omega$  or an element of  $\mathbb{R}^d$ . Note that we can consider the probability measure induced by the random vector  $\boldsymbol{\xi}$ , if we choose the corresponding Borel  $\sigma$ -algebra  $\mathcal{B}$  on  $\Xi$ . Thus, we can see  $\mathcal{Q}$  as a set of probability measures defined over  $(\Xi, \mathcal{B})$ , so we write  $\mathcal{Q} = \mathcal{Q}(\Xi)$ . We define the uncertainty set  $\mathcal{P}$  for the probability vector  $\mathbf{p} \in \mathbb{R}^{|\mathcal{I}|}$ , with  $|\mathcal{I}|$  being the number of partitions, as the intersection of  $\Theta$  and the set defined by constraint (3.1c). The support set  $\Theta$ , which includes the order cone constraints on the probability masses  $\mathbf{p}$ , is given by:

$$\Theta = \{\mathbf{p} \in \mathbb{R}^{|\mathcal{I}|} : \langle \mathbf{e}, \mathbf{p} \rangle = 1, \mathbf{p} \in \mathbb{R}_+^{|\mathcal{I}|}, \mathbf{p} \in \mathcal{C}\} \quad (3.2)$$

where  $\mathcal{C}$  is a proper (convex, closed, full and pointed) cone. Hence,  $\Theta$  is a convex compact set.

In constraint (3.1d),  $C$  is the optimal transport cost defined as

$$C(P, Q) = \inf_{\pi \in \Pi(P, Q)} \left\{ \int c(x, y) \pi(dx, dy) \right\}$$

where  $\Pi(P, Q)$  is the set of all joint distributions of  $x$  and  $y$  with marginals  $P$  and  $Q$ , and  $c$  is a measurable cost function with  $c(x, y)$  representing the cost of moving a unit of mass from location  $x$  to location  $y$ . We assume that this cost function  $c$  is a non-negative jointly convex lower semicontinuous function such that if  $x = y$ , then  $c(x, y) = 0$ . In the remainder of the thesis we take for granted that we have existence and uniqueness of the optimal transport problem (see, for example, Theorem 4.1 in [133]). Note that if the cost function  $c$  is given by a norm, we recover the 1-Wasserstein metric introduced in Definition 2.2 of Section 2.3.

In problem (POC),  $\rho$  and  $\varepsilon$  are non-negative parameters, to be tuned by the decision maker, which control the size of the ambiguity set defined by equations (3.1b)–(3.1f).

We represent this set as  $\mathcal{U}_{\rho, \varepsilon}(\hat{Q})$ , where  $\hat{Q}$  is a nominal distribution expressed in terms of  $\hat{\mathbf{p}}$  and  $\hat{Q}_i$  as

$$\hat{Q} = \sum_{i \in \mathcal{I}} \hat{p}_i \hat{Q}_i \quad (3.3)$$

where

$$\hat{p}_i = \frac{N_i}{N + |I'|} \quad (3.4)$$

and

$$\hat{Q}_i = \frac{1}{N_i} \sum_{j=1}^{N_i} \delta_{\hat{\boldsymbol{\xi}}_j^i} \quad (3.5)$$

Further,  $I' = \{i \in \mathcal{I} \text{ such that partition } i \text{ does not contain any data from the sample}\}$ ,  $\hat{\boldsymbol{\xi}}_j^i \in \{\hat{\boldsymbol{\xi}}_1^i, \dots, \hat{\boldsymbol{\xi}}_{N_i}^i\}$  and  $N_i$  is the number of atoms in region  $\Xi_i$ . Here we set  $N_i = 1$  and

$\hat{\xi}_1^i := \arg \sup_{\xi \in \Xi_i} f(\mathbf{x}, \xi)$  for those  $i \in I'$ . Implicitly, we assume that this supremum is attained. We remark that this modeling choice protects the decision maker in those cases where there is a total absence of information on the conditional distributions  $Q_i, i \in I'$ . Indeed, by introducing the “artificial” data point  $\hat{\xi}_1^i := \arg \sup_{\xi \in \Xi_i} f(\mathbf{x}, \xi)$  in a partition  $\Xi_i$  with no samples, we are considering the worst-case form that the true conditional distribution  $Q_i$  could possibly take, that is, a Dirac distribution supported on  $\hat{\xi}_1^i$ .

Finally, we note that the ambiguity set defined by constraints (3.1b)–(3.1f) is unequivocally determined by specifying the partitions  $\Xi_i, i \in \mathcal{I}$ , the nominal distribution  $\hat{Q}$ , the budgets  $\rho$  and  $\varepsilon$ , and the order cone constraints  $\mathbf{p} \in \mathcal{C}$  in (3.2). In fact, if these constraints are removed and we set  $\rho = \varepsilon = 0$ , then we have  $p_i = \hat{p}_i$  and  $Q_i = \hat{Q}_i, \forall i$ , and therefore,  $Q = \hat{Q}$ .

The following theorem shows that problem (POC) can be reformulated as a single-level problem.

**Theorem 3.1 (Reformulation based on strong duality).** *For any non-negative values of parameters  $\varepsilon, \rho$ , problem (POC) is equivalent to the following:*

$$\begin{aligned}
 (POC-0) \quad & \inf_{\mathbf{x}, \lambda, \mu, \eta, \tilde{\mathbf{p}}, \theta, \mathbf{t}} \lambda \rho + \eta + \theta \varepsilon + \lambda \tilde{c}_{\tilde{\mathbf{p}}}^* \left( \frac{\left( \frac{1}{N_i} \sum_{j=1}^{N_i} t_{i,j} \right)_{i \in \mathcal{I}} + \boldsymbol{\mu} - \eta \mathbf{e} + \tilde{\mathbf{p}}}{\lambda} \right) \\
 \text{s.t.} \quad & t_{i,j} \geq \sup_{\xi \in \Xi_i} \left[ f(\mathbf{x}, \xi) - \theta c(\xi, \hat{\xi}_j^i) \right], \forall i \in \mathcal{I}, j \leq N_i \\
 & \mathbf{x} \in X, \lambda \geq 0, \boldsymbol{\mu} \in \mathbb{R}_+^{|\mathcal{I}|}, \eta \in \mathbb{R}, \tilde{\mathbf{p}} \in \mathcal{C}^*, \theta \geq 0 \\
 & t_{i,j} \in \mathbb{R}, \forall i \in \mathcal{I}, j \leq N_i
 \end{aligned} \tag{3.6}$$

where  $\tilde{c}_{\tilde{\mathbf{p}}}^*(\cdot)$  is the convex conjugate function of  $\tilde{c}_{\tilde{\mathbf{p}}}(\cdot) := \tilde{c}(\cdot, \tilde{\mathbf{p}})$ , with  $\tilde{\mathbf{p}}$  fixed, and  $\left( \frac{1}{N_i} \sum_{j=1}^{N_i} t_{i,j} \right)_{i \in \mathcal{I}}$  is the vector with the  $|\mathcal{I}|$  components  $\frac{1}{N_i} \sum_{j=1}^{N_i} t_{i,j}$ .

Moreover, in the case that the cost function  $\tilde{c}(\cdot, \cdot)$  is given by a norm, we have  $\tilde{c}_{\tilde{\mathbf{p}}}(\mathbf{p}) = \|\mathbf{p} - \tilde{\mathbf{p}}\|$ . The next corollary deals with this particular case.

**Corollary 3.1.** *If the cost functions  $c(\cdot, \cdot)$  and  $\tilde{c}(\cdot, \cdot)$  are given by norms, then for any non-negative values of parameters  $\varepsilon, \rho$ , problem (POC) is equivalent to the following one*

$$\begin{aligned}
 (POC-1) \quad & \inf_{\mathbf{x}, \lambda, \mu, \eta, \tilde{\mathbf{p}}, \theta, \mathbf{t}} \lambda \rho + \eta + \theta \varepsilon + \sum_{i \in \mathcal{I}} \hat{p}_i \left( \frac{1}{N_i} \sum_{j=1}^{N_i} t_{i,j} + \mu_i - \eta + \tilde{p}_i \right) \\
 \text{s.t.} \quad & t_{i,j} \geq \sup_{\xi \in \Xi_i} \left[ f(\mathbf{x}, \xi) - \theta \left\| \xi - \hat{\xi}_j^i \right\| \right], \forall i \in \mathcal{I}, \forall j \leq N_i
 \end{aligned} \tag{3.7}$$

$$\left\| \left( \frac{1}{N_i} \sum_{j=1}^{N_i} t_{i,j} + \mu_i - \eta + \tilde{p}_i \right)_{i \in \mathcal{I}} \right\|_* \leq \lambda$$

$$\mathbf{x} \in X, \lambda \geq 0, \boldsymbol{\mu} \in \mathbb{R}_+^{|\mathcal{I}|}, \eta \in \mathbb{R}, \tilde{\mathbf{p}} \in \mathcal{C}^*, \theta \geq 0$$

$$t_{i,j} \in \mathbb{R}, \forall i \in \mathcal{I}, \forall j \leq N_i$$

**Remark 3.1.** Our data-driven DRO framework (POC) can be easily understood as a generalization of other popular DRO approaches. To see this, first we need to remove the order cone constraints on the probabilities associated with each subregion into which the support  $\Xi$  has been partitioned, that is, the condition  $\mathbf{p} \in \mathcal{C}$ , and then proceed as indicated below:

1. If we set  $\varepsilon = 0$ ,  $|\mathcal{I}| = N$ , with every partition containing a single and different data point from the sample, and use a  $\phi$ -divergence to build the cost function, i.e.,  $\tilde{c}_{\tilde{\mathbf{p}}}(\mathbf{p}) = \sum_{i \in \mathcal{I}} \hat{p}_i \phi\left(\frac{p_i}{\hat{p}_i}\right)$  and hence,  $c_{\tilde{\mathbf{p}}}^*(\mathbf{s}) = \sum_{i \in \mathcal{I}} \hat{p}_i \phi^*(s_i)$ , then our data-driven DRO approach boils down to that of [11] and [13].
2. On the contrary, if we set  $|\mathcal{I}| = 1$  and  $C$  is given by the 1-Wasserstein metric, we get the model of [99].

Finally, we remark that constraint (3.6) for each  $i \in \mathcal{I}'$  is equivalent (under the assumptions we make on the transportation cost function) to  $t_{i,1} \geq \sup_{\xi \in \Xi_i} f(x, \xi)$ .

### 3.1.2 Tractable reformulations

In this section we provide *nice* reformulations of our DRO model (POC) under mild assumptions. For this purpose, we make use of the theoretical foundations laid out in [99]. Likewise, some extensions to our model, such as the extension to two-stage stochastic programming problems, are omitted here for brevity and because they can be easily derived in a similar way as done in [99] for the data-driven DRO approach they develop.

We start our theoretical development with the following assumption.

**Assumption 3.1.** We consider that  $\Xi_i$ , for each  $i \in \mathcal{I}$ , is a closed convex set, and that  $f(\mathbf{x}, \boldsymbol{\xi}) := \max_{k \leq K} g_k(\mathbf{x}, \boldsymbol{\xi})$ , with  $g_k$ , for each  $k \leq K$ , being a proper, concave and upper semicontinuous function with respect to  $\boldsymbol{\xi}$  (for any fixed value of  $\mathbf{x} \in X$ ) and not identically  $\infty$  on  $\Xi_i$ .

Theorem 3.2 below provides a tractable reformulation of problem (POC-1) as a finite convex problem. For ease of notation, we suppress the dependence on the variable  $\mathbf{x}$  (bearing in mind that this dependence occurs through functions  $g_k$ ,  $k \leq K$ ).

**Theorem 3.2.** *If Assumption 3.1 holds and if we choose a norm (in  $\mathbb{R}^d$ ) as the transportation cost function  $c$ , then for any values of  $\rho$  and  $\varepsilon$ , problem (POC-1) is equivalent to the following finite convex problem:*

$$\begin{aligned}
 (POC-1') \quad & \inf_{\mathbf{x}, \lambda, \eta, \boldsymbol{\mu}, \tilde{\mathbf{p}}, \mathbf{z}_{ijk}, \mathbf{v}_{ijk}, \theta, \mathbf{t}} \lambda \rho + \eta + \theta \varepsilon + \sum_{i \in \mathcal{I}} \hat{p}_i \left( \frac{1}{N_i} \sum_{j=1}^{N_i} t_{i,j} + \mu_i - \eta + \tilde{p}_i \right) \\
 \text{s.t. } & t_{i,j} \geq [-g_k]^*(\mathbf{z}_{ijk} - \mathbf{v}_{ijk}) + S_{\Xi_i}(\mathbf{v}_{ijk}) - \langle \mathbf{z}_{ijk}, \hat{\boldsymbol{\xi}}_j^i \rangle \\
 & \forall i \in \mathcal{I}, \forall j \leq N_i, \forall k \leq K \\
 & \|\mathbf{z}_{ijk}\|_* \leq \theta, \forall i \in \mathcal{I}, \forall j \leq N_i, \forall k \leq K \\
 & \left\| \left( \frac{1}{N_i} \sum_{j=1}^{N_i} t_{i,j} + \mu_i - \eta + \tilde{p}_i \right) \right\|_{i \in \mathcal{I}} \leq \lambda \\
 & \mathbf{x} \in X, \lambda \geq 0, \theta \geq 0, \eta \in \mathbb{R}, \boldsymbol{\mu} \in \mathbb{R}_+^{[I]}, \tilde{\mathbf{p}} \in \mathcal{C}^*, \\
 & \mathbf{z}_{ijk}, \mathbf{v}_{ijk} \in \mathbb{R}^d, \forall i \in \mathcal{I}, \forall j \leq N_i, \forall k \leq K \\
 & t_{i,j} \in \mathbb{R}, \forall i \in \mathcal{I}, \forall j \leq N_i
 \end{aligned}$$

where  $[-g_k]^*(\mathbf{z}_{ijk} - \mathbf{v}_{ijk})$  is the conjugate function of  $-g_k$  evaluated at  $\mathbf{z}_{ijk} - \mathbf{v}_{ijk}$  and  $S_{\Xi_i}$  is the support function of  $\Xi_i$ .

We note that Assumption 3.1 covers the particular case where functions  $g_k$ ,  $k \leq K$ , are affine and, as a result,  $f$  is convex piecewise linear. The single-item newsvendor problem, which we illustrate in Section 3.2.1, constitutes a well-known example of this case.

### 3.1.3 Separable objective function

Now we extend the results presented above to a class of objective functions which are additively separable with respect to the dimension  $d$ . We assume here that  $\boldsymbol{\xi} = (\boldsymbol{\xi}_1, \dots, \boldsymbol{\xi}_d)$ , where  $\boldsymbol{\xi}_l \in \mathbb{R}^p$ , for each  $l = 1, \dots, d$ . Furthermore, we consider the separable norm  $\|\boldsymbol{\xi}\|_d := \sum_{l=1}^d \|\boldsymbol{\xi}_l\|$  associated with the base norm  $\|\cdot\|$  (on  $\mathbb{R}^p$ ). Finally, we assume that the function  $f$  is given as follows:

$$f(\mathbf{x}, \boldsymbol{\xi}) = \sum_{l=1}^d \max_{k \leq K} g_{lk}(\mathbf{x}, \boldsymbol{\xi}_l) \quad (3.8)$$

In this case, the complexity of the resulting DRO problem is linear with respect to the number  $N$  of samples. The multi-item newsvendor problem, which we illustrate in Section 3.2.1, constitutes a popular example of this case.

**Theorem 3.3.** *If  $f(\mathbf{x}, \boldsymbol{\xi}) = \sum_{l=1}^d \max_{k \leq K} g_{lk}(\mathbf{x}, \boldsymbol{\xi}_l)$ ,  $\{g_{lk}\}_{k \leq K}$  satisfy Assumption 3.1 for all  $l \leq d$ , and  $\Xi_i$ , for each  $i \in \mathcal{I}$ , is given by the Cartesian product of closed*

convex sets (that is,  $\Xi_i := \prod_{l=1}^d D_l^i$ , with  $D_l^i$  a closed convex set), and if we choose the norm  $\|\cdot\|_d$  as the transportation cost function  $c$ , then for any values of  $\rho$  and  $\varepsilon$ , problem (POC) is equivalent to the following finite convex problem:

$$(POC-2) \quad \inf_{\mathbf{x}, \lambda, \eta, \boldsymbol{\mu}, \tilde{\mathbf{p}}, \mathbf{z}_{ijkl}, \mathbf{v}_{ijkl}, \theta, \boldsymbol{\omega}} \lambda \rho + \eta + \theta \varepsilon + \sum_{i \in \mathcal{I}} \hat{p}_i \left( \frac{1}{N_i} \sum_{j=1}^{N_i} \sum_{l=1}^d \omega_{ijl} + \mu_i - \eta + \tilde{p}_i \right) \quad (3.9)$$

$$s.t. \omega_{ijl} \geq [-g_{lk}]^* (\mathbf{z}_{ijkl} - \mathbf{v}_{ijkl}) + S_{D_l^i}(\mathbf{v}_{ijkl}) - \langle \mathbf{z}_{ijkl}, \hat{\boldsymbol{\xi}}_{jl}^i \rangle, \quad (3.10)$$

$$\forall i \in \mathcal{I}, \forall j \leq N_i, \forall k \leq K, \forall l \leq d$$

$$\|\mathbf{z}_{ijkl}\|_* \leq \theta, \forall i \in \mathcal{I}, \forall j \leq N_i, \forall k \leq K, \forall l \leq d \quad (3.11)$$

$$\left\| \left( \frac{1}{N_i} \sum_{j=1}^{N_i} \sum_{l=1}^d \omega_{ijl} + \mu_i - \eta + \tilde{p}_i \right) \right\|_* \leq \lambda \quad (3.12)$$

$$\mathbf{x} \in X, \lambda \geq 0, \theta \geq 0, \eta \in \mathbb{R}, \boldsymbol{\mu} \in \mathbb{R}_+^{|\mathcal{I}|}, \tilde{\mathbf{p}} \in \mathcal{C}^*, \quad (3.13)$$

$$\omega_{ijl} \in \mathbb{R}, \forall i \in \mathcal{I}, \forall j \leq N_i, \forall l \leq d \quad (3.14)$$

$$\mathbf{z}_{ijkl}, \mathbf{v}_{ijkl} \in \mathbb{R}^p, \forall i \in \mathcal{I}, \forall j \leq N_i, \forall k \leq K, \forall l \leq d \quad (3.15)$$

### 3.1.4 Order cone constraints

To account for *a-priori* knowledge about the probability distribution of the random parameter vector  $\boldsymbol{\xi}$  (for example, the decision maker may have some information about the shape of this distribution), we propose to convey this knowledge using order constraints on the probability masses  $p_i$  associated with each subregion  $\Xi_i$  into which the support  $\Xi$  of  $\boldsymbol{\xi}$  is partitioned. These order constraints are based on order cones, which, in turn, can be represented in the form of graphs. We can build order cones from graphs that allow for the comparison of all probabilities  $p_i$ . In that case, we say that the graph, and the associated cone, establish a *total order*. If, on the contrary, the graph only allows some of those probabilities to be compared, we talk about *partial order*. For more details about order cones we refer the reader to [104].

Below, we present some common choices of order cones.

- *Simple order cone (monotonicity):*

$$\mathcal{C} = \{p \in \mathbb{R}^{|\mathcal{I}|} : p_1 \geq \dots \geq p_{|\mathcal{I}|}\}$$

- *Tree order cone:*

$$\mathcal{C} = \{p \in \mathbb{R}^{|\mathcal{I}|} : p_i \geq p_{|\mathcal{I}|}, i = 1, \dots, |\mathcal{I}| - 1\}$$

- *Star-shaped cone (decrease on average):*

$$\mathcal{C} = \left\{ p \in \mathbb{R}^{|\mathcal{I}|} : p_1 \geq \frac{p_1 + p_2}{2} \geq \dots \geq \frac{p_1 + \dots + p_{|\mathcal{I}|}}{|\mathcal{I}|} \right\}$$

- *Umbrella cone (unimodality):*

$$\mathcal{C} = \{ p \in \mathbb{R}^{|\mathcal{I}|} : p_1 \leq p_2 \leq \dots \leq p_m \geq p_{m+1} \geq \dots \geq p_{|\mathcal{I}|} \}$$

An order cone is a *polyhedral convex cone* and as such, can be algebraically expressed in the form  $\mathcal{C} = \{ \mathbf{p} \in \mathbb{R}^{|\mathcal{I}|} : \mathbf{A}\mathbf{p} \geq 0 \}$ , with  $\mathbf{A}$  being a matrix of appropriate dimensions. Its dual  $\mathcal{C}^*$  can, therefore, be easily computed as  $\mathcal{C}^* = \{ \tilde{\mathbf{p}} = \mathbf{A}^\top \boldsymbol{\nu} : \boldsymbol{\nu} \geq \mathbf{0} \}$  (see, for instance, Corollary 3.12.9 in [126]). Notwithstanding, our DRO approach can be equally applied under other types of support sets, as long as the problem

$$\sup_{\mathbf{p} \in \Theta} \left[ \left\langle \mathbf{p}, \left( \frac{1}{N_i} \sum_{j=1}^{N_i} t_{i,j} \right)_{i \in \mathcal{I}} \right\rangle - \lambda \tilde{c}(\mathbf{p}, \hat{\mathbf{p}}) \right] \quad (3.16)$$

admits a strong dual (we refer the interested reader to [13] for a list of types of support sets under which strong duality holds).

As compared to other approaches available in the technical literature, order cones provide a straightforward way of encoding modality information in the ambiguity set of the DRO problem. For instance, [70] indirectly introduces multi-modality information by imposing first and second moment conditions on the different ambiguous components of a mixture with known weights. Their approach, however, results in a semidefinite program. Unlike [70], the authors in [89] explicitly incorporate modality information into their ambiguity set through moment and generalized unimodal constraints. Nonetheless, they still need to solve a semidefinite program and their DRO approach overlooks the data-driven nature of those constraints. In [36], they construct an ambiguity set made up of those absolutely continuous probability distributions whose density function is bounded by some bands with a certain confidence level. Their approach can be used to impose monotonicity or unimodality of the probability distributions, but can only be applied to the univariate case.

Beyond modality, the order cone constraints on the partition probabilities that characterize our DRO approach equip the decision maker with a versatile and intuitive framework to exploit information on the shape of the ambiguous probability distribution. For example, as we do in the numerical experiments in this chapter, we can construct an order cone that constrains the ratios among the partition probabilities, which can be seen as a discrete approximation of encoding “derivative” information on the ambiguous probability distribution (if this admits a density function). Likewise, other order cones

could be used to bestow some sense of “convexity” on this distribution.

### 3.1.5 On convergence and out-of-sample performance guarantees

In this section, we show that our DRO approach (POC) naturally inherits the convergence and performance guarantees of that introduced in [99]. Following [99], the training data sample,  $\{\hat{\xi}^i\}_{i=1}^N \subseteq \Xi$ , can be seen as a random vector governed by the probability distribution  $\mathbb{Q}^N := \mathbb{Q} \times \cdots \times \mathbb{Q}$  ( $N$  times) supported on  $\Xi^N$  (with the respective product  $\sigma$ -algebra). Ideally, we strive to develop a method capable of identifying a highly reliable data-driven solution with a certificate as low as possible.

The data-driven DRO approach that we propose to address the problem defined by (2.1) accounts for the uncertainty about the true data-generating distribution  $\mathbb{Q}$ , while taking advantage of some a-priori *order* information that the decision maker may have on some probabilities induced by  $\mathbb{Q}$  over a partition of the support set  $\Xi$ .

Below, we claim that the pair  $(\hat{\mathbf{x}}_N, \hat{J}_N)$  provided by our distributionally robust optimization problem (POC) features performance guarantees in line with those discussed in [99]. More specifically, for a suitable choice of the ambiguity set, the optimal value  $\hat{J}_N$  of problem (POC) constitutes a certificate of the type (2.4) providing a confidence level  $1 - \beta$  on the out-of-sample performance of the data-driven solution  $\hat{\mathbf{x}}_N$ . This can be formally stated under some assumptions about the underlying true conditional probability distributions.

To this end, we first provide probabilistic guarantees on the partition probabilities  $p_i$ ,  $\forall i \leq |\mathcal{I}|$ . In this vein, note that the empirical probability  $\hat{p}_i$ , defined as in Equation (3.4), can be modeled as a binomial distribution with success probability  $p_i^*$ , divided by the total number of trials. Consequently, by the Strong Law of Large Numbers (SLLN),  $\hat{p}_i$  converges to  $p_i^*$  almost surely.

Now suppose that we choose a  $\phi$ -divergence as  $\tilde{c}$ , where  $\phi$  is a twice continuously differentiable function around 1 with  $\phi''(1) > 0$ . Then, take  $\beta_p > 0$ . If we choose as  $\rho$  the value

$$\rho(\beta_p) := (\phi''(1)/(2N))\chi_{|\mathcal{I}|-1, 1-\beta_p}^2 \quad (3.17)$$

we get a confidence set of level  $1 - \beta_p$  on the true partition probabilities  $\mathbf{p}^*$  (see [13] and [11]).

If, alternatively, we choose the total variation distance as  $\tilde{c}$ , we can use Equation (19) in [68] to take  $\rho$  as

$$\rho(\beta_p) := (|\mathcal{I}|/\sqrt{N})(2 + \sqrt{2 \log(|\mathcal{I}|/\beta_p)}) \quad (3.18)$$

and obtain a confidence set of level  $1 - \beta_p$  on  $\mathbf{p}^*$ .

Next we establish a concentration tail inequality of the probability weighted Wasser-

stein metric of order 1 between each conditional distribution and its respective true conditional distribution. For this purpose, we first need to make the following assumption:

**| Assumption 3.2 (Light-tailed Conditional Distributions).** *For each  $i \in \mathcal{I}$ , there exist  $a_i, \gamma_i \in \mathbb{R}$ , with  $a_i > 1$  and  $\gamma_i > 0$  such that*

$$\mathbb{E}_{\mathbb{Q}_i} [\exp(\gamma_i \|\boldsymbol{\xi}\|^{a_i})] = \int_{\Xi_i} \exp(\gamma_i \|\boldsymbol{\xi}\|^{a_i}) \mathbb{Q}_i(d\boldsymbol{\xi}) < \infty. \quad (3.19)$$

The following theorem provides a tail concentration inequality for the weighted sum of the Wasserstein metrics of order 1 between the true and empirical conditional distributions.

**| Theorem 3.4 (Concentration Inequality for Conditional Distributions ).** *If Assumption 3.2 holds, for each  $i \in \mathcal{I}$ , given  $\beta_i \in (0, 1]$  we have that  $\forall N_i \geq 1$ ,  $\dim(\boldsymbol{\xi}) \neq 2$  and for all  $\varepsilon > \sum_{i \in \mathcal{I}} p_i \varepsilon_{N_i}(\beta_i)$ , for any values  $p_i, i \in \mathcal{I}$  such that  $p_i \geq 0$  and  $\sum_{i \in \mathcal{I}} p_i = 1$ , the following holds*

$$\mathbb{P} \left[ \sum_{i \in \mathcal{I}} p_i \mathcal{W}(\mathbb{Q}_i, \hat{\mathbb{Q}}_i) \leq \varepsilon \right] \geq 1 - \sum_{i \in \mathcal{I}} \beta_i \quad (3.20)$$

where

$$\varepsilon_{N_i}(\beta_i) := \begin{cases} \left( \frac{\log(B_i \beta_i^{-1})}{C_i N_i} \right)^{1/\max\{\dim(\boldsymbol{\xi}), 2\}} & \text{if } N_i \geq \frac{\log(B_i \beta_i^{-1})}{C_i}, \\ \left( \frac{\log(B_i \beta_i^{-1})}{C_i N_i} \right)^{1/a_i} & \text{if } N_i < \frac{\log(B_i \beta_i^{-1})}{C_i}. \end{cases} \quad (3.21)$$

Theorem 3.4 sets the probabilistic bound  $\sum_{i \in \mathcal{I}} p_i \varepsilon_{N_i}(\beta_i)$  on the weighted Wasserstein metric of order 1 between each conditional distribution and its respective true conditional distribution, with at least confidence level  $1 - \sum_{i \in \mathcal{I}} \beta_i$ . We remark that, if the partitions are compact, stronger results like those in Theorem 2 of [79] could be used to choose the radii of the Wasserstein balls. More specifically, the result in Theorem 2 of [79] depends on the diameter of the compact support set (i.e., the maximum distance between two elements of that set). The result stated in our theorem, in contrast, is valid for unbounded partitions, as it only requires the true conditional distribution associated with each partition be light-tailed. The next theorem states the finite-sample guarantee performance of the proposed DRO method we develop in this thesis:

**| Theorem 3.5 (Finite sample guarantee).** *Suppose that Assumption 3.2 holds and that we have chosen as  $\rho$  the value given by Equation (3.17) or (3.18). Then, the finite sample guarantee (2.4) holds with at least confidence level  $(1 - \beta_p)(1 - \sum_{i \in \mathcal{I}} \beta_i)$ .*

**| Remark 3.2.** *In practice, proper values for the hyperparameters  $\varepsilon$  and  $\rho$  are set by way of data-driven procedures like bootstrapping or cross-validation, as we illustrate in*



the numerical experiments in Section 3.2.2 (see also [35], [40], [99], [122], and [136] for more examples). These procedures allow the decision maker to tune those parameters as a function of the sample size  $N$  in order to get reliable decisions without giving up too much on out-of-sample performance. Following this line, and as noted in Remark 5 in [87], the requirement to include the true distribution inside the ambiguity set is only a sufficient, but not necessary condition to ensure a finite sample guarantee. Indeed, this guarantee can be sustained even if the parameters of the ambiguity set are reduced below the lowest values for which the ambiguity set represents a confidence set for the true distribution.

Furthermore, recall that the partition probabilities  $\mathbf{p}$  belong to the support set  $\Theta$  defined by the order cone constraints. Since we assume that these constraints are coherent with the true distribution  $\mathbb{Q}$ , we do not need to explore those probability measures  $Q$  in the Wasserstein ball  $\mathbb{B}_{\rho_N(\beta)}$  that do not comply with them. Consider, for example, the case in which the worst-case distribution in the ball  $\mathbb{B}_{\rho_N(\beta)}$  does not satisfy the order cone constraints. One could expect, therefore, that, in practice, our approach could benefit from this fact to produce a data-driven solution  $\hat{\mathbf{x}}_N$  as reliable as that given by the method of [99], but with a tighter certificate  $\hat{J}_N$ . This is precisely what we observe in the numerical experiments that we present in Section 3.2.1 and Section 3.2.2.

We conclude this section with some remarks on the convergence and asymptotic consistency of our DRO approach: We have that, as the number  $N$  of samples grows to infinity,

$$(\hat{\mathbf{x}}_N, \hat{J}_N) \rightarrow (\mathbf{x}^*, J^*) \quad (3.22)$$

where  $\mathbf{x}^*$  (resp.  $J^*$ ) is an optimizer (resp. the optimal solution value) of problem defined by (2.1).

Indeed, assume that Theorem 3.6 in [99] holds, then take a confidence level  $1 - \beta$ , and choose  $\varepsilon$  and  $\rho$  by way of Theorem 3.4 and Equations (3.17) (or (3.18)), respectively. When  $N$  grows to infinity, we have, on the one hand, that the conditional distributions converge (in the Wasserstein metric) to their respective true conditional distributions and the probability weights converge a.s. by the SLLN to their respective true values. Therefore, both  $\varepsilon$  and  $\rho$  tend to zero as  $N$  increases to infinity. Consequently, our ambiguity set only contains the empirical distribution  $\hat{Q}_N$ , which converges almost surely to the true distribution  $\mathbb{Q}$ .

## 3.2 Numerical experiments

The purpose of this section is to provide additional insights into the computational aspects and the performance guarantees of our proposed distributionally robust opti-

mization scheme with order cone constraints. For this purpose, we consider two test instances: the (single and multi-item) newsvendor problem and the problem of a strategic firm competing *à la Cournot* in a market, which will be discussed in detail in Section 3.2.1 and Section 3.2.2, respectively. These two problems have been intentionally selected, because they are qualitatively different when addressed by the standard Wasserstein-metric-based DRO approach proposed in [99]. In effect, the former features an objective function  $f(\mathbf{x}, \boldsymbol{\xi})$  whose Lipschitz constant with respect to  $\boldsymbol{\xi}$  is independent of the decision  $\mathbf{x}$ . Consequently, as per Remark 6.7 in [99], the standard Wasserstein-metric-based DRO approach renders the same minimizer for this problem as the sample average approximation, whenever the support of the uncertainty  $\boldsymbol{\xi}$  is unbounded. This is, in contrast, not true for the problem of a strategic firm competing *à la Cournot* in a market, which is characterized by an objective function with a Lipschitz constant over  $\boldsymbol{\xi}$  that is a function of  $\mathbf{x}$ . This allows us to highlight the differences of our approach with regard to [99] in two distinct settings.

All the numerical experiments have been implemented in Python. The optimization problems have been built using Pyomo and solved with CPLEX 12.10 on a PC with Windows 10 and a CPU Intel (R) Core i7-8550U clocking at 1.80 GHz and with 8 GB of RAM. The statistical methods that have been used for the numerical experiments have been coded by means of the module Scikit-learn (see [114]). In what follows we provide some implementation details regarding the proposed model. The numerical experiments have been designed under the following assumptions:

1. *A-priori information.* Given a fixed and known partition of the sample space  $\Xi$ , we can construct an order cone that is consistent with the true probability distribution. That is, the probability masses that the true distribution assigns to each partition verify the order cone constraints. In practice, this a-priori information is determined by the nature of the problem and the random phenomena, and is assumed to be known by the decision maker based on experience and expert knowledge. Furthermore, in the case that the decision maker has no full certainty about the a-priori information, s/he may resort to statistical hypothesis testing to assess the confidence that the partition probabilities belong to a given order cone (see, for instance, [23] and references therein).

In our numerical experiments, we specifically apply the following approach: Given a fixed number of partitions (later we explain how the partition set is obtained), we consider that the decision maker knows a total order between the probability masses associated with each of the regions into which the sample space  $\Xi$  is split. Furthermore, s/he also knows their ratios approximately, within a certain tolerance (which, in the subsequent experiments, we set to 0.1).

For instance, suppose we have three partitions with (true) probability masses of

$p_1^* = 0.6$ ,  $p_2^* = 0.3$  and  $p_3^* = 0.1$ . The decision maker only knows their relative ratios with a tolerance error of 0.1, that is:

$$\begin{aligned} p_1 &\geq (0.6/0.3 - 0.1)p_2 \\ p_2 &\geq (0.3/0.1 - 0.1)p_3 \end{aligned}$$

This way, we get the following order cone constraints:

$$\begin{aligned} p_1 &\geq 1.9p_2 \\ p_2 &\geq 2.9p_3 \end{aligned}$$

2. *Support set  $\Xi$ .* The support set is the Cartesian product of closed intervals (that is, an hypercube, whose size is indicated in each example) and, therefore, is a closed convex set.
3. *True distribution.* For simulation and analysis, the data-generating distribution is approximated by a certain number of data points (15 000 in the newsvendor setting and 10 000 in the problem of the Cournot producer) drawn from a mixture of three normal distributions, whose characteristics are specified in each of the two examples we consider in the following subsections. Furthermore, those data points that fall outside the support set  $\Xi$  are discarded.
4. *Construction of partitions  $\Xi_i$ ,  $i = 1, \dots, |\mathcal{I}|$ :* In order to construct the partitions, we proceed as follows.
  - (a) *Clustering phase:* Firstly, we employ the  $K$ -means clustering technique to group the total number of data points that approximate the true data distribution into  $K$  clusters. The number  $K$  of clusters is decided upon using the well-known Elbow's method (see, for example, [41]). It is based on the value of the average distortion produced by different values of  $K$ . If  $K$  increases, the average distortion will decrease and the improvement in average distortion will diminish. The value of  $K$  at which the improvement in distortion decreases the most is called the *elbow*. At this value of  $K$ , we should stop dividing the data into further clusters and choose this value as the number of clusters. In addition, we assign a label to identify each of the  $K$  clusters. In all the numerical experiments that are presented next, the true data-generating distribution is constructed as a mixture of three (univariate or multivariate) normal distributions. We assume that the decision maker has a good estimate of the number of components of this mixture and thus, we consider, for example, four clusters, i.e.,  $K = 4$ .

- (b) *Decision-tree classifier phase:* Once all the clusters have been labelled, we use the aforementioned total number of data points to train a decision-tree multi-classifier with a maximum number of leafs equal to  $K$ . The tree will be then used to allocate new data points into one of the  $K$  clusters, which, in effect, is equivalent to having a partition of the support set in  $K$  disjoint regions.
5. *Comparative analysis:* We compare three different data-driven approaches to address the solution to problem  $\inf_{\mathbf{x} \in X} \mathbb{E}_{\mathbb{Q}} [f(\mathbf{x}, \boldsymbol{\xi})]$ , namely, our approach (DROC), the one of [99] (DROW) and the sample average approximation (SAA). Recall that we denote  $x^* \in \arg \min_{\mathbf{x} \in X} \mathbb{E}_{\mathbb{Q}} [f(\mathbf{x}, \boldsymbol{\xi})]$  and  $J^* = \mathbb{E}_{\mathbb{Q}} [f(x^*, \boldsymbol{\xi})]$ , which, in practice, are unknown to the decision maker, but, for analysis purposes, we estimate using the total number of data points that approximate the true data-generating distribution. Moreover, in all numerical experiments, we consider the 1-norm as the functions  $c$  and  $\tilde{c}$ . To compare the three data-driven approaches we consider, we use two performance metrics, specifically, the *out-of-sample performance* of the data-driven solution (which we also refer to as its *actual expected cost*) and its *out-of-sample disappointment*. The former is given by  $\mathbb{E}_{\mathbb{Q}} [f(\hat{\mathbf{x}}_N^m, \boldsymbol{\xi})]$ , while the latter is calculated as  $\mathbb{E}_{\mathbb{Q}} [f(\hat{\mathbf{x}}_N^m, \boldsymbol{\xi})] - \hat{J}_N^m$ , where  $m = \{\text{DROC}, \text{DROW}, \text{SAA}\}$  and  $\hat{J}_N^m$  is the objective function value yielded by the data-driven optimization problem solved by method  $m$ . We stress that a negative out-of-sample disappointment represents a favourable outcome. As  $\mathbb{E}_{\mathbb{Q}} [f(\hat{\mathbf{x}}_N^m, \boldsymbol{\xi})]$  and  $\hat{J}_N^m$  are random variables (they are direct functions of the sample data), we conduct a certain number of runs, each with an independent sample of size  $N$ . This way we can provide (visual) estimates of the expected value and variability of the out-of-sample performance and disappointment for several values of the sample size  $N$ . These estimates are illustrated in the form of box plots in a series of figures. In these figures, the dotted black horizontal line corresponds to either solution  $x^*$  or to its associated optimal cost  $J^*$  with complete information (i.e., without ambiguity about the true data distribution).

For the sole purpose of conducting a comparison as fairly as possible, parameters  $\varepsilon$  and  $\rho$  in both DROC and DROW are tuned so that the underlying true distribution of the data belongs to the corresponding ambiguity set with, at least, a pre-fixed confidence level of probability. In the case of the newsvendor examples, we guarantee this by trial and error for simplicity. In practice, however, these parameters should be calibrated by way of a (statistical) procedure that uses the data available to the decision maker, for example, through cross-validation or bootstrapping, as we illustrate in Section 3.2.2 with the problem of a strategic firm competing *à la Cournot* in a market for a homogeneous product.

Finally, we stress that, in our approach, caution should be exercised when selecting  $\varepsilon$  and  $\rho$ , as they should be such that problem (POC) has at least one feasible solution. This is not guaranteed in the case that the empirical distribution  $\hat{Q}$  does not satisfy the order cone constraints on the probability masses associated with each subregion  $\Xi_i$  of the support set  $\Xi$ . Intuitively, in this case, optimization problem (POC) must have enough “budget” (i.e.,  $\varepsilon$  and  $\rho$  must be high enough) to “transport” the empirical distribution to another one that complies with the a-priori information. In other words, the ambiguity set of problem (POC) must be sufficiently large to contain at least one probability distribution that assigns probability masses verifying the order cone constraints to the partitions.

### 3.2.1 Application I. Newsvendor problems

In this section, we illustrate the proposed DRO formulation on the popular *newsvendor* problem (also known as the *newsboy problem*). Many extensions and variants of this problem have been considered since it was first posed in the 50s (see, for example, the work in [5], [39], [61], [111], and references therein). According to [110],

*The newsboy problem is probably the most studied stochastic inventory model in inventory control theory and the one with most extensions in recent years. This problem reflects many real-life situations and is often used to aid decision making in both manufacturing and retailing. It is particularly important for items with significant demand uncertainty and large over-stocking and under-stocking costs.*

#### The single-item newsvendor problem

In the single-item newsvendor model, the decision maker has to plan the inventory level for a certain product before the random demand  $\xi$  for that product is realized, facing both holding and backorder costs. The newsvendor problem can be formulated as

$$\inf_{x \geq 0} \mathbb{E}_Q[h(x - \xi)^+ + b(\xi - x)^+]$$

where  $x$  is the order quantity, and  $b, h > 0$  are the unit holding cost and the unit backorder cost, respectively. It is known that the solution to the single-item newsvendor problem is equivalent to that of a quantile regression problem, where the goal is to estimate the quantile  $b/(b + h)$  of the distribution of the uncertainty  $y$ , with  $h$  and  $b$  being the unit holding and backorder costs, respectively. Here we have assumed that  $h = 4$  and  $b = 2$ .

The demand for the item (unknown to the decision maker) is assumed to follow a mixture (with weights  $\omega_1 = 0.1$ ,  $\omega_2 = 0.35$  and  $\omega_3 = 0.55$ ) of the three normal distributions  $\mathcal{N}_1(0.2, 0.05)$ ,  $\mathcal{N}_1(0.5, 0.1)$ , and  $\mathcal{N}_1(0.8, 0.05)$ , truncated over the unit interval  $[0, 1]$ . Figure 3.1a provides a visual illustration of the resulting mixture. Recall

that, in the numerical experiments that follow, we have used 15 000 samples drawn from this mixture of Gaussian distributions to approximate the true distribution of the item demand and to partition its support set  $[0, 1]$  into four regions, based on the two-phase procedure we have previously described. In fact, what we show in Figure 3.1a is the histogram of those 15 000 data points and its corresponding kernel density estimate.

For the sole purpose of conducting a comparison as fairly as possible, parameters  $\varepsilon$  and  $\rho$  in both DROC and DROW are tuned so that the underlying true distribution of the data belongs to the corresponding ambiguity set with at least 95% of probability. We check whether this condition holds or not *a posteriori* (by trial and error), by counting the number of runs (out of the one thousand we perform) for which the out-of-sample disappointment is negative.

The values we have used for the parameters  $\varepsilon$  and  $\rho$  in DROC and DROW are collated in Table 3.1. We insist that these parameters have been adjusted so that at most 50 out of the 1000 runs we have conducted for each sample size  $N$  deliver a positive out-of-sample disappointment (that is, to achieve and maintain a similar level of reliability for the data-driven solutions given by DROC and DROW). As expected, therefore, the values of both  $\varepsilon$  and  $\rho$  decrease as the sample size  $N$  grows.

Table 3.1: Single-item newsvendor problem: Values for parameters  $\varepsilon, \rho$  in DROC and  $\rho$  in DROW.

| $N$ | DROC          |        | DROW   |
|-----|---------------|--------|--------|
|     | $\varepsilon$ | $\rho$ | $\rho$ |
| 2   | 0.9           | 0.9    | 1      |
| 5   | 0.8           | 0.8    | 0.9    |
| 10  | 0.7           | 0.7    | 0.8    |
| 20  | 0.4           | 0.6    | 0.6    |
| 50  | 0.15          | 0.25   | 0.4    |
| 100 | 0.1           | 0.2    | 0.25   |
| 200 | 0.01          | 0.15   | 0.05   |

Figures 3.1b, 3.1c, and 3.1d show the box plots corresponding to the order quantity, the out-of-sample disappointment and the actual expected cost delivered by each of the considered data-driven approaches for various sample sizes. The shaded areas have been obtained by joining the whiskers of the box plots, while the associated solid lines link their medians. Interestingly, whereas the medians of the order quantity estimators provided by SAA are very close to the optimal one  $x^*$ , their high variability results in (large) disappointment with very high probability. On the contrary, the median of the order quantity delivered by DROW is significantly far from the optimal one (with complete information) for small sample sizes, but it manages to keep the out-of-sample disappointment below zero in return. To do so, however, DROW tends to produce costly (overconservative) solutions on average, as inferred from their actual expected cost in

Figure 3.1d. In plain words, DROW pays quite a lot to ensure a highly reliable/robust order quantity. The proposed approach DDRO, however, is able to leverage the a-priori information on the partition probabilities  $(p_i)_{i=1}^{|I|}$  to substantially reduce the cost to pay for reliable data-driven solutions, especially for small sample sizes. Intuitively, this information enables DROC to identify highly reliable solutions that are myopically deemed as non-reliable and, therefore, discarded by DROW. Logically, this is contingent on the quality of the a-priori information that is supplied to DROC in the form of order cone constraints on  $(p_i)_{i=1}^{|I|}$ .

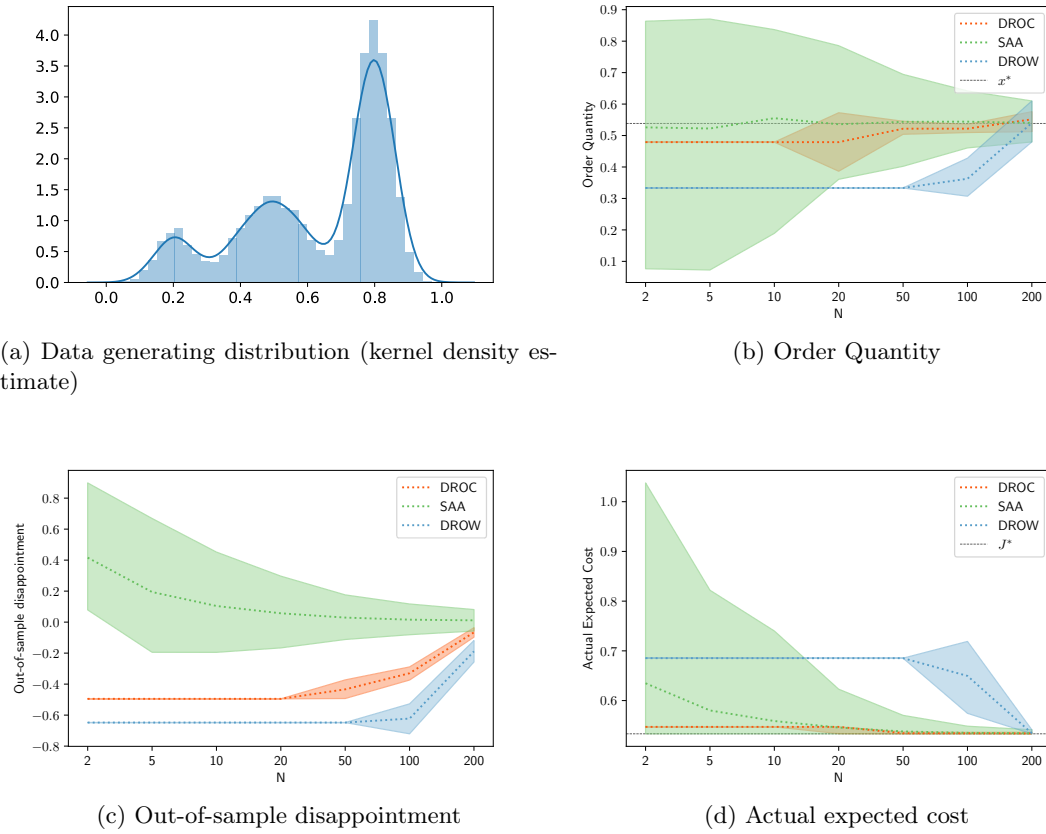


Figure 3.1: Single-item newsvendor problem: (Approximate) true data-generating distribution, order quantity and performance metrics

### The multi-item newsvendor problem

In this section, we carry out an analysis similar to that of Subsection 3.2.1, but for the multi-item newsvendor problem, which can be formulated as follows:

$$\inf_{\mathbf{x} \geq \mathbf{0}} \mathbb{E}_Q \sum_{l=1}^d [h_l(x_l - \xi_l)^+ + b_l(\xi_l - x_l)^+]$$

where  $x_l$  is the order quantity for the  $l$ -th item,  $Q$  is the joint probability distribution governing the demands for the  $d$  items, and  $b_l, h_l > 0$  are the unit holding cost and the unit backorder cost for the  $l$ -th item, respectively.

To illustrate our approach in a higher dimensional setting, we consider twenty items, i.e.,  $d = 20$ . We consider the following parameters:  $h_1 = \dots = h_{10} = 2, h_{11} = \dots = h_{20} = 4, b_1 = \dots = b_{10} = 4$ ; and  $b_{11} = \dots = b_{20} = 2$ . The demands for the twenty items are assumed to follow a mixture of three multivariate normal distributions  $\mathcal{N}_{20}(\boldsymbol{\mu}_1, \Sigma_1)$ ,  $\mathcal{N}_{20}(\boldsymbol{\mu}_2, \Sigma_2)$ , and  $\mathcal{N}_{20}(\boldsymbol{\mu}_3, \Sigma_3)$ , where  $\boldsymbol{\mu}_1 = [3, \dots, 3] \in \mathbb{R}^{20}$ ,  $\Sigma_1 = \text{diag}(1, \dots, 1) \in \mathbb{R}^{20 \times 20}$ ;  $\boldsymbol{\mu}_2 = [5, \dots, 5] \in \mathbb{R}^{20}$ ,  $\Sigma_2 = \text{diag}(0.5, \dots, 0.5) \in \mathbb{R}^{20 \times 20}$ ; and  $\boldsymbol{\mu}_3 = [7, \dots, 7] \in \mathbb{R}^{20}$ ,  $\Sigma_3 = \text{diag}(0.1, \dots, 0.1) \in \mathbb{R}^{20 \times 20}$ . The weights of the mixture are  $\omega_1 = 0.1$ ,  $\omega_2 = 0.65$  and  $\omega_3 = 0.25$ , respectively. Furthermore, the mixture has been truncated on the hypercube  $[0, 10]^{20}$ .

The values we have used for the parameters  $\varepsilon$  and  $\rho$  in DROC and DROW are collated in Table 3.2.

Table 3.2: Multi-item newsvendor problem: Values for parameters  $\varepsilon, \rho$  in DROC and  $\rho$  in DROW

| $N$ | DROC          |        | DROW   |
|-----|---------------|--------|--------|
|     | $\varepsilon$ | $\rho$ | $\rho$ |
| 2   | 5             | 2      | 60     |
| 5   | 5             | 2      | 50     |
| 10  | 4.5           | 1.5    | 40     |
| 20  | 4             | 1      | 20     |
| 50  | 2.5           | 0.6    | 10     |
| 100 | 1.75          | 0.5    | 8      |
| 200 | 1.25          | 0.35   | 4      |

Again, for a meaningful and fair comparison, these parameters have been tuned by trial and error in such a way that at most 50 out of the 1000 runs we have carried out for each sample size  $N$  yield a positive out-of-sample disappointment. The values for the parameters, which we need to this end, diminish as we gain more information (i.e., as the sample size  $N$  grows). Note that, for small sample sizes, for which the available data provide very little information about their true distribution, a great deal of robustness is required to produce highly reliable data-driven solutions. Consequently,



it is little wonder that the selected values for  $\rho$  in DROC are equal to two, which is the maximum value that the total variation distance between  $P$  and  $\hat{P}$  can take on. In

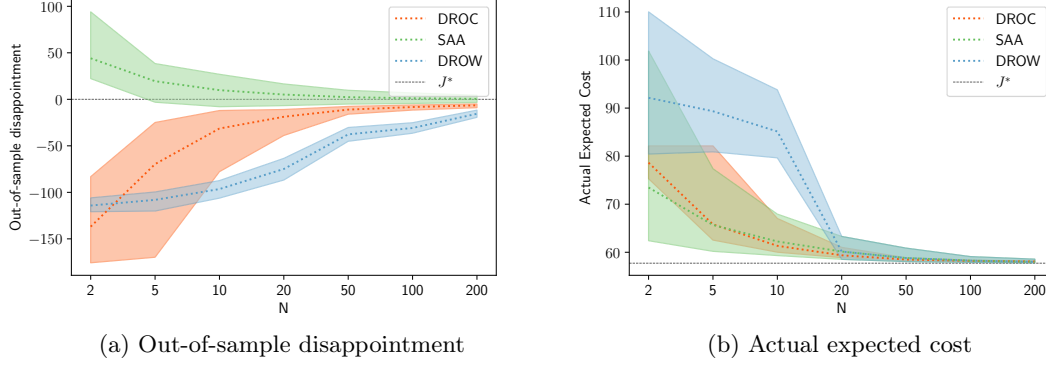


Figure 3.2: Multi-item newsvendor problem: Performance metrics

the same fashion as in the case of the previous example of the single-item newsvendor problem, Figures 3.2a and 3.2b show, for various sample sizes, the box plots pertaining to the out-of-sample disappointment and the actual expected cost associated with each of the considered data-driven approaches, in that order. The results conveyed by these figures confirm our initial conclusions: The ability of our approach DROC to exploit a-priori knowledge of the order among some partition probabilities permits identifying solutions that perform noticeably better out of sample with the same level of confidence. We underline that, in terms of the out-of-sample disappointment, the decision maker seeks a data-driven method  $m$  that renders an estimate  $\hat{J}_N^m$  that results in a positive surprise (i.e., negative disappointment) with a high probability, but that is as close as possible to the cost with full information  $J^*$ . Consequently, the large negative out-of-sample disappointment that the solutions given by DROW feature can be attributed to its over-conservativeness.

In terms of computational time, solving DROC for this instance of the multi-item newsvendor problem, with 20 items, four partitions and a sample size of 200, takes less than a second with CPLEX 12.10 running on a Windows 10 PC with a CPU Intel (R) Core i7-8550U clocking at 1.80 GHz and 8 GB of RAM.

### 3.2.2 Application II. The problem of a strategic firm competing *à la Cournot* in a market

In this section, we illustrate the proposed DRO formulation considering the problem of a strategic firm competing *à la Cournot* in a market for an undifferentiated product. This could be the case, for instance, of the electricity market (see, e.g., [60, Ch. 3] and

[119]). Suppose the firm can produce up to one per-unit amount of product at a cost given by  $a_2x^2 + a_1x + a_0$ , where  $x$  is the per-unit amount of product eventually produced and  $a_0, a_1$  and  $a_2$  are *known* parameters taking values in  $\mathbb{R}^+$ . Furthermore, assume an inverse residual demand function in the form  $\lambda = \alpha - \beta x$ , where  $\lambda$  is the market clearing price for the product, and  $\alpha, \beta \in \mathbb{R}^+$  are *unknown and uncertain* parameters. The firm seeks, therefore, to minimize its cost  $(a_2x^2 + a_1x + a_0) - \lambda x$  subject to  $x \in [0, 1]$ . After some basic manipulation, the problem of the firm can be posed as

$$\inf_{x \in [0, 1]} \mathbb{E}_Q[(-x)\xi + x^2]$$

where  $\xi = \frac{\alpha - a_1}{\beta + a_2}$ .

The most interesting feature of this example is that, unlike in the aforementioned newsvendor problems, the Lipschitz constant of the objective function  $f(x, \xi) := (-x)\xi + x^2$  with respect to  $\xi$  is dependent on the decision variable  $\mathbf{x}$ .

We consider that  $\xi$  follows a (true) probability distribution given by 10 000 points sampled from a mixture of three Gaussian distributions with variances all equal to 0.3 and means  $\mu_1 = 0, \mu_2 = 1.2$  and  $\mu_3 = 2.5$ . The weights of the mixture are  $\omega_1 = 0.5, \omega_2 = 0.2$  and  $\omega_3 = 0.3$ . Furthermore, the mixture has been truncated over the interval  $[-1.8, 3]$ . Figure 3.3a plots the kernel estimate of the data-generating distribution.

As in the previous Newsvendor problem experiments, we have divided the support  $[-1.8, 3]$  into four partitions, using the procedure described at the beginning of Section 3.2. However, in a different way to what we did in the newsvendor examples, here we select parameters  $\varepsilon$  and  $\rho$  following a procedure that solely relies on the available data, similarly to what is done in [99].

Essentially, given a desired confidence level  $(1 - \beta)$  for the finite-sample guarantee (set to 0.85 in our numerical experiments), we need to *estimate*, using the data sample available only, the parameters  $\varepsilon$  and  $\rho$  that deliver, at least, this confidence level while yielding the best out-of-sample performance. To this end, we use bootstrapping. The estimator of those parameters is denoted as  $\text{param}_N^m(\beta)$ , underlining that the number and type of parameters to be estimated depend on the method  $m$ . The estimation procedure is carried out as follows for each sample of size  $N$  (in the experiments we have considered 300 independent data samples for each size  $N$ ):

1. We construct  $kboot$  resamples of size  $N$  (with replacement), each playing the role of a different training dataset. Moreover, take those data points that have not been resampled to form a validation dataset (one per resample of size  $N$ ). In our experiments below, we have considered  $kboot = 50$ .
2. For each resample  $k = 1, \dots, kboot$  and each candidate value for  $\text{param}$ , get a DRO solution from method  $j$  with parameter (or pair of parameters)  $\text{param}$  on

the  $k$ -th resample. The resulting optimal decision is denoted as  $\hat{x}_N^{j,k}(param)$  and its associated objective value as  $\hat{J}_N^{j,k}(param)$ . Subsequently, we compute the out-of-sample performance  $J(\hat{x}_N^{j,k}(param))$  of the data-driven solution  $\hat{x}_N^{j,k}(param)$  over the  $k$ -th validation dataset.

3. From among the candidate values for  $param$  such that  $\hat{J}_N^{j,k}(param)$  exceeds the value  $J(\hat{x}_N^{j,k}(param))$  in at least  $(1 - \beta) \times kboot$  different resamples, take the one with the lowest  $\frac{\sum_{k=1}^{kboot} J(\hat{x}_N^{j,k}(param))}{kboot}$  (that is, with the highest out-of-sample performance averaged over the  $kboot$  resamples).
4. Finally, compute the solution given by method  $j$  with parameter  $param_N^{\beta,j}$ ,  $\hat{x}_N^j := \hat{x}_N^j(param_N^{\beta,j})$  and the respective certificate  $\hat{J}_N^j := \hat{J}_N^j(param_N^{\beta,j})$ .

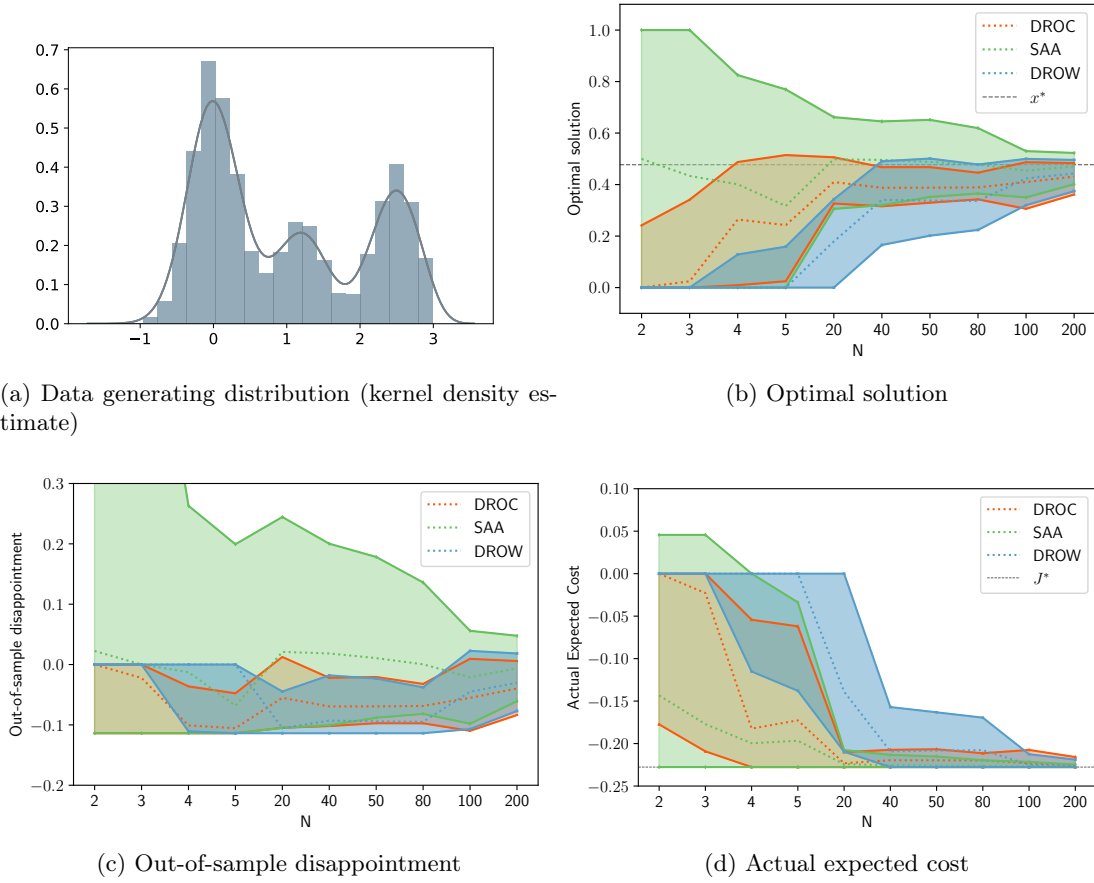


Figure 3.3: Strategic firm problem: (Approximate) true data-generating distribution, optimal solution and performance metrics

As for the newsvendor examples, Figures 3.3b, 3.3c, and 3.3d show, for various sample sizes, the box plots pertaining to the optimal decision, the out-of-sample disappointment and the actual expected cost associated with each of the considered data-driven

approaches, in that order. Once again, the results conveyed by these figures confirm our previous conclusions: Our approach DROC is able to leverage a-priori knowledge of the order among some partition probabilities to deliver solutions that perform significantly better out of sample for the same level of confidence. Furthermore, we see that the decision computed by the proposed method DROC converges to the true optimal solution (with complete information) faster than the solutions provided by the other methods.

### 3.3 Summary

In this chapter, we have presented a novel framework for data-driven distributionally robust optimization (DRO) based on optimal transport theory in combination with order cone constraints to leverage *a-priori* information on the true data-generating distribution. More specifically, motivated by the reported over-conservativeness of the traditional DRO approach based on the Wasserstein metric, we have formulated an ambiguity set able to incorporate information about the order among the probabilities that the true distribution of the problem's uncertain parameters assigns to the events within a partition of its support set. Our approach can accommodate a wide range of shape information (such as that related to monotonicity or multi-modality) in a practical and intuitive way. Moreover, we have shown that, under mild assumptions, the resulting DRO problem can be, in fact, reformulated as a finite convex problem where the a-priori information (expressed through the order cone constraints) are cast as linear constraints as opposed to the more computationally challenging formulations that exist in the literature. Furthermore, our approach is supported by theoretical performance guarantees and is capable of turning the provided information into solutions with increased reliability and improved performance, as illustrated by the numerical experiments we have prepared based on the well-known newsvendor problem and the problem of a strategic firm competing *à la Cournot* in a market for a homogeneous product.

## Chapter 4

# Conditional stochastic programs: A distributionally robust solution approach based on probability trimmings

### Contents

|            |  |           |
|------------|--|-----------|
| <b>4.1</b> | <b>Methodology and theoretical foundations . . . . .</b>   | <b>43</b> |
| 4.1.1      | Preliminaries and motivation . . . . .   | 43        |
| 4.1.2      | The Partial Mass Transportation Problem and Trimmings . .  | 45        |
| 4.1.3      | Tractable reformulation of the partial mass transportation problem . . . . .   | 50        |
| 4.1.4      | Finite sample guarantee and asymptotic consistency . . . . .   | 51        |
|            | Case $\mathbb{Q}(\tilde{\Xi}) = \alpha > 0$ . Applications in data-driven decision making under contaminated samples . . . . . | 52        |
|            | The case of unknown $\mathbb{Q}(\tilde{\Xi}) = \alpha > 0$ . . . . .   | 55        |
|            | The case $\mathbb{Q} \ll \lambda^d$ and $\mathbb{Q}(\tilde{\Xi}) = \alpha = 0$ . . . . .                                       | 56        |
| <b>4.2</b> | <b>Application I. Newsvendor problem . . . . .</b>   | <b>60</b> |
| <b>4.3</b> | <b>Application II. Portfolio allocation problem . . . . .</b>  | <b>65</b> |
| 4.3.1      | Case $\mathbb{Q}(\tilde{\Xi}) = \alpha = 0$ . . . . .  | 65        |
| 4.3.2      | Case $\mathbb{Q}(\tilde{\Xi}) = \alpha > 0$ . . . . .  | 70        |
| <b>4.4</b> | <b>Application III. Optimal Power Flow problem . . . . .</b>   | <b>74</b> |
| 4.4.1      | Introduction . . . . .   | 74        |
| 4.4.2      | DC-OPF under uncertainty: Mathematical Formulation . . .   | 76        |
|            | Variables and constraints . . . . .  | 77        |

|   |           |
|---|-----------|
| Dealing with uncertainty in the DC-OPF problem: Joint chance constraints, Distributionally Robust Optimization and contextual information . . . . . | 78        |
| 4.4.3 A tractable and conservative <b>CVaR</b> -based approximation of the distributionally robust joint chance constraints . . . . .               | 80        |
| 4.4.4 An exact tractable reformulation of the worst-case expected cost  | 81        |
| 4.4.5 Numerical results . . . . .   | 83        |
| 4.4.6 Evaluation of the out-of-sample performance via re-optimization   | 84        |
| 4.4.7 A 118-bus case study . . . . .  | 85        |
| Medium wind penetration case . . . . .  | 86        |
| High wind penetration case . . . . .  | 89        |
| <b>4.5 Summary . . . . .</b>  | <b>93</b> |

In this chapter, we propose a general framework for data-driven distributionally robust optimization with conditional information that relies on two related tools, namely, the *optimal mass transport theory* and the concept of *trimming of a probability measure*. We first introduce some preliminaries that help motivate our proposal and then lay out the theoretical foundations that support it, which can also be found in our work [53]. Finally, we discuss some computational experiments and applications of the proposed framework. For ease of reading, all the proofs of the theoretical results that are presented next have been moved to the appendices.

## 4.1 Methodology and theoretical foundations

This section develops the theoretical basis that underpins our proposal. Before that, though, we begin with some preliminaries that will serve us to build and motivate the distributionally robust optimization approach we propose to address conditional stochastic programs.

### 4.1.1 Preliminaries and motivation

We start this section by providing a generic formulation of a *conditional stochastic program*. For this purpose, let  $\mathbf{x} \in X \subseteq \mathbb{R}^{d_{\mathbf{x}}}$  be the decision variable vector and  $\mathbf{y}$ , with support set  $\Xi_{\mathbf{y}} \subseteq \mathbb{R}^{d_{\mathbf{y}}}$ , the random vector that models the uncertainty affecting the value of the decision. Let  $\mathbf{z}$ , with support set  $\Xi_{\mathbf{z}} \subseteq \mathbb{R}^{d_{\mathbf{z}}}$ , be the (random) feature vector and denote the objective function to be minimized as  $f(\mathbf{x}, \boldsymbol{\xi})$ , where  $\boldsymbol{\xi} := (\mathbf{z}, \mathbf{y})$ .

Given a new piece of information in the form of the event  $\boldsymbol{\xi} \in \tilde{\Xi}$ , the decision maker seeks to compute the optimal decision that minimizes the (true) conditional expected cost:

$$J^* := \inf_{\mathbf{x} \in X} \mathbb{E}_{\mathbb{Q}} [f(\mathbf{x}, \boldsymbol{\xi}) \mid \boldsymbol{\xi} \in \tilde{\Xi}] = \inf_{\mathbf{x} \in X} \mathbb{E}_{\mathbb{Q}_{\tilde{\Xi}}} [f(\mathbf{x}, \boldsymbol{\xi})] \quad (4.1)$$

where  $\mathbb{Q}$  is the true joint distribution of  $\boldsymbol{\xi} := (\mathbf{z}, \mathbf{y})$  with support set  $\Xi \subseteq \mathbb{R}^{d_z + d_y}$  and  $\mathbb{Q}_{\Xi}$  is the associated true distribution of  $\boldsymbol{\xi}$  *conditional on*  $\boldsymbol{\xi} \in \Xi$ . Hence, we implicitly assume that  $\mathbb{Q}_{\Xi}$  is a regular conditional distribution and that the conditional expectation (4.1) is well defined.

An example of  $\tilde{\Xi}$  would be  $\tilde{\Xi} := \{\boldsymbol{\xi} = (\mathbf{z}, \mathbf{y}) \in \Xi : \mathbf{z} \in \mathcal{Z}\}$ , with  $\mathcal{Z} \subseteq \Xi_{\mathbf{z}}$  being an uncertainty set built from the information on the features. We note that this definition includes the case in which  $\mathcal{Z}$  reduces to a singleton  $\mathbf{z}^*$  representing a particular realization of the features.

Unfortunately, when it comes to solving the *conditional stochastic program* (4.1), neither the true distribution  $\mathbb{Q}$  nor —even less so— the conditional one  $\mathbb{Q}_{\Xi}$  are generally known to the decision maker. Actually, the decision maker typically counts only on a data sample consisting of  $N$  observations  $\hat{\boldsymbol{\xi}}_i := (\hat{\mathbf{z}}_i, \hat{\mathbf{y}}_i)$  for  $i = 1, \dots, N$ , which we assume are i.i.d. Therefore, the solution to problem (4.1) *per se* is, in practice, out of reach and the best the decision maker can do is to approximate the solution to (4.1) with some (probabilistic) performance guarantees. Within this context, *Distributionally Robust Optimization* (DRO) emerges as a powerful modeling framework to achieve that goal. In brief, the DRO approach aims to find a decision  $\mathbf{x} \in X$  that is *robust* against all *conditional* probability distributions that are somehow *plausible* given the information at the decision maker's disposal. This is mathematically stated as follows:

$$\inf_{\mathbf{x} \in X} \sup_{Q_{\Xi} \in \hat{\mathcal{U}}_N} \mathbb{E}_{Q_{\Xi}} [f(\mathbf{x}, \boldsymbol{\xi})] \quad (4.2)$$

where  $\hat{\mathcal{U}}_N$  is a so-called *ambiguity set* that contains all those plausible conditional distributions. This ambiguity set must be built from the available information on  $\boldsymbol{\xi}$ , which, in our case, comprises the  $N$  observations  $\{\hat{\boldsymbol{\xi}}_i\}_{i=1}^N$ . The subscript  $N$  in  $\hat{\mathcal{U}}_N$  is intended to underline this issue. Furthermore, the condition  $Q_{\Xi}(\tilde{\Xi}) = 1$  for all  $Q_{\Xi} \in \hat{\mathcal{U}}_N$  is implicit in the construction of that set. In our setup, however, problem (4.2) poses a major challenge, which has to do with the fact that the observations  $\{\hat{\boldsymbol{\xi}}_i\}_{i=1}^N$  pertain to the true *joint* distribution  $\mathbb{Q}$ , and *not* to the conditional one  $\mathbb{Q}_{\Xi}$ . Consequently, we need to build an ambiguity set  $\hat{\mathcal{U}}_N$  for the plausible *conditional* distributions from the limited joint information on  $\mathbb{Q}$  provided by the data  $\{\hat{\boldsymbol{\xi}}_i\}_{i=1}^N$ .

At this point, we should note that there are several approaches in the technical literature to handle the conditional stochastic optimization problem (4.1) for the particular case in which  $\tilde{\Xi}$  is defined as  $\tilde{\Xi} := \{\boldsymbol{\xi} = (\mathbf{z}, \mathbf{y}) \in \Xi : \mathbf{z} = \mathbf{z}^*\}$ . For example, the authors of [17] approximate (4.1) by the following conditional estimate

$$\inf_{\mathbf{x} \in X} \sum_{i=1}^N w_N^i(\mathbf{z}^*) f(\mathbf{x}, (\mathbf{z}^*, \hat{\mathbf{y}}_i)) \quad (4.3)$$

where  $w_N^i(\mathbf{z}^*)$  is a weight function that can be given by various non-parametric machine learning methods such as  $K$ -nearest neighbors, kernel regression, CART, and random forests. Formulation (4.3) can be naturally interpreted as a (conditional) Sample-Average-Approximation (SAA) of problem (4.1).

The authors in [18] extend the work in [17] to accommodate the setting in which the outcome of the uncertainty  $\mathbf{y}$  may be contingent on the taken decision  $\mathbf{x}$ . For this purpose, they work with an enriched data set comprising observations of the uncertainty  $\mathbf{y}$ , the decision  $\mathbf{x}$  and the covariates  $\mathbf{z}$ , and allow the weights in (4.3) to depend on  $\mathbf{x}$  too. Besides, they add terms to the objective function of (4.3) to penalize estimates of its variance and bias. The case in which the weight function (4.3) is given by the Nadaraya-Watson (NW) kernel regression estimator is considered in [74, 112]. In [112], in addition, they leverage techniques from moderate deviations theory to design a regularization scheme that reduces the optimistic bias of the NW approximation and to provide insight into its out-of-sample performance. The work in [22] focuses on conditional estimators (4.3) where the weights are provided by the NW or KNN method. They use DRO, based on the relative entropy distance for discrete distributions to get decisions from (4.3) that perform well on a large portion of resamples *bootstrapped* from the empirical distribution of the available data set.

Finally, the authors in [19] provide a robustified version of the conditional estimator (4.3), which takes the following form

$$\inf_{\mathbf{x} \in X} \sum_{i=1}^N w_N^i(\mathbf{z}^*) \sup_{\mathbf{y} \in \mathcal{U}_N^i} [f(\mathbf{x}, (\mathbf{z}^*, \mathbf{y}))] \quad (4.4)$$

where  $\mathcal{U}_N^i := \{\mathbf{y} \in \Xi_{\mathbf{y}} : \|\mathbf{y} - \hat{\mathbf{y}}_i\|_p \leq \varepsilon_N\}$ . This problem can be seen as a robust SAA method capable of exploiting side information and has also been used in [20, 21].

In our case, however, we follow a different path to address the conditional stochastic optimization problem (4.1) by way of (4.2). More precisely, we leverage the notion of *trimmings of a distribution* and the related theory of *partial mass transportation*.

#### 4.1.2 The Partial Mass Transportation Problem and Trimmings

This section introduces some concepts about trimmings and the partial mass transportation problem that help us construct the ambiguity set  $\hat{\mathcal{U}}_N$  in (4.2) from the sample data  $\{\hat{\xi}_i\}_{i=1}^N$ . For simplicity, we restrict ourselves to probability measures defined in  $\mathbb{R}^d$ .

If  $\mathbb{Q}(\Xi) = \alpha > 0$  (our analysis, though, will also cover the case  $\alpha = 0$  later in Section 4.1.4), problem (4.1) can be recast as

$$J^* := \inf_{\mathbf{x} \in X} \frac{1}{\alpha} \mathbb{E}_{\mathbb{Q}} [f(\mathbf{x}, \xi) \mathbb{I}_{\Xi}(\xi)] \quad (4.5)$$



which only requires that  $\mathbb{E}_{\mathbb{Q}} [|f(\mathbf{x}, \boldsymbol{\xi}) \mathbb{I}_{\Xi}(\boldsymbol{\xi})|] < \infty$  for all  $\mathbf{x} \in X$  (see [67, Eq. 6.2]).

Now we introduce the notion of a *trimming* of a distribution, which is at the core of our proposed DRO framework.

**Definition 4.1 ((1- $\alpha$ )-trimmings, Definition 1.1 from [10]).** Given  $0 \leq \alpha \leq 1$  and probability measures  $P, Q \in \mathbb{R}^d$ , we say that  $Q$  is an  $(1 - \alpha)$ -trimming of  $P$  if  $Q$  is absolutely continuous with respect to  $P$ , and the Radon-Nikodym derivative satisfies  $\frac{dQ}{dP} \leq \frac{1}{\alpha}$ . The set of all  $(1 - \alpha)$ -trimmings (or trimming set of level  $1 - \alpha$ ) of  $P$  will be denoted by  $\mathcal{R}_{1-\alpha}(P)$ .

**Remark 4.1.** As extreme cases, we have that for  $\alpha = 1$ ,  $\mathcal{R}_0(P)$  is just  $P$ , while, for  $\alpha = 0$ ,  $\mathcal{R}_1(P)$  is the set of all probability measures absolutely continuous with respect to  $P$ . Given a probability  $P$  on  $\mathbb{R}^d$ , if  $\alpha_1 \leq \alpha_2$ , then  $\mathcal{R}_{1-\alpha_2}(P) \subset \mathcal{R}_{1-\alpha_1}(P)$ . Especially useful is the fact that a trimming set is a convex set, which is, besides, compact under the topology of weak convergence. We refer the reader to [3, Proposition 2.7] for other interesting properties about the set  $\mathcal{R}_{1-\alpha}(P)$ .

For ease of understanding, we provide below an example of a  $(1 - \alpha)$ -empirical trimmings set.

**Example 4.1.** Consider the empirical joint measure  $\hat{\mathbb{Q}}_N := \sum_{i=1}^3 \delta_{\hat{\boldsymbol{\xi}}_i} = \frac{1}{3}(\delta_{(1,0)} + \delta_{(0,5)} + \delta_{(2,3)})$  ( $N = 3$ ). If  $\alpha = 0.5$ , then  $\frac{1}{N\alpha} = \frac{1}{3 \cdot 0.5} = \frac{2}{3}$ . Therefore, the 0.5-trimmings set of  $\hat{\mathbb{Q}}_N$  (see Figure 4.1) is given by

$$\mathcal{R}_{0.5}(\hat{\mathbb{Q}}_N) := \left\{ \sum_{i=1}^3 b_i \delta_{\hat{\boldsymbol{\xi}}_i} : 0 \leq b_i \leq \frac{2}{3}, \forall i = 1, \dots, 3; \sum_{i=1}^3 b_i = 1 \right\}$$

The following statements hold thus true:

$$\begin{aligned} \hat{\mathbb{Q}}_N &= \frac{1}{3}(\delta_{(1,0)} + \delta_{(0,5)} + \delta_{(2,3)}) \in \mathcal{R}_{0.5}(\hat{\mathbb{Q}}_N) \\ \mathcal{Q} &= \frac{2}{3}\delta_{(1,0)} + \frac{1}{3}\delta_{(0,5)} \in \mathcal{R}_{0.5}(\hat{\mathbb{Q}}_N) \\ \mathcal{P} &= \delta_{(1,0)} \notin \mathcal{R}_{0.5}(\hat{\mathbb{Q}}_N) \text{ (the trimming must retain one point and a half at least)} \\ \mathcal{S} &= \frac{2}{3}\delta_{(1,0)} + \frac{1}{6}\delta_{(0,5)} + \frac{1}{6}\delta_{(2,3)} \in \mathcal{R}_{0.5}(\hat{\mathbb{Q}}_N) \\ \mathcal{V} &= \frac{3}{4}\delta_{(1,0)} + \frac{1}{12}\delta_{(0,5)} + \frac{2}{12}\delta_{(2,3)} \notin \mathcal{R}_{0.5}(\hat{\mathbb{Q}}_N) \text{ (because } b_1 > 2/3) \end{aligned}$$

Consider now the following minimization problem:

$$\inf_{Q \in \mathcal{R}_{1-\alpha}(P)} D(Q, R) \quad (4.6)$$

where  $D$  is a probability metric.

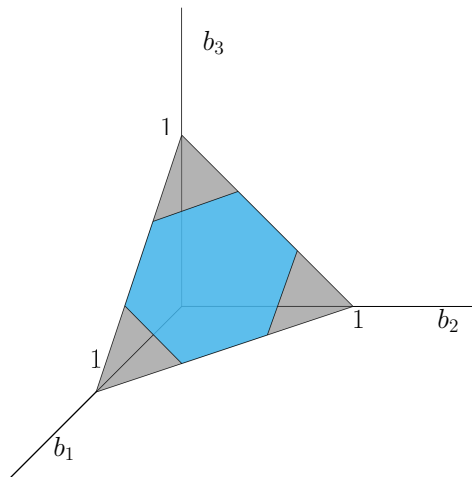


Figure 4.1: Probability simplex (in blue) corresponding to the trimming set  $\mathcal{R}_{0.5}(\hat{\mathbb{Q}}_N)$

Problem (4.6) is known as the  $(D, 1 - \alpha)$ -*partial (or incomplete) mass problem* [10]. While there is a variety of probability metrics we could choose from to play the role of  $D$  in (4.6), here we work with the space  $\mathcal{P}_p(\mathbb{R}^d)$  of probability distributions supported on  $\mathbb{R}^d$  with finite  $p$ -th moment and restrict ourselves to the  $p$ -Wasserstein metric,  $\mathcal{W}_p$ , for its tractability and theoretical advantages. In such a case (i.e., when  $D = \mathcal{W}_p$ ), problem (4.6) is referred to as a partial mass *transportation* problem and interpolates between the classical optimal mass transportation problem (when  $\alpha = 1$ ) and the random quantization problem (when  $\alpha = 0$ ).

Intuitively, the partial optimal transport problem goes as follows. We have an excess of offer of a certain quantity of mass at origin (supply) and a mass that needs to be satisfied at destination (demand), so that it is not necessary to serve all the mass (demand =  $\alpha \times$  supply). In other words, some  $(1 - \alpha)$ -fraction of the mass at origin can be left non-served. The goal is to perform this task at the cheapest transportation cost. If we represent the demand at destination by a target probability distribution  $R$ , we can model the supply at origin as  $\frac{P}{\alpha}$ , where  $P$  is another probability distribution and the mass required at destination is  $\alpha$  times the mass at origin. This way, a *partial optimal transportation plan* is a probability measure  $\Pi$  on  $\mathbb{R}^d \times \mathbb{R}^d$  with first marginal in  $\mathcal{R}_{1-\alpha}(P)$  and with second marginal equal to  $R$ , which solves the following cost minimization problem:

$$\mathcal{W}_p(\mathcal{R}_{1-\alpha}(P), R) := \min_{Q \in \mathcal{R}_{1-\alpha}(P)} \mathcal{W}_p(Q, R)$$

The following lemma allows us to characterize the connection between the joint distribution  $\mathbb{Q}$  and the conditional distribution  $\mathbb{Q}_{\Xi}$  in problem (4.1) above in terms of the partial mass problem.

**Lemma 4.1.** *Let  $Q$  be a probability on  $\mathbb{R}^d$  such that  $Q(\tilde{\Xi}) = \alpha > 0$  and let  $Q_{\Xi}$*

be the  $Q$ -conditional probability distribution given the event  $\xi \in \tilde{\Xi}$ . Also, assume that for a given probability metric  $D$ ,  $\mathcal{R}_{1-\alpha}(Q)$  is closed for  $D$  over an appropriate set of probability distributions. Then,  $Q_{\tilde{\Xi}}$  is the unique distribution that satisfies  $Q_{\tilde{\Xi}}(\tilde{\Xi}) = 1$  and  $D(\mathcal{R}_{1-\alpha}(Q), Q_{\tilde{\Xi}}) = 0$ .

By way of Lemma (4.1), we can reformulate Problem (4.1) as follows:

$$\inf_{\mathbf{x} \in X} \sup_{Q_{\tilde{\Xi}}} \mathbb{E}_{Q_{\tilde{\Xi}}} [f(\mathbf{x}, \xi)] \quad (4.7a)$$

$$\text{s.t. } \mathcal{W}_p^p(\mathcal{R}_{1-\alpha}(\mathbb{Q}), Q_{\tilde{\Xi}}) = 0 \quad (4.7b)$$

$$Q_{\tilde{\Xi}}(\tilde{\Xi}) = 1 \quad (4.7c)$$

which now presents a form which is much more suited to our purpose, that is, to get to the DRO-type of problem (4.2) we propose. The change, nonetheless, has been essentially cosmetic, because problem (4.7) still relies on the true *joint* distribution  $\mathbb{Q}$  and therefore, is of no use in practice as it stands right now. To make it practical, we need to rewrite it not in terms of the unknown  $\mathbb{Q}$ , but in terms of the information available to the decision maker, i.e., the sample data  $\{\hat{\xi}_i\}_{i=1}^N$ . For that purpose, it seems sensible and natural to replace  $\mathbb{Q}$  in (4.7b) with its best approximation taken directly from the data, namely, the empirical measure of the sample,  $\hat{\mathbb{Q}}_N$ . Logically, to accommodate the approximation, we will need to introduce a *budget*  $\tilde{\rho}$  in equation (4.7b), that is,

$$(P) \quad \inf_{\mathbf{x} \in X} \sup_{Q_{\tilde{\Xi}}} \mathbb{E}_{Q_{\tilde{\Xi}}} [f(\mathbf{x}, \xi)] \quad (4.8a)$$

$$\text{s.t. } \mathcal{W}_p^p(\mathcal{R}_{1-\alpha}(\hat{\mathbb{Q}}_N), Q_{\tilde{\Xi}}) \leq \tilde{\rho} \quad (4.8b)$$

$$Q_{\tilde{\Xi}}(\tilde{\Xi}) = 1 \quad (4.8c)$$

Hereinafter we will use  $\hat{\mathcal{U}}_N(\alpha, \tilde{\rho})$  to denote the ambiguity set defined by constraints (4.8b)–(4.8c). Under certain conditions, this uncertainty set enjoys nice topological properties, as we state in Proposition B.1.3 in Appendix B.

Now we define what we call the *minimum transportation budget*, which plays an important role in the selection of budget  $\tilde{\rho}$  in problem (P).

**Definition 4.2 (Minimum transportation budget).** Given  $\alpha > 0$  in problem (P), the minimum transportation budget, which we denote as  $\underline{\epsilon}_{N\alpha}$ , is the  $p$ -Wasserstein distance between the set  $\mathcal{P}_p(\tilde{\Xi})$  and the  $(1 - \alpha)$ -trimming of the empirical distribution  $\hat{\mathbb{Q}}_N$  that is the closest to that set, i.e.,  $\inf\{\mathcal{W}_p(P, Q) : P \in \mathcal{R}_{1-\alpha}(\hat{\mathbb{Q}}_N), Q \in \mathcal{P}_p(\tilde{\Xi})\}$ , which is given by

$$\underline{\epsilon}_{N\alpha} = \left( \frac{1}{N\alpha} \sum_{k=1}^{\lfloor N\alpha \rfloor} \text{dist}(\xi_{k:N}, \tilde{\Xi})^p + \left( 1 - \frac{\lfloor N\alpha \rfloor}{N\alpha} \right) \text{dist}(\xi_{\lfloor N\alpha \rfloor+1:N}, \tilde{\Xi})^p \right)^{\frac{1}{p}} \quad (4.9)$$

where  $\xi_{k:N}$  is the  $k$ -th nearest data point from the sample to set  $\tilde{\Xi}$  and  $\text{dist}(\xi_j, \tilde{\Xi}) := \inf_{\xi \in \tilde{\Xi}} \text{dist}(\xi_j, \xi) = \inf_{\xi \in \tilde{\Xi}} \|\xi_j - \xi\|$ . If  $\alpha = 0$ , then  $\epsilon_{N0} = \text{dist}(\xi_{1:N}, \tilde{\Xi})$ .

Importantly, the minimum transportation budget to the power of  $p$ , i.e.,  $\epsilon_{N\alpha}^p$ , is the minimum value of  $\tilde{\rho}$  in (P) for this problem to be feasible. Furthermore,  $\epsilon_{N\alpha}$  is random, because it depends on the available data sample, but realizes before the decision  $\mathbf{x}$  is to be made. It constitutes, therefore, input data to problem (P).

We note that, if the random vector  $\mathbf{y}$  takes values in a set that is independent of the feature vector  $\mathbf{z}$ , i.e., for all  $\mathbf{z}^* \in \Xi_{\mathbf{z}}$ ,  $\{\mathbf{y} \in \Xi_{\mathbf{y}} : \xi = (\mathbf{z}^*, \mathbf{y}) \in \Xi\} = \Xi_{\mathbf{y}}$ , then  $\text{dist}(\xi_j, \tilde{\Xi}) = \inf_{\xi \in \tilde{\Xi}} \|\xi_j - \xi\| = \inf_{\xi=(\mathbf{z}, \mathbf{y}) \in \tilde{\Xi}} \|\mathbf{z}_j - \mathbf{z}\|$ .

Furthermore, in what follows, we assume that  $\text{dist}(\xi_j, \tilde{\Xi})$  (interpreted as a random variable) conditional on  $\xi_j \notin \tilde{\Xi}$  has a continuous distribution function. This ensures that, in the case  $\mathbb{Q}(\tilde{\Xi}) = 0$ , which we study in Section 4.1.4, there will be exactly  $K$  nearest data points to  $\tilde{\Xi}$  with probability one.

Next we present an interesting result, which deals with the inner supremum of problem (P) and adds more meaning to this problem by linking it to an alternative formulation more in the style of the Wasserstein data-driven DRO approach proposed in [99], where, however, no side information is taken into account. In fact, the distributionally robust approach to conditional stochastic optimization that is proposed in [107] is based on this alternative formulation (see Proposition A.4 in that work)<sup>1</sup>.

**Proposition 4.1.** *Given  $N \geq 1$ ,  $\mathbb{Q}(\tilde{\Xi}) = \alpha > 0$ , and any positive value of  $\tilde{\rho}$ , problem (SP2) is a relaxation of (SP1), where (SP1) and (SP2) are given by*

$$(\text{SP1}) \quad \begin{cases} \sup_Q & \mathbb{E}_Q [f(\mathbf{x}, \xi) \mid \xi \in \tilde{\Xi}] \\ \text{s.t.} & \mathcal{W}_p^p(Q, \hat{\mathbb{Q}}_N) \leq \tilde{\rho} \cdot \alpha \\ & Q(\tilde{\Xi}) = \alpha \end{cases}, \quad (\text{SP2}) \quad \begin{cases} \sup_{Q_{\tilde{\Xi}}} & \mathbb{E}_{Q_{\tilde{\Xi}}} [f(\mathbf{x}, \xi)] \\ \text{s.t.} & \mathcal{W}_p^p(\mathcal{R}_{1-\alpha}(\hat{\mathbb{Q}}_N), Q_{\tilde{\Xi}}) \leq \tilde{\rho} \\ & Q_{\tilde{\Xi}}(\tilde{\Xi}) = 1 \end{cases}$$

and where by “relaxation” it is meant that any solution  $Q$  feasible in (SP1) can be mapped into a solution  $Q_{\tilde{\Xi}}$  feasible in (SP2) with the same objective function value.

Moreover, if  $\hat{\mathbb{Q}}_N(\tilde{\Xi}) = 0$  or  $\alpha = 1$ , then (SP1) and (SP2) are equivalent.

Among other things, Proposition 4.1 reveals that parameter  $\tilde{\rho}$  in problem (SP2), and hence in problem (P), can be understood as a cost budget *per unit of transported mass*. Likewise, parameter  $\alpha$  can be interpreted as the minimum amount of mass (in per unit) of the empirical distribution  $\hat{\mathbb{Q}}_N$  that must be transported to the support  $\tilde{\Xi}$ . This interpretation of parameters  $\tilde{\rho}$  and  $\alpha$  will be useful to follow the rationale behind the DRO solution approaches that we develop later on.

On the other hand, despite the connection between problems (SP1) and (SP2) that Proposition 4.1 unveils, the latter is qualitatively more amenable to further generaliza-

<sup>1</sup>Proposition 4.1 in our work published in [53] predates the release of preprint [107].

tion and analysis. Examples of this are given by the relevant cases  $\alpha = 0$ , for which problem (SP1) is *ill-posed*, while problem (SP2) is not, and  $\alpha$  unknown, for which the use of trimming sets in (SP2) allows for a more straightforward treatment. We will deal with both cases in Section 4.1.4. Before that, we provide an implementable reformulation of the proposed DRO problem (P).

### 4.1.3 Tractable reformulation of the partial mass transportation problem

In this section, we put the proposed DRO problem (P) in a form more suited to tackle its computational implementation and solution. For this purpose, we first need to introduce a technical result whereby we characterize the trimming sets of an empirical probability measure.

**Lemma 4.2.** *Consider the sample data  $\{\hat{\xi}_i\}_{i=1}^N$  and their associated empirical measure  $\hat{\mathbb{Q}}_N = \frac{1}{N} \sum_{i=1}^N \delta_{\hat{\xi}_i}$ . If  $\alpha > 0$ , the set of all  $(1 - \alpha)$ -trimmings of  $\hat{\mathbb{Q}}_N$  is given by all probability distributions in the form  $\sum_{i=1}^N b_i \delta_{\hat{\xi}_i}$  such that  $0 \leq b_i \leq \frac{1}{N\alpha}$ ,  $\forall i = 1, \dots, N$ , and  $\sum_{i=1}^N b_i = 1$ . Furthermore, if  $\alpha = 0$ , the set  $\mathcal{R}_{1-\alpha}(\hat{\mathbb{Q}}_N)$  of  $(1 - \alpha)$ -trimmings of  $\hat{\mathbb{Q}}_N$  becomes  $\mathcal{R}_1(\hat{\mathbb{Q}}_N) = \{\sum_{i=1}^N b_i \delta_{\hat{\xi}_i} \text{ such that } b_i \geq 0, \forall i = 1, \dots, N, \text{ and } \sum_{i=1}^N b_i = 1\}$ .*

In short, Lemma 4.2 tells us that trimming a data sample of size  $N$  with level  $1 - \alpha$  involves reweighting the empirical distribution of such data by giving a new weight less than or equal to  $\frac{1}{N\alpha}$  to each data point.

Therefore, Applying Lemma 4.2, we can recast constraint  $\mathcal{W}_p^p(\mathcal{R}_{1-\alpha}(\hat{\mathbb{Q}}_N), Q_{\Xi}) \leq \tilde{\rho}$  in problem (P) as

$$\begin{aligned} \min_{b_i, \forall i \leq N} \mathcal{W}_p \left( \sum_{i=1}^N b_i \delta_{\hat{\xi}_i}, Q_{\Xi} \right) &\leq \tilde{\rho}^{1/p} \\ \text{s.t. } 0 &\leq b_i \leq \frac{1}{N\alpha}, \quad \forall i \leq N \\ \sum_{i=1}^N b_i &= 1 \end{aligned}$$

We are now ready to introduce the main result of this section.

**Theorem 4.1 (Reformulation based on strong duality).** *For  $\alpha > 0$  and any value of  $\tilde{\rho} \geq \underline{\epsilon}_{N\alpha}^p$ , subproblem (SP2) is equivalent to the following one:*

$$\begin{aligned} (\text{SP2}') \quad &\inf_{\lambda \geq 0; \bar{\mu}_i, \forall i \leq N; \theta \in \mathbb{R}} \lambda \tilde{\rho} + \theta + \frac{1}{N\alpha} \sum_{i=1}^N \bar{\mu}_i \\ \text{s.t. } &\bar{\mu}_i + \theta \geq \sup_{(\mathbf{z}, \mathbf{y}) \in \tilde{\Xi}} (f(\mathbf{x}, (\mathbf{z}, \mathbf{y})) - \lambda \|(\mathbf{z}, \mathbf{y}) - (\hat{\mathbf{z}}_i, \hat{\mathbf{y}}_i)\|^p), \quad \forall i \leq N \end{aligned}$$

$$\bar{\mu}_i \geq 0, \quad \forall i \leq N$$

Surely the most important takeaway message of Theorem 4.1 is that problem (P) is *as tractable as* the standard Wasserstein-metric-based DRO formulation proposed in [87] and [99]. In these two papers, conditions under which the inner supremum in (SP2') can be recast in a more tractable form are provided. As an example, in Appendix B.1.4 we provide a more refined reformulation of (SP2'), whereby the problems we solve in this chapter can be directly handled.

In the following section, we show that problem (P) works, under certain conditions, as a statistically meaningful surrogate decision-making model for the target conditional stochastic program (4.1).

#### 4.1.4 Finite sample guarantee and asymptotic consistency

Next we argue that the worst-case optimal expected cost provided by problem (P) for a fixed sample size  $N$  and a suitable choice of parameters  $(\alpha, \tilde{\rho})$  (dependent on  $N$ ) leads to an upper confidence bound on the out-of-sample performance attained by the optimizers of (P) (*finite sample guarantee*) and that those optimizers almost surely converge to an optimizer of the true optimal expected cost as  $N$  grows to infinity (*asymptotic consistency*). Recall that we say that a data-driven method built to address problem (4.1) enjoys a *finite sample guarantee*, if it produces pairs  $(\hat{\mathbf{x}}_N, \hat{J}_N)$  satisfying a relation in the form

$$\mathbb{Q}^N \left[ \mathbb{E}_{\mathbb{Q}}[f(\hat{\mathbf{x}}_N, \boldsymbol{\xi}) \mid \boldsymbol{\xi} \in \tilde{\Xi}] \leq \hat{J}_N \right] \geq 1 - \beta \quad (4.10)$$

Our analysis relies on the lemma below, which immediately follows from setting  $P_1 := \hat{\mathbb{Q}}_N, Q := \mathbb{Q}_{\tilde{\Xi}}, P_2 := \mathbb{Q}$  in Lemma 3.13 on probability trimmings in [1].

**Lemma 4.3.** *Assume that  $\mathbb{Q}_{\tilde{\Xi}}, \mathbb{Q} \in \mathcal{P}_p(\mathbb{R}^d)$ , and take  $p \geq 1$ , then*

$$\mathcal{W}_p(\mathcal{R}_{1-\alpha}(\hat{\mathbb{Q}}_N), \mathbb{Q}_{\tilde{\Xi}}) \leq \mathcal{W}_p(\mathcal{R}_{1-\alpha}(\mathbb{Q}), \mathbb{Q}_{\tilde{\Xi}}) + \frac{1}{\alpha^{1/p}} \mathcal{W}_p(\hat{\mathbb{Q}}_N, \mathbb{Q}) \quad (4.11)$$

We notice that the term  $\mathcal{W}_p(\mathcal{R}_{1-\alpha}(\mathbb{Q}), \mathbb{Q}_{\tilde{\Xi}})$  in (4.11) is not random and depends exclusively on the true distributions  $\mathbb{Q}_{\tilde{\Xi}}, \mathbb{Q}$ , and the trimming level  $\alpha$ . It is, therefore, independent of the data sample (unlike the other two terms involved). Inequality (4.11) reveals an interesting trade-off. On the one hand, the distance  $\mathcal{W}_p(\mathcal{R}_{1-\alpha}(\mathbb{Q}), \mathbb{Q}_{\tilde{\Xi}})$  diminishes as  $\alpha$  decreases to zero, because the trimming set  $\mathcal{R}_{1-\alpha}(\mathbb{Q})$  grows in size. On the other, the term  $\frac{1}{\alpha^{1/p}} \mathcal{W}_p(\hat{\mathbb{Q}}_N, \mathbb{Q})$  becomes larger as  $\alpha$  approaches zero. As we will see later on, controlling this trade-off is key to endowing problem (P) with performance guarantees. To this end, we will make use of the assumption and proposition below.

**Assumption 4.1.** *The true joint probability distribution  $\mathbb{Q}$  is light-tailed, i.e., there exists a constant  $a > p \geq 1$  such that  $\mathbb{E}_{\mathbb{Q}}[\exp(\|\xi\|^a)] < \infty$ .*

**Proposition 4.2 (Concentration tail inequality).** *Suppose that Assumption 4.1 holds. Then, there are constants  $c, C > 0$  such that, for all  $\epsilon > 0, \alpha > 0$ , and  $N \geq 1$ , it holds*

$$\mathbb{Q}^N \left[ \mathcal{W}_p \left( \mathcal{R}_{1-\alpha}(\hat{\mathbb{Q}}_N), \mathbb{Q}_{\Xi} \right) \geq \mathcal{W}_p(\mathcal{R}_{1-\alpha}(\mathbb{Q}), \mathbb{Q}_{\Xi}) + \epsilon \right] \leq \beta_{p,\epsilon,\alpha}(N) \quad (4.12)$$

where

$$\beta_{p,\epsilon,\alpha}(N) = \mathbb{I}_{\{\epsilon \leq 1/\alpha^{1/p}\}} C \begin{cases} \exp(-cN \alpha^2 \epsilon^{2p}) & \text{if } p > d/2, \\ \exp(-cN (\alpha \epsilon^p / \log(2 + 1/\alpha \epsilon^p))^2) & \text{if } p = d/2, \\ \exp(-cN \alpha^{d/p} \epsilon^d) & \text{if } p \in [1, d/2), d > 2 \end{cases} \quad (4.13)$$

$$+ C \exp(-cN \alpha^{a/p} \epsilon^a) \mathbb{I}_{\{\epsilon > 1/\alpha^{1/p}\}}$$

with  $d = d_{\mathbf{z}} + d_{\mathbf{y}}$ .

Assuming  $p \neq d/2$ , if we equate  $\beta$  to  $\beta_{p,\epsilon,\alpha}(N)$  and solving for  $\epsilon$  we get:

$$\epsilon_{N,p,\alpha}(\beta) := \begin{cases} \left( \frac{\log(C\beta^{-1})}{cN} \right)^{1/2p} \frac{1}{\alpha^{1/p}} & \text{if } N \geq \frac{\log(C\beta^{-1})}{c}, \quad p > d/2, \\ \left( \frac{\log(C\beta^{-1})}{cN} \right)^{1/d} \frac{1}{\alpha^{1/p}} & \text{if } N \geq \frac{\log(C\beta^{-1})}{c}, \quad p \in [1, d/2), d > 2 \\ \left( \frac{\log(C\beta^{-1})}{cN} \right)^{1/a} \frac{1}{\alpha^{1/p}} & \text{if } N < \frac{\log(C\beta^{-1})}{c} \end{cases} \quad (4.14)$$

In what follows, we distinguish three general setups that may appear in the real-life use of Conditional Stochastic Optimization, namely, the case  $\mathbb{Q}(\tilde{\Xi}) = \alpha > 0$  with  $\alpha$  known, the case  $\mathbb{Q}(\tilde{\Xi}) = \alpha > 0$  with  $\alpha$  unknown, and the case  $\mathbb{Q} \ll \lambda^d$  with  $\mathbb{Q}(\tilde{\Xi}) = \alpha = 0$ .

**Case  $\mathbb{Q}(\tilde{\Xi}) = \alpha > 0$ . Applications in data-driven decision making under contaminated samples**

When  $\mathbb{Q}(\tilde{\Xi}) = \alpha > 0$  and *known*, we can solve the following DRO problem:

$$(P_{(\alpha, \tilde{\rho}_N)}) \inf_{\mathbf{x} \in X} \sup_{Q_{\Xi}} \mathbb{E}_{Q_{\Xi}}[f(\mathbf{x}, \xi)] \quad (4.15a)$$

$$\text{s.t. } \mathcal{W}_p(\mathcal{R}_{1-\alpha}(\hat{\mathbb{Q}}_N), Q_{\Xi}) \leq \tilde{\rho}_N \quad (4.15b)$$

$$Q_{\Xi}(\tilde{\Xi}) = 1 \quad (4.15c)$$

As we show below, problem  $P_{(\alpha, \tilde{\rho}_N)}$  enjoys a finite sample guarantee and produces solutions that are asymptotically consistent, i.e., that converge to the true solution

(under complete information) given by problem (4.1). This is somewhat hinted at by the connection between problems (SP1) and (SP2) highlighted in Proposition 4.1.

**Theorem 4.2 (Case  $\alpha > 0$ : Finite sample guarantee).** *Suppose that the assumptions of Proposition 4.2 hold and take  $p \neq d/2$ . Given  $N \geq 1$  and  $\alpha > 0$ , choose  $\beta \in (0, 1)$ , and determine  $\epsilon_{N,p,\alpha}(\beta)$  through (4.14). Then, for all  $\tilde{\rho}_N \geq \max(\epsilon_{N,p,\alpha}^p(\beta), \underline{\epsilon}_{N\alpha}^p)$ , where  $\underline{\epsilon}_{N\alpha}^p$  is the minimum transportation budget as in Definition 4.2, the pair  $(\hat{\mathbf{x}}_N, \hat{J}_N)$  that is solution to problem  $(P_{(\alpha, \tilde{\rho}_N)})$  enjoys the finite sample guarantee (4.10).*

We point out that, in the case  $\alpha > 0$ , data points may fall into the set  $\tilde{\Xi}$ . Logically, the contribution of these points to the minimum transportation budget  $\underline{\epsilon}_{N\alpha}^p$  is null and their order (the way their tie is broken) is irrelevant to our purpose.

Now we state that the solutions of the distributionally robust optimization problem  $(P_{(\alpha, \tilde{\rho}_N)})$  converge to the solution of the target conditional stochastic program (4.1) as  $N$  increases, for a careful choice of the budget  $\tilde{\rho}_N$ . This result is underpinned by the fact that, under that selection of  $\tilde{\rho}_N$ , any distribution in  $\hat{\mathcal{U}}_N(\alpha, \tilde{\rho}_N)$  converges to the true conditional distribution  $\mathbb{Q}_{\tilde{\Xi}}$ . This is formally stated in the following lemma.

**Lemma 4.4 (Case  $\alpha > 0$ : Convergence of conditional distributions).** *Suppose that the assumptions of Proposition 4.2 hold. Choose a sequence  $\beta_N \in (0, 1)$ ,  $N \in \mathbb{N}$ , such that  $\sum_{N=1}^{\infty} \beta_N < \infty$  and  $\lim_{N \rightarrow \infty} \epsilon_{N,p,\alpha}(\beta_N) \rightarrow 0$ . Then,*

$$\mathcal{W}_p(Q_{\tilde{\Xi}}^N, \mathbb{Q}_{\tilde{\Xi}}) \rightarrow 0 \text{ a.s.}$$

for any sequence  $Q_{\tilde{\Xi}}^N$ ,  $N \in \mathbb{N}$ , such that  $Q_{\tilde{\Xi}}^N \in \hat{\mathcal{U}}_N(\alpha, \tilde{\rho}_N)$  with  $\tilde{\rho}_N = \max(\epsilon_{N,p,\alpha}^p(\beta_N), \underline{\epsilon}_{N\alpha}^p)$ .

Once the convergence of  $Q_{\tilde{\Xi}}^N$  to the true conditional distribution  $\mathbb{Q}_{\tilde{\Xi}}$  in the  $p$ -Wasserstein metric has been established by the previous lemma, the following asymptotic consistency result, which is analogous to that of [99, Theorem 3.6], can also be derived.

**Theorem 4.3 (Asymptotic consistency).** *Consider that the conditions of Theorem 4.2 hold. Take a sequence  $\tilde{\rho}_N$  as in Lemma 4.4. Then, we have*

- (i) *If for any fixed value  $\mathbf{x} \in X$ ,  $f(\mathbf{x}, \boldsymbol{\xi})$  is continuous in  $\boldsymbol{\xi}$  and there is  $L \geq 0$  such that  $|f(\mathbf{x}, \boldsymbol{\xi})| \leq L(1 + \|\boldsymbol{\xi}\|^p)$  for all  $\mathbf{x} \in X$  and  $\boldsymbol{\xi} \in \tilde{\Xi}$ , then we have that  $\hat{J}_N \rightarrow J^*$  almost surely when  $N$  grows to infinity.*
- (ii) *If the assumptions in (i) are satisfied,  $f(\mathbf{x}, \boldsymbol{\xi})$  is lower semicontinuous on  $X$  for any fixed  $\boldsymbol{\xi} \in \tilde{\Xi}$ , and the feasible set  $X$  is closed, then we have that any accumulation point of the sequence  $\{\hat{\mathbf{x}}_N\}_N$  is almost surely an optimal solution of problem (4.1).*

In the following remark, we show how problem  $P_{(\alpha, \tilde{\rho}_N)}$  can be used to make distributionally robust decisions in a context where the data available to the decision maker



is contaminated.

**| Remark 4.2 (Data-driven decision-making under contaminated samples).**

Suppose that the dataset  $\hat{\xi}_i := (\hat{\mathbf{z}}_i, \hat{\mathbf{y}}_i)$  for  $i = 1, \dots, N$  is composed of correct and contaminated samples. The decision maker only knows that a sample is correct with probability  $\alpha$  and contaminated with probability  $1 - \alpha$ , but does not know which type each sample belongs to. Thus, the data have been generated from a mixture distribution given by  $P = \alpha Q^* + (1 - \alpha)R$ , where  $Q^*$  is the correct distribution and  $R$  a contamination.

In our context, this is equivalent to stating that  $Q^* \in \mathcal{R}_{1-\alpha}(P)$ , which, in turn, can be formulated as  $\mathcal{W}_p(\mathcal{R}_{1-\alpha}(P), Q^*) = 0$ . Since we only have limited information on  $P$  in the form of the empirical distribution  $\hat{P}_N$ , we propose to solve problem  $P_{(\alpha, \tilde{\rho}_N)}$ , that is,

$$\inf_{\mathbf{x} \in X} \sup_Q \mathbb{E}_Q[f(\mathbf{x}, \xi)] \quad (4.16a)$$

$$s.t. \mathcal{W}_p^p(\mathcal{R}_{1-\alpha}(\hat{P}_N), Q) \leq \tilde{\rho}_N \quad (4.16b)$$

where we have assumed that the correct distribution  $Q^*$ , the contamination  $R$  and the data-generating distribution  $P$  are all supported on  $\Xi$ .

The decision maker can profit from the finite sample guarantee that the solution to problem (4.16a)–(4.16b) satisfies as per Theorem 4.2, with  $\tilde{\rho}_N \geq \epsilon_{N,p,\alpha}^p(\beta)$ ,  $\beta \in (0, 1)$ , since  $\epsilon_{N,\alpha}^p = 0$  in this case. Furthermore, if we choose a summable sequence of  $\beta_N \in (0, 1)$ ,  $N \in \mathbb{N}$ , such that  $\lim_{N \rightarrow \infty} \epsilon_N(\beta_N) = 0$ , then we have that

$$P^\infty \left( \lim_{N \rightarrow \infty} \mathcal{W}_p \left( \mathcal{R}_{1-\alpha}(\hat{P}_N), Q^* \right) = 0 \right) = 1 \quad (4.17)$$

In plain words, for  $N$  large enough, the decision vector  $\mathbf{x}$  is being optimized by way of problem (4.16a)–(4.16b) over the “smallest” ambiguity set that almost surely contains the correct distribution  $Q^*$  of the data (in the absence of any other information on  $Q^*$ ). In fact, this means our DRO approach deals with contaminated samples in a way that is distinctly more convenient than that of [34] and [58]. Essentially, they suggest optimizing over a 1-Wasserstein ball centered at  $\hat{P}_N$  of radius  $\tilde{\rho}$ , that is,

$$\inf_{\mathbf{x} \in X} \sup_Q \mathbb{E}_Q[f(\mathbf{x}, \xi)] \quad (4.18a)$$

$$s.t. \mathcal{W}_1(\hat{P}_N, Q) \leq \tilde{\rho} \quad (4.18b)$$

under the argument that for  $\rho$  sufficiently large, the Wasserstein ball contains the true distribution of the data  $Q^*$  with a certain confidence level. For instance, the author of [58] uses the triangle inequality and the convexity property of the Wasserstein distance to establish that  $\mathcal{W}_1(\hat{P}_N, Q^*) \leq \mathcal{W}_1(\hat{P}_N, P) + (1 - \alpha)\mathcal{W}_1(R, Q^*)$ , so that the extra budget

$(1 - \alpha)\mathcal{W}_1(R, Q^*)$  would ensure that  $Q^*$  is within the Wasserstein ball with a given confidence level (a similar argument is made in [34]). In practice, though, this extra budget as such cannot be computed, because neither the correct distribution  $Q^*$  nor the contamination  $R$  are known to the decision maker. However, our approach naturally encodes it in the ambiguity set (4.16b). Indeed, for  $N$  large enough, result (4.17) tells us that the correct distribution  $Q^*$  belongs, almost surely, to the  $(1 - \alpha)$ -trimming set of the empirical distribution  $\hat{P}_N$ . It follows precisely from this and Proposition B.2, in Appendix B, page 106, that  $\mathcal{W}_p(\hat{P}_N, Q^*) \rightarrow \mathcal{W}_p(\alpha Q^* + (1 - \alpha)R, Q^*) \leq \alpha \mathcal{W}_p(Q^*, Q^*) + (1 - \alpha)\mathcal{W}_p(R, Q^*)$ , i.e.,  $\mathcal{W}_p(\hat{P}_N, Q^*) \leq (1 - \alpha)\mathcal{W}_p(R, Q^*)$ .

In short, our approach offers probabilistic guarantees in the finite-sample regime and, in the asymptotic one, naturally exploits all the information we have on  $Q^*$ , namely,  $Q^* \in \mathcal{R}_{1-\alpha}(P)$ , to robustify the decision  $\mathbf{x}$  under contamination.

**The case of unknown  $\mathbb{Q}(\tilde{\Xi}) = \alpha > 0$ .**

In this section, we discuss how we can use the proposed DRO approach to deal with the case in which  $\mathbb{Q}(\tilde{\Xi}) = \alpha > 0$  is unknown. For this purpose, we first introduce a proposition that will allow us to design a distributionally robust strategy to tackle problem (4.1) by means of problem (P).

**Proposition 4.3.** *Suppose that  $\mathbb{Q}(\tilde{\Xi}) = \alpha > 0$ . Take  $0 < \alpha' < \alpha$  and any positive value of  $\tilde{\rho}$ . Given  $N \geq 1$ , the following problem*

$$\begin{aligned} (\text{SP3}) \quad & \sup_{Q_{\tilde{\Xi}}} \mathbb{E}_{Q_{\tilde{\Xi}}} [f(\mathbf{x}, \boldsymbol{\xi})] \\ & \text{s.t. } \mathcal{W}_p^p(\mathcal{R}_{1-\alpha'}(\hat{Q}_N), Q_{\tilde{\Xi}}) \leq \tilde{\rho} \\ & Q_{\tilde{\Xi}}(\tilde{\Xi}) = 1 \end{aligned}$$

is either fully equivalent to (SP2), if  $\frac{1}{N} \geq \alpha$  or a relaxation otherwise.

Based on Proposition 4.3, we could use the following two-step *safe* strategy to handle the case of unknown  $\mathbb{Q}(\tilde{\Xi}) = \alpha > 0$ :

1. First, solve the following uncertainty quantification problem (see [63, 99] for further details),

$$\alpha_N := \inf_{Q \in \mathbb{B}_{\epsilon_N}(\hat{Q}_N)} Q(\boldsymbol{\xi} \in \tilde{\Xi}) = 1 - \sup_{Q \in \mathbb{B}_{\epsilon_N}(\hat{Q}_N)} Q(\boldsymbol{\xi} \notin \tilde{\Xi}) \quad (4.20)$$

where the radius  $\epsilon_N$  of the Wasserstein ball has been chosen so that  $\alpha_N$  represents the minimum probability that the joint true distribution  $\mathbb{Q}$  of the data assigns to the event  $\boldsymbol{\xi} \in \tilde{\Xi}$  with confidence  $1 - \beta_N$ ,  $\beta_N \in (0, 1)$ .

2. Next, solve problem  $(P_{(\alpha_N, \tilde{\rho}_N)})$ , that is,

$$\inf_{\mathbf{x} \in X} \sup_{Q_{\Xi}} \mathbb{E}_{Q_{\Xi}} [f(\mathbf{x}, \boldsymbol{\xi})] \quad (4.21a)$$

$$\text{s.t. } \mathcal{W}_p^p(\mathcal{R}_{1-\alpha_N}(\hat{\mathbb{Q}}_N), Q_{\Xi}) \leq \tilde{\rho}_N \quad (4.21b)$$

$$Q_{\Xi}(\tilde{\Xi}) = 1 \quad (4.21c)$$

with  $\tilde{\rho}_N \geq \epsilon_N^p(\beta_N)/\alpha_N$ .

Now suppose that  $\mathbb{Q} \in \mathbb{B}_{\epsilon_N(\beta_N)}(\hat{\mathbb{Q}}_N)$  and therefore,  $\alpha_N \leq \alpha$  (this is a random event that occurs with probability at least  $1 - \beta_N$ ). According to Lemma 4.3, we have

$$\begin{aligned} \alpha^{1/p} \mathcal{W}_p(\mathcal{R}_{1-\alpha}(\hat{\mathbb{Q}}_N), Q_{\Xi}) &\leq \mathcal{W}_p(\hat{\mathbb{Q}}_N, \mathbb{Q}) \leq \epsilon_N(\beta_N) \\ \mathcal{W}_p^p(\mathcal{R}_{1-\alpha_N}(\hat{\mathbb{Q}}_N), Q_{\Xi}) &\leq \mathcal{W}_p^p(\mathcal{R}_{1-\alpha}(\hat{\mathbb{Q}}_N), Q_{\Xi}) \leq \frac{\epsilon_N^p(\beta_N)}{\alpha} \leq \frac{\epsilon_N^p(\beta_N)}{\alpha_N} = \tilde{\rho}_N \end{aligned}$$

Hence,  $Q_{\Xi} \in \hat{\mathcal{U}}_N(\alpha_N, \tilde{\rho}_N)$  with probability at least  $1 - \beta_N$ . In other words, the two-step procedure here described does not degrade the reliability of the DRO solution. Furthermore, the minimum transportation budget  $\epsilon_{N\alpha_N}$  that makes problem  $(P_{(\alpha_N, \tilde{\rho}_N)})$  feasible is always zero here, if the event  $\boldsymbol{\xi} \in \tilde{\Xi}$  has been observed at least once. This is so because the uncertainty quantification problem of step 1 ensures that  $\alpha_N$  is lower than or equal to the fraction of training data points falling in  $\tilde{\Xi}$ . Moreover, when  $N$  grows to infinity, this uncertainty quantification problem reduces to computing such a fraction of points, which, by the Strong Law of Large Numbers converges to the real  $\alpha$ , i.e.,  $\alpha_N \rightarrow \alpha$  with probability one. Therefore, in the asymptotic regime, this case resembles that of known  $\alpha > 0$ .

**Remark 4.3.** We notice, however, that, in practice, setting  $\tilde{\rho}_N \geq \epsilon_N^p(\beta_N)/\alpha_N$  may result in too large budgets  $\tilde{\rho}_N$ , and thus, in overly conservative solutions, because, as  $\epsilon_N$  is increased,  $\alpha_N$  decreases to zero. For this reason, in Section 4.3.2, we provide an alternative data-driven procedure to address the case  $\alpha > 0$ , in which we simply set  $\alpha_N = \hat{\mathbb{Q}}_N(\tilde{\Xi})$  in problem  $(P_{(\alpha_N, \tilde{\rho}_N)})$  and use the data to tune parameter  $\tilde{\rho}_N$ .

**The case  $\mathbb{Q} \ll \lambda^d$  and  $Q(\tilde{\Xi}) = \alpha = 0$ .**

Suppose that the true joint distribution  $\mathbb{Q}$  governing the random vector  $\boldsymbol{\xi} := (\mathbf{z}, \mathbf{y})$  admits a density function with respect to the Lebesgue measure  $\lambda^d$ , with  $d = d_{\mathbf{z}} + d_{\mathbf{y}}$ . Without loss of generality, consider the event  $\boldsymbol{\xi} \in \tilde{\Xi}$ , where  $\tilde{\Xi}$  is defined as  $\tilde{\Xi} = \{\boldsymbol{\xi} = (\mathbf{z}, \mathbf{y}) \in \Xi : \mathbf{z} = \mathbf{z}^*\}$ . This means that  $Q(\tilde{\Xi}) = \alpha = 0$ .

Therefore, our focus in this case is on the particular variant of problem (4.1) given

by

$$J^* := \inf_{\mathbf{x} \in X} \mathbb{E}_{\mathbb{Q}} [f(\mathbf{x}, (\mathbf{z}, \mathbf{y})) \mid \mathbf{z} = \mathbf{z}^*] \quad (4.22)$$

Problem (4.22) has become a central object of study in what has recently come to be known as *Prescriptive Stochastic Programming or Conditional Stochastic Optimization*, (see, e.g., [9, 18, 19, 17, 22, 45, 112, 121], all of which have been discussed in Chapters 2 and 4). Devising a DRO approach to problem (4.22) using the standard Wasserstein ball  $\mathcal{W}_p(\hat{\mathbb{Q}}_N, Q) \leq \varepsilon$  is of no use here, because any point from the support of  $\hat{\mathbb{Q}}_N$  with an *arbitrarily small* mass can be transported to the set  $\tilde{\Xi}$  at an arbitrarily small cost in terms of  $\mathcal{W}_p(\hat{\mathbb{Q}}_N, Q)$ . This way, one could always place this arbitrarily small particle at a point  $(\mathbf{z}^*, \mathbf{y}') \in \arg \max_{(\mathbf{z}, \mathbf{y}) \in \tilde{\Xi}} f(\mathbf{x}, (\mathbf{z}, \mathbf{y}))$ . In contrast, problem (P), which is based on *partial* mass transportation, offers a richer framework to seek for a distributional robust solution to (4.22). To see this, consider again the inequality (4.11). If we could set  $\alpha = 0$ , the term  $\mathcal{W}_p(\mathcal{R}_{1-\alpha}(\mathbb{Q}), \mathbb{Q}_{\Xi})$  would vanish, because we could take random variables  $\boldsymbol{\xi} \sim \mathbb{Q}_{\Xi}$ ,  $\boldsymbol{\xi}_m \sim \mathbb{Q}_m \in \mathcal{R}_1(\mathbb{Q})$ ,  $m \in \mathbb{N}$ , such that  $\mathcal{W}_p(\mathbb{Q}_m, \mathbb{Q}_{\Xi}) \rightarrow 0$ . Unfortunately, fixing  $\alpha$  to zero is not a real option due to the term  $\frac{1}{\alpha^{1/p}} \mathcal{W}_p(\hat{\mathbb{Q}}_N, Q)$  in the inequality. Therefore, what we propose instead is to solve a sequence of optimization problems in the form

$$(\mathbf{P}_{(\alpha_N, \tilde{\rho}_N)}) \quad \inf_{\mathbf{x} \in X} \sup_{Q_{\Xi}} \mathbb{E}_{Q_{\Xi}} [f(\mathbf{x}, \boldsymbol{\xi})] \quad (4.23a)$$

$$\text{s.t. } \mathcal{W}_p^p(\mathcal{R}_{1-\alpha_N}(\hat{\mathbb{Q}}_N), Q_{\Xi}) \leq \tilde{\rho}_N \quad (4.23b)$$

$$Q_{\Xi}(\tilde{\Xi}) = 1 \quad (4.23c)$$

with both  $\alpha_N$  and  $\tilde{\rho}_N$  tending to zero appropriately as  $N$  increases. Next we show that, under certain conditions, problem  $(\mathbf{P}_{(\alpha_N, \tilde{\rho}_N)})$  enjoys a finite sample guarantee and is asymptotically consistent.

**| Assumption 4.2 (Condition (3.6) from [57]).** Let  $B(\mathbf{z}^*, r) := \{\mathbf{z} \in \Xi_{\mathbf{z}} : \|\mathbf{z} - \mathbf{z}^*\| \leq r\}$  denote the closed ball in  $\mathbb{R}^{d_{\mathbf{z}}}$  with center  $\mathbf{z}^*$  and radius  $r$ . The random vector  $\boldsymbol{\xi} := (\mathbf{z}, \mathbf{y})$  has a joint density  $\phi$  that verifies the following for some  $r_0 > 0$ .

1. It admits uniformly for  $r \in [0, r_0]$  and  $\mathbf{y} \in \mathbb{R}^{d_{\mathbf{y}}}$  the following expansion:

$$\phi(\mathbf{z}^* + r \mathbf{u}, \mathbf{y}) = \phi(\mathbf{z}^*, \mathbf{y}) [1 + r \langle \mathbf{u}, \ell_1(\mathbf{y}) \rangle + O(r^2 \ell_2(\mathbf{y}))] \quad (4.24)$$

where  $\mathbf{u} \in \mathbb{R}^{d_{\mathbf{z}}}$  with  $\|\mathbf{u}\| = 1$ , and where  $\ell_1 : \mathbb{R}^{d_{\mathbf{y}}} \rightarrow \mathbb{R}^{d_{\mathbf{z}}}$  and  $\ell_2 : \mathbb{R}^{d_{\mathbf{y}}} \rightarrow \mathbb{R}$  satisfy  $\int (\|\ell_1(\mathbf{y})\|^2 + |\ell_2(\mathbf{y})|^2) \phi(\mathbf{z}^*, \mathbf{y}) d\mathbf{y} < \infty$ .

2. The marginal density of  $\mathbf{z}$  is bounded away from zero in  $B(\mathbf{z}^*, r_0)$ .

**| Assumption 4.3 (Regularity and boundedness).** We assume that

1. There exists  $\tilde{C} > 0$  and  $r_0 > 0$  such that  $\mathbb{P}(\|\mathbf{z}^* - \mathbf{z}\| \leq r) \geq \tilde{C}r^{d_{\mathbf{z}}}$ , for all  $0 < r \leq r_0$ .
2. The uncertainty  $\mathbf{y}$  is bounded, that is,  $\|\mathbf{y}\| \leq M$  a.s. for some constant  $M > 0$ .

We note that Assumption 4.3.1 is automatically implied by Assumption 4.2, but we explicitly state it here for ease of readability. Furthermore, under the boundedness condition established in Assumption 4.3.2, Assumption 2 is satisfied, for example, by a twice differentiable joint density  $\phi(\mathbf{z}, \mathbf{y})$  with continuous and bounded partial derivatives in  $B(\mathbf{z}^*, r) \times \Xi_{\mathbf{y}}$  and bounded away from zero in that set. These are standard regularity conditions in the technical literature on kernel density estimation and regression [112].

**Theorem 4.4 (Case  $\alpha = 0$ : Finite sample guarantee).** Suppose that Assumptions 4.2, 4.3 and those of Proposition 4.2 hold. Set  $\alpha_0 := \tilde{C}r_0^{d_{\mathbf{z}}}$ . Given  $N \geq 1$ , choose  $\alpha_N \in (0, \alpha_0]$ ,  $\beta \in (0, 1)$ , and determine  $\epsilon_{N,p,\alpha_N}(\beta)$  through (4.14).

Then, for all

$$\tilde{\rho}_N \geq \max \left[ \left( \epsilon_{N,p,\alpha_N}(\beta) + O \left( \alpha_N^{\min\{1, 2/p\}/d_{\mathbf{z}}} \right) \right)^p, \epsilon_{N,\alpha_N}^p \right] \quad (4.25)$$

we have that the pair  $(\hat{\mathbf{x}}_N, \hat{J}_N)$  delivered by problem  $(P_{(\alpha_N, \tilde{\rho}_N)})$  with parameters  $\tilde{\rho}_N$  and  $\alpha_N$  enjoys the finite sample guarantee (4.10).

**Remark 4.4.** There are conditions on the smoothness of the true joint distribution  $\mathbb{Q}$  around  $\mathbf{z} = \mathbf{z}^*$ , other than those stated in Assumptions 4.2 and 4.3, for which we can also upper bound the distance  $\mathcal{W}_p(\mathcal{R}_{1-\alpha}(\mathbb{Q}), \mathbb{Q}_{\Xi})$ . We provide below two examples of these conditions, which have been invoked in [83, 84] and [19], respectively, and neither of which requires the boundedness of the uncertainty  $\mathbf{y}$ .

**Example 4.2.** Suppose that the true data-generating model is given by  $\mathbf{y} = f^*(\mathbf{z}) + \mathbf{e}$ , where  $f^*(\mathbf{z}) := \mathbb{E}[\mathbf{y} \mid \mathbf{z} = \mathbf{z}^*]$  is the regression function and  $\mathbf{e}$  is a zero-mean random error. Furthermore, suppose that Assumption 4.3.1 holds and there exists a positive constant  $L$  such that  $\|f^*(\mathbf{z}') - f^*(\mathbf{z})\| \leq L\|\mathbf{z}' - \mathbf{z}\|$ , for all  $0 \leq \|\mathbf{z}' - \mathbf{z}\| \leq r_0$ .

Take  $\alpha(r) = \tilde{C}r^{d_{\mathbf{z}}}$ , for all  $0 < r \leq r_0$  and set  $\alpha_0 := \alpha(r_0)$ . With abuse of notation, we can write for any event within  $B(\mathbf{z}^*, r) \times \Xi_{\mathbf{y}}$

$$\mathbb{Q}_{B(\mathbf{z}^*, r) \times \Xi_{\mathbf{y}}}(d\mathbf{z}, d\mathbf{y}) = \frac{1}{\mathbb{P}(B(\mathbf{z}^*, r))} \mathbb{Q}(d\mathbf{z}, d\mathbf{y}) = \frac{1}{\mathbb{Q}_{\mathbf{z}}(B(\mathbf{z}^*, r))} \mathbb{Q}_{\mathbf{z}=\mathbf{z}'}(d\mathbf{y}) \mathbb{Q}_{\mathbf{z}}(d\mathbf{z}')$$

where  $\mathbb{Q}_{\mathbf{z}}$  is the probability law of the feature vector  $\mathbf{z}$  and  $\mathbb{Q}_{\mathbf{z}=\mathbf{z}'}$  is the conditional measure of  $\mathbb{Q}$  given that  $\mathbf{z} = \mathbf{z}'$ .

Since  $\mathbb{Q}_{B(\mathbf{z}^*, r) \times \Xi_{\mathbf{y}}} \in \mathcal{R}_{1-\alpha(r)}(\mathbb{Q})$  for all  $0 < r \leq r_0$ , by the convexity of the Wasserstein distance, we have

$$\mathcal{W}_p(\mathcal{R}_{1-\alpha}(\mathbb{Q}), \mathbb{Q}_{\Xi}) \leq \mathcal{W}_p(\mathbb{Q}_{B(\mathbf{z}^*, r) \times \Xi_{\mathbf{y}}}, \mathbb{Q}_{\Xi})$$

$$\begin{aligned}
&\leq \int_{B(\mathbf{z}^*, r)} [\|\mathbf{z}' - \mathbf{z}^*\| + \mathcal{W}_p(\mathbb{Q}_{\mathbf{z}=\mathbf{z}'}, \mathbb{Q}_{\Xi})] \frac{\mathbb{Q}_{\mathbf{z}}(d\mathbf{z}')}{\mathbb{Q}_{\mathbf{z}}(B(\mathbf{z}^*, r))} \\
&= \int_{B(\mathbf{z}^*, r)} [\|\mathbf{z}' - \mathbf{z}^*\| + \mathcal{W}_p(f^*(\mathbf{z}') + \mathbf{e}, f^*(\mathbf{z}^*) + \mathbf{e})] \frac{\mathbb{Q}_{\mathbf{z}}(d\mathbf{z}')}{\mathbb{Q}_{\mathbf{z}}(B(\mathbf{z}^*, r))} \\
&\leq \int_{B(\mathbf{z}^*, r)} [\|\mathbf{z}' - \mathbf{z}^*\| + \|f^*(\mathbf{z}') - f^*(\mathbf{z}^*)\|] \frac{\mathbb{Q}_{\mathbf{z}}(d\mathbf{z}')}{\mathbb{Q}_{\mathbf{z}}(B(\mathbf{z}^*, r))} \\
&\leq (1 + L) \int_{B(\mathbf{z}^*, r)} \|\mathbf{z}' - \mathbf{z}^*\| \frac{\mathbb{Q}_{\mathbf{z}}(d\mathbf{z}')}{\mathbb{Q}_{\mathbf{z}}(B(\mathbf{z}^*, r))} = (1 + L)O(r) = O(\alpha^{1/d_{\mathbf{z}}})
\end{aligned}$$

for all  $0 < \alpha \leq \alpha_0$ .

**Example 4.3.** Take  $p = 1$ . Suppose that there exists a positive constant  $L$  such that

$$\mathcal{W}_1(\mathbb{Q}_{\mathbf{z}=\mathbf{z}'}, \mathbb{Q}_{\mathbf{z}=\mathbf{z}^*}) \leq L\|\mathbf{z}' - \mathbf{z}^*\|, \text{ for all } 0 \leq \|\mathbf{z}' - \mathbf{z}\| \leq r_0$$

and that Assumption 4.3.1 holds. Following a line of reasoning that is parallel to that of the previous example, we also get

$$\mathcal{W}_1(\mathcal{R}_{1-\alpha}(\mathbb{Q}), \mathbb{Q}_{\Xi}) = O(\alpha^{1/d_{\mathbf{z}}}) \text{ for all } 0 < \alpha \leq \alpha_0, \text{ with } \alpha_0 := \alpha(r_0).$$

Equation (4.25) and Examples 4.2 and 4.3 reveal that our finite sample guarantee is affected by the *curse of dimensionality*. Recently, powerful ideas to break this curse have been introduced in [62] under the standard Wasserstein-metric-based DRO scheme. In our setup, however, we also need distributional robustness against the (uncertain) error incurred when inferring conditional information from a sample of the true *joint* distribution. This implies increasing the robustness budget in our approach by an amount linked to the term  $\mathcal{W}_p(\mathcal{R}_{1-\alpha}(\mathbb{Q}), \mathbb{Q}_{\Xi})$ . Consequently, we might need stronger assumptions on the data-generating model to break the dependence of this term with the dimension of the feature vector and thus extend the ideas in [62] to the realm of conditional stochastic optimization.

Now we state the conditions under which the sequence of problems  $(P_{(\alpha_N, \tilde{\rho}_N)})$ ,  $N \rightarrow \infty$ , is asymptotically consistent.

**Lemma 4.5 (Convergence of conditional distributions).** Suppose that the support  $\Xi$  of the true joint distribution  $\mathbb{Q}$  is compact and that Assumptions 4.2 and 4.3.1 hold. Take  $(\alpha_N, \tilde{\rho}_N)$  such that  $\alpha_N \rightarrow 0$ ,  $\frac{N\alpha_N^2}{\log(N)} \rightarrow \infty$ , and  $\tilde{\rho}_N \downarrow \epsilon_{N\alpha_N}^p$ , where  $\epsilon_{N\alpha_N}$  is the minimum transportation budget as in Definition 4.2. Then, we have that

$$\mathcal{W}_p(Q_{\Xi}^N, \mathbb{Q}_{\Xi}) \rightarrow 0 \text{ a.s.}$$

where  $Q_{\Xi}^N$  is any distribution from the ambiguity set  $\hat{\mathcal{U}}_N(\alpha_N, \tilde{\rho}_N)$ .

**Remark 4.5.** The compactness of the support set  $\Xi$  is assumed here just to simplify

the proof. In fact, in Appendix B, Section B.3, we use results from nearest neighbors to show that the convergence of conditional distributions can be attained under the less restrictive condition  $\frac{N\alpha_N}{\log(N)} \rightarrow \infty$  even in some cases for which the uncertainty  $\mathbf{y}$  and the feature vector  $\mathbf{z}$  are unbounded. In addition, we also make use of those results to demonstrate that distributionally robust versions of some local nonparametric predictive methods, such as Nadaraya-Watson kernel regression and  $K$ -nearest neighbors, naturally emerge from our approach.

**Remark 4.6.** The convergence of conditional distributions allows us to establish an asymptotic consistency result analogous to that of Theorem 4.3, by simply replacing “Theorem 4.2”, “ $\tilde{\rho}_N$ ” and “Lemma 4.4” with “Theorem 4.4”, “ $(\alpha_N, \tilde{\rho}_N)$ ” and “Lemma 4.5”, respectively.

**Remark 4.7.** Suppose that the event  $\tilde{\Xi}$  on which we condition problem (4.1) is given by  $\tilde{\Xi} := \{\boldsymbol{\xi} = (\mathbf{z}_1, \mathbf{z}_2, \mathbf{y}) \in \Xi : \mathbf{z}_1 = \mathbf{z}_1^*, \mathbf{z}_2 \in \mathcal{Z}_2\}$ , with  $\mathbb{Q}(\tilde{\Xi}) = 0$  and  $\mathbb{P}(\mathbf{z}_2 \in \mathcal{Z}_2) > 0$ . Let  $\mathbb{Q}_{\mathcal{Z}_2}$  be the probability measure of  $(\mathbf{z}_1, \mathbf{y})$  conditional on  $\mathbf{z}_2 \in \mathcal{Z}_2$ . If we have that there is  $\tilde{C} > 0$  and  $r_0 > 0$  such that  $\mathbb{P}(\|\mathbf{z}_1^* - \mathbf{z}_1\| \leq r) \geq \tilde{C}r^{d_{\mathbf{z}_1}}$ , for all  $0 < r \leq r_0$ , and that  $\mathbb{Q}_{\mathcal{Z}_2}$  satisfies the smoothness condition invoked in either Theorem 4.4, Example 4.2 or Example 4.3, then the analysis in this section extends to that type of event by setting  $\alpha(r) = \tilde{C}r^{d_{\mathbf{z}_1}} \cdot \mathbb{P}(\mathbf{z}_2 \in \mathcal{Z}_2)$  and noticing that  $\mathbb{Q}_{B(\mathbf{z}_1^*, r) \times \mathcal{Z}_2 \times \Xi_{\mathbf{y}}} \in \mathcal{R}_{1-\alpha(r)}(\mathbb{Q})$ ,  $0 < r \leq r_0$ , where  $\mathbb{Q}_{B(\mathbf{z}_1^*, r) \times \mathcal{Z}_2 \times \Xi_{\mathbf{y}}}$  is the probability measure of  $(\mathbf{z}_1, \mathbf{z}_2, \mathbf{y})$  conditional on  $(\mathbf{z}_1, \mathbf{z}_2, \mathbf{y}) \in B(\mathbf{z}_1^*, r) \times \mathcal{Z}_2 \times \Xi_{\mathbf{y}}$ .

In the following sections, we discuss three applications of the methodology introduced in Section 4.1 in order to provide additional insights into the computational aspects and the performance guarantees of the DRO framework with side information that we propose. First, we consider the newsvendor problem, which was first introduced in Chapter 3, Section 3.2.1, but considering side information. As a second application, we consider a portfolio allocation problem considering some side information under two scenarios: the case  $\mathbb{Q}(\tilde{\Xi}) = \alpha = 0$  and the case  $\mathbb{Q}(\tilde{\Xi}) = \alpha > 0$ . Finally, the application of the proposed methodology to a chance-constrained DRO problem is illustrated using a real-life problem within the realm of power systems, the so-called *Optimal Power Flow* problem. In all the numerical experiments, we take the  $p$ -norm with  $p = 1$  and, accordingly, we use the Wasserstein distance of order 1. This choice allows us to recast all the optimization problems that we solve as linear programs.

## 4.2 Application I. Newsvendor problem

The newsvendor problem with side information has received a lot of attention lately (see, for example, [9, 77] and references therein). For the particular instance of this problem that we analyze next, we have considered  $h = 1$  and  $b = 10$ . Further-



more, the true joint distribution of the data  $\hat{\xi}_i := (\hat{z}_i, \hat{y}_i)$ ,  $i = 1, \dots, N$  is assumed to follow a mixture (with equal weights) of two normal bivariate distributions with means  $\mu_1 = [0.6, 0.75]^\top$ ,  $\mu_2 = [0.5, -0.75]^\top$  and covariance matrices  $\Sigma_1 = \begin{bmatrix} 0.5 & 0 \\ 0 & 0.01 \end{bmatrix}$ ,  $\Sigma_2 = \begin{bmatrix} 0.0001 & 0 \\ 0 & 0.1 \end{bmatrix}$ , respectively. Therefore, the support set of this distribution is the whole space  $\mathbb{R}^{d_z+d_y}$ , with  $d_z = d_y = 1$ . In addition, we consider as  $\mathcal{Z}$  the singleton  $\{z^* = 0.44\}$ , with  $\tilde{\Xi}$  being the real line  $\mathbb{R}$  as a result. Figure (4.2a) shows a heat map of the true joint distribution, together with a kernel estimate of the probability density function of the random variable  $y$  conditional on  $z^*$ . Moreover, the white dotted curve in the figure corresponds to the optimal order quantity as a function of the feature  $z$ . Note that this curve is highly nonlinear around the context  $z^*$ . Also, the demand may be negative, which, in the context of the newsvendor problem, can be interpreted as items being returned to the stores due to, for example, some quality defect.

We compare five data-driven approaches to address the solution to this problem: A Sample Average Approximation method based on a local predictive technique, in particular, the  $K_N$  nearest neighbors, which we refer to as “KNN” (see [17] for further details); this very same local predictive method followed by a standard Wasserstein-metric-based DRO approach to robustify it, as suggested in [19, Section 5], which we call “KNNDRO”; the robustified KNN method (4.4), also proposed in [19], which we term “KNNROBUST”; and our approach, i.e., problem  $P_{(\alpha_N, \tilde{\rho}_N)}$  with  $\alpha_N = K_N/N$ , which we denote “DROTRIMM”. The rule  $\alpha_N = K_N/N$  is a natural choice that guarantees a fair comparison of the four methods and turns DROTRIMM into a distributionally robust  $K_N$ -nearest neighbors as per Corollary B.3 in Appendix B.3. Finally, the fifth method we analyze is the machine learning algorithm proposed in [9], which was especially designed for the newsvendor problem with features. In this algorithm, a polynomial mapping between the optimal order quantity (i.e., the optimal quantile) and the covariates is presumed. We denote this latter approach as ML from “Machine Learning”.

We clarify that KNNDRO uses the  $K_N$  nearest neighbors projected onto the set  $\tilde{\Xi}$  as the nominal “empirical” distribution that is used as the center of the Wasserstein ball in [99]. Indeed, the newsvendor problem features an objective function with a Lipschitz constant with respect to the uncertainty that is independent of the decision  $\mathbf{x}$ . Consequently, as per [99, Remark 6.7], KNNDRO renders the same minimizer for this problem as that of KNN whenever the support set  $\tilde{\Xi}$  is equal to the whole space. This is, in contrast, not true for the portfolio allocation problem which will be considered in Section 4.3, which has an objective function with a Lipschitz constant with regard to the uncertainty that depends on the decision  $\mathbf{x}$ .

We consider a series of different values for the size  $N$  of the sample data. Unless



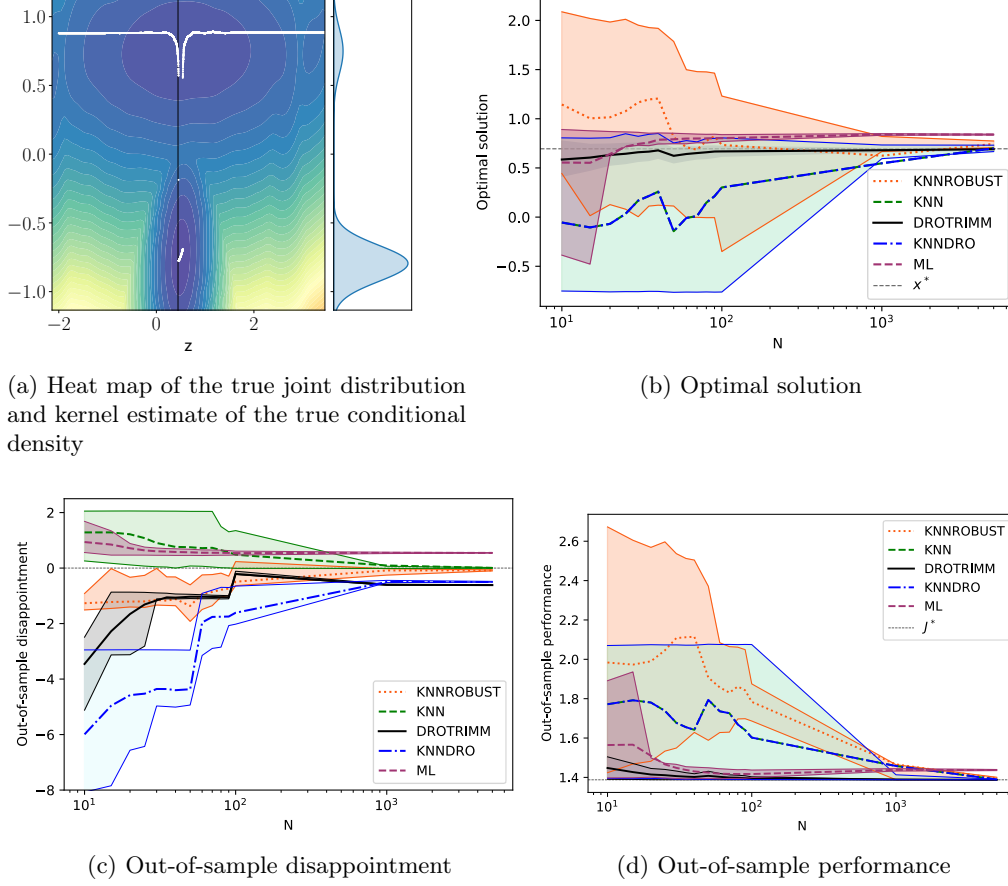


Figure 4.2: Newsvendor problem with features: True distributions, quantile estimate and performance metrics

stated otherwise in the text, for each  $N$ , we choose as the number of neighbors,  $K_N$ , the value  $\lfloor N/\log(N+1) \rfloor$ , where  $\lfloor \cdot \rfloor$  stands for the floor function.

We estimate  $\mathbf{x}^* \in \arg \min_{\mathbf{x} \in X} \mathbb{E}_{Q_{\Xi}} [f(\mathbf{x}, \boldsymbol{\xi})]$  and  $J^* = \mathbb{E}_{Q_{\Xi}} [f(\mathbf{x}^*, \boldsymbol{\xi})]$  using a discrete proxy of the true conditional distribution  $Q_{\Xi}$ . In this newsvendor problem, said proxy is made up of 1085 data points, resulting from applying the KNN method (with the logarithmic rule) to 10 000 samples from the true data-generating joint distribution.

To compare the five data-driven approaches under consideration, we use two performance metrics, specifically, the *out-of-sample performance* of the data-driven solution and its *out-of-sample disappointment*. The former is given by  $J = \mathbb{E}_{Q_{\Xi}} [f(\hat{\mathbf{x}}_N^m, \boldsymbol{\xi})]$ , while the latter is calculated as  $J - \hat{J}_N^m$ , where  $m = \{\text{KNNROBUST}, \text{DROTRIMM}, \text{KNNNDRO}, \text{KNN}\}$  and  $\hat{J}_N^m$  is the objective function value yielded by the data-driven optimization

problem solved by method  $m$ . We note that a negative out-of-sample disappointment represents a favorable outcome.

Since  $\mathbb{E}_{\mathbb{Q}_{\Xi}}[f(\hat{\mathbf{x}}_N^m, \boldsymbol{\xi})]$  and  $\hat{J}_N^m$  are functions of the sample data, we conduct a certain number of runs (400 for this instance of the newsvendor problem) for every  $N$ , each run with an independent sample of size  $N$ .

This way we can get (visual) estimates of the out-of-sample performance and disappointment for several values of the sample size  $N$  for different independent runs. These estimates are illustrated in the form of box plots in a series of figures, where the dotted black horizontal line corresponds to either the optimal solution  $\mathbf{x}^*$  or to its associated optimal cost  $J^*$  with complete information.

As is customary in practice, we use a data-driven procedure to tune the robustness parameter of each method. In particular, for a desired value of reliability  $1 - \beta \in (0, 1)$  (in our numerical experiments, we set  $\beta$  to 0.15), and for each method  $j$ , where  $j = \{\text{KNNROBUST}, \text{KNNDRO}, \text{DROTRIMM}\}$ , we aim for the value of the robustness parameter for which the estimate of the objective value  $\hat{J}_N^j$  given by method  $j$  provides an upper  $(1 - \beta)$ -confidence bound on the out-of-sample performance of its respective optimal solution (see Equation (4.10)), while delivering the best out-of-sample performance. As the optimal robustness parameter is unknown and depends on the available data sample, we need to derive an estimator  $param_N^{\beta, j}$  that is also a function of the training data. We construct  $param_N^{\beta, j}$  and the corresponding reliability-driven solution as follows:

1. We generate  $kboot$  resamples (with replacement) of size  $N$ , each playing the role of a different training set. In our experiments we set  $kboot = 50$ . Moreover, we build a validation dataset determining the  $K_{N_{val}}$ -neighbors of the  $N_{val}$  data points of the original sample of size  $N$  that have not been used to form the training set.
2. For each resample  $k = 1, \dots, kboot$  and each candidate value for  $param$ , we compute a solution by method  $j$  with parameter  $param$  on the  $k$ -th resample. The resulting optimal decision is denoted as  $\hat{x}_N^{j, k}(param)$  and its corresponding objective value as  $\hat{J}_N^{j, k}(param)$ . Thereafter, we calculate the out-of-sample performance  $J(\hat{x}_N^{j, k}(param))$  of the data-driven solution  $\hat{x}_N^{j, k}(param)$  over the validation set.
3. From among the candidate values for  $param$  such that  $\hat{J}_N^{j, k}(param)$  exceeds the value  $J(\hat{x}_N^{j, k}(param))$  in at least  $(1 - \beta) \times kboot$  different resamples, we take as  $param_N^{\beta, j}$  the one yielding the best out-of-sample performance averaged over the  $kboot$  validation datasets.
4. Finally, we compute the solution given by method  $j$  with parameter  $param_N^{\beta, j}$ ,  $\hat{x}_N^j := \hat{x}_N^j(param_N^{\beta, j})$  and the respective certificate  $\hat{J}_N^j := \hat{J}_N^j(param_N^{\beta, j})$ .

Recall that, in our approach DROTRIMM, the robustness parameter  $\tilde{\rho}_N$  must be greater than or equal to the minimum transportation budget to the power of  $p$ , that is,  $\varepsilon_{N\alpha_N}^p$  (we point out again that we have taken  $p = 1$ ). Hence, if we decompose  $\tilde{\rho}_N$  as  $\tilde{\rho}_N = \varepsilon_{N\alpha_N}^p + \Delta\tilde{\rho}_N$ , what one really needs to tune in DROTRIMM is the budget excess  $\Delta\tilde{\rho}_N$ . Furthermore, for the same amount of budget  $\Delta\tilde{\rho}_N$ , our approach will lead to more robust decisions  $\mathbf{x}$  than KNNDRO, because the worst-case distribution in KNNDRO is also feasible in DROTRIMM. Consequently, in practice, the tuning of one of these methods could guide the tuning of the other.

All the simulations have been run on a Linux-based server using up to 116 CPUs running in parallel, each clocking at 2.6 GHz with 4 GB of RAM. We have employed Gurobi 9.0 under Pyomo 5.2 to solve the associated linear programs.

The set of candidate values from which the robustness parameters in methods KNNROBUST, KNNDRO and DROTRIMM have been selected is the discrete set composed of the thirty linearly spaced numbers between 0 and 2. Last but not least, the degree of the polynomial used by ML is tuned in a way analogous to how the robustness parameters of KNNROBUST, KNNDRO and DROTRIMM are tuned using the bootstrapping procedure described above. Nevertheless, we have only considered polynomial mappings up to the fourth degree.

Figures (4.2b), (4.2c), and (4.2d) illustrate the box plots corresponding to the quantile estimators (i.e., the optimal solution of the problem), the out-of-sample disappointment and the out-of-sample performance delivered by each of the considered data-driven approaches for various sample sizes and runs, in that order. The shaded color areas have been obtained by joining the 15th and 85th percentiles of the box plots, while the associated bold colored lines link their means. The true optimal quantile (with complete information) and its out-of-sample performance are also depicted in Figures (4.2c) and (4.2b), respectively, using black dotted lines.

Interestingly, whereas the quantile estimators provided by DROTRIMM, KNNDRO and KNNROBUST all lead to negative out-of-sample disappointment in general, KNNDRO and KNNROBUST exhibit substantially worse out-of-sample performance both in expectation and volatility. Recall that KNNDRO delivers the same solutions provided by KNN for this problem. Its behavior is, therefore, influenced by the bias introduced by the  $K$ -nearest neighbors estimation, which is particularly notorious for small-size samples in this case, given the shape of the true conditional density, see Figure (4.2a). Actually, for some runs, the  $K$ -nearest neighbors, and hence KNNDRO, lead to negative quantile estimates, while the true one is positive and greater than 0.5. By construction, both KNNDRO and KNNROBUST are mainly affected by the estimation error of the conditional probability distribution incurred by the local predictive method. On the contrary, our approach DROTRIMM offers a natural protection against this error and a richer spectrum of data-driven solutions. Indeed, DROTRIMM is able to identify

solutions that lead to a better out-of-sample performance with a negative out-of-sample disappointment.

Finally, both ML and DROTRIMM exhibit a notorious stable behavior against the randomness of the sample. The order quantity provided by the former, however, does not converge to the true optimal one, because the relationship between the true optimal order and the feature  $z$  is far from being polynomial. Note that ML is a *global* method that seeks to learn the optimal order quantity for *all* possible contexts by using a polynomial up to the fourth degree. However, the (true) optimal order curve (that is, the white line in Figure 4.2a) is highly nonlinear within a neighborhood of the context  $z^* = 0.44$ , but practically constant outside of it.

### 4.3 Application II. Portfolio allocation problem

#### 4.3.1 Case $\mathbb{Q}(\tilde{\Xi}) = \alpha = 0$

We consider in this section an instance of the portfolio optimization problem that is based on that used in [18] and [22]. The instance corresponds to a single-stage portfolio optimization problem in which we wish to find an allocation of a fixed budget to six different assets. Thus,  $\mathbf{x} \in \mathbb{R}_+^6$  denotes the decision variable vector, that is, the asset allocations, and their uncertain return is represented by  $\mathbf{y} \in \mathbb{R}^6$ . In practice, these uncertain returns may be influenced by a set of features. First, the decision maker observes auxiliary covariates and later, selects the portfolio. We consider three different covariates that can potentially impact the returns and that we denote as  $\mathbf{z} = (z_1, z_2, z_3)$ . The decision maker wishes to leverage this side information to improve his/her decision-making process in which the goal is to maximize the expected value of the return while minimizing the Conditional Value-at-Risk (**CVaR**) of the portfolio, that is, the risk that the loss  $(-\langle \mathbf{x}, \mathbf{y} \rangle)^+ := \max(-\langle \mathbf{x}, \mathbf{y} \rangle, 0)$  is large. Using the reformulation of the **CVaR** (see [118] and [22]) and introducing the auxiliary variable  $\beta'$ , the decision maker aims to solve the following optimization problem given the value of the covariate  $\mathbf{z}^* = (1000, 0.01, 5)$  in the numerical experiments):

$$\min_{(\mathbf{x}, \beta') \in X} \mathbb{E} \left[ \beta' + \frac{1}{\delta} (-\langle \mathbf{x}, \mathbf{y} \rangle - \beta')^+ - \lambda \langle \mathbf{x}, \mathbf{y} \rangle \mid \mathbf{z} = \mathbf{z}^* \right] \quad (4.26)$$

where the feasible set of decision variables of the problem, that is,  $X$  is equal to  $\{(\mathbf{x}, \beta') \in \mathbb{R}_+^6 \times \mathbb{R} : \sum_{j=1}^6 x_j = 1\}$ . We set  $\delta = 0.5$  and  $\lambda = 0.1$  to simulate an investor with a moderate level of risk aversion. The parameter  $\lambda \in \mathbb{R}_+$  serves to tradeoff between risk and return, and  $\delta$  refers to the  $(1-\delta)$ -quantile of the loss distribution. We take the same marginal distributions for the covariates as in Section 5.2 of [22], i.e.,  $z_1 \rightsquigarrow \mathcal{N}(1000, 50)$ ,  $z_2 \rightsquigarrow \mathcal{N}(0.02, 0.01)$  and  $\log(z_3) \rightsquigarrow \mathcal{N}(0, 1)$ . Furthermore, we follow their approach to

construct the joint true distribution of the covariates and the asset returns. In particular, we take

$$\mathbf{y}/(\mathbf{z} = (z_1, z_2, z_3)) \rightsquigarrow \mathcal{N}_6(\boldsymbol{\mu} + 0.1 \cdot (z_1 - 1000) \cdot \mathbf{v}_1 + 1000 \cdot z_2 \cdot \mathbf{v}_2 + 10 \cdot \log(z_3 + 1) \cdot \mathbf{v}_3, \boldsymbol{\Sigma})$$

with  $\mathbf{v}_1 = (1, 1, 1, 1, 1, 1)^\top$ ,  $\mathbf{v}_2 = (4, 1, 1, 1, 1, 1)^\top$ ,  $\mathbf{v}_3 = (1, 1, 1, 1, 1, 1)^\top$ , and with  $\boldsymbol{\mu}, \boldsymbol{\Sigma}^{1/2}$  given in [22, 54].

We employ the analytic form of the conditional distribution  $\mathbb{Q}_{\Xi}$  provided above to construct a 10 000-data-point approximation. We use this approximation to assess the out-of-sample performance of the data-driven methods KNN, KNNROBUST, KNNDRO and DROTRIMM, all of which were introduced in the single-item newsvendor problem of Section 4.2. Similarly to the case of the single-item newsvendor problem, we consider a series of different values for the size  $N$  of the sample data. Unless stated otherwise in the text, for each  $N$ , the number of neighbors,  $K_N$ , is chosen among the values  $\lfloor N/\log(N+1) \rfloor$ ,  $\lfloor N^{0.9} \rfloor$  and  $\lfloor \sqrt{N} \rfloor$  to assess the impact of the number of neighbors on the out-of-sample performance of the four methods we compare.

Note that, unlike in [22], not all the features affect equally all the asset returns. Moreover, feature  $z_3$  is log-normal and therefore, Assumption 4.1 does not hold. Nonetheless, as we show below, DROTRIMM performs satisfactorily, which reveals that the conditions we derive to guarantee that our approach performs well are sufficient, but not necessary. Indeed, the condition  $\mathbb{Q}_{\Xi} \in \hat{\mathcal{U}}_N(\alpha_N, \tilde{\rho}_N)$  is not required to ensure performance guarantees [62, 87]. For all the methods, we have standardized the covariates  $\mathbf{z}$  and the asset returns  $\mathbf{y}$  using their means and variances. In all the simulations, the robustness parameter each method uses (i.e.,  $\varepsilon_N$  in KNNROBUST, the radius of the Wassertein ball,  $\rho_N$ , in KNNDRO, and the budget excess  $\Delta\tilde{\rho}_N$  in DROTRIMM) has been chosen from the discrete set  $\{b \cdot 10^c : b \in \{0, \dots, 9\}, c \in \{-2, -1, 0\}\}$ , following the data-driven tuning procedure described for the newsvendor problem in Section 4.2.

Similarly to the case of the single-item newsvendor problem, Figure 4.3 shows, for various sample sizes and 200 runs, the box plots pertaining to the out-of-sample disappointment and performance associated with each of the considered data-driven approaches. Each of the three pairs of subplots at the top of the figure has been obtained with a different rule to determine the number  $K_N$  of nearest neighbors. Increasing this number seems to have a positive effect on the convergence speed of all the methods for this instance, although KNNROBUST (and KNNDRO to a lesser extent) has some trouble ensuring the desired reliability level, with the 85% line above 0 for the largest values of  $N$  we represent. In contrast, DROTRIMM manages to keep the disappointment negative. This is, in addition, accompanied by an important improvement of the out-of-sample performance (in line with the criterion for selecting the best portfolio that we have established). In fact, DROTRIMM produces boxplots that appear to be

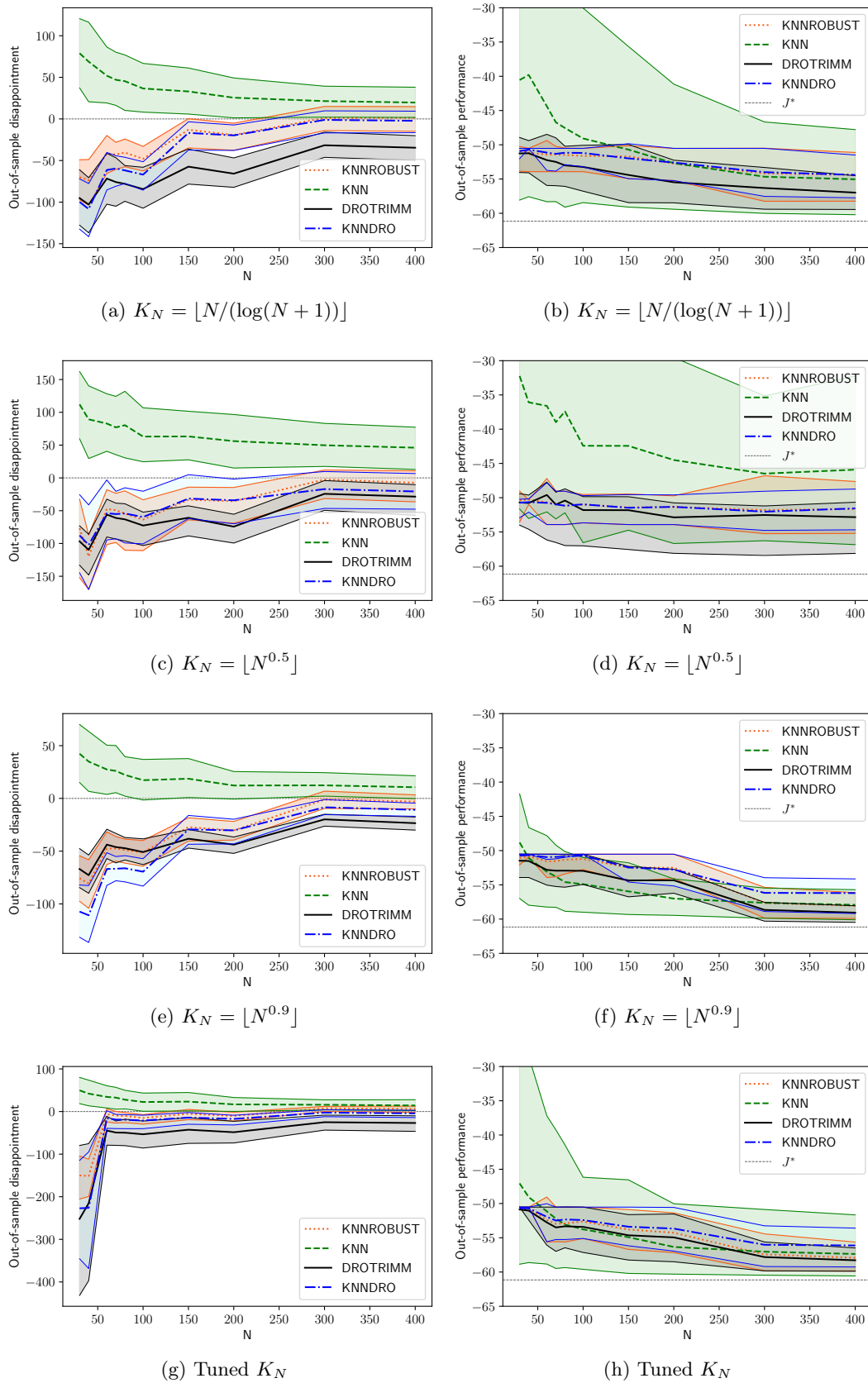


Figure 4.3: Portfolio problem with features: Performance metrics

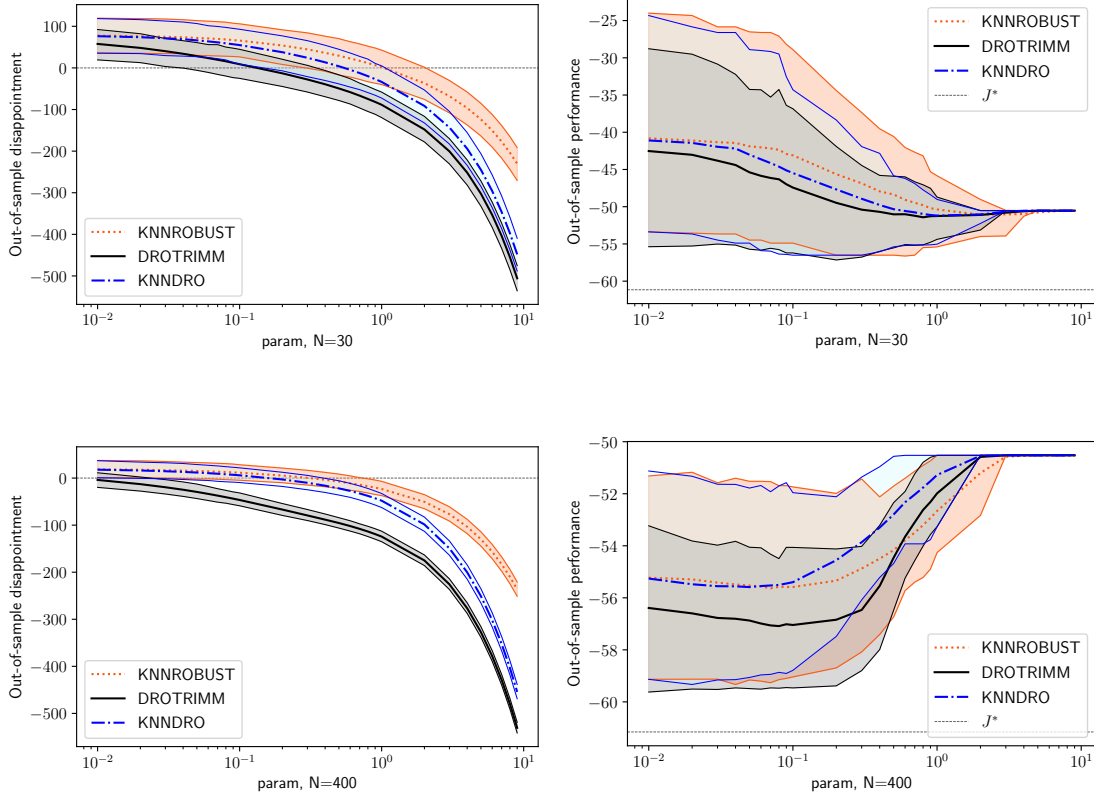


Figure 4.4: Impact of the robustness parameter with 200 training samples,  $K_N = \lfloor N/(\log(N+1)) \rfloor$  and  $\delta = 0.5$ ,  $\lambda = 0.1$

shifted downward, i.e., in the direction of better objective function values. On the other hand, the KNN method substantially improves its performance by employing a larger number of neighbors. However, it is way too optimistic in any case.

The results shown in the pair of subplots at the bottom of Figure 4.3 correspond to a number  $K_N$  of neighbors that has been tuned jointly with the robustness parameter and for each method independently. For this purpose, we have selected the best value of  $K_N$  for each approach from the discrete set  $\{N^{0.1}, N^{0.2}, \dots, N^{0.9}\}$  following the bootstrapping-based procedure previously described. The data-driven tuning of the number  $K_N$  of neighbors appears not to have a major effect on the performance of the different methods, especially in comparative terms. We do observe that the out-of-sample performance of KNNROBUST and KNNDRO is slightly improved on average. This improvement in cost performance is, however, accompanied by an increase in the number of sample sizes for which these methods do not satisfy the reliability requirement, particularly in the case of KNNROBUST and small sample sizes.

To facilitate the analysis of the results shown in Figure 4.3, we also provide Figure 4.4, which illustrates the (random) performance of the methods KNNROBUST,



DROTRIMM and KNNDRO as a function of their respective robustness parameter, estimated over 200 independent runs. Again, the shaded areas cover the 15th and 85th percentiles, while the bold colored lines correspond to the average performance. The various plots are obtained for  $N = 30$  and  $N = 400$ , with the number of neighbours given by the logarithmic rule. These plots are especially informative, because they are independent of the specific validation procedure used to tune the robustness parameters of the methods and thus, provide insight into the potential of each method to identify good solutions. Note that the out-of-sample performance of all the three methods stabilizes around the same value as their respective robustness parameters grow large enough. This phenomenon is analogous to that discussed in [99, Section 7.1]. However, the value we observe here does not correspond to the “equally weighted portfolio,” because we have standardized the data on the asset returns. As a result, the “robust portfolio” that delivers this out-of-sample performance depends on and is solely driven by the standard deviations of the different assets. Very interestingly, DROTRIMM is able to uncover portfolios whose out-of-sample performance features a better mean-variance trade-off, in general. Furthermore, it requires a smaller value of the robustness parameter to guarantee reliability. All this is more evident (and useful) for the case  $N = 400$ , as we explain next. When  $N = 30$ , all the considered methods need large values of their robustness parameter to ensure reliability, so they all tend to operate close to the “robust portfolio” we mentioned above. DROTRIMM can certainly afford lower values of  $\Delta\tilde{\rho}$  in an attempt to improve performance, but this proves not to be that profitable for such a small sample size, for which the robust portfolio performs very well. As  $N$  increases, the robust portfolio loses its appeal, since its performance gradually becomes comparatively worse. DROTRIMM is then able to identify portfolios that perform significantly better in expectation, while providing an estimate of their return such that the desired reliability is guaranteed. For their part, KNNDRO and KNNROBUST are also able to discover solutions with an actual average cost lower than that of the robust portfolio (albeit with a worse expectation and a higher variance than those given by DROTRIMM). However, they are more prone to overestimate their returns.

Finally, we study the behavior of the different methods under other contexts. For this, we consider several values of  $N$ , one random data sample for each  $N$ , and 200 different contexts  $\mathbf{z}^*$  sampled from the marginal distributions of the features. The performance metrics (i.e., the out-of-sample disappointment and performance) are plotted in Figures 4.5a and 4.5b, respectively, under an optimal selection of the robustness parameters (that is, for each method we use the value of the robustness parameter that, while ensuring a negative disappointment, delivers the best out-of-sample performance). We observe that DROTRIMM systematically performs better, with an actual cost averaged over the 200 contexts that is lower irrespective of the sample size.



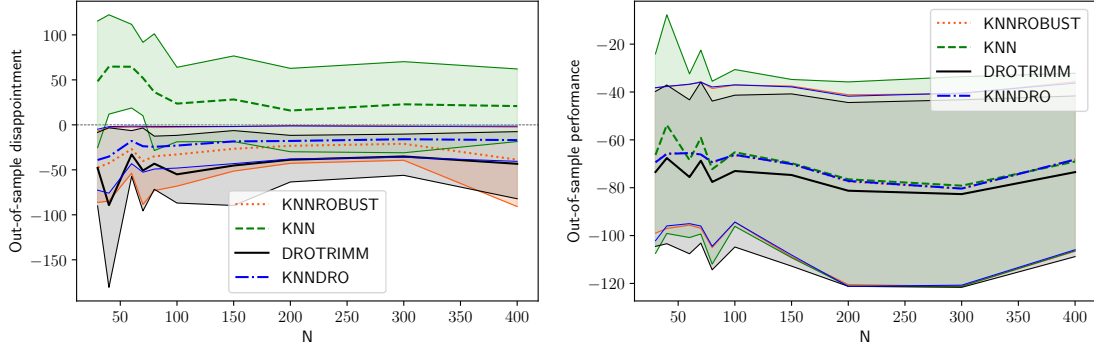


Figure 4.5: Portfolio problem with features: Varying context under an optimal selection of the robustness parameters,  $K_N = \lfloor N/(\log(N+1)) \rfloor$  and  $\delta = 0.5$ ,  $\lambda = 0.1$

#### 4.3.2 Case $\mathbb{Q}(\tilde{\Xi}) = \alpha > 0$

In this section, we present and discuss some numerical results for the case  $\mathbb{Q}(\tilde{\Xi}) = \alpha > 0$ . For this purpose, we use the same portfolio allocation problem described before. To this end, we assume instead that the feature vector lives in an uncertainty set  $\mathcal{Z}$  such that  $\mathbb{Q}(\tilde{\Xi}) > 0$ . In particular, we consider  $\mathcal{Z} := \{\mathbf{z} \in \mathbb{R}^3 : \|\tilde{\mathbf{z}}\|_\infty \leq r\}$ , with  $\tilde{\mathbf{z}}$  being the standardized feature vector. Thus, we have that  $\tilde{\Xi}$  is given by

$$\tilde{\Xi} := \{(\mathbf{z}, \mathbf{y}) \in \mathbb{R}^{3+6} : \|\tilde{\mathbf{z}}\|_\infty \leq r\}$$

We take  $r = 0.6$  for the simulation experiments.

We draw 50 000 samples from the true joint data-generating distribution through the explicit form of  $\mathbf{y}/\mathbf{z}$  given in Section 4.3.1. We then use the conditional empirical distribution made up of those samples falling within  $\tilde{\Xi}$ , specifically, 7306 data points, as a proxy of the true conditional distribution  $\mathbb{Q}_{\tilde{\Xi}}$ . Consequently, we have that  $\mathbb{Q}(\tilde{\Xi}) \approx 0.14612$ . We wish to solve the following optimization problem

$$\min_{(\mathbf{x}, \beta') \in X} \mathbb{E} \left[ \beta' + \frac{1}{\delta} (-\langle \mathbf{x}, \mathbf{y} \rangle - \beta')^+ - \lambda \langle \mathbf{x}, \mathbf{y} \rangle \mid (\mathbf{z}, \mathbf{y}) \in \tilde{\Xi} \right] \quad (4.27)$$

with the rest of the parameters being equal to the values taken in the instance  $\alpha = 0$ .

We also compare here four data-driven approaches to solve problem (4.27), namely:

- Our two approaches, i.e., problem  $P_{(\alpha, \tilde{\rho}_N)}$  with  $\alpha := \mathbb{Q}(\tilde{\Xi})$  called “DROTRIMM1” and problem  $P_{(\alpha_N, \tilde{\rho}_N)}$ , where  $\alpha_N := \hat{\mathbb{Q}}_N(\tilde{\Xi})$  is an estimate of  $\alpha$ . We refer to this approach as “DROTRIMM2.” In principle, this would be the natural approach that a decision-maker with no knowledge of  $\alpha$  would use.
- A sample average approximation (SAA) method that works with the samples

falling in  $\tilde{\Xi}$ .

- The aforementioned SAA method followed by a standard Wasserstein-metric-based DRO approach to robustify it, which we call “SAADRO”.

As in the previous numerical experiments, we employ a similar bootstrapping procedure based on the available data sample to tune the robustness parameter that each method  $j$ , with  $j \in \{\text{DROMTRIMM1}, \text{DROTRIMM2}, \text{SAADRO}\}$ , uses. More specifically, for each  $j \in \{\text{DROMTRIMM1}, \text{DROTRIMM2}, \text{SAADRO}\}$  and a given value of reliability  $1 - \beta \in (0, 1)$  (in our numerical experiments, we set  $\beta$  to 0.15), we seek an estimator  $param_N^{\beta,j}$  that leads to the best out-of-sample performance, while guaranteeing the desired level of confidence  $1 - \beta$ . For each sample of size  $N$ , we use the following algorithm to derive  $param_N^{\beta,j}$  and the corresponding portfolio solution:

1. We construct  $kboot$  resamples (with replacement) of size  $N$ , each playing the role of a different training dataset. In our experiments we use  $kboot = 50$ . Moreover, we build a validation dataset (per resample) from those data points from the original sample of size  $N$  that fall in  $\tilde{\Xi}$ , but which have not been involved in the resample. We only consider resamples from which we can build a validation set of at least one data point. Furthermore, unlike DROTRIMM1 and DROTRIMM2, SAADRO can only be implemented if we have at least one data point falling within  $\tilde{\Xi}$  in the training set (the same occurs with SAA). Thus, we implicitly assume that the source sample has no fewer than two data points in  $\tilde{\Xi}$ .
2. For each resample  $k = 1, \dots, kboot$  and each candidate value for  $param$  (taken from the discrete set  $\{b \cdot 10^c : b \in \{0, \dots, 9\}, c \in \{-3, -2, -1, 0\}\}$ ), we compute a solution by method  $j$  with parameter  $param$  on the  $k$ -th resample. The resulting optimal decision is denoted as  $\hat{x}_N^{j,k}(param)$  and its corresponding objective value as  $\hat{J}_N^{j,k}(param)$ . Thereafter, we calculate the out-of-sample performance  $J(\hat{x}_N^{j,k}(param))$  of the data-driven solution  $\hat{x}_N^{j,k}(param)$  over the validation dataset.
3. From among the candidate values for  $param$  such that  $\hat{J}_N^{j,k}(param)$  exceeds the value  $J(\hat{x}_N^{j,k}(param))$  in at least  $(1 - \beta) \times kboot$  different resamples, we take as  $param_N^{\beta,j}$  the one yielding the best cost performance averaged over the  $kboot$  resamples.
4. Finally, we compute the solution given by method  $j$  with parameter  $param_N^{\beta,j}$ ,  $\hat{x}_N^j := \hat{x}_N^j(param_N^{\beta,j})$  and the respective certificate  $\hat{J}_N^j := \hat{J}_N^j(param_N^{\beta,j})$ .

Figure 4.6 shows the box plots pertaining to the out-of-sample disappointment and performance associated with each of the considered data-driven approaches for various sample sizes. The box plots have been obtained from 200 independent runs per sample

size  $N$ . The SAA method provides portfolios that, in expectation, perform reasonably well, especially when the sample size is large enough. However, SAA definitely fails to ensure the desired level of reliability. As for the three approaches that incorporate robustness in the decision-making, DROTRIMM1 and DROTRIMM2 seem to systematically identify reliable portfolios with a better expected performance than those given by SAADRO.

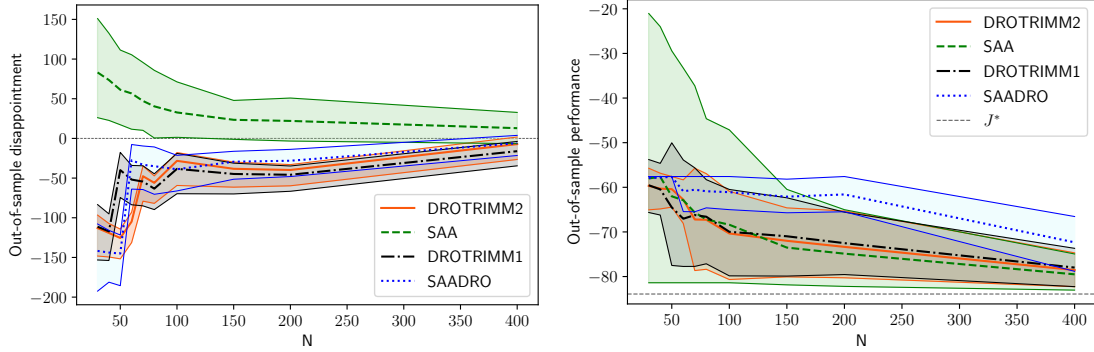


Figure 4.6: Portfolio problem with features: Performance metrics. Case  $\alpha > 0$  and  $\delta = 0.5$ ,  $\lambda = 0.1$

To investigate the ability of SAADRO, DROTRIMM1 and DROTRIMM2 to identify good portfolios, we provide Figure 4.7, which is analogous to Figure 4.4 in the case  $\alpha = 0$ . Observe that both DROTRIMM1 and DROTRIMM2 guarantee reliability for smaller values of their robustness parameter than SAADRO. This gives the former a competitive advantage over the latter, essentially because it appears that a better out-of-sample performance (in expectation) is, in general, aligned with a lower distributional robustness (this finding is consistent with the fact that the unreliable SAA solution performs fairly well in terms of the weighted mean-risk asset returns). To be more precise, taking a small sample size  $N$  (say 50) and an equal value for each of their robustness parameters, DROTRIMM1 and DROTRIMM2 deliver portfolios with an actual expected cost (and variance) that is lower than or approximately equal to that of the portfolios provided by SAADRO. They do so for any value of their robustness parameter. Furthermore, when  $N$  is increased, even though there exists a range of values of the robustness parameter for which SAADRO also identifies portfolios with a good performance out of sample, these are discarded by the method because they do not comply with the reliability specification. For instance, take  $N = 400$ . SAADRO needs a radius larger than 0.2-0.3 to ensure reliability. However, for these values of the Wasserstein-ball radius, the portfolios given by SAADRO result in an actual expected cost above  $-70$ . On the other hand, DROTRIMM2 guarantees reliability with a value of its robustness parameter above 0.003-0.004, for which, in addition, it provides solutions

with an actual expected cost below  $-77$ .

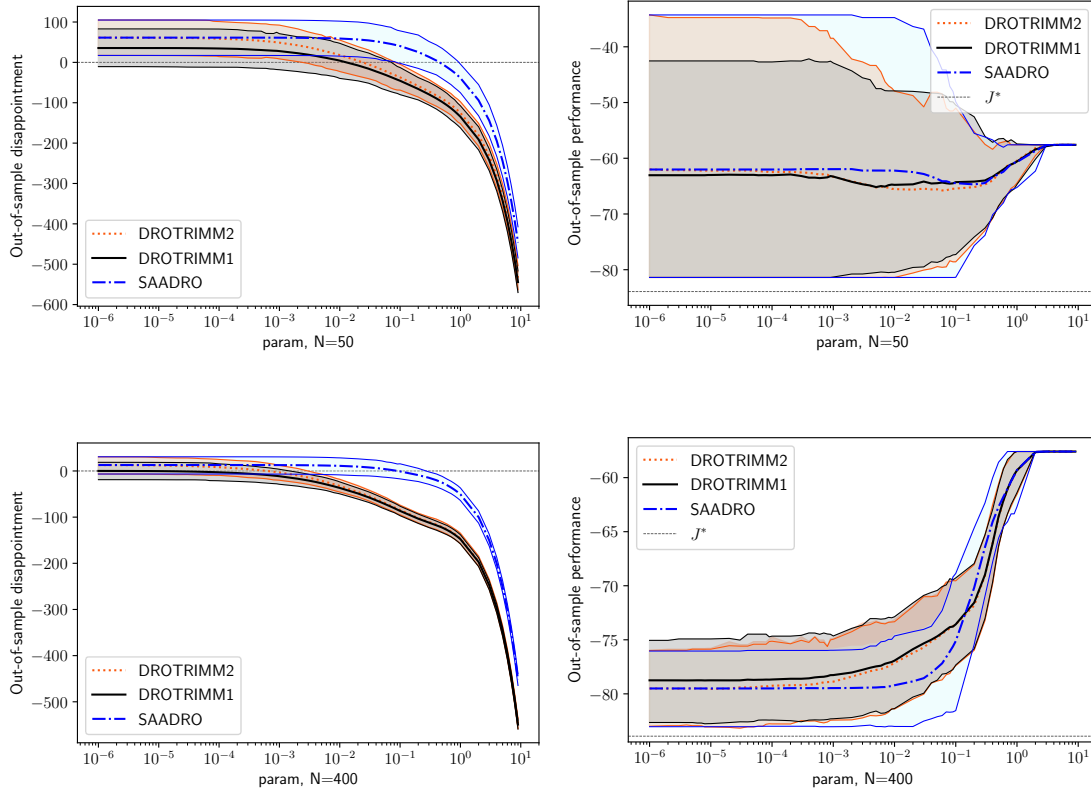


Figure 4.7: Case  $\alpha > 0$ , impact of the robustness parameter with 200 training samples and  $\delta = 0.5, \lambda = 0.1$

To further support this finding, we conclude this section with Figure 4.8, which is similar to Figure 4.6. However, Figure 4.8 has been obtained through a different experiment, in which the value of the robustness parameter that each method uses has been *optimally* selected from the previously indicated discrete set. In other words, the results shown in that figure are those a decision-maker would obtain in the hypothetical case that the true conditional distribution  $\mathbb{Q}_{\Xi}$  could be used to tune the robustness parameters of the DRO methods. Therefore, these results correspond to the best solutions that can be obtained from SAADRO, DROTRIMM1 and DROTRIMM2, and confirm that our approaches (especially, DROTRIMM2) can potentially identify portfolios that significantly outperform those delivered by SAADRO under the same reliability requirement.

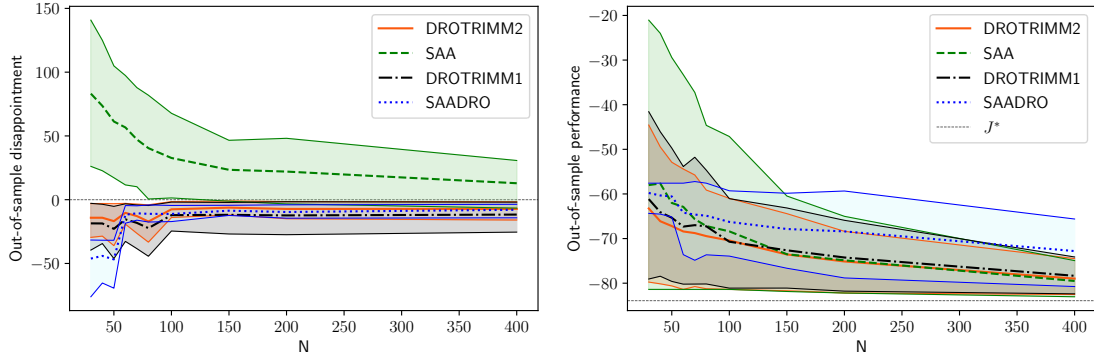


Figure 4.8: Portfolio problem with features: Performance metrics under an optimal selection of the robustness parameters. Case  $\alpha > 0$  and  $\delta = 0.5$ ,  $\lambda = 0.1$

## 4.4 Application III. Optimal Power Flow problem

This section discusses a realistic application of the proposed methodology taken from the realm of Power Systems, namely, the *Optimal Power Flow problem* with side/contextual information. The content of this section is, therefore, based on our preprint [51]. This application requires a more extensive treatment and elaboration than the two applications presented above for the following two reasons at least: (i) The Optimal Power Flow problem is modeled as a chance-constrained program, and hence, it does not fall into the standard conditional stochastic program that is introduced in Section 4.1; (ii) The Optimal Power Flow problem translates into a much more intricate optimization model that requires more notation and a more careful presentation than in the cases of the newsvendor and the portfolio allocation problems considered before.

### 4.4.1 Introduction

The Optimal Power Flow (OPF) is a fundamental problem in power system operations. Traditionally, the goal of the OPF problem is to minimize the cost of the power generation dispatch that supplies the electricity demand while complying with some physical and engineering constraints. The growing penetration of electricity generation sources like wind and solar power, which are intrinsically uncertain, has led power system engineers to account for randomness in OPF analyses. Hence, the OPF is to be formulated today as an optimization problem *under uncertainty*.

A common way to cope with uncertainty in the constraints of an optimization problem and, in particular, of an OPF model is by way of the so-called *chance constraints*, which allow the modeler to impose the constraint satisfaction with a certain probability only. Accordingly, chance-constrained Optimal Power Flow models (CC-OPF) have

been developed to control the violation probability of, for instance, line and generation capacity limits. In particular, references [80, 113, 134] consider *joint* chance constraints, by which the system operator enforces that all constraints must simultaneously hold with a probability greater than or equal to  $1 - \epsilon$ , where  $\epsilon \in (0, 1)$  is a pre-fixed acceptable tolerance of dispatch infeasibility. This is opposed to *single* chance constraints, whereby the probability of constraint satisfaction is imposed on each constraint of the OPF separately. Although single chance constraints have been considered in the technical literature of CC-OPF models due to their attractive tractability properties (see, e.g., [25, 96, 141] and references therein), satisfying all multiple individual constraints does not provide strong guarantees on the security of the entire power system, and leads to costly and over-conservative dispatch solutions to achieve joint feasibility [113].

In any case, one of the main challenges in solving CC-OPF problems is that the underlying probability distribution of the random variables affecting the OPF constraints is generally unknown. In fact, in practice, only past historical observations of those variables are available to the system operator. No wonder, therefore, that a variety of distributionally robust chance-constrained Optimal Power Flow (DRCC-OPF) models has been proposed in recent years. These models seek the optimal dispatch such that all the model constraints are satisfied with a pre-fixed confidence level for all the probability distributions within the ambiguity set built either by means of the Wasserstein metric ([6, 7, 108, 144]), using moments ([88, 90, 91, 137]), or by way of a discrete probability distribution with probability masses and locations varying within a box [78].

Unfortunately, the consideration of *joint* chance constraints in a DRO framework ([73, 138]) renders an intractable optimization problem in general. For this reason, researchers have considered DRO-OPF models based on the well-known conservative, but far more tractable approximation given by the concept of *Conditional-Value-at-Risk* (**CVaR**), see, for example, [6, 78, 108] and references therein. The authors in [37] show that the **CVaR** offers a tight convex approximation of the chance constraints under the DRO framework, which justifies its popularity. On top of that, by way of **CVaR**-based chance-constraints, the power system operator can control not only the violation probability, but also the violation magnitude, which can be important from the standpoint of power system operations.

Yet another key issue is that the system operator is able to not only tune the robustness of the resulting chance-constrained DRO model by adjusting the probability of constraint violation, but also, and very importantly, through the specificity degree of the ambiguity set. In this regard, some previous approaches have focused on producing more meaningful ambiguity sets by incorporating structural information on the underlying true probability distribution (see, e.g., [6, 7, 88, 90, 91]). More specifically, the authors in [6] introduce a DRCC-OPF model with single chance constraints, which considers all distributions within a Wasserstein ball that conform to a given copula-

based dependence structure among the random variables. The authors in [7] propose an iterative algorithm for solving a bilinear exact reformulation of a DRCC-OPF model with single chance constraints considering some support information. The authors of [88] and [90] provide and study DRCC-OPF models where some moment and unimodality information is included in the ambiguity set. Both aforementioned approaches are extended in [91] to allow for misspecified modes.

Ideally, one would like to have the smallest ambiguity set that contains the true data-generating distribution. In this vein, if we have some information on the true distribution, we should use it to discard all those other distributions that do not conform with that information from the ambiguity set. As mentioned already in this thesis, that information can be, for example, some dependence structure via copulas [6], support information [7] or shape information (such as unimodality) [90, 91]. Our aim is also to leverage a more informed ambiguity set, but, for the first time to our knowledge, that information refers to a given *context*. As discussed in this chapter, this side/contextual information is related to outcomes of random variables that may have predictive power on the OPF's uncertainties. Accordingly, we make use of an ambiguity set that accounts for the possible dependence between the uncertainties and these explanatory variables. Thus, the contextual information allows us to discard implausible distributions.

More specifically, in the work we present here, we exploit the contextual information provided by the *point forecasts* of those uncertainties. Indeed, it is well known in the energy forecasting community that the power forecast error of a wind farm highly depends on the wind power forecast itself [31, 56]. Within the context of DRCC-OPF, this means that the wind power point forecast constitutes valuable information to build a proper ambiguity set for the wind power forecast error.

To our knowledge, this work is thus the first to tackle a chance constraint system with a distributionally robust approach *that accounts for contextual information*.

The reader is referred to Appendix C.1 for a complete list of the notation we use to formulate the DRCC-OPF model we introduce next.

#### 4.4.2 DC-OPF under uncertainty: Mathematical Formulation

The DC-OPF problem under uncertainty is formulated here as a distributionally robust version of the *joint* chance-constrained DC-OPF model described in [113], where we have also accounted for the procurement of reserve capacity and its associated cost, as in [90, 91]. Nonetheless, unlike in [113], where the generators' cost functions are assumed to be quadratic, here we model those costs as convex piecewise linear functions. Furthermore, there exists a number of different variants of the distributionally robust chance-constrained DC-OPF problem (e.g., [7, 78, 90, 108, 144]), which essentially differ in the treatment of the chance constraints (single vs. joint), the cost structure of generators that is assumed, and the ambiguity set used. What makes our formulation



unique among those variants is its ability to exploit contextual information.

### Variables and constraints

Consider a power system with a set  $\mathcal{L}$  of transmission lines, a set  $\mathcal{B}$  of buses, a set  $\mathcal{W}$  of wind power plants (or, more generally, weather-dependent renewable generators), and a set  $\mathcal{G}$  of conventional generators (i.e., dispatchable units that are not weather-dependent). For ease of formulation, power loads are assumed to be deterministic. Next we introduce each of the main components of the DC-OPF problem.

1. *Wind power plants.* For each wind power plant  $m \in \mathcal{W}$ , the random power output is modeled as  $f_m + \omega_m$ , where  $f_m$  is the predicted power output and  $\omega_m$  is the (random) wind forecast error at wind power plant  $m$ . We denote the system-wise aggregate wind power forecast error as  $\Omega$ , i.e.,  $\Omega := \sum_{m \in \mathcal{W}} \omega_m = \langle \mathbf{1}, \boldsymbol{\omega} \rangle$ . Let  $\mathbf{f} := (f_m)_{m \in \mathcal{W}}, \boldsymbol{\omega} := (\omega_m)_{m \in \mathcal{W}}$  be the array of predicted power outputs and wind power prediction errors, respectively.
2. *Generators:* For each  $j \in \mathcal{G}$ , the actual power output of generator  $j$ ,  $\tilde{g}_j(\boldsymbol{\omega})$ , is expressed as the sum of the scheduled generation,  $g_j$ , and the (random) adjusted power  $\tilde{r}_j(\boldsymbol{\omega})$  (also known as *deployed reserve*). As customary, we assume an affine control policy to counterbalance the wind forecast errors by deploying generators' reserves [25], that is,

$$\tilde{g}_j(\boldsymbol{\omega}) := g_j + \tilde{r}_j(\boldsymbol{\omega}) = g_j - \beta_j \Omega = g_j - \beta_j \langle \mathbf{1}, \boldsymbol{\omega} \rangle, \quad \forall j \in \mathcal{G} \quad (4.28)$$

where  $\beta_j$  is the participation factor of generator  $j$ . Denote by  $\boldsymbol{\beta} := (\beta_j)_{j \in \mathcal{G}}, \mathbf{g} := (g_j)_{j \in \mathcal{G}}$  the array of non-negative participation factors and scheduled generation, respectively. Let  $\tilde{\mathbf{g}}(\boldsymbol{\omega}) := (\tilde{g}_j(\boldsymbol{\omega}))_{j \in \mathcal{G}}, \tilde{\mathbf{r}}(\boldsymbol{\omega}) := (\tilde{r}_j(\boldsymbol{\omega}))_{j \in \mathcal{G}} = (-\beta_j \langle \mathbf{1}, \boldsymbol{\omega} \rangle)_{j \in \mathcal{G}}$  be the array of actual power outputs and deployed reserves, in that order.

The following constraints determine the provision of reserve capacities:

$$-\mathbf{r}^D \leq \tilde{\mathbf{r}}(\boldsymbol{\omega}) \leq \mathbf{r}^U \quad (4.29)$$

with  $\mathbf{r}^D, \mathbf{r}^U$  being the arrays of downward and upward reserve capacity provided by the generators, respectively.

Naturally, the following technical constraints, which link the generation dispatches and the provision of reserve capacities, must hold:

$$\mathbf{g} + \mathbf{r}^U \leq \mathbf{g}^{\max}, \quad (4.30)$$

$$\mathbf{g} - \mathbf{r}^D \geq \mathbf{g}^{\min} \quad (4.31)$$



where  $\mathbf{g}^{\min}, \mathbf{g}^{\max}$  are the arrays of minimum and maximum power output of the generators, respectively.

3. *Network constraints.* The total power generation must equal the total system demand (*power balance constraint*), that is,

$$\langle \mathbf{1}, \tilde{\mathbf{g}}(\boldsymbol{\omega}) \rangle + \langle \mathbf{1}, \mathbf{f} + \boldsymbol{\omega} \rangle = \langle \mathbf{1}, \mathbf{L} \rangle \quad (4.32)$$

where  $\mathbf{L} := (L_b)_{b \in \mathcal{B}}$  denotes the array of nodal loads. Using (4.28), Eq. (4.32) is equivalent to:

$$\langle \mathbf{1}, \mathbf{g} \rangle + \langle \mathbf{1}, \mathbf{f} \rangle = \langle \mathbf{1}, \mathbf{L} \rangle \quad (4.33)$$

$$\langle \mathbf{1}, \boldsymbol{\beta} \rangle = 1, \quad \boldsymbol{\beta} \geq \mathbf{0} \quad (4.34)$$

which guarantee the power balance both in the dispatch and the real-time stages, respectively.

Finally, we assume that the power flow through the lines is given by a linear function of the nodal power injections, that is,  $\mathbf{M}^{\mathcal{G}}(\tilde{\mathbf{g}}(\boldsymbol{\omega})) + \mathbf{M}^{\mathcal{W}}(\mathbf{f} + \boldsymbol{\omega}) - \mathbf{M}^{\mathcal{B}}\mathbf{L}$ , based on the DC power flow approximation, where  $\mathbf{M}^{\mathcal{G}}, \mathbf{M}^{\mathcal{W}}$  and  $\mathbf{M}^{\mathcal{B}}$  are the matrix for generators, wind plants and loads given by the DC power transfer distribution factors [128], in that order. Hence, the constraints

$$-\mathbf{Cap} \leq \mathbf{M}^{\mathcal{G}}(\tilde{\mathbf{g}}(\boldsymbol{\omega})) + \mathbf{M}^{\mathcal{W}}(\mathbf{f} + \boldsymbol{\omega}) - \mathbf{M}^{\mathcal{B}}\mathbf{L} \leq \mathbf{Cap} \quad (4.35)$$

enforce the transmission capacity limits where  $\mathbf{Cap} := (\text{Cap}_\ell)_{\ell \in \mathcal{L}}$  denotes the array of transmission line capacities.

### Dealing with uncertainty in the DC-OPF problem: Joint chance constraints, Distributionally Robust Optimization and contextual information

In practice, it is often the case that the random vector of forecast errors  $\boldsymbol{\omega}$  shows some statistical dependence on some features/covariates, which we can model, in general, by some random vector  $\mathbf{z}$ . In fact, the forecast wind power output  $\mathbf{f}$  serves in itself as an obvious explanatory random vector for the subsequent forecast error  $\boldsymbol{\omega}$ . In this approach, we want to exploit this statistical dependence to identify a better power generation dispatch and provision of reserve capacity.

Let  $\mathbf{z} := (z_m)_{m \in \mathcal{W}}$  be the random vector modeling the features and let  $\mathbb{Q}$  be the probability measure of the joint distribution of  $\mathbf{z}$  and  $\boldsymbol{\omega}$ , which is supported on  $\Xi$ . For convenience, we define  $\boldsymbol{\xi} := (\mathbf{z}, \boldsymbol{\omega})$ . Given the array of forecast wind power outputs,  $\mathbf{f} := (f_m)_{m \in \mathcal{W}}$ , set the contextual information  $\boldsymbol{\xi} := (\mathbf{z}, \boldsymbol{\omega}) \in \tilde{\Xi}$  defined by the event  $(\mathbf{z} = \mathbf{f}; \boldsymbol{\omega} \in \tilde{\Xi}_{\boldsymbol{\omega}})$ , with  $\tilde{\Xi}_{\boldsymbol{\omega}}$  being the support of  $\boldsymbol{\omega}$  conditional on  $\mathbf{z} = \mathbf{f}$ . The errors

of forecasting the power output of a wind farm are naturally bounded. Their lower bound is the forecast value itself, while their upper bound is given by the difference of the capacity of the wind farm and the predicted value. Therefore,  $\tilde{\Xi}_\omega$  is the hypercube  $\prod_{m \in \mathcal{W}} [-f_m, \bar{C}_m - f_m]$ , where  $\bar{C}_m$  represents the capacity of wind farm  $m$ . Note that the optimal dispatch is, therefore, parametrized on the predicted wind power outputs  $\{f_m\}_{m \in \mathcal{W}}$ .

In real life, however, neither the joint distribution  $\mathbb{Q}$ , nor the conditional one  $\mathbb{Q}_{\omega/\mathbf{z}=\mathbf{f}}$ , are known. The system operator only has access to a finite set of samples of size  $N$  (i.e. the training set) of the true joint distribution  $\mathbb{Q}$ , which we denote as  $\hat{\Xi}_\omega^N := \{\hat{\xi}_i\}_{i=1}^N = \{(\hat{\mathbf{z}}_i, \hat{\omega}_i)\}_{i=1}^N$ . In our context,  $\hat{\Xi}_\omega^N$  is made up of  $N$  past observations of the predicted wind power outputs and their associated errors. Hence, the system operator needs to infer or construct a proxy of  $\mathbb{Q}_{\omega/\mathbf{z}=\mathbf{f}}$  from the sample  $\hat{\Xi}_\omega^N$ , so that this proxy can be used to compute a reliable and cost-efficient OPF solution. However, the limited information that  $\hat{\Xi}_\omega^N$  conveys on  $\mathbb{Q}_{\omega/\mathbf{z}=\mathbf{f}}$  makes this inference process ambiguous, and as such, we propose employing the following *distributionally robust* chance-constrained OPF model to protect the system operator's decision against this ambiguity:

$$\min_{\mathbf{x} \in X} \sup_{Q_{\Xi} \in \hat{\mathcal{U}}} \mathbb{E}_{Q_{\Xi}} [C(\tilde{\mathbf{g}}(\omega)) + \langle \mathbf{c}^D, \mathbf{r}^D \rangle + \langle \mathbf{c}^U, \mathbf{r}^U \rangle] \quad (4.36)$$

$$\text{s.t.} \quad \inf_{Q_{\Xi} \in \hat{\mathcal{U}}} Q_{\Xi} \left( \begin{array}{c} -\mathbf{r}^D \leq \tilde{\mathbf{r}}(\omega) \leq \mathbf{r}^U \\ -\mathbf{Cap} \leq \mathbf{M}^G(\tilde{\mathbf{g}}(\omega)) + \mathbf{M}^W(\mathbf{f} + \omega) - \mathbf{M}^B \mathbf{L} \leq \mathbf{Cap} \end{array} \right) \geq 1 - \epsilon \quad (4.37)$$

where  $X$  stands for the deterministic feasible set for the array of decision variables  $\mathbf{x} = (\mathbf{g}, \beta, \mathbf{r}^D, \mathbf{r}^U)$  defined by the constraints (4.30), (4.31), (4.33) and (4.34).

The set  $\hat{\mathcal{U}}$  in (4.36)-(4.37) represents an ambiguity set for the true conditional distribution  $\mathbb{Q}_{\omega/\mathbf{z}=\mathbf{f}}$ . Here we use the ambiguity set based on probability trimmings and optimal transport introduced in Section 4.1.2 (taking  $p = 1$ ),  $\hat{\mathcal{U}}_N(\alpha, \rho)$ , which allows us to exploit the side information within a DRO framework in a fully data-driven sense.

Objective function (4.36) minimizes the expected total operational cost over the worst-case distribution from the ambiguity set  $\hat{\mathcal{U}}$ , including the sum of the (random) total generation cost,  $C(\tilde{\mathbf{g}}(\omega))$ , and the (deterministic) cost of providing up- and down-reserve capacities,  $\langle \mathbf{c}^D, \mathbf{r}^D \rangle + \langle \mathbf{c}^U, \mathbf{r}^U \rangle$ . The total generation cost function  $C(\cdot)$  is assumed to be given by the sum of  $|\mathcal{G}|$  convex piecewise linear cost functions with  $S_j$  pieces/blocks, i.e.,  $C(\tilde{\mathbf{g}}(\omega)) := \sum_{j \in \mathcal{G}} \max_{s=1, \dots, S_j} \{m_{js} \tilde{g}_j(\omega) + n_{js}\}$ , where  $m_{js}, n_{js}$  stand for the slope and the intercept of the  $s$ -th piece for generator  $j$ , respectively. Parameters  $\mathbf{c}^D, \mathbf{c}^U$  are the arrays of downward and upward reserve procurement cost of the generators, respectively. Note that the wind power production cost is assumed to be zero. Finally, constraint (4.37) establishes a tolerance  $\epsilon$  in terms of the joint violation probability of the OPF constraints under any conditional probability distribution  $Q_{\Xi}$  for  $\omega$  (given some contextual information) within the ambiguity set  $\hat{\mathcal{U}}$ .

**Remark 4.8.** *In order to account for contextual information in a DRO framework, the authors in [19] propose an ambiguity set  $\widehat{\mathcal{U}}$  different to the one based on probability trimmings (i.e.,  $\widehat{\mathcal{U}}_N(\alpha, \rho)$ ) that we advocate here. More specifically, the ambiguity set they suggest is a Wasserstein ball centered at the discrete distribution supported on the  $\widehat{\omega}$ -coordinates of the  $K$  data points from the sample  $\widehat{\Xi}_\omega^N$  that are the closest to  $\widetilde{\Xi}$ . This results in the data-driven decision-making model KNNDRO that we have described and analyzed in the previous two applications. For completeness, in Appendix C.5, we use an example based on a small three-node system to illustrate that our trimmings-based ambiguity set also delivers better dispatch solutions in terms of expected cost and system reliability than the one introduced in [19] (i.e., KNNDRO) for the DC-OPF problem under uncertainty.*

In the next section, we introduce a tractable and conservative approximation of the distributionally robust joint chance constraints (4.37) using the notion of Conditional Value-at-Risk (**CVaR**). As previously mentioned, the use of the **CVaR** to this end in the context of the chance-constrained distributionally robust OPF is very popular in the technical literature (see, for example, the recent publications [78] and [108]).

#### 4.4.3 A tractable and conservative CVaR-based approximation of the distributionally robust joint chance constraints

The distributionally robust joint chance constraints (4.37) can be written equivalently as the following distributionally robust single chance constraint, where we have already replaced the generic  $\widehat{\mathcal{U}}$  with the ambiguity set based on probability trimmings  $\widehat{\mathcal{U}}_N(\alpha, \rho)$ :

$$\inf_{Q_{\Xi} \in \widehat{\mathcal{U}}_N(\alpha, \rho)} Q_{\Xi} \left( \max_{k \leq K} \langle \mathbf{a}_{1k}, \boldsymbol{\omega} \rangle + a_{2k} \leq 0 \right) \geq 1 - \epsilon \quad (4.38)$$

Functions  $\langle \mathbf{a}_{1k}, \boldsymbol{\omega} \rangle + a_{2k}$ ,  $k \leq K$ , represent the OPF constraints involved in the joint chance constraint (4.37) expressed as inequalities lower than or equal to zero. These constraints are all linear with respect to the random vector  $\boldsymbol{\omega}$ .

In practice, the system operator is not only concerned about the joint violation of the OPF constraints, but also about the magnitude of this violation. Indeed, the joint chance constraint (4.38) does not offer guarantees on how positive  $\max_{k \leq K} \langle \mathbf{a}_{1k}, \boldsymbol{\omega} \rangle + a_{2k}$  is. This is one of the main reasons to adopt a risk-averse approach to handle the joint chance constraint via the well-known concept of Conditional-Value-at-Risk (**CVaR**), which quantifies the conditional expectation of  $\max_{k \leq K} \langle \mathbf{a}_{1k}, \boldsymbol{\omega} \rangle + a_{2k}$  on its right  $\epsilon$ -tail.

In lieu of (4.38), we consider the following tractable (convex) approximation:

$$\sup_{Q_{\Xi} \in \widehat{\mathcal{U}}_N(\alpha, \rho)} Q_{\Xi} - \mathbf{CVaR}_{\epsilon} \left( \max_{k \leq K} \langle \mathbf{a}_{1k}, \boldsymbol{\omega} \rangle + a_{2k} \right) \leq 0 \quad (4.39)$$

which is, in addition, conservative, because (4.39) implies (4.38).

Constraint (4.39) can be equivalently reformulated as follows [108], [118]:

$$\inf_{\tau \in \mathbb{R}} \left\{ \tau + \frac{1}{\epsilon} \sup_{Q_{\Xi} \in \hat{\mathcal{U}}_N(\alpha, \rho)} \mathbb{E}_{Q_{\Xi}} \left[ \left( \max_{k \leq K} \langle \mathbf{a}_{1k}, \boldsymbol{\omega} \rangle + a_{2k} - \tau \right)^+ \right] \right\} \leq 0 \quad (4.40)$$

The next proposition states a tractable reformulation of (4.40). For ease of exposition, we first need to recast function  $(\max_{k \leq K} \langle \mathbf{a}_{1k}, \boldsymbol{\omega} \rangle + a_{2k} - \tau)^+$  as

$$\left( \max_{k \leq K} \langle \mathbf{a}_{1k}, \boldsymbol{\omega} \rangle + a_{2k} - \tau \right)^+ := \max_{k \leq K+1} \langle \mathbf{a}_{1k}, \boldsymbol{\omega} \rangle + a'_{2k} \quad (4.41)$$

where  $a'_{2k} = a_{2k} - \tau$  for  $k \leq K$ ,  $\mathbf{a}_{1K+1} = \mathbf{0}$  and  $a'_{2K+1} = 0$ .

**Proposition 4.4 (Reformulation of the CVaR-based distributionally robust joint chance constraints).** *Set  $\alpha > 0$ . Then, for any value of  $\rho \geq \underline{\epsilon}_{N\alpha}$ , the CVaR-based distributionally robust joint chance constraints defined by (4.39) can be equivalently reformulated as follows:*

$$\inf_{\tau \in \mathbb{R}, \lambda_2 \geq 0, \mu_i \geq 0, \theta_2 \in \mathbb{R}, \boldsymbol{\gamma}_{ik}, \mathbf{v}_{ik}} \left\{ \tau + \frac{1}{\epsilon} \left[ \lambda_2 \rho + \theta_2 + \frac{1}{N\alpha} \sum_{i=1}^N \mu_i \right] \right\} \leq 0 \quad (4.42a)$$

$$\begin{aligned} \text{s.t. } \mu_i + \theta_2 + \lambda_2 \|\mathbf{z}^* - \hat{\mathbf{z}}_i\|_1 &\geq a'_{2k} + S_{\Xi_{\omega}}(\mathbf{v}_{ik}) \\ &- \langle \boldsymbol{\gamma}_{ik}, \hat{\boldsymbol{\omega}}_i \rangle, \forall i \leq N, \forall k \leq K+1 \end{aligned} \quad (4.42b)$$

$$\boldsymbol{\gamma}_{ik} - \mathbf{v}_{ik} = -\mathbf{a}_{1k}, \forall i \leq N, \forall k \leq K+1 \quad (4.42c)$$

$$\|\boldsymbol{\gamma}_{ik}\|_{\infty} \leq \lambda_2, \forall i \leq N, \forall k \leq K+1 \quad (4.42d)$$

where  $S_{\Xi_{\omega}}(\cdot)$  stands for the support function of  $\Xi_{\omega}$  and  $\langle \cdot, \cdot \rangle$  represents the dot product (see Appendix C.1.5).

Once we have reformulated the CVaR-based distributionally robust joint chance constraint (4.39), we only need to reformulate the DRO problem defined by the inner supremum in (4.36). Since this requires a careful and independent analysis, we consider it in the following section.

#### 4.4.4 An exact tractable reformulation of the worst-case expected cost

In what follows, we provide an exact and tractable reformulation of the objective function (4.36) as a continuous linear program.

The term

$$C(\tilde{\mathbf{g}}(\boldsymbol{\omega})) = \sum_{j \in \mathcal{G}} \max_{s=1, \dots, S_j} \left\{ m_{js} [g_j - \beta_j \Omega] + n_{js} \right\} \quad (4.43)$$

is a sum of a maximum of *univariate* linear functions in terms of  $\Omega$ , which is, moreover, convex in  $\Omega$ . This observation is key to reformulating (4.36) in a tractable way. In fact, the ambiguity set  $\hat{\mathcal{U}}_N(\alpha, \rho)$  for the worst-case probability distribution in the inner supremum of (4.36) can be equivalently replaced with the following one, which is also expressed in terms of  $\Omega$  only:

$$\hat{\mathcal{U}}_N^\Omega(\alpha, \rho) := \{P_{\Xi_\Omega} : W_1(\mathcal{R}_{1-\alpha}(\hat{\mathbb{P}}_N), P_{\Xi_\Omega}) \leq \rho, P_{\Xi_\Omega}(\Xi_\Omega) = 1\} \quad (4.44)$$

where  $\hat{\mathbb{P}}_N := \frac{1}{N} \sum_{i=1}^N \delta_{(\hat{\mathbf{z}}_i, \hat{\Omega}_i)}$  is the empirical distribution supported on the samples  $(\hat{\mathbf{z}}, \hat{\Omega}_i), i = 1, \dots, N$ ; and  $\Xi_\Omega$  stands for the event

$$(\mathbf{z} = \mathbf{f}; \Omega \in [\underline{\Omega}, \bar{\Omega}]), \text{ with } [\underline{\Omega}, \bar{\Omega}] = \left[ -\sum_{m \in \mathcal{W}} f_m, \sum_{m \in \mathcal{W}} (\bar{C}_m - f_m) \right]$$

The interval  $[\underline{\Omega}, \bar{\Omega}]$  is the conditional support for the random variable  $\Omega$  (that is, the support set for the system-wise aggregate wind power forecast error, given the predicted power outputs of the wind farms). Essentially, what we have done above is to map the original probability space for the random vector  $(\mathbf{z}, \boldsymbol{\omega})$  onto a new probability space for the random vector  $(\mathbf{z}, \Omega)$  by the linear map  $\boldsymbol{\omega} \mapsto \sum_{m \in \mathcal{W}} \omega_m$ ,  $\Omega = \sum_{m \in \mathcal{W}} \omega_m$ , which leaves the objective cost function unaltered. In doing so, the inner supremum in (4.36) can be fully recast in terms of  $\Omega$  only as follows:

$$\sup_{P_{\Xi_\Omega} \in \hat{\mathcal{U}}_N^\Omega(\alpha, \rho)} \mathbb{E}_{P_{\Xi_\Omega}} \left[ \sum_{j \in \mathcal{G}} \max_{s=1, \dots, S_j} \left\{ m_{js} [g_j - \beta_j \Omega] + n_{js} \right\} + \langle \mathbf{c}^D, \mathbf{r}^D \rangle + \langle \mathbf{c}^U, \mathbf{r}^U \rangle \right] \quad (4.45)$$

The proposition below presents a tractable reformulation of (4.45) as a continuous linear program.

**Proposition 4.5 (LP reduction of the worst-case expected cost).** *Set  $\alpha > 0$  and assume that  $\|(\mathbf{z}, \Omega)\| := \|\mathbf{z}\| + |\Omega|$  for some norm  $\|\cdot\|$  in  $\mathbb{R}^{d_z}$ . Then, for any value of  $\rho \geq \epsilon_{N\alpha}$ , the DRO problem defined by (4.45) can be reformulated as the following continuous linear program:*

$$\inf_{\lambda \geq 0; \theta \in \mathbb{R}; \bar{\mu}_i, t_i, \underline{t}_{ij}, \bar{t}_{ij}, \hat{t}_{ij} \forall i \leq N, \forall j \in \mathcal{G}} \lambda \rho + \theta + \frac{1}{N\alpha} \sum_{i=1}^N \bar{\mu}_i + \langle \mathbf{c}^D, \mathbf{r}^D \rangle + \langle \mathbf{c}^U, \mathbf{r}^U \rangle \quad (4.46a)$$

$$s.t. \bar{\mu}_i + \theta + \lambda \|\mathbf{z}^* - \hat{\mathbf{z}}_i\| \geq t_i, \forall i \leq N \quad (4.46b)$$

$$t_i \geq \sum_{j \in \mathcal{G}} \underline{t}_{ij} - \lambda(\underline{\Omega} - \hat{\Omega}_i), \forall i \in \underline{I} \quad (4.46c)$$

$$t_i \geq \sum_{j \in \mathcal{G}} \bar{t}_{ij} - \lambda(\bar{\Omega} - \hat{\Omega}_i), \forall i \in \bar{I} \quad (4.46d)$$

$$t_i \geq \sum_{j \in \mathcal{G}} \underline{t}_{ij} + \lambda(\underline{\Omega} - \widehat{\Omega}_i), \quad \forall i \in \bar{I} \quad (4.46e)$$

$$t_i \geq \sum_{j \in \mathcal{G}} \bar{t}_{ij} + \lambda(\bar{\Omega} - \widehat{\Omega}_i), \quad \forall i \in \bar{I} \quad (4.46f)$$

$$t_i \geq \sum_{j \in \mathcal{G}} \bar{t}_{ij} - \lambda(\bar{\Omega} - \widehat{\Omega}_i), \quad \forall i \in I \quad (4.46g)$$

$$t_i \geq \sum_{j \in \mathcal{G}} \underline{t}_{ij} + \lambda(\underline{\Omega} - \widehat{\Omega}_i), \quad \forall i \in I \quad (4.46h)$$

$$t_i \geq \sum_{j \in \mathcal{G}} \widehat{t}_{ij}, \quad \forall i \in I \quad (4.46i)$$

$$\bar{\mu}_i \geq 0, \quad \forall i \leq N \quad (4.46j)$$

$$\underline{t}_{ij} \geq m_{js} [g_j - \beta_j \underline{\Omega}] + n_{js}, \quad \forall i \leq N, \quad \forall j \in \mathcal{G}, \quad \forall s \leq S_j \quad (4.46k)$$

$$\bar{t}_{ij} \geq m_{js} [g_j - \beta_j \bar{\Omega}] + n_{js}, \quad \forall i \leq N, \quad \forall j \in \mathcal{G}, \quad \forall s \leq S_j \quad (4.46l)$$

$$\widehat{t}_{ij} \geq m_{js} [g_j - \beta_j \widehat{\Omega}_i] + n_{js}, \quad \forall i \leq N, \quad \forall j \in \mathcal{G}, \quad \forall s \leq S_j \quad (4.46m)$$

where  $\underline{I} := \{i \in \{1, \dots, N\} : \widehat{\Omega}_i < \underline{\Omega}\}$ ,  $I := \{i \in \{1, \dots, N\} : \widehat{\Omega}_i \in [\underline{\Omega}, \bar{\Omega}]\}$ , and  $\bar{I} := \{i \in \{1, \dots, N\} : \widehat{\Omega}_i > \bar{\Omega}\}$ .

#### 4.4.5 Numerical results

In this section, we present and discuss results from a series of numerical experiments that have been run on a modified version of the IEEE 118-bus system considered in [78]. All the data and codes needed to reproduce those experiments are available for download in the GITHUB repository [52]. The experiments have been carried out on a Linux-based server using up to 13200 CPUs running in parallel, each clocking at 2.6 GHz with 200 GB of RAM. We have employed CPLEX 20.1.0 under DOpplex Python Modeling API to solve the associated continuous linear programs with the barrier algorithm using up to 22 threads. In addition, we have set the CPLEX parameter `preprocessing_dual` to 1.

We solve the CC-DRO OPF problem (4.36)–(4.37) using the **CVaR**-based approximation stated in Section 4.4.3, but with different ambiguity sets, namely: (i) The ambiguity set based on probability trimmings, introduced in [53], which we refer to as DROTRIMM; and (ii) a Wasserstein ball centered at the empirical distribution supported on the  $\widehat{\omega}$ -coordinates of the  $N$  samples in  $\widehat{\Xi}_\omega^N$ , i.e.,  $\{(\widehat{\mathbf{z}}_i, \widehat{\omega}_i)\}_{i=1}^N$ . This leads to the distributionally robust chance-constrained OPF model proposed in [108], which we call DROW. Importantly, this is a DRO model that fully ignores the contextual information, since the center of the Wasserstein ball it uses is made up of *all* past samples of wind power forecast errors (regardless of the current wind power point predictions). Roughly speaking, DROTRIMM also works with all the past  $N$  samples of wind power forecast errors, but only those that lead to the worst-case conditional distribution of

the prediction errors are moved onto the conditional support. However, this movement must entail a transportation cost smaller than or equal to a given budget  $\rho$  and the computation of that cost is directly contingent on the current context (that is, the current wind power point forecasts).

In addition, we benchmark the previous two distributionally robust methods with an alternative approach that is commonly used in the technical literature for solving optimization problems with probabilistic constraints, known as the *scenario approach*, but adapted to account for contextual information. We have taken the required adaptation from [86, Chapter 4], which, in our setting, involves solving a DC-OPF problem in which the uncertain constraints are enforced for the wind power forecast errors associated with the  $K$  samples nearest to the context. We refer to this adaptation of the popular scenario approach as SCENA.

The training data consist of a set of pairs  $\{(\widehat{\mathbf{z}}_i, \widehat{\boldsymbol{\omega}}_i)\}_{i=1}^N$ , from which we can directly obtain the collection of pairs  $\{(\widehat{\mathbf{z}}_i, \widehat{\Omega}_i)\}_{i=1}^N$ , where  $\widehat{\Omega}_i := \sum_{m \in \mathcal{W}} \widehat{\omega}_{i,m}$ . For ease of computation and to simplify the analysis below, we have considered the same radius or transportation budget for the two ambiguity sets in both the objective and the chance constraints of the DRO OPF problem (4.36)–(4.37).

#### 4.4.6 Evaluation of the out-of-sample performance via re-optimization

Given a context (in the form of point forecasts of the power outputs of the wind farms), a training dataset, and a robustness parameter  $\rho$ , each method *met* (either DROTRIMM or DROW in our case) provides a forward generation dispatch and reserve capacity provision  $\mathbf{y}^{met} := (\mathbf{g}, \mathbf{r}^D, \mathbf{r}^U)$ . To evaluate the actual or out-of-sample performance of that  $\mathbf{y}^{met}$ , we draw a sample of wind power forecast errors  $\widehat{\boldsymbol{\omega}}_j$  from a test dataset, and the vector of recourse variables  $\mathbf{r}$  (that is, the real-time power adjustments) is computed by solving the deterministic Optimal Power Flow available in Appendix C.4. In this deterministic OPF problem, wind spillage (with a cost equal to 0) and involuntary load curtailment (with a cost equal to \$500/MWh) are considered as feasible recourse actions, aside from the deployment of reserves by generators. In this way, the *out-of-sample performance* of a method *met*, which produces the forward dispatch  $\mathbf{y}^{met}$ ,  $J(\mathbf{y}^{met})$ , is computed by the empirical out-of-sample cost averaged over the test set formed by a certain number of samples of  $\mathbb{Q}_{\boldsymbol{\omega}/\mathbf{z}=\mathbf{f}}$ . In addition, in order to measure the *reliability* of a solution (that is, if  $\mathbf{y}^{met}$  is feasible or not in real time), the *violation probability* is estimated over the test set. In this estimation, we count as a violation every time a recourse action involving load curtailment or wind spillage is to be taken in real time to restore the power balance. This is equivalent to counting (over the test set) the number of times a constraint is violated in the original affine-policy-based OPF model.



#### 4.4.7 A 118-bus case study

As previously mentioned, we consider a modified version of the IEEE 118-bus system used in [78]. The system includes 54 conventional generators and eight wind power plants that we have added and placed at buses 2, 16, 33, 37, 55, 67, 83, and 116. In addition, the piecewise linear cost functions of all generators are comprised of three pieces or blocks. All the data pertaining to the network, generators, and transmission lines are available at the GITHUB repository [52].

We analyze two scenarios, which differ in the level of wind power penetration in the system. Below we explain how we have generated samples for the joint distribution of the wind power forecast and its error at each wind power plant. The so generated samples are also available online at the GITHUB repository [52]:

1. Let  $\tilde{f}_m$  be the per-unit point forecast of the power output at wind plant  $m \in \mathcal{W}$ . A sample of  $\tilde{f}_m$ , for all  $m \in \mathcal{W}$ , is randomly drawn with replacement from a collection of 16 694 p.u. wind power data recorded in several zones and made available by the Global Energy Forecasting Competition 2014 [76]. We have selected zones 1, 2, 3, 4, 5, 6, 9, and 10 of the aforementioned data set and assigned them to the eight wind power plants located at buses 16, 116, 83, 2, 55, 67, 33, and 37, respectively.
2. For each wind farm  $m \in \mathcal{W}$ , we have assumed that the (nominal, normalized) random variable  $W_m$ , which represents the nominal actual power generated at wind plant  $m$ , follows a Beta distribution with mean  $\tilde{f}_m$  and standard deviation  $\sigma$ . This standard deviation depends on both physical parameters and the quality of the forecasting model, following the model proposed in [56]. For simplicity, in all numerical experiments, given  $\tilde{f}_m$ , we determine  $\sigma$  as the value of the following function  $\sigma(\tilde{f}_m) := 0.2\tilde{f}_m + 0.02$ , empirically obtained in [56] for the case of a lead time of six hours. Therefore, the actual wind power production, and hence, the forecast error are conditional on the forecast power output issued. More specifically, the forecast error is given as the difference of a realization  $\widehat{W}_m$  of the r.v.  $W_m \sim \text{Beta}(A, B)$  and the point forecast  $\tilde{f}_m$ , where  $A, B > 0$  are the solution (if it exists) of the following system of non-linear equations:

$$\tilde{f}_m = \frac{A}{A+B} \tag{4.47a}$$

$$\sigma^2(\tilde{f}_m) = \frac{AB}{(A+B)^2(A+B+1)} \tag{4.47b}$$

To ensure that this non-linear system of equations has a solution, the samples  $\tilde{f}_m$  from the dataset that are less than or equal to 0.05 p.u. are set to 0.05, and the ones greater than or equal to 0.95 p.u. are set to 0.95. Thus, for each wind power



plant, the per-unit point forecast lies in the interval  $[0.05, 0.95]$ .

3. To work with MW, we multiply the per-unit realized power output and the point prediction  $\tilde{f}_m$  by the wind plant capacity  $\bar{C}_m$ , thus getting a pair of predicted power output and its error  $(\bar{C}_m \tilde{f}_m, \bar{C}_m(\tilde{W}_m - \tilde{f}_m))$  for wind farm  $m$ .
4. Steps 1, 2 and 3 are repeated  $N$  times to get the desired sample size.

Each independent run in our simulations involves repeating the above process.

Finally, the test set used to compute the out-of-sample performance of a data-driven solution via re-optimization (i.e., the actual probability of violating the uncertain OPF constraints and the actual expected operational cost) is constructed by drawing 1000 samples from the wind-power-data generating model based on the beta distribution presented above, with mean equal to the point prediction acting as the selected context. Therefore, this test set constitutes a discrete approximation of the forecast error distribution conditional on a given context, which will be specified later. Importantly, the shape and size of the ambiguity set that DROTRIMM uses is to be changed with the sample size  $N$  (which is indicative of the amount of information on the joint distribution of  $(\mathbf{z}, \boldsymbol{\omega})$  we have). Consequently, the trimming level  $\alpha$  defining this set is to be dependent on  $N$ . Accordingly, we have set  $\alpha_N := K_N/N$ , where  $K_N$  is the number of nearest neighbors used by SCENA. We have specifically taken  $K_N := \lfloor N^{0.9} \rfloor$  so that the resulting  $\alpha_N$  is consistent with the convergence results in Lemma 4.5. Again, both  $\alpha$  and  $K$  have been augmented with the subscript  $N$  to make their dependence on the sample size explicit.

### Medium wind penetration case

In this case, all eight wind farms in the system have a capacity of 200 MW and the context is given by  $\mathbf{z}^* = 180 \cdot \mathbf{1}$  MW, that is, the point forecast is 180 MW for all the wind power plants. Hence, the level of wind power penetration in the system (i.e., the ratio of the predicted system-wise wind power production to the total system demand) is approximately 63%.

Figures 4.9 and 4.10 illustrate the box plots corresponding to the total downward and upward reserve capacity that is scheduled, the violation probability and the expected cost delivered out of sample by SCENA, DROTRIMM and DROW as a function of their corresponding robustness parameter for sample sizes  $N = 100$  and  $N = 300$ , respectively. Naturally, the results provided by SCENA do not change along the  $x$ -axis in the plots, because this method is not based on *distributional robustness*. The box plots have been obtained from 200 independent runs for each sample size. We have set  $\epsilon = 0.1$ . The robustness parameter of DROW is the radius of the Wasserstein ball,

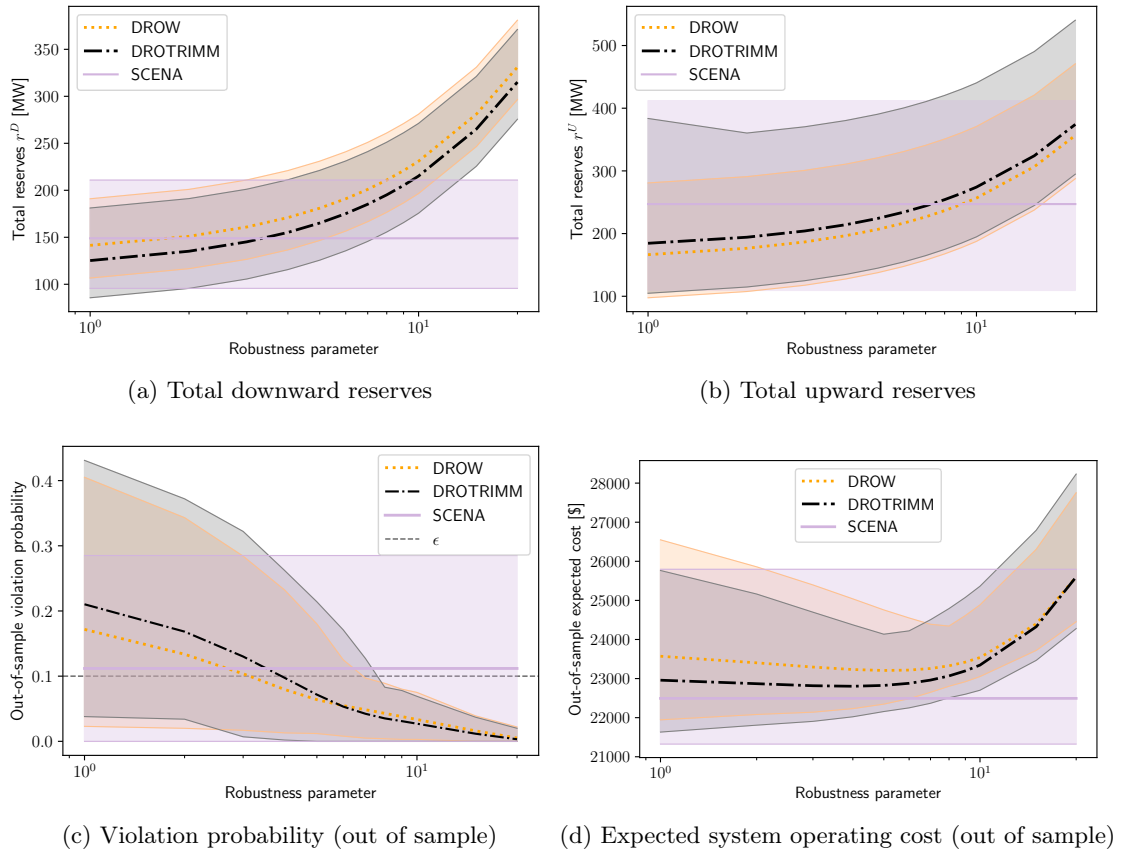


Figure 4.9: Medium level of wind penetration,  $N = 100$  and  $\epsilon = 0.1$ : Total downward and upward reserve capacity and performance metrics

while the robustness parameter for DROTRIMM is the budget excess over the minimum transportation budget (see Definition 4.2) .

The color-shaded areas have been obtained by joining the minimum and maximum edge cases of the box plots, while the associated bold colored lines link their means. These figures allow us to check which of the methods provides the most cost-efficient dispatch solutions on average without exceeding the threshold  $\epsilon$ . As expected, the reliability of the OPF solution given by DROW and DROTRIMM increases as the value of their robustness parameter is augmented, because more reserve capacity is procured. In turn, as more reserve capacity is scheduled, the magnitude and frequency of expensive load shedding events tend to diminish, which explains why the expected system operating cost may also decrease with the robustness parameter. This justifies the use of Distributionally Robust Optimization to tackle the chance-constrained OPF problem. However, when said parameter reaches a large enough value, the expected cost starts to grow quickly, because the cost of procuring additional reserve capacity no longer compensates for the cost savings entailed by the reduction in the amount of

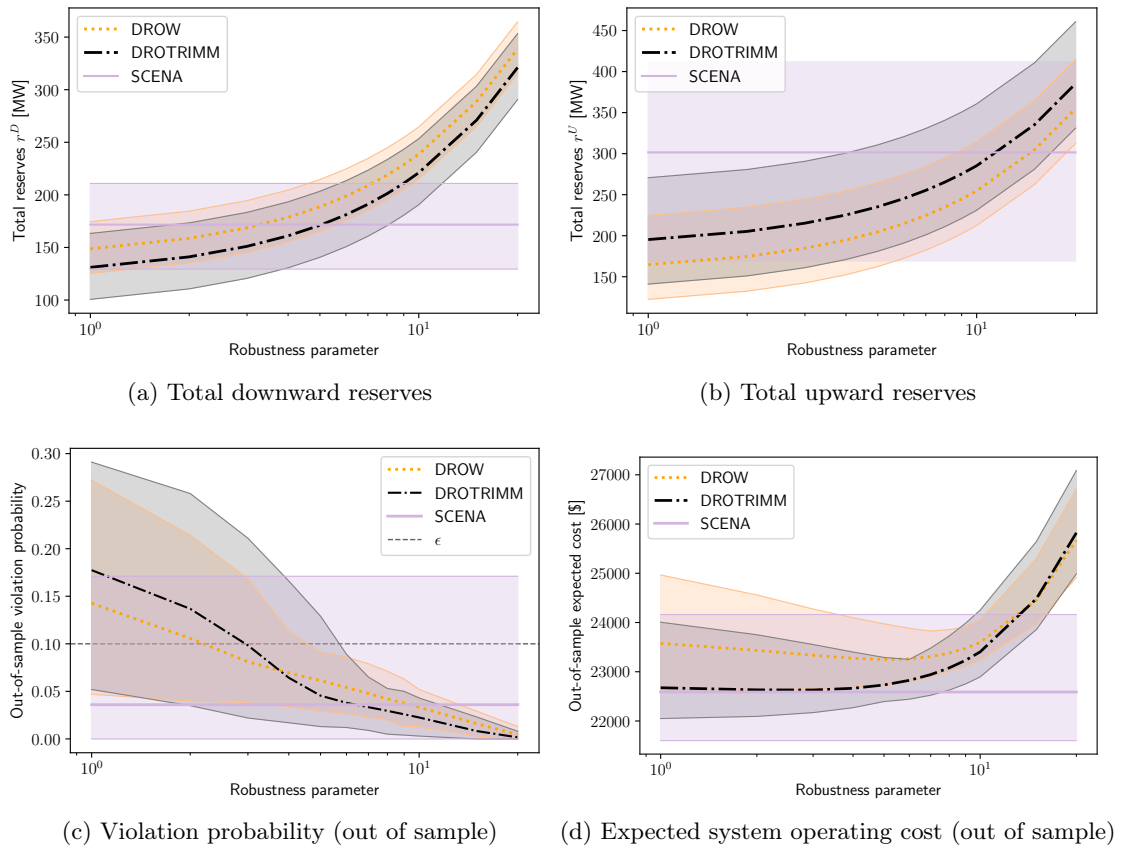


Figure 4.10: Medium level of wind penetration,  $N = 300$  and  $\epsilon = 0.1$ : Total downward and upward reserve capacity and performance metrics

curtailed load.

While SCENA provides OPF solutions that are competitive in terms of expected cost, these solutions do not comply with the specified reliability threshold in many of the runs. In addition, the performance of the OPF solutions obtained from SCENA exhibit a high variability, which is clearly due to the fact that this method is a non-robust approach and as such, is highly negatively affected by the uncertainty associated with the conditional inference it must perform.

On the other hand, when comparing DROW and DROTRIMM, whereas the former needs a lower value of the robustness parameter to attain the desired level of solution reliability, DROTRIMM gets to identify OPF solutions that are also reliable, while systematically cheaper on average. This phenomenon becomes even more evident when we increase the sample size  $N$  from 100 to 300. Indeed, a richer joint data sample contains more information on the statistical dependence of the wind power forecast error on the associated point prediction, which our approach manages to take advantage of. To give some numbers, if we just consider the range of values for the robustness parameters for which the violation probability is kept below the tolerance  $\epsilon$ , the average expected cost savings of DROTRIMM with respect to DROW go from 0.82%, when  $N = 100$  to 1.82%, when  $N = 300$ . From a technical point of view, DROW tends to produce OPF solutions with a higher cost because it underestimates the amount of upward reserve capacity that should be procured, clearly because this method is oblivious to the context and therefore, plans for the *marginal* distribution of the wind power forecast errors and not for the conditional one.

To elaborate further on the differences among the three methods, Table 4.1 includes the maximum, average, and minimum out-of-sample expected cost<sup>2</sup> under the value of the robustness parameter that is *optimal* for methods DROW and DROTRIMM, i.e., which leads to reliable OPF solutions with the minimum average expected cost for each of these two approaches. The standard deviation of this cost is also provided in the last row of Table 4.1. When  $N = 100$ , the (exacerbated) robustness of DROW produces OPF solutions with low average cost and variance, although DROTRIMM manages to find OPF solutions that are more economical in expectation. When  $N = 300$ , DROTRIMM clearly beats DROW on all metrics, because the excessive robustness of DROW (which is the result of ignoring the context) no longer pays off. Again, SCENA provides the cheapest OPF solutions on average, but these are useless because they do not satisfy the reliability requirement.

### High wind penetration case

In this alternative setting, all the wind farms have a capacity of 250 MW and the context is given by  $\mathbf{z}^* = 225 \cdot \mathbf{1}$  MW. Hence, the level of wind power penetration in the system

<sup>2</sup>These statistics are computed over the 200 independent runs.

Table 4.1: Medium level of wind penetration, summary data for total expected cost [\$] under the optimal value of the robustness parameter for methods DROW and DROTRIMM.

|     | $N = 100$ |       |       | $N = 300$ |       |       |
|-----|-----------|-------|-------|-----------|-------|-------|
|     | DROTRIMM  | DROW  | SCENA | DROTRIMM  | DROW  | SCENA |
| max | 24790     | 24388 | 25795 | 23250     | 23972 | 24160 |
| avg | 23068     | 23258 | 22493 | 22826     | 23250 | 22588 |
| min | 22511     | 22649 | 21325 | 22443     | 22728 | 21602 |
| std | 300       | 295   | 742   | 159       | 246   | 496   |

is approximately 80%.

Figures 4.11 and 4.12, and Table 4.2 are analogous to Figures 4.9 and 4.10, and Table 4.1 of the previous case, respectively. The higher level of wind power penetration in this new instance implies a higher level of uncertainty in the system. This accentuates the difference in performance between DROW and DROTRIMM when  $N = 100$ , that is, in a small sample regime. More specifically, the relative difference between the out-of-sample average expected cost achieved by DROW and DROTRIMM increases from 0.82% in the previous case to 2.27% in this new one. It is true, though, that DROW offers reliable OPF solutions with the lowest variance in expected cost when  $N = 100$ , provided that its robustness parameter is optimally tuned, see Table 4.2. However, its superiority in this respect ends when  $N$  grows to 300, at which point DROTRIMM provides the most cost-efficient OPF solutions in every respect<sup>3</sup>. Again the reason for this difference in performance has to do with the different provision of upward and downward reserve capacity that DROW and DROTRIMM prescribe.

For its part, the SCENA method keeps on providing cheap, but unreliable OPF solutions under a higher level of wind power penetration. In fact, the variability in cost, violation probability and reserves of the OPF solutions given by this method is remarkably high in contrast with that of DROTRIMM and DROW, even higher than in the case of a medium wind power penetration level (compare the range of the box plots in Figure 4.12).

We conclude this section with a remark on computational time. DROTRIMM and DROW have the same complexity (essentially, the number of constraints grows linearly with the sample size  $N$ ). The continuous linear program that results from tackling the chance-constrained DRO OPF problem by way of DROTRIMM and the **CVaR** approximation takes around 15 minutes to be solved on average, for a sample size  $N = 300$ , using CPLEX 20.1.0 on a Linux-based server with 22 CPUs clocking at 2.6 GHz and 200 GB of RAM in total.

<sup>3</sup>Note in Table 4.2 that, while the standard deviation of the expected cost is a bit higher under DROTRIMM than under DROW when  $N = 300$ , the maximum and minimum values reached by the expected cost under each method reveals that DROTRIMM produces a distribution of the expected cost displaced towards cheaper OPF solutions.

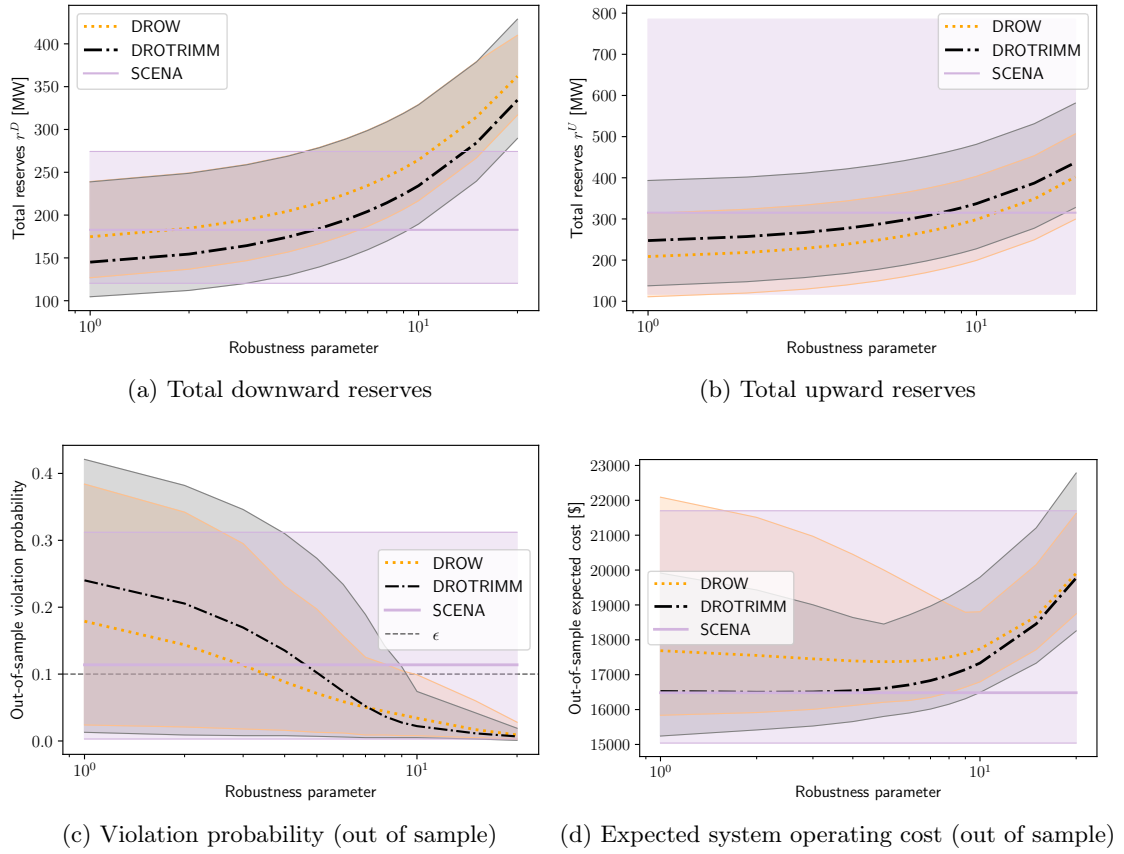


Figure 4.11: High level of wind penetration,  $N = 100$  and  $\epsilon = 0.1$ : Total downward and upward reserve capacity and performance metrics

Table 4.2: High level of wind penetration, summary data for total expected cost [\$] under the optimal value of the robustness parameter for DROW and DROTRIMM.

|     | $N = 100$ |       |       | $N = 300$ |       |       |
|-----|-----------|-------|-------|-----------|-------|-------|
|     | DROTRIMM  | DROW  | SCENA | DROTRIMM  | DROW  | SCENA |
| max | 19794     | 18804 | 21699 | 18101     | 18255 | 19863 |
| avg | 17334     | 17737 | 16483 | 17025     | 17274 | 16779 |
| min | 16490     | 16795 | 15040 | 16486     | 16703 | 15639 |
| std | 488       | 381   | 1093  | 296       | 292   | 767   |

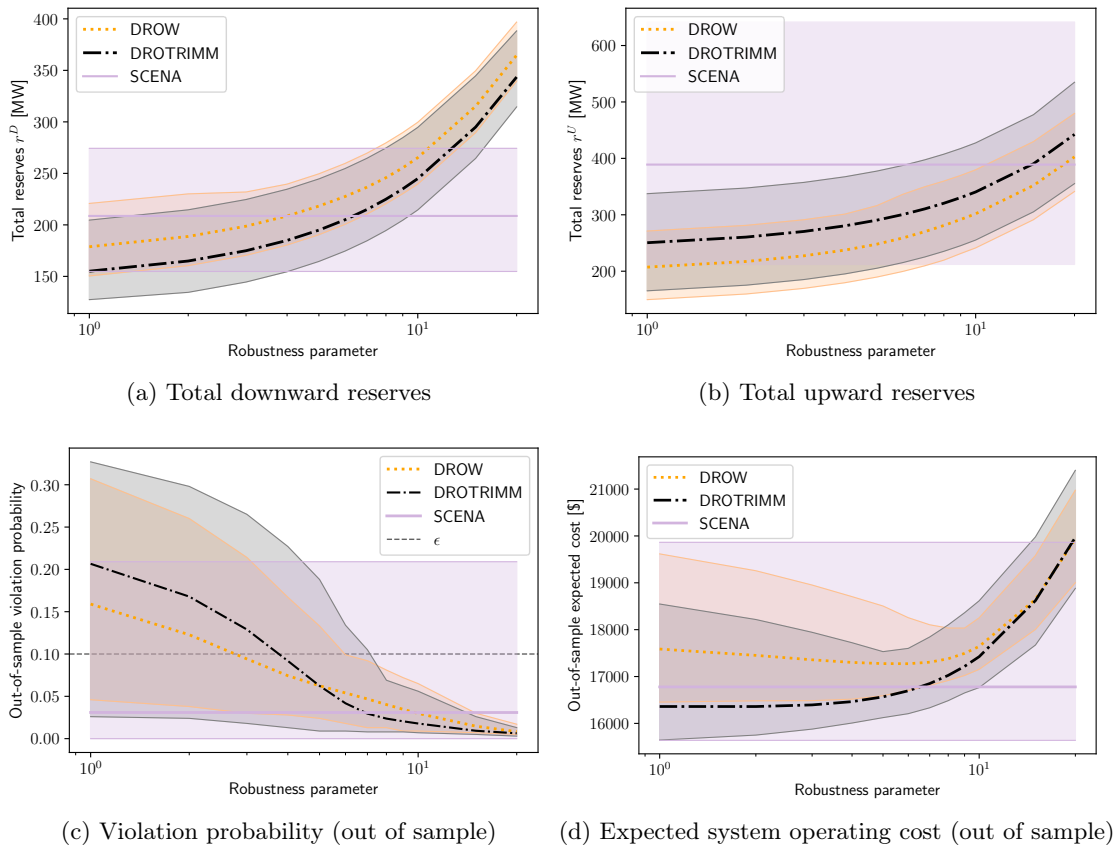


Figure 4.12: High level of wind penetration,  $N = 300$  and  $\epsilon = 0.1$ : Total downward and upward reserve capacity and performance metrics

## 4.5 Summary

In this chapter, we have exploited the connection between probability trimmings and partial mass transportation to provide an easy, but powerful and novel way to extend the standard Wasserstein-metric-based DRO to the case of *conditional* stochastic programs. Our approach produces decisions that are distributionally robust against the uncertainty in the whole process of inferring the conditional probability measure of the random parameters from a finite sample coming from the true joint data-generating distribution. Through a series of numerical experiments built on the single-item newsvendor problem and a portfolio allocation problem, we have demonstrated that our method attains notably better out-of-sample performance than some existing alternatives. We have supported these empirical findings with theoretical analysis, showing that our approach enjoys attractive performance guarantees. Finally, we have developed a distributionally robust chance-constrained OPF model that is able to exploit contextual information through an ambiguity set based on probability trimmings. We have provided a reformulation of this model as a continuous linear program using the well-known **CVaR** approximation. By way of a series of numerical experiments conducted on a modified 118-bus power network with wind uncertainty, we have shown that, by exploiting the statistical dependence between the point forecast of the wind power outputs and its associated forecast error, our approach can identify dispatch solutions that, while satisfying the required system reliability, lead to costs savings of up to several percentage points with respect to the OPF solutions provided by an alternative DRO method that ignores said statistical dependence.



## Chapter 5

# Conclusions and future work

In this closing chapter, the thesis is summarized, conclusions are drawn and directions for future research are given.

### 5.1 Summary and conclusions

Many real-world decision-making problems involve data parameters that are random and noisy. It is customary to formulate these problems as mathematical optimization programs under uncertainty, whose parameters are treated as random variables. Disregarding this uncertainty may lead to infeasible/suboptimal decisions. Today's decision makers not only collect observations of the uncertainties directly affecting their decision-making processes, but also gather some prior information about the data-generating distribution of the uncertainty. This prior information is used by the decision maker to prescribe a more accurate set of potential probability distributions, the so-called *ambiguity set* in distributionally robust optimization. The prior information that has been studied in this thesis can take the form of:

- *Structural information*, which can be given by some expert knowledge of the optimization problem to solve. This structural information could be, for example, shape information such as multimodality.
- *Conditional information* given in terms of a generic measurable event. This event may convey some conditional information which serves as contextual/side information and could be delivered by some *covariates* (also known as exogenous variables, features or attributes).

In this thesis, we have developed several distributionally robust optimization models by making use of tools of convex analysis, probability theory, statistics, and optimization under uncertainty. The contents of this thesis are included in the published papers [55], [53] and the preprint [51]:

- In our paper [55], we present a novel framework for data-driven distributionally robust optimization (DRO) based on optimal transport theory in combination with order cone constraints to leverage a-priori information on the true data-generating distribution. Motivated by the reported over-conservativeness of the traditional DRO approach based on the Wasserstein metric, we formulate an ambiguity set able to incorporate information about the order among the probabilities that the true distribution of the problem's uncertain parameters assigns to some subregions of its support set. Our approach can practically and intuitively accommodate a wide range of shape information (such as that related to monotonicity or multimodality). Moreover, under mild assumptions, the resulting distributionally robust optimization problem can be, in fact, reformulated as a finite convex problem where the a-priori information (expressed through the order cone constraints) is cast as linear constraints as opposed to the more computationally challenging formulations that exist in the literature. Our approach is supported by theoretical performance guarantees and is capable of turning the information provided into solutions with increased reliability and improved performance, as illustrated by our numerical experiments. These are based on the well-known newsvendor problem and the problem of a strategic firm competing *à la Cournot* in a market for a homogeneous product.
- In the article [53], we exploit the connection between probability trimmings and partial mass transportation to provide an easy, but powerful and novel way to extend the standard Wasserstein-metric-based DRO to the case of conditional stochastic programs. Our approach produces decisions that are distributionally robust against the uncertainty in the whole process of inferring the conditional probability measure of the random parameters from a finite sample taken from the true *joint* data-generating distribution. Through a series of numerical experiments built on the single-item newsvendor problem and a portfolio allocation problem, we demonstrate that our method attains a notably better out-of-sample performance than some existing alternatives. We support these empirical findings with theoretical analysis, showing that our approach enjoys attractive performance guarantees.
- In our preprint [51], we develop a distributionally robust chance-constrained Optimal Power Flow (OPF) model that is able to exploit contextual information through an ambiguity set based on probability trimmings. We provide a reformulation of this model as a continuous linear program using the well known **CVaR** approximation. By way of a series of numerical experiments conducted on a modified 118-bus power network with wind uncertainty, we show that, by exploiting the statistical dependence between the point forecast of the wind power outputs

and its associated forecast error, our approach can identify *reliable* dispatch solutions that are significantly cheaper than those provided by an alternative DRO method that ignores said statistical dependence.

## 5.2 Directions for future research

Directions for future research resulting from the study carried out in this thesis are listed below:

1. The development of decomposition methods to solve large-scale distributionally robust optimization programs based on the Wasserstein metric.
2. Theoretical analysis is required to investigate if, and under which conditions, it is possible to break the dependence of the finite-sample guarantees on the uncertainty dimension in the realm of conditional stochastic optimization.
3. Further research is needed into how to properly extend the use of probability trimmings to conditional multi-stage stochastic programs.
4. Data-driven schemes for appropriately tuning the robustness parameter in our distributionally robust chance-constrained OPF model in accordance with the risk preferences of the system operator (for instance, by resorting to cross-validation or bootstrapping) must be studied. Moreover, this model has to be extended to account for intertemporal constraints, which, among other factors, will involve adapting our probability-trimming-based ambiguity set to deal with stochastic processes and time series data.

# Appendix A

## Proofs of Chapter 3

### Contents

---

|     |                                  |     |
|-----|----------------------------------|-----|
| A.1 | Proof of Theorem 3.1 . . . . .   | 97  |
| A.2 | Proof of Corollary 3.1 . . . . . | 100 |
| A.3 | Proof of Theorem 3.2 . . . . .   | 101 |
| A.4 | Proof of Theorem 3.3 . . . . .   | 101 |
| A.5 | Proof of Theorem 3.4 . . . . .   | 101 |
| A.6 | Proof of Theorem 3.5 . . . . .   | 102 |

---

### A.1 Proof of Theorem 3.1

Recall that we have assumed that regions  $\Xi_i$  are disjoint. Thus, using the law of total probability, we can rewrite problem (POC) as follows:

$$\inf_{\mathbf{x} \in X} \sup_{\mathbf{p} \in \mathcal{P}} G(\mathbf{x}, \mathbf{p}) \quad (\text{A.1})$$

where we have considered the subproblem (SPOC):

$$(\text{SPOC}) \quad G(\mathbf{x}, \mathbf{p}) = \sup_{Q_i \in \mathcal{Q}_i, \forall i} \sum_{i \in \mathcal{I}} p_i \mathbb{E}_{Q_i} [f(\mathbf{x}, \boldsymbol{\xi})] \quad (\text{A.2a})$$

$$\text{s.t.} \quad \sum_{i \in \mathcal{I}} p_i C(Q_i, \hat{Q}_i) \leq \varepsilon \quad (\text{A.2b})$$

The probability distribution  $\hat{Q}_i$  is defined as  $\hat{Q}_i = \frac{1}{N_i} \sum_{j=1}^{N_i} \delta_{\hat{\boldsymbol{\xi}}_j^i}$ , with  $N_i$  being the number of data points in  $\Xi_i$  and  $\hat{\boldsymbol{\xi}}_j^i \in \{\hat{\boldsymbol{\xi}}_1^i, \dots, \hat{\boldsymbol{\xi}}_{N_i}^i\}$ .

Note that the structure of problem (A.1) does not fit in the general ambiguity set proposed in [38].

Equivalently, we can recast the subproblem (SPOC) as

$$(\text{SPOC}) = \left\{ \begin{array}{ll} \sup_{Q_i \in \mathcal{Q}_i, \Pi_i, \forall i} \sum_{i \in \mathcal{I}} p_i \int_{\Xi_i} f(\mathbf{x}, \boldsymbol{\xi}) Q_i(d\boldsymbol{\xi}) & \\ \text{s.t.} & \sum_{i \in \mathcal{I}} p_i \int_{\Xi_i^2} c(\boldsymbol{\xi}, \boldsymbol{\xi}') \Pi_i(d\boldsymbol{\xi}, d\boldsymbol{\xi}') \leq \varepsilon \\ & \left\{ \begin{array}{l} \forall i, \Pi_i \text{ is a joint distribution of } \boldsymbol{\xi} \text{ and } \boldsymbol{\xi}' \\ \text{with marginals } Q_i \text{ and } \hat{Q}_i, \text{ respectively} \end{array} \right. \end{array} \right. \quad (\text{A.3})$$

$$= \left\{ \begin{array}{ll} \sup_{\tilde{Q}_j^i, \forall i \in \mathcal{I}, j \leq N_i} \sum_{i \in \mathcal{I}} \frac{p_i}{N_i} \sum_{j=1}^{N_i} \int_{\Xi_i} f(\mathbf{x}, \boldsymbol{\xi}) \tilde{Q}_j^i(d\boldsymbol{\xi}) & \\ \text{s.t.} & \sum_{i \in \mathcal{I}} \frac{p_i}{N_i} \sum_{j=1}^{N_i} \int_{\Xi_i} c(\boldsymbol{\xi}, \hat{\boldsymbol{\xi}}_j^i) \tilde{Q}_j^i(d\boldsymbol{\xi}) \leq \varepsilon \\ & \int_{\Xi_i} \tilde{Q}_j^i(d\boldsymbol{\xi}) = 1, \forall i \in \mathcal{I}, j \leq N_i \end{array} \right. \quad (\text{A.4})$$

where reformulation (A.4) follows on from the fact that the marginal distribution of  $\boldsymbol{\xi}'$  is the discrete uniform distribution supported on points  $\hat{\boldsymbol{\xi}}_j^i, j = 1, \dots, N_i$ . Thus,  $\Pi_i$  is completely determined by the conditional distributions  $\tilde{Q}_j^i = \Pi_i(\boldsymbol{\xi}, \boldsymbol{\xi}' | \boldsymbol{\xi}' = \hat{\boldsymbol{\xi}}_j^i), \forall i \leq N_i$ , that is,  $\Pi_i(d\boldsymbol{\xi}, d\boldsymbol{\xi}') = \frac{1}{N_i} \sum_{j=1}^{N_i} \delta_{\hat{\boldsymbol{\xi}}_j^i}(d\boldsymbol{\xi}') \tilde{Q}_j^i(d\boldsymbol{\xi})$  [99].

The mathematical program (A.4) constitutes a generalized moment problem over the normalized measures  $\tilde{Q}_j^i$ , for which strong duality holds (see, for example, [123]). We can, therefore, dualize the  $\varepsilon$ -budget constraint on the transport cost, thus obtaining:

$$\inf_{\theta \geq 0} \sup_{\tilde{Q}_j^i, \forall i \in \mathcal{I}, j \leq N_i} \theta \varepsilon + \sum_{i \in \mathcal{I}} \frac{p_i}{N_i} \sum_{j=1}^{N_i} \int_{\Xi_i} [f(\mathbf{x}, \boldsymbol{\xi}) - \theta c(\boldsymbol{\xi}, \hat{\boldsymbol{\xi}}_j^i)] \tilde{Q}_j^i(d\boldsymbol{\xi}) \quad (\text{A.5})$$

$$\text{s.t.} \quad \int_{\Xi_i} \tilde{Q}_j^i(d\boldsymbol{\xi}) = 1, \forall i \in \mathcal{I}, j \leq N_i \quad (\text{A.6})$$

$$= \inf_{\theta \geq 0} \theta \varepsilon + \sum_{i \in \mathcal{I}} \frac{p_i}{N_i} \sum_{j=1}^{N_i} \sup_{\tilde{Q}_j^i} \int_{\Xi_i} [f(\mathbf{x}, \boldsymbol{\xi}) - \theta c(\boldsymbol{\xi}, \hat{\boldsymbol{\xi}}_j^i)] \tilde{Q}_j^i(d\boldsymbol{\xi}) \quad (\text{A.7})$$

$$\text{s.t.} \quad \int_{\Xi_i} \tilde{Q}_j^i(d\boldsymbol{\xi}) = 1, \forall i \in \mathcal{I}, j \leq N_i \quad (\text{A.8})$$

$$= \inf_{\theta \geq 0} \theta \varepsilon + \sum_{i \in \mathcal{I}} \frac{p_i}{N_i} \sum_{j=1}^{N_i} \sup_{\boldsymbol{\xi} \in \Xi_i} [f(\mathbf{x}, \boldsymbol{\xi}) - \theta c(\boldsymbol{\xi}, \hat{\boldsymbol{\xi}}_j^i)] \quad (\text{A.9})$$

$$= \inf_{\theta, t_{i,j}, \forall i \in \mathcal{I}, j \leq N_i} \theta \varepsilon + \sum_{i \in \mathcal{I}} \frac{p_i}{N_i} \sum_{j=1}^{N_i} t_{i,j} \quad (\text{A.10})$$

$$\text{s.t.} \quad t_{i,j} \geq \sup_{\boldsymbol{\xi} \in \Xi_i} [f(\mathbf{x}, \boldsymbol{\xi}) - \theta c(\boldsymbol{\xi}, \hat{\boldsymbol{\xi}}_j^i)], \forall i \in \mathcal{I}, j \leq N_i \quad (\text{A.11})$$

$$\theta \geq 0 \quad (\text{A.12})$$

where the second equality derives from the fact that we can choose a Dirac distribution supported on  $\Xi_i$  as  $\tilde{Q}_j^i$ .

Now, dualizing the  $\rho$ -budget constraint on the transport cost in the inner supremum of problem (A.1), we obtain:

$$\inf_{\lambda \geq 0} \lambda \rho + \sup_{\mathbf{p} \in \Theta} [G(\mathbf{x}, \mathbf{p}) - \lambda \tilde{c}(\mathbf{p}, \hat{\mathbf{p}})] \quad (\text{A.13})$$

Thus,

$$\inf_{\lambda \geq 0} \lambda \rho + \sup_{\mathbf{p} \in \Theta} [G(\mathbf{x}, \mathbf{p}) - \lambda \tilde{c}(\mathbf{p}, \hat{\mathbf{p}})] \quad (\text{A.14})$$

$$= \inf_{\lambda \geq 0} \lambda \rho + \sup_{\mathbf{p} \in \Theta} \left[ \inf_{\theta \geq 0, (t_{i,j}) \text{ s.t. (A.11)}} \theta \varepsilon + \sum_{i \in \mathcal{I}} \frac{p_i}{N_i} \sum_{j=1}^{N_i} t_{i,j} - \lambda \tilde{c}(\mathbf{p}, \hat{\mathbf{p}}) \right] \quad (\text{A.15})$$

Since function  $\theta \varepsilon + \sum_{i \in \mathcal{I}} \frac{p_i}{N_i} \sum_{j=1}^{N_i} t_{i,j} - \lambda \tilde{c}(\mathbf{p}, \hat{\mathbf{p}})$  is upper semicontinuous and concave in  $\mathbf{p}$  on the compact convex set  $\Theta$  (recall that  $\tilde{c}$  is nonnegative, lower semicontinuous, and convex in  $\mathbf{p}$ ), and linear in  $\theta$  and  $t_{i,j}$  on the convex set defined by  $\theta \geq 0$  and (A.11), we can apply Sion's min-max theorem ([127]) and in this way, interchange the innermost infimum with the outer supremum. Then, by merging the two infima, we arrive at

$$\begin{aligned} & \inf_{\lambda \geq 0, \theta \geq 0, (t_{i,j})} \lambda \rho + \theta \varepsilon + \sup_{\mathbf{p} \in \Theta} \left[ \sum_{i \in \mathcal{I}} \frac{p_i}{N_i} \sum_{j=1}^{N_i} t_{i,j} - \lambda \tilde{c}(\mathbf{p}, \hat{\mathbf{p}}) \right] \\ & \text{s.t. } t_{i,j} \geq \sup_{\boldsymbol{\xi} \in \Xi_i} \left[ f(\mathbf{x}, \boldsymbol{\xi}) - \theta_i c(\boldsymbol{\xi}, \hat{\boldsymbol{\xi}}_j^i) \right], \forall i \in \mathcal{I}, j \leq N_i \end{aligned}$$

We focus now on the inner supremum,

$$\sup_{\mathbf{p} \in \Theta} \left[ \left\langle \mathbf{p}, \left( \frac{1}{N_i} \sum_{j=1}^{N_i} t_{i,j} \right)_{i \in \mathcal{I}} \right\rangle - \lambda \tilde{c}(\mathbf{p}, \hat{\mathbf{p}}) \right] \quad (\text{A.16})$$

where we have written  $\sum_{i \in \mathcal{I}} \frac{p_i}{N_i} \sum_{j=1}^{N_i} t_{i,j}$  as  $\left\langle \mathbf{p}, \left( \frac{1}{N_i} \sum_{j=1}^{N_i} t_{i,j} \right)_{i \in \mathcal{I}} \right\rangle$ . This is a concave maximization problem (be aware that  $\langle \mathbf{p}, \mathbf{H}(\mathbf{x}) \rangle - \lambda \tilde{c}(\mathbf{p}, \hat{\mathbf{p}})$  is a concave function with respect to  $\mathbf{p}$  and  $\Theta$  is a convex compact set; furthermore, notice that we have  $\mathbf{H}(\mathbf{x}) = \left( \frac{1}{N_i} \sum_{j=1}^{N_i} t_{i,j} \right)_{i \in \mathcal{I}}$  in our particular case). Consequently, strong duality holds if a Slater condition is satisfied, that is, if there exists a point  $\mathbf{p}^* \in \text{relint}(\mathbb{R}_+^{|\mathcal{I}|})$  such that  $\langle \mathbf{e}, \mathbf{p}^* \rangle = 1$ , and  $\mathbf{p}^* \in \text{int}(\mathcal{C})$  (see, for example, [32]). Using a standard duality argument, we dualize the constraints  $\mathbf{p} \in \mathbb{R}_+^{|\mathcal{I}|}$ ,  $\langle \mathbf{e}, \mathbf{p} \rangle = 1$  and  $\mathbf{p} \in \mathcal{C}$ , with associated multipliers

$\boldsymbol{\mu} \in \mathbb{R}_+^{|\mathcal{I}|}$ ,  $\eta \in \mathbb{R}$  and  $\tilde{\mathbf{p}} \in \mathcal{C}^*$ , respectively. Thus, we obtain the following problem:

$$\begin{aligned} & \inf_{\eta \in \mathbb{R}, \boldsymbol{\mu} \in \mathbb{R}_+^{|\mathcal{I}|}, \tilde{\mathbf{p}} \in \mathcal{C}^*} \sup_{\mathbf{p}} \left\{ \left\langle \mathbf{p}, \left( \frac{1}{N_i} \sum_{j=1}^{N_i} t_{i,j} \right)_{i \in \mathcal{I}} \right\rangle - \lambda \tilde{c}(\mathbf{p}, \hat{\mathbf{p}}) + \langle \boldsymbol{\mu}, \mathbf{p} \rangle + \eta(1 - \langle \mathbf{e}, \mathbf{p} \rangle) + \langle \tilde{\mathbf{p}}, \mathbf{p} \rangle \right\} = \\ & \inf_{\eta \in \mathbb{R}, \boldsymbol{\mu} \in \mathbb{R}_+^{|\mathcal{I}|}, \tilde{\mathbf{p}} \in \mathcal{C}^*} \eta + \sup_{\mathbf{p}} \left\{ \left\langle \mathbf{p}, \left( \frac{1}{N_i} \sum_{j=1}^{N_i} t_{i,j} \right)_{i \in \mathcal{I}} + \boldsymbol{\mu} - \eta \mathbf{e} + \tilde{\mathbf{p}} \right\rangle - \lambda \tilde{c}(\mathbf{p}, \hat{\mathbf{p}}) \right\} = \\ & \inf_{\eta \in \mathbb{R}, \boldsymbol{\mu} \in \mathbb{R}_+^{|\mathcal{I}|}, \tilde{\mathbf{p}} \in \mathcal{C}^*} \eta + \lambda \sup_{\mathbf{p}} \left\{ \left\langle \mathbf{p}, \frac{\left( \frac{1}{N_i} \sum_{j=1}^{N_i} t_{i,j} \right)_{i \in \mathcal{I}} + \boldsymbol{\mu} - \eta \mathbf{e} + \tilde{\mathbf{p}}}{\lambda} \right\rangle - \tilde{c}(\mathbf{p}, \hat{\mathbf{p}}) \right\} = \\ & \inf_{\eta \in \mathbb{R}, \boldsymbol{\mu} \in \mathbb{R}_+^{|\mathcal{I}|}, \tilde{\mathbf{p}} \in \mathcal{C}^*} \eta + \lambda \tilde{c}_{\hat{\mathbf{p}}}^* \left( \frac{\left( \frac{1}{N_i} \sum_{j=1}^{N_i} t_{i,j} \right)_{i \in \mathcal{I}} + \boldsymbol{\mu} - \eta \mathbf{e} + \tilde{\mathbf{p}}}{\lambda} \right) \end{aligned}$$

where  $\tilde{c}_{\hat{\mathbf{p}}}^*(\cdot)$  is the convex conjugate function of  $\tilde{c}(\cdot, \hat{\mathbf{p}})$ , with  $\hat{\mathbf{p}}$  fixed.

Therefore, problem (A.1) can be equivalently reformulated as follows:

$$\begin{aligned} (\text{POC-0}) \quad & \inf_{\mathbf{x}, \lambda, \boldsymbol{\mu}, \eta, \tilde{\mathbf{p}}, \theta, \mathbf{t}} \lambda \rho + \eta + \theta \varepsilon + \lambda \tilde{c}_{\hat{\mathbf{p}}}^* \left( \frac{\left( \frac{1}{N_i} \sum_{j=1}^{N_i} t_{i,j} \right)_{i \in \mathcal{I}} + \boldsymbol{\mu} - \eta \mathbf{e} + \tilde{\mathbf{p}}}{\lambda} \right) \\ & \text{s.t. } t_{i,j} \geq \sup_{\boldsymbol{\xi} \in \Xi_i} \left[ f(\mathbf{x}, \boldsymbol{\xi}) - \theta c(\boldsymbol{\xi}, \hat{\boldsymbol{\xi}}_j^i) \right], \forall i \in \mathcal{I}, j \leq N_i \quad (\text{A.17}) \\ & \mathbf{x} \in X, \lambda \geq 0, \boldsymbol{\mu} \in \mathbb{R}_+^{|\mathcal{I}|}, \eta \in \mathbb{R}, \tilde{\mathbf{p}} \in \mathcal{C}^*, \theta \geq 0 \\ & t_{i,j} \in \mathbb{R}, \forall i \in \mathcal{I}, j \leq N_i \end{aligned}$$

□

## A.2 Proof of Corollary 3.1

We use the following Lemma to put problem (POC-0) in a better shape.

**Lemma A.1.** *Let  $\tilde{c}_{\hat{\mathbf{p}}}(\mathbf{p}) = \|\mathbf{p} - \hat{\mathbf{p}}\|$ , where  $\hat{\mathbf{p}} \in \mathbb{R}^{|\mathcal{I}|}$  is a fixed vector and  $\|\cdot\|$  a norm in  $\mathbb{R}^{|\mathcal{I}|}$ . Then, it holds that the convex conjugate function of  $\tilde{c}_{\hat{\mathbf{p}}}(\mathbf{p})$  is as follows*

$$\tilde{c}_{\hat{\mathbf{p}}}^*(\mathbf{s}) = \begin{cases} \sum_{i \in \mathcal{I}} \hat{p}_i s_i & \text{if } \|\mathbf{s}\|_* \leq 1 \\ \infty & \text{if } \|\mathbf{s}\|_* > 1 \end{cases}$$

*Proof.* The claim of the Lemma follows from Proposition 5.1.4. (vii) and Example 5.1.2 (b) of [97]. □

Therefore, problem (POC-0) reduces to

$$\begin{aligned}
 \text{(POC-1)} \quad & \inf_{\mathbf{x}, \lambda, \boldsymbol{\mu}, \eta, \tilde{\mathbf{p}}, \theta, \mathbf{t}} \lambda \rho + \eta + \theta \varepsilon + \sum_{i \in \mathcal{I}} \hat{p}_i \left( \frac{1}{N_i} \sum_{j=1}^{N_i} t_{i,j} + \mu_i - \eta + \tilde{p}_i \right) \\
 \text{s.t. } & t_{i,j} \geq \sup_{\boldsymbol{\xi} \in \Xi_i} \left[ f(\mathbf{x}, \boldsymbol{\xi}) - \theta \left\| \boldsymbol{\xi} - \hat{\boldsymbol{\xi}}_j^i \right\| \right], \forall i \in \mathcal{I}, \forall j \leq N_i \\
 & \left\| \left( \frac{1}{N_i} \sum_{j=1}^{N_i} t_{i,j} + \mu_i - \eta + \tilde{p}_i \right) \right\|_{i \in \mathcal{I}} \leq \lambda \\
 & \mathbf{x} \in X, \lambda \geq 0, \boldsymbol{\mu} \in \mathbb{R}_+^{|\mathcal{I}|}, \eta \in \mathbb{R}, \tilde{\mathbf{p}} \in \mathcal{C}^*, \theta \geq 0 \\
 & t_{i,j} \in \mathbb{R}, \forall i \in \mathcal{I}, \forall j \leq N_i
 \end{aligned} \tag{A.18}$$

□

### A.3 Proof of Theorem 3.2

In essence, the complexity of problem (POC-1) depends on our ability to reformulate the supremum in constraint (3.7) in a tractable manner. This is possible under Assumption 3.1, following similar steps to those in the proof of Theorem 4.2 in [99], to which we refer. □

### A.4 Proof of Theorem 3.3

The proof runs in a similar way to that of Theorem 6.1 in [99]. □

### A.5 Proof of Theorem 3.4

Given Assumption 3.2, for all  $i \in \mathcal{I}$ , we deduce from Theorem 2 in [59] that

$$\mathbb{P} \left[ \mathcal{W}(\mathbb{Q}_i, \hat{Q}_i) \leq \varepsilon_{N_i}(\beta_i) \right] \geq 1 - \beta_i.$$

Thus, we have that

$$\mathbb{P} \left[ \sum_{i \in \mathcal{I}} p_i \mathcal{W}(\mathbb{Q}_i, \hat{Q}_i) \leq \sum_{i \in \mathcal{I}} p_i \varepsilon_{N_i}(\beta_i) \right] \geq \mathbb{P} \left[ \bigcap_{i \in \mathcal{I}} \left( p_i \mathcal{W}(\mathbb{Q}_i, \hat{Q}_i) \leq p_i \varepsilon_{N_i}(\beta_i) \right) \right] \tag{A.19}$$

$$= 1 - \mathbb{P} \left[ \bigcup_{i \in \mathcal{I}} \left( p_i \mathcal{W}(\mathbb{Q}_i, \hat{Q}_i) > p_i \varepsilon_{N_i}(\beta_i) \right) \right] \tag{A.20}$$

$$\geq 1 - \sum_{i \in \mathcal{I}} \mathbb{P} \left[ p_i \mathcal{W}(\mathbb{Q}_i, \hat{Q}_i) > p_i \varepsilon_{N_i}(\beta_i) \right] \tag{A.21}$$



$$\geq 1 - \sum_{i \in \mathcal{I}} \beta_i \quad (\text{A.22})$$

□

## A.6 Proof of Theorem 3.5

The claim follows from Theorem 3.4 and Equations (3.17) and (3.18), which imply that  $\mathbb{P}(\mathbb{Q} \in \mathcal{U}_{\rho, \varepsilon}(\hat{Q})) \geq (1 - \beta_p)(1 - \sum_{i \in \mathcal{I}} \beta_i)$ . Hence,

$$\mathbb{E}_{\mathbb{Q}}[f(\hat{\mathbf{x}}_N, \boldsymbol{\xi})] \leq \sup_{Q \in \mathcal{U}_{\rho, \varepsilon}(\hat{Q})} \mathbb{E}_Q[f(\hat{\mathbf{x}}_N, \boldsymbol{\xi})] = \hat{J}_N$$

with probability at least  $(1 - \beta_p)(1 - \sum_{i \in \mathcal{I}} \beta_i)$ . □

# Appendix B

## Additional material to Chapter 4

### Contents

---

|  |            |
|--|------------|
| <b>B.1 Complementary definitions and technical results . . . . .</b> | <b>104</b> |
| B.1.1 Concepts from measure theory and the Wasserstein metric . .    | 104        |
| B.1.2 Concepts and technical results from probability trimmings . .  | 105        |
| B.1.3 Topological properties of the ambiguity set . . . . .          | 106        |
| B.1.4 Tractable reformulation and maximizer of problem (SP2) . .     | 107        |
| <b>B.2 Proofs of Chapter 4 . . . . .</b>                             | <b>108</b> |
| B.2.1 Proof of Lemma 4.1 . . . . .                                   | 108        |
| B.2.2 Proof of Proposition 4.1 . . . . .                             | 109        |
| B.2.3 Proof of Theorem 4.1 . . . . .                                 | 110        |
| B.2.4 Proof of Proposition 4.2 . . . . .                             | 112        |
| B.2.5 Proof of Theorem 4.2 . . . . .                                 | 112        |
| B.2.6 Proof of Lemma 4.4 . . . . .                                   | 113        |
| B.2.7 Proof of Theorem 4.3 . . . . .                                 | 114        |
| B.2.8 Proof of Proposition 4.3 . . . . .                             | 114        |
| B.2.9 Proof of Theorem 4.4 . . . . .                                 | 114        |
| B.2.10 Proof of Lemma 4.5 . . . . .                                  | 115        |
| <b>B.3 Asymptotic consistency under a nearest neighbors lens . .</b> | <b>115</b> |

---

This appendix contains some additional material of interest related to the DRO framework we propose to handle conditional stochastic programs in Chapter 4. First, we state some complementary theoretical results. Second, we include all the proofs of Chapter 4. Finally, we use tools from nearest neighbors to show that our DRO approach is asymptotically consistent under assumptions slightly different than those made in the main text of this dissertation in Section 4.1.4 (some of these assumptions are less restrictive).

**Notation.** The recession cone of a non-empty set  $A \subseteq \mathbb{R}^d$  is given by  $\{\mathbf{b} \in \mathbb{R}^d / \mathbf{a} + \lambda \mathbf{b} \in A, \forall \mathbf{a} \in A, \forall \lambda \geq 0\}$ .

## B.1 Complementary definitions and technical results

This section contains some theoretical results which are complementary to the theory developed in the manuscript. First, we introduce a few preliminary concepts and definitions about measure theory and the Wasserstein metric. Second, we present some definitions and technical results related to probability trimmings. Third, we state the topological properties of the ambiguity set  $\hat{\mathcal{U}}_N(\alpha, \tilde{\rho})$  in problem (P). Finally, we introduce a tractable reformulation of our DRO approach, which is similar to that in [87, Theorem 8].

### B.1.1 Concepts from measure theory and the Wasserstein metric

This section compiles some definitions and results from the measure theory that underpins our research. It starts with concepts related to the weak convergence of measures and compactness. Subsequently, some known facts in connection with the topology generated by the Wasserstein metric  $\mathcal{W}_p$  are presented. We denote the set of all Borel probability measures supported on  $\mathcal{X}$  as  $\mathcal{P}(\mathcal{X})$ . Although some of the following concepts and results are still true in the more general setting of Polish spaces, we restrict ourselves here to  $\mathcal{X} \subseteq \mathbb{R}^d$ . Similarly, we denote the  $p$ -Wasserstein space as  $\mathcal{P}_p(\mathcal{X})$ , that is, the set of all Borel probability measures supported on  $\mathcal{X}$  with a finite  $p$ -th moment. It is well known that the  $p$ -Wasserstein metric defines a metric in  $\mathcal{P}_p(\mathcal{X})$  [132, Theorem 7.3].

**Definition B.1 (Weak convergence of probability measures).** *Given a sequence of probability measures  $\{Q_N\}_N \subseteq \mathcal{P}(\mathcal{X})$ , we say that it converges weakly to  $Q$  if*

$$\lim_{N \rightarrow \infty} \int_{\mathcal{X}} \ell(\boldsymbol{\xi}) Q_N(d\boldsymbol{\xi}) = \int_{\mathcal{X}} \ell(\boldsymbol{\xi}) Q(d\boldsymbol{\xi}) \quad (\text{B.1})$$

for all bounded and continuous function  $\ell$  on  $\mathcal{X}$ .

**Definition B.2 (Tightness).** *A given set  $\mathcal{K} \subseteq \mathcal{P}(\mathcal{X})$  is tight if for all  $\varepsilon > 0$ , there is a compact set  $X_\varepsilon \subset \mathcal{X}$  such that  $\inf_{Q \in \mathcal{K}} Q(X_\varepsilon) > 1 - \varepsilon$ . If  $\mathcal{K}$  reduces to a singleton, then we refer to the “tightness of a probability measure”.*

**Definition B.3 (Closed sets).** *A given set  $\mathcal{K} \subseteq \mathcal{P}(\mathcal{X})$  is closed (under the topology of weak convergence) if for all sequence  $\{Q_N\}_N \subset \mathcal{K}$  such that  $Q_N$  converges weakly to  $Q$ , we have  $Q \in \mathcal{K}$ .*

The following theorem, which is known as Prokhorov’s Theorem, connects the notions of weak compactness and tightness.

**Theorem B.1 (Prokhorov's Theorem).** *A set  $\mathcal{K} \subseteq \mathcal{P}(\mathcal{X})$  is tight if and only if the closure of  $\mathcal{K}$  is weakly compact in  $\mathcal{P}(\mathcal{X})$ .*

**Definition B.4 (Weak compactness).** *A set  $\mathcal{K} \subseteq \mathcal{P}(\mathcal{X})$  is weakly compact if for all sequence of probability measures  $\{Q_N\}_N \subset \mathcal{K}$ , there exists a subsequence  $\{Q_{N'}\}_{N'}$  that converges weakly to  $Q \in \mathcal{K}$ .*

**Definition B.5 ( $p$ -uniform integrability).** *A set  $\mathcal{K} \subseteq \mathcal{P}(\mathcal{X})$  is said to have  $p$ -uniformly integrable moments if*

$$\lim_{t \rightarrow \infty} \int_{\{\|\xi\|/t > 1\}} \|\xi\|^p Q(d\xi) = 0 \text{ uniformly w.r.t. } Q \in \mathcal{K} \quad (\text{B.2})$$

Finally, we introduce a proposition that connects some of the aforementioned concepts with the Wasserstein metric. More concretely, this proposition establishes the topological properties of the Wasserstein space.

**Proposition B.1.** *Given  $p \geq 1$  and  $\mathcal{X} \subseteq \mathbb{R}^d$  a closed set, we have:  $\mathcal{P}_p(\mathcal{X})$  endowed with  $\mathcal{W}_p$  is a Polish space. A closed set  $\mathcal{K} \subseteq \mathcal{P}_p(\mathcal{X})$  is weakly compact if and only if it has  $p$ -uniformly integrable moments (and hence tight). Specifically, given a sequence of probability measures  $\{Q_N\}_N \subseteq \mathcal{P}_p(\mathcal{X})$ , the following statements are equivalent:*

1.  $\mathcal{W}_p(Q_N, Q) \rightarrow 0$ .
2.  $Q_N$  converges weakly to  $Q$  and  $\{Q_N\}_N$  has  $p$ -uniformly integrable moments.
3.  $Q_N$  converges weakly to  $Q$  and the following holds

$$\int_{\mathcal{X}} \|\xi\|^p Q_N(d\xi) \xrightarrow{N \rightarrow \infty} \int_{\mathcal{X}} \|\xi\|^p Q(d\xi).$$

4. For any  $L > 0$  and any continuous function  $\ell : \mathcal{X} \rightarrow \mathbb{R}$  such that verifies  $|\ell(\xi)| \leq L(1 + \|\xi\|^p)$  for all  $\xi$ , the following holds

$$\int_{\mathcal{X}} \ell(\xi) Q_N(d\xi) \xrightarrow{N \rightarrow \infty} \int_{\mathcal{X}} \ell(\xi) Q(d\xi).$$

**Remark B.1.** *Proposition B.1 compiles results from Prop. 7.1.5 in [4] and Th. 7.12 in [132]. It implies that the topology generated by  $\mathcal{W}_p$  and the weak topology do coincide on any subset  $\mathcal{K}$  which has  $p$ -uniformly integrable moments. We note that assertion 2 in Proposition B.1 is reduced to weak convergence if  $\mathcal{X}$  is a compact set (see, for example, [109, Corollary 2.2.2]).*

## B.1.2 Concepts and technical results from probability trimmings

This section compiles some definitions and technical results which complement the theoretical core of this chapter.

**Definition B.6 (Contamination of a distribution).** Given two probabilities  $P, Q$  on  $\mathbb{R}^d$ , we say that  $P$  is a  $(1 - \alpha)$ -contaminated version of  $Q$ , if  $P = \alpha Q + (1 - \alpha)R$ , where  $R$  is some probability. A  $(1 - \alpha)$ -contamination neighbourhood of  $Q$  is the set of all  $(1 - \alpha)$ -contaminated versions of  $Q$  and will be denoted as  $\mathcal{F}_{1-\alpha}(Q)$ .

**Proposition B.2 (Section 2.2. from [2] and p.18 in [1]).** Let  $P, Q$  be probabilities on  $\mathbb{R}^d$  and  $\alpha \in (0, 1]$ , then

$$Q \in \mathcal{R}_{1-\alpha}(P) \iff P = \alpha Q + (1 - \alpha)R \iff P \in \mathcal{F}_{1-\alpha}(Q) \quad (\text{B.3})$$

for some probability  $R$ . Moreover, if  $D$  is a probability metric such that  $\mathcal{R}_{1-\alpha}(P)$  is closed for  $D$  over an appropriate set of probability distributions, then (B.3) is equivalent to  $D(Q, \mathcal{R}_{1-\alpha}(P)) = 0$ .

**Remark B.2.** As a particular case, if we consider  $D = \mathcal{W}_p$  over the set of probability distributions with finite  $p$ -th moment,  $\mathcal{P}_p$ , we have that, if  $P, Q \in \mathcal{P}_p$ , then  $Q \in \mathcal{R}_{1-\alpha}(P)$  if and only if  $\mathcal{W}_p(Q, \mathcal{R}_{1-\alpha}(P)) = 0$ .

### B.1.3 Topological properties of the ambiguity set

The following proposition formally establishes the topological properties of our ambiguity set:

**Proposition B.3.** Given  $\mathbb{Q} \in \mathcal{P}_p(\mathbb{R}^d)$ ,  $\alpha > 0$ , and  $\tilde{\rho} \geq \epsilon_{N\alpha}^p$ , the ambiguity set of problem (P),  $\hat{\mathcal{U}}_N(\alpha, \tilde{\rho})$ , is non-empty, tight, weakly compact, and  $p$ -uniformly integrable.

*Proof.* Proof The set  $\hat{\mathcal{U}}_N(\alpha, \tilde{\rho})$  is non-empty, because  $\tilde{\rho} \geq \epsilon_{N\alpha}^p$ . We can equivalently rewrite  $\hat{\mathcal{U}}_N(\alpha, \tilde{\rho})$  as

$$\left\{ Q_{\Xi} \in \mathcal{P}(\Xi) : \mathcal{W}_p^p(R, Q_{\Xi}) \leq \tilde{\rho} \text{ for some } R \in \mathcal{R}_{1-\alpha}(\hat{\mathbb{Q}}_N) \right\}.$$

If  $\alpha > 0$ , then the trimming set  $\mathcal{R}_{1-\alpha}(\hat{\mathbb{Q}}_N)$  is tight and weakly compact, see [33, Lemmas 2 and 3]. Furthermore,  $\hat{\mathcal{U}}_N(\alpha, \tilde{\rho})$  is a subset of

$$\mathcal{K} := \left\{ Q_{\Xi} \in \mathcal{P}(\mathbb{R}^d) : \mathcal{W}_p^p(R, Q_{\Xi}) \leq \tilde{\rho} \text{ for some } R \in \mathcal{R}_{1-\alpha}(\hat{\mathbb{Q}}_N) \right\}$$

which is tight and weakly compact by [115, Proposition 3]. The tightness of  $\hat{\mathcal{U}}_N(\alpha, \tilde{\rho})$  is trivially guaranteed, since any subset of a tight set is also tight. Hence, by Prokhorov's theorem, to demonstrate that  $\hat{\mathcal{U}}_N(\alpha, \tilde{\rho})$  is also weakly compact, it suffices to show that it is closed. For this purpose, let  $\{Q_{\Xi}^N\}_N$  be a sequence of probability measures in  $\hat{\mathcal{U}}_N(\alpha, \tilde{\rho})$  that converges weakly to  $Q$ . We need to show that  $Q$  is in  $\hat{\mathcal{U}}_N(\alpha, \tilde{\rho})$  too. In turn, since  $\hat{\mathcal{U}}_N(\alpha, \tilde{\rho})$  is a subset of  $\mathcal{K}$ , which is closed, this boils down to proving that the weak limit satisfies the condition  $Q \in \mathcal{P}(\Xi)$ , that is,  $Q(\Xi) = 1$ . Given that the

sequence  $\{Q_{\Xi}^N\}_N$  converges weakly to  $Q$  and the support set  $\tilde{\Xi}$  is closed, Portmanteau's theorem (see [26, Theorem 2.1]) tells us that  $\limsup_{N \rightarrow \infty} Q_{\Xi}^N(\tilde{\Xi}) = 1 \leq Q(\tilde{\Xi})$ . This implies that  $Q(\tilde{\Xi}) = 1$ .

Finally, the  $p$ -uniform integrability of our ambiguity set follows from Proposition B.1. To apply this proposition, we only need to check whether any distribution of  $\hat{\mathcal{U}}_N(\alpha, \tilde{\rho})$  has a finite  $p$ -th moment. From [3] (see p. 363 for the case  $p = 2$ , although the proof works similarly for any  $p \geq 1$ ), we know that  $\mathcal{R}_{1-\alpha}(\hat{\mathbb{Q}}_N) \subset \mathcal{P}_p(\mathbb{R}^d)$  if  $\mathbb{Q} \in \mathcal{P}_p(\mathbb{R}^d)$ . Now, assume that there is a distribution  $Q_{\Xi}$  in  $\hat{\mathcal{U}}_N(\alpha, \tilde{\rho})$  that does not have a finite  $p$ -th moment. If this were the case, we would have  $\mathcal{W}_p(Q_{\Xi}, R) = \infty$  for some  $R \in \mathcal{R}_{1-\alpha}(\hat{\mathbb{Q}}_N)$ , which is in contradiction with the fact that  $\mathcal{W}_p(Q_{\Xi}, R)$  must be less or equal to a finite  $\tilde{\rho}^{1/p}$ .  $\square$

#### B.1.4 Tractable reformulation and maximizer of problem (SP2)

Next we provide a more manageable reformulation of problem (SP2), which can be used directly to address the decision-making problems considered in our numerical experiments. However, we omit its proof, as it runs in parallel with that of [99, Theorem 4.2] and [87, Theorem 8]. See also [143]. Said reformulation relies on the following assumption.

**Assumption B.1.** *The region  $\tilde{\Xi}$  is a closed convex set, and  $f(\mathbf{x}, \boldsymbol{\xi}) := \max_{k \leq K} g_k(\mathbf{x}, \boldsymbol{\xi})$ , with  $g_k$ , for each  $k \leq K$ , being a proper, concave and upper semicontinuous function with respect to  $\boldsymbol{\xi}$  (for any fixed value of  $\mathbf{x} \in X$ ) and not identically  $\infty$  on  $\tilde{\Xi}$ .*

**Theorem B.2.** *Let  $p, q \geq 1$  such that  $\frac{1}{p} + \frac{1}{q} = 1$ . If Assumption B.1 holds, then for any value of  $\tilde{\rho} \geq \underline{\epsilon}_{N\alpha}^p$ , subproblem (SP2) is equivalent to the following finite convex problem:*

$$\begin{aligned}
 (\text{SP2}'') \quad & \inf_{\lambda, \bar{\mu}_i, \theta, \mathbf{v}_{ik}, \mathbf{v}'_{ik}, \mathbf{w}_{ik}, \mathbf{w}'_{ik}} \lambda \tilde{\rho} + \theta + \frac{1}{N\alpha} \sum_{i=1}^N \bar{\mu}_i \\
 & \text{s.t. } \bar{\mu}_i \geq [-g_k]^*((\mathbf{v}_{ik}, \mathbf{w}_{ik}) - (\mathbf{v}'_{ik}, \mathbf{w}'_{ik})) \\
 & \quad + S_{\tilde{\Xi}}((\mathbf{v}'_{ik}, \mathbf{w}'_{ik})) - \langle (\mathbf{v}_{ik}, \mathbf{w}_{ik}), (\hat{\mathbf{z}}_i, \hat{\mathbf{y}}_i) \rangle \\
 & \quad + \varphi(q)\lambda \left\| \frac{(\mathbf{v}_{ik}, \mathbf{w}_{ik})}{\lambda} \right\|_*^q - \theta, \quad \forall i \leq N, \forall k \leq K \\
 & \quad \lambda \geq 0 \\
 & \quad \bar{\mu}_i \geq 0, \quad \forall i \leq N
 \end{aligned}$$

where  $[-g_k]^*((\mathbf{v}_{ik}, \mathbf{w}_{ik}) - (\mathbf{v}'_{ik}, \mathbf{w}'_{ik}))$  is the conjugate function of  $-g_k$  evaluated at  $(\mathbf{v}_{ik}, \mathbf{w}_{ik}) - (\mathbf{v}'_{ik}, \mathbf{w}'_{ik})$  and  $S_{\tilde{\Xi}}$  is the support function of  $\tilde{\Xi}$ . Moreover,  $\varphi(q) = (q-1)^{q-1}/q^q$  if  $q > 1$ ,

and  $\varphi(1) = 1$ . If  $\lambda = 0$ , then  $0 \left\| \frac{(\mathbf{v}_{ik}, \mathbf{w}_{ik})}{0} \right\|_*^q := \lim_{\lambda \downarrow 0} \lambda \left\| \frac{(\mathbf{v}_{ik}, \mathbf{w}_{ik})}{\lambda} \right\|_*^q$ .

In problem (SP2''), we have suppressed the dependence of functions  $g_k$  on  $\mathbf{x}$  for ease of notation.

The following theorem serves to construct a maximizer (i.e., a worst-case distribution) of problem (SP2). Again, we omit its proof, as it is analogous to the proof of [99, Theorem 4.4] and [87, Theorem 9].

**Theorem B.3 (Worst-case distributions).** *Under the assumptions of Theorem B.2, the worst-case expectation in (SP2) is equal to the optimal objective value of the following finite convex optimization problem*

$$\left\{ \begin{array}{ll} \sup_{\gamma_{ik}, \mathbf{q}_{ik}} & \sum_{i=1}^N \sum_{k=1}^K \gamma_{ik} g_k(\hat{\xi}_i - \frac{\mathbf{q}_{ik}}{\gamma_{ik}}) \\ \text{s.t.} & \sum_{i=1}^N \sum_{k=1}^K \gamma_{ik} \left\| \frac{\mathbf{q}_{ik}}{\gamma_{ik}} \right\|^p \leq \tilde{\rho} \\ & \sum_{i=1}^N \sum_{k=1}^K \gamma_{ik} = 1 \\ & \sum_{k=1}^K \gamma_{ik} \leq \frac{1}{N\alpha} \quad \forall i \leq N \\ & \gamma_{ik} \geq 0 \quad \forall i \leq N, \quad \forall k \leq K \\ & \hat{\xi}_i - \frac{\mathbf{q}_{ik}}{\gamma_{ik}} \in \tilde{\Xi} \quad \forall i \leq N, \quad \forall k \leq K \end{array} \right.$$

where  $0g_k(\hat{\xi}_i - \frac{\mathbf{q}_{ik}}{0})$  is interpreted as the value which makes the function  $\gamma_{ik}g_k(\hat{\xi}_i - \frac{\mathbf{q}_{ik}}{\gamma_{ik}})$  upper semicontinuous at  $(\mathbf{q}_{ik}, \gamma_{ik}) = (\mathbf{q}_{ik}, 0)$ . Also, the constraint  $\hat{\xi}_i - \mathbf{q}_{ik}/0 \in \tilde{\Xi}$  means that  $\mathbf{q}_{ik}$  is in the recession cone of  $\tilde{\Xi}$ , and  $0 \|\mathbf{q}_{ik}/0\|^p$  is understood as  $\lim_{\gamma_{ik} \downarrow 0} \gamma_{ik} \|\mathbf{q}_{ik}/\gamma_{ik}\|^p$ .

Moreover, if we assume that  $p > 1$  or that  $\tilde{\Xi}$  is bounded (with  $p \geq 1$ ), then if  $(\gamma_{ik}^*, \mathbf{q}_{ik}^*)$  maximizes the problem above, we have that the discrete probability distribution  $Q_{\tilde{\Xi}}$  defined as

$$Q_{\tilde{\Xi}} = \sum_{i=1}^N \sum_{k=1}^K \gamma_{ik}^* \delta_{\xi_{ik}^*}$$

where  $\xi_{ik}^* := \hat{\xi}_i - \frac{\mathbf{q}_{ik}^*}{\gamma_{ik}^*} \in \tilde{\Xi}$ , represents a maximizer of the worst-case expectation problem.

## B.2 Proofs of Chapter 4

### B.2.1 Proof of Lemma 4.1

We will prove the lemma by contradiction. Suppose there are two different probability distributions  $Q_{\tilde{\Xi}}$  and  $Q'_{\tilde{\Xi}}$  such that

$$D(\mathcal{R}_{1-\alpha}(Q), Q_{\tilde{\Xi}}) = D(\mathcal{R}_{1-\alpha}(Q), Q'_{\tilde{\Xi}}) = 0$$

and  $Q_{\Xi}(\tilde{\Xi}) = Q'_{\Xi}(\tilde{\Xi}) = 1$ .

Because  $D(\mathcal{R}_{1-\alpha}(Q), Q_{\Xi}) = D(\mathcal{R}_{1-\alpha}(Q), Q'_{\Xi}) = 0$ , we know by Proposition B.2 above that  $Q_{\Xi}, Q'_{\Xi} \in \mathcal{R}_{1-\alpha}(Q)$ . Therefore, again applying Proposition B.2, we have

$$Q = \alpha Q_{\Xi} + (1 - \alpha)R$$

$$Q = \alpha Q'_{\Xi} + (1 - \alpha)R'$$

for some probabilities  $R$  and  $R'$  with  $R(\tilde{\Xi}) = R'(\tilde{\Xi}) = 0$ .

Since, by hypothesis,  $Q_{\Xi}$  and  $Q'_{\Xi}$  are different, there must exist an event  $A \subset \tilde{\Xi}$  such that  $Q_{\Xi}(A) \neq Q'_{\Xi}(A)$ . We take that event and compute  $Q(A)$  as follows:

$$Q(A) = \alpha Q_{\Xi}(A) + (1 - \alpha)R(A) = \alpha Q'_{\Xi}(A) + (1 - \alpha)R'(A),$$

which renders a contradiction given that  $R(A) = R'(A) = 0$ .

□

### B.2.2 Proof of Proposition 4.1

First of all, we need the following preliminary results:

**Corollary B.1 (Corollary 3.12 from [1]).** *Given two probabilities  $P, Q \in \mathcal{P}_p(\mathbb{R}^d)$  and  $\alpha \in (0, 1)$ , there exists  $P_{1-\alpha} \in \mathcal{F}_{1-\alpha}(Q)$  such that  $P_{1-\alpha} = \alpha Q + (1 - \alpha)R_{1-\alpha}$  for some  $R_{1-\alpha} \in \mathcal{R}_{\alpha}(P)$  and  $\mathcal{W}_p(P, P_{1-\alpha}) = \min_{R \in \mathcal{F}_{1-\alpha}(Q)} \mathcal{W}_p(P, R)$ .*

**Proposition B.4 (Proposition 3.14 from [1]).** *Take  $P, Q \in \mathcal{P}_p(\mathbb{R}^d)$ . If  $\alpha \in (0, 1)$ , then*

$$\mathcal{W}_p^p(P, \mathcal{F}_{1-\alpha}(Q)) = \alpha \mathcal{W}_p^p(\mathcal{R}_{1-\alpha}(P), Q)$$

Moreover, if  $\hat{P}_{1-\alpha} \in \mathcal{R}_{1-\alpha}(P)$  is such that  $\mathcal{W}_p(\hat{P}_{1-\alpha}, Q) = \mathcal{W}_p(\mathcal{R}_{1-\alpha}(P), Q)$ , then if we construct the probability measure  $\tilde{P}_{1-\alpha} = \frac{1}{1-\alpha} \left( P - \alpha \hat{P}_{1-\alpha} \right)$ , we have that  $P_{1-\alpha} := \alpha Q + (1 - \alpha)\tilde{P}_{1-\alpha} \in \mathcal{F}_{1-\alpha}(Q)$  and  $\mathcal{W}_p(P, P_{1-\alpha}) = \mathcal{W}_p(P, \mathcal{F}_{1-\alpha}(Q))$ .

We begin by proving the first claim of Proposition 4.1.

We show that every feasible solution of (SP1) can be mapped into a feasible solution of (SP2) with the same objective function value. To this end, take  $Q$  as a feasible solution of (SP1) and let  $Q_{\Xi}$  be the  $Q$ -conditional probability measure given  $\xi \in \tilde{\Xi}$ . Take  $\hat{Q}_N$  and  $Q_{\Xi}$  as the two probabilities in Corollary B.1 with  $\alpha \in (0, 1)$ . There exists  $Q_{1-\alpha} \in \mathcal{F}_{1-\alpha}(Q_{\Xi})$  such that  $Q_{1-\alpha} = \alpha Q_{\Xi} + (1 - \alpha)\tilde{Q}_{1-\alpha}$ , with  $\tilde{Q}_{1-\alpha} \in \mathcal{R}_{\alpha}(\hat{Q}_N)$  and  $\mathcal{W}_p(\hat{Q}_N, Q_{1-\alpha}) = \mathcal{W}_p(\hat{Q}_N, \mathcal{F}_{1-\alpha}(Q_{\Xi}))$ . Furthermore, it automatically follows from Proposition B.4 that

$$\mathcal{W}_p^p(\hat{Q}_N, \mathcal{F}_{1-\alpha}(Q_{\Xi})) = \alpha \mathcal{W}_p^p(\mathcal{R}_{1-\alpha}(\hat{Q}_N), Q_{\Xi})$$



Since  $Q \in \mathcal{F}_{1-\alpha}(Q_{\Xi})$ , we deduce that  $\mathcal{W}_p^p(\hat{Q}_N, \mathcal{F}_{1-\alpha}(Q_{\Xi})) \leq \mathcal{W}_p^p(\hat{Q}_N, Q) \leq \tilde{\rho} \cdot \alpha$ . Hence, it holds that  $\mathcal{W}_p^p(\mathcal{R}_{1-\alpha}(\hat{Q}_N), Q_{\Xi}) \leq \tilde{\rho}$ . In other words,  $Q_{\Xi}$  is feasible in (SP2). Besides, since  $Q_{\Xi}$  is the  $Q$ -conditional probability measure given  $\xi \in \Xi$ , we have that

$$\mathbb{E}_Q [f(\mathbf{x}, \xi) \mid \xi \in \Xi] = \frac{1}{\alpha} \mathbb{E}_Q [f(\mathbf{x}, \xi) \mathbb{I}_{\Xi}(\xi)] = \mathbb{E}_{Q_{\Xi}} [f(\mathbf{x}, \xi)] \text{ a.s.}$$

Next we prove the second claim of the proposition. To this end, first we show that, if  $\hat{Q}_N(\Xi) = 0$ , then every feasible solution of (SP2) can also be mapped into a feasible solution of (SP1) with the same objective function value. To this end, take  $Q_{\Xi}$  feasible in (SP2) and consider  $\hat{Q}_{1-\alpha} \in \mathcal{R}_{1-\alpha}(\hat{Q}_N)$  such that  $\mathcal{W}_p(\hat{Q}_{1-\alpha}, Q_{\Xi}) = \mathcal{W}_p(\mathcal{R}_{1-\alpha}(\hat{Q}_N), Q_{\Xi})$ . Fix  $\tilde{Q}_{1-\alpha} = \frac{1}{1-\alpha}(\hat{Q}_N - \alpha \hat{Q}_{1-\alpha})$ . By Proposition B.4, we have

$$Q_{1-\alpha} = \alpha Q_{\Xi} + (1-\alpha) \tilde{Q}_{1-\alpha} = \alpha Q_{\Xi} + \hat{Q}_N - \alpha \hat{Q}_{1-\alpha} \in \mathcal{F}_{1-\alpha}(Q_{\Xi})$$

Hence,  $Q_{1-\alpha}(\Xi) = \alpha$ , because  $\hat{Q}_N(\Xi)$  gives zero measure to  $\Xi$  and so does any of its  $(1-\alpha)$ -trimmings. Besides, we have that

$$\mathcal{W}_p^p(\hat{Q}_N, Q_{1-\alpha}) = \mathcal{W}_p^p(\hat{Q}_N, \mathcal{F}_{1-\alpha}(Q_{\Xi})) = \alpha \mathcal{W}_p^p(\mathcal{R}_{1-\alpha}(\hat{Q}_N), Q_{\Xi}) \leq \alpha \tilde{\rho}.$$

Therefore,  $Q_{1-\alpha}$  is feasible in (SP1) and  $Q_{\Xi}$  is the  $Q_{1-\alpha}$ -conditional probability measure given  $\xi \in \Xi$ .

Finally, if  $\alpha = 1$ , then  $\mathcal{R}_{1-\alpha}(\hat{Q}_N) = \hat{Q}_N$ ,  $\mathbb{E}_Q [f(\mathbf{x}, \xi) \mid \xi \in \Xi] = \mathbb{E}_Q [f(\mathbf{x}, \xi)]$  and the mapping is direct, namely,  $Q = Q_{\Xi}$ .  $\square$

### B.2.3 Proof of Theorem 4.1

Thanks to Lemma 4.2, the subproblem (SP2) can be written equivalently as follows:

$$\begin{aligned} \text{(SP2)} \quad & \sup_{Q_{\Xi}; \mathbf{b} \in \Delta(\alpha_N)} \mathbb{E}_{Q_{\Xi}} [f(\mathbf{x}, \xi)] \\ \text{s.t.} \quad & Q_{\Xi}(\Xi) = 1 \\ & \mathcal{W}_p \left( \sum_{i=1}^N b_i \delta_{\hat{\xi}_i}, Q_{\Xi} \right) \leq \tilde{\rho}^{1/p} \end{aligned}$$

where  $\Delta(\alpha_N)$  stands for the set of constraints  $\{0 \leq b_i \leq \frac{1}{N\alpha_N}, \forall i \leq N, \sum_{i=1}^N b_i = 1\}$ .

which, in turn, can be reformulated as

$$\begin{aligned}
 & \left\{ \begin{array}{l} \sup_{Q_{\Xi}; \Pi; \mathbf{b} \in \Delta(\alpha_N)} \int_{\Xi} f(\mathbf{x}, (\mathbf{z}, \mathbf{y})) Q_{\Xi}(d\mathbf{z}, d\mathbf{y}) \\ \text{s.t.} \\ \int_{\Xi} Q_{\Xi}(d\mathbf{z}, d\mathbf{y}) = 1 \\ \left( \int_{\Xi \times \Xi} \|(\mathbf{z}, \mathbf{y}) - (\mathbf{z}, \mathbf{y}')\|^p \Pi(d(\mathbf{z}, \mathbf{y}), d(\mathbf{z}, \mathbf{y}')) \right)^{1/p} \leq \tilde{\rho}^{1/p} \\ \left\{ \begin{array}{l} \Pi \text{ is a joint distribution of } (\mathbf{z}, \mathbf{y}) \text{ and } (\mathbf{z}, \mathbf{y}') \\ \text{with marginals } Q_{\Xi} \text{ and } \sum_{i=1}^N b_i \delta_{\hat{\mathbf{z}}_i}, \text{ respectively} \end{array} \right. \end{array} \right. \quad (\text{B.4}) \\
 & = \left\{ \begin{array}{l} \sup_{Q_{\Xi}^i; \mathbf{b} \in \Delta(\alpha_N)} \sum_{i=1}^N b_i \int_{\Xi} f(\mathbf{x}, (\mathbf{z}, \mathbf{y})) Q_{\Xi}^i(d\mathbf{z}, d\mathbf{y}) \\ \text{s.t.} \\ \int_{\Xi} Q_{\Xi}^i(d\mathbf{z}, d\mathbf{y}) = 1, \forall i \leq N \\ \sum_{i=1}^N b_i \int_{\Xi} \|(\mathbf{z}, \mathbf{y}) - (\hat{\mathbf{z}}_i, \hat{\mathbf{y}}_i)\|^p Q_{\Xi}^i(d\mathbf{z}, d\mathbf{y}) \leq \tilde{\rho} \end{array} \right. \quad (\text{B.5})
 \end{aligned}$$

where reformulation (B.5) follows from the fact that the marginal distribution of  $(\mathbf{z}, \mathbf{y})'$  is the discrete distribution supported on points  $(\hat{\mathbf{z}}_i, \hat{\mathbf{y}}_i)$ , with probability masses  $b_i$ ,  $i = 1, \dots, N$ . Thus,  $\Pi$  is completely determined by the conditional distributions  $Q_{\Xi}^i$  of  $(\mathbf{z}, \mathbf{y})$  given  $(\mathbf{z}, \mathbf{y})' = (\hat{\mathbf{z}}_i, \hat{\mathbf{y}}_i)$ ,  $i = 1, \dots, N$ , that is,

$$\Pi(d(\mathbf{z}, \mathbf{y}), d(\mathbf{z}, \mathbf{y}')) = \sum_{i=1}^N b_i \delta_{(\hat{\mathbf{z}}_i, \hat{\mathbf{y}}_i)}(d(\mathbf{z}, \mathbf{y}')) Q_{\Xi}^i(d(\mathbf{z}, \mathbf{y}))$$

Now we split up the supremum into two:

$$\sup_{\mathbf{b} \in \Delta(\alpha_N)} \sup_{Q_{\Xi}^i, \forall i \leq N} \sum_{i=1}^N b_i \int_{\Xi} f(\mathbf{x}, (\mathbf{z}, \mathbf{y})) Q_{\Xi}^i(d\mathbf{z}, d\mathbf{y}) \quad (\text{B.6a})$$

$$\text{s.t.} \int_{\Xi} Q_{\Xi}^i(d\mathbf{z}, d\mathbf{y}) = 1, \forall i \leq N \quad (\text{B.6b})$$

$$\sum_{i=1}^N b_i \int_{\Xi} \|(\mathbf{z}, \mathbf{y}) - (\hat{\mathbf{z}}_i, \hat{\mathbf{y}}_i)\|^p Q_{\Xi}^i(d\mathbf{z}, d\mathbf{y}) \leq \tilde{\rho} \quad (\text{B.6c})$$

If we set  $\lambda$  as the dual variable of constraint (B.6c), then using standard duality arguments, we can equivalently rewrite the inner supremum as

$$\sup_{\mathbf{b} \in \Delta(\alpha_N)} \inf_{\lambda \geq 0} \sup_{Q_{\Xi}^i, \forall i \leq N} \lambda \tilde{\rho} + \sum_{i=1}^N b_i \int_{\Xi} (f(\mathbf{x}, (\mathbf{z}, \mathbf{y})) - \lambda \|(\mathbf{z}, \mathbf{y}) - (\hat{\mathbf{z}}_i, \hat{\mathbf{y}}_i)\|^p) Q_{\Xi}^i(d\mathbf{z}, d\mathbf{y}) \quad (\text{B.7})$$

$$\text{s.t. } \int_{\tilde{\Xi}} Q_{\tilde{\Xi}}^i(d\mathbf{z}, d\mathbf{y}) = 1, \quad \forall i \leq N \quad (\text{B.8})$$

$$= \sup_{\mathbf{b} \in \Delta(\alpha_N)} \inf_{\lambda \geq 0} \lambda \tilde{\rho} + \sum_{i=1}^N b_i \sup_{(\mathbf{z}, \mathbf{y}) \in \tilde{\Xi}} (f(\mathbf{x}, (\mathbf{z}, \mathbf{y})) - \lambda \|(\mathbf{z}, \mathbf{y}) - (\hat{\mathbf{z}}_i, \hat{\mathbf{y}}_i)\|^p) \quad (\text{B.9})$$

$$= \inf_{\lambda \geq 0} \sup_{\mathbf{b} \in \Delta(\alpha_N)} \lambda \tilde{\rho} + \sum_{i=1}^N b_i \sup_{(\mathbf{z}, \mathbf{y}) \in \tilde{\Xi}} (f(\mathbf{x}, (\mathbf{z}, \mathbf{y})) - \lambda \|(\mathbf{z}, \mathbf{y}) - (\hat{\mathbf{z}}_i, \hat{\mathbf{y}}_i)\|^p) \quad (\text{B.10})$$

$$= \inf_{\lambda \geq 0; \bar{\mu}_i, \forall i \leq N; \theta \in \mathbb{R}} \lambda \tilde{\rho} + \theta + \frac{1}{N\alpha} \sum_{i=1}^N \bar{\mu}_i \quad (\text{B.11})$$

$$\text{s.t. } \bar{\mu}_i + \theta \geq \sup_{(\mathbf{z}, \mathbf{y}) \in \tilde{\Xi}} (f(\mathbf{x}, (\mathbf{z}, \mathbf{y})) - \lambda \|(\mathbf{z}, \mathbf{y}) - (\hat{\mathbf{z}}_i, \hat{\mathbf{y}}_i)\|^p), \quad \forall i \leq N \quad (\text{B.12})$$

$$\bar{\mu}_i \geq 0, \quad \forall i \leq N \quad (\text{B.13})$$

where we have swapped the supremum and the infimum in (B.9) by appealing to Sion's min-max theorem [127], given that the objective function in (B.9) is linear in the  $b_i, i = 1, \dots, N$ , over a compact convex set, and a positively weighted sum of convex functions in  $\lambda$ .  $\square$

**Remark B.3 (Limiting case  $\alpha = 0$ ).** If  $\alpha = 0$ ,  $\mathcal{R}_1(\hat{\mathbb{Q}}_N) = \{\sum_{i=1}^N b_i \delta_{\hat{\xi}_i} \text{ such that } b_i \geq 0, \forall i = 1, \dots, N, \text{ and } \sum_{i=1}^N b_i = 1\}$ . Therefore, dual variables  $\bar{\mu}_i, \forall i \leq N$ , do not appear in (B.11)–(B.13) in this case. Similarly, if  $\frac{1}{N\alpha} \geq 1$ , the constraints  $b_i \leq \frac{1}{N\alpha}, \forall i \leq N$ , become redundant and hence we can set  $\bar{\mu}_i = 0, \forall i \leq N$ .

## B.2.4 Proof of Proposition 4.2

Because of Lemma 4.3 we have

$$\mathbb{Q}^N \left( \mathcal{W}_p(\mathcal{R}_{1-\alpha}(\hat{\mathbb{Q}}_N), \mathbb{Q}_{\tilde{\Xi}}) - \mathcal{W}_p(\mathcal{R}_{1-\alpha}(\mathbb{Q}), \mathbb{Q}_{\tilde{\Xi}}) \geq \epsilon \right) \leq \mathbb{Q}^N \left( \mathcal{W}_p^p(\hat{\mathbb{Q}}_N, \mathbb{Q}) \geq \alpha \epsilon^p \right)$$

where the right-hand side of this inequality is upper bounded by (4.13) according to [59, Theorem 2].  $\square$

## B.2.5 Proof of Theorem 4.2

For problem  $(P_{(\alpha, \tilde{\rho}_N)})$  to be feasible, we must have  $\tilde{\rho}_N \geq \epsilon_{N\alpha}^p$ . Furthermore,

$$\mathcal{W}_p(\mathcal{R}_{1-\alpha}(\mathbb{Q}), \mathbb{Q}_{\tilde{\Xi}}) = 0$$

in (4.12) because of Lemma 4.1. Hence, Proposition 4.2 ensures that

$$\mathbb{Q}^N \left( \mathbb{Q}_{\tilde{\Xi}} \in \hat{\mathcal{U}}_N(\alpha, \tilde{\rho}_N) \right) \geq 1 - \beta \text{ for any } \tilde{\rho}_N \geq \epsilon_{N,p,\alpha}^p(\beta)$$

It follows then

$$\mathbb{E}_{\mathbb{Q}}[f(\hat{\mathbf{x}}_N, \boldsymbol{\xi}) \mid \boldsymbol{\xi} \in \tilde{\Xi}] = \mathbb{E}_{\mathbb{Q}_{\tilde{\Xi}}}[f(\hat{\mathbf{x}}_N, \boldsymbol{\xi})] \leq \hat{J}_N := \sup_{Q_{\tilde{\Xi}}} \left\{ \mathbb{E}_{Q_{\tilde{\Xi}}}[f(\hat{\mathbf{x}}_N, \boldsymbol{\xi})] : Q_{\tilde{\Xi}} \in \hat{\mathcal{U}}_N(\alpha, \tilde{\rho}_N) \right\}$$

with probability at least  $1 - \beta$ .  $\square$

### B.2.6 Proof of Lemma 4.4

Take  $N$  large enough and let  $\hat{Q}_{N/\tilde{\Xi}}$  be the conditional probability distribution of  $\hat{Q}_N$  given  $\boldsymbol{\xi} \in \tilde{\Xi}$ . We have

$$\mathcal{W}_p(Q_{\tilde{\Xi}}^N, \mathbb{Q}_{\tilde{\Xi}}) \leq \mathcal{W}_p(Q_{\tilde{\Xi}}^N, \hat{Q}_{N/\tilde{\Xi}}) + \mathcal{W}_p(\hat{Q}_{N/\tilde{\Xi}}, \mathbb{Q}_{\tilde{\Xi}})$$

We show that the two terms on the right-hand side of the above inequality vanish with probability one as  $N$  grows to infinity. We start with  $\mathcal{W}_p(\hat{Q}_{N/\tilde{\Xi}}, \mathbb{Q}_{\tilde{\Xi}})$ .

Let  $I$  denote the subset of observations  $\hat{\boldsymbol{\xi}}_i := (\hat{\mathbf{z}}_i, \hat{\mathbf{y}}_i)$  for  $i = 1, \dots, N$ , such that  $\hat{\boldsymbol{\xi}}_i \in \tilde{\Xi}$ . It follows from the Strong Law of Large Numbers that  $\hat{Q}_N(\tilde{\Xi}) = \frac{|I|}{N} = \alpha_N \rightarrow \alpha$  almost surely. Besides, since the sequence  $\beta_N, N \in \mathbb{N}$  is summable and  $\lim_{N \rightarrow \infty} \epsilon_N(\beta_N) \rightarrow 0$ , the Borel-Cantelli Lemma and Proposition 4.2 implies

$$\mathcal{W}_p(\mathcal{R}_{1-\alpha}(\hat{Q}_N), \mathbb{Q}_{\tilde{\Xi}}) \rightarrow 0 \text{ a.s.}$$

Then, from Lemma 4.1, we deduce that  $\mathcal{W}_p(\hat{Q}_{N/\tilde{\Xi}}, \mathbb{Q}_{\tilde{\Xi}}) \rightarrow 0$  with probability one.

We can deal with the term  $\mathcal{W}_p(Q_{\tilde{\Xi}}^N, \hat{Q}_{N/\tilde{\Xi}})$  in a similar fashion, except for the subtle difference that, in this case, we require  $\tilde{\rho}_N = \max(\epsilon_{N,p,\alpha}^p(\beta_N), \underline{\epsilon}_{N\alpha}^p)$ , so that, for all  $N \in \mathbb{N}$ , problem  $P_{(\alpha, \tilde{\rho}_N)}$  delivers a feasible  $Q_{\tilde{\Xi}}^N$  in the sequence. Hence, in order to prove that  $\mathcal{W}_p(Q_{\tilde{\Xi}}^N, \hat{Q}_{N/\tilde{\Xi}}) \rightarrow 0$  almost surely, we need to show that  $\lim_{N \rightarrow \infty} \underline{\epsilon}_{N\alpha} = 0$  with probability one. This is something that can be directly deduced from the definition of  $\underline{\epsilon}_{N\alpha}$ , namely,

$$\underline{\epsilon}_{N\alpha}^p := \mathcal{W}_p^p(\mathcal{R}_{1-\alpha}(\hat{Q}_N), \mathcal{P}_p(\tilde{\Xi})) = \min_{Q' \in \mathcal{P}_p(\tilde{\Xi})} \mathcal{W}_p^p(\mathcal{R}_{1-\alpha}(\hat{Q}_N), Q') \quad (\text{B.14})$$

$$\leq \mathcal{W}_p^p(\mathcal{R}_{1-\alpha}(\hat{Q}_N), \mathbb{Q}_{\tilde{\Xi}}) \rightarrow 0 \text{ a.s.} \quad (\text{B.15})$$

$\square$

**Remark B.4.** Note that, by Equation (4.9) in Definition 4.2, we have that  $\underline{\epsilon}_{N\alpha} > 0$  if and only if

$$\lceil N\alpha \rceil > |I| \Leftrightarrow \frac{\lceil N\alpha \rceil}{N} > \frac{|I|}{N} = \alpha_N = \hat{Q}_N(\tilde{\Xi}) \Leftrightarrow \alpha > \alpha_N$$

### B.2.7 Proof of Theorem 4.3

We omit the proof, because it is essentially the same as the one in [99, Theorem 3.6], except that, since we are working with  $p \geq 1$ , we additionally require that  $f(\mathbf{x}, \boldsymbol{\xi})$  be continuous in  $\boldsymbol{\xi}$  so that we can make use of Theorem 7.12 from [132].  $\square$

### B.2.8 Proof of Proposition 4.3

The proof of the proposition is trivial and directly follows from the fact that  $\mathcal{R}_{1-\alpha}(\widehat{\mathbb{Q}}_N) \subset \mathcal{R}_{1-\alpha'}(\widehat{\mathbb{Q}}_N)$ , if  $\alpha' \leq \alpha$ , and that  $\mathcal{R}_{1-\alpha}(\widehat{\mathbb{Q}}_N) = \mathcal{R}_{1-\alpha'}(\widehat{\mathbb{Q}}_N)$  if, besides,  $\frac{1}{N\alpha} \geq 1$ .  $\square$

### B.2.9 Proof of Theorem 4.4

For problem  $(P_{(\alpha_N, \tilde{\rho}_N)})$  to be feasible, we need  $\tilde{\rho}_N \geq \epsilon_{N\alpha_N}^p$ .

The proof essentially relies on upper bounding the term  $\mathcal{W}_p(\mathcal{R}_{1-\alpha}(\mathbb{Q}), \mathbb{Q}_{\Xi})$  that appears in Equation (4.12) of Proposition 4.2. To that end, define  $\alpha(r) = \tilde{C}r^{d_{\mathbf{z}}}$ , for all  $0 < r \leq r_0$ . Set  $\alpha_0 := \alpha(r_0)$ . Let  $\mathbb{Q}_{B(\mathbf{z}^*, r) \times \Xi_{\mathbf{y}}}$  be the probability measure of  $(\mathbf{z}, \mathbf{y})$  conditional on  $(\mathbf{z}, \mathbf{y}) \in B(\mathbf{z}^*, r) \times \Xi_{\mathbf{y}}$  and let  $\mathbb{Q}_{B(\mathbf{z}^*, r)}$  be its  $\mathbf{y}$ -marginal. Note that, by Assumption 4.3.1,  $\mathbb{Q}_{B(\mathbf{z}^*, r) \times \Xi_{\mathbf{y}}} \in \mathcal{R}_{1-\alpha(r)}(\mathbb{Q})$  provided that  $0 < r \leq r_0$ .

Furthermore, according to Theorem 3.5.2 in [57], there exists a positive constant  $A$  such that

$$\text{Hell}(\mathbb{Q}_{B(\mathbf{z}^*, r)}, \mathbb{Q}_{\Xi}) \leq Ar^2$$

uniformly for  $0 < r < r_0$ , where Hell stands for *Hellinger distance*.

From Equation (5.1) in [120] and Assumption 4.3.2 we know that

$$\mathcal{W}_p(\mathbb{Q}_{B(\mathbf{z}^*, r)}, \mathbb{Q}_{\Xi}) \leq M^{\frac{p-1}{p}} \mathcal{W}_1(\mathbb{Q}_{B(\mathbf{z}^*, r)}, \mathbb{Q}_{\Xi})^{1/p}$$

In turn, from [65] we have that  $\mathcal{W}_1(\mathbb{Q}_{B(\mathbf{z}^*, r)}, \mathbb{Q}_{\Xi}) \leq M \cdot \text{Hell}(\mathbb{Q}_{B(\mathbf{z}^*, r)}, \mathbb{Q}_{\Xi})$ . Hence,

$$\begin{aligned} \mathcal{W}_p^p(\mathbb{Q}_{B(\mathbf{z}^*, r)}, \mathbb{Q}_{\Xi}) &\leq M^p \text{Hell}(\mathbb{Q}_{B(\mathbf{z}^*, r)}, \mathbb{Q}_{\Xi}) \\ \mathcal{W}_p(\mathbb{Q}_{B(\mathbf{z}^*, r)}, \mathbb{Q}_{\Xi}) &\leq MA^{1/p} r^{2/p}, \quad 0 < r \leq r_0 \end{aligned}$$

Thus,

$$\mathcal{W}_p(\mathbb{Q}_{B(\mathbf{z}^*, r) \times \Xi_{\mathbf{y}}}, \mathbb{Q}_{\Xi}) \leq r + MA^{1/p} r^{2/p}, \quad 0 < r \leq r_0$$

Since  $\mathbb{Q}_{B(\mathbf{z}^*, r) \times \Xi_{\mathbf{y}}} \in \mathcal{R}_{1-\alpha(r)}(\mathbb{Q})$  for all  $0 < r \leq r_0$ , it holds

$$\mathcal{W}_p(\mathcal{R}_{1-\alpha(r)}(\mathbb{Q}), \mathbb{Q}_{\Xi}) \leq \mathcal{W}_p(\mathbb{Q}_{B(\mathbf{z}^*, r) \times \Xi_{\mathbf{y}}}, \mathbb{Q}_{\Xi}) \leq r + MA^{1/p} r^{2/p}$$

which we can express in terms of  $\alpha$  as

$$\mathcal{W}_p(\mathcal{R}_{1-\alpha}(\mathbb{Q}), \mathbb{Q}_{\Xi}) \leq \frac{\alpha^{1/d_{\mathbf{z}}}}{\tilde{C}^{1/d_{\mathbf{z}}}} + A^{1/p} M \frac{\alpha^{2/(pd_{\mathbf{z}})}}{\tilde{C}^{2/(pd_{\mathbf{z}})}}$$

$$\mathcal{W}_p(\mathcal{R}_{1-\alpha}(\mathbb{Q}), \mathbb{Q}_{\Xi}) = O\left(\alpha^{\min\{1, 2/p\}/d_{\mathbf{z}}}\right)$$

provided that  $0 < \alpha \leq \alpha_0$ .  $\square$

### B.2.10 Proof of Lemma 4.5

First, we need to provide conditions under which  $\mathcal{W}_p(\mathcal{R}_{1-\alpha}(\hat{\mathbb{Q}}_N), \mathbb{Q}_{\Xi}) \rightarrow 0$  a.s. Since  $\Xi$  is compact and  $\mathcal{W}_{p-1}(\mathcal{R}_{1-\alpha}(\hat{\mathbb{Q}}_N), \mathbb{Q}_{\Xi}) \leq \mathcal{W}_p(\mathcal{R}_{1-\alpha}(\hat{\mathbb{Q}}_N), \mathbb{Q}_{\Xi})$ , we can take  $p > d/2$  and  $\alpha_N$  such that  $\frac{N\alpha_N^2}{\log(N)} \rightarrow \infty$ , so that the probabilities (4.12) becomes summable over  $N$  for any arbitrarily small  $\epsilon$ . In this way, we can choose a sequence  $\beta_N \in (0, 1)$ ,  $N \in \mathbb{N}$ , such that  $\sum_{N=1}^{\infty} \beta_N < \infty$  and  $\lim_{N \rightarrow \infty} \epsilon_{N,p,\alpha_N}(\beta_N) \rightarrow 0$ . With this choice, we have

$$\begin{aligned} \mathbb{Q}^{\infty} \left[ \lim_{N \rightarrow \infty} \mathcal{W}_p(\mathcal{R}_{1-\alpha_N}(\hat{\mathbb{Q}}_N), \mathbb{Q}_{\Xi}) - \mathcal{W}_p(\mathcal{R}_{1-\alpha_N}(\mathbb{Q}), \mathbb{Q}_{\Xi}) = 0 \right] \\ = \mathbb{Q}^{\infty} \left[ \lim_{N \rightarrow \infty} \mathcal{W}_p(\mathcal{R}_{1-\alpha_N}(\hat{\mathbb{Q}}_N), \mathbb{Q}_{\Xi}) = 0 \right] = 1 \end{aligned}$$

because  $\mathcal{W}_p(\mathcal{R}_{1-\alpha_N}(\mathbb{Q}), \mathbb{Q}_{\Xi}) = O\left(\alpha_N^{2/pd_{\mathbf{z}}}\right) \rightarrow 0$  for  $\alpha_N \rightarrow 0$ .

Since,  $\mathbb{Q}_{\Xi} \in \mathcal{R}_{1-\alpha_N}(\hat{\mathbb{Q}}_N)$  a.s. in the limit and, by definition,  $\mathbb{Q}_{\Xi}(\tilde{\Xi}) = 1$ , we have that  $\mathbb{Q}_{\Xi} \in \hat{\mathcal{U}}_N(\alpha_N, \tilde{\rho}_N)$  for  $N$  sufficiently large, with both  $\alpha_N, \tilde{\rho}_N \rightarrow 0$ .

For its part, because  $Q_{\Xi}^N \in \hat{\mathcal{U}}_N(\alpha_N, \tilde{\rho}_N)$ , this means that  $\mathcal{W}_p(\mathcal{R}_{1-\alpha_N}(\hat{\mathbb{Q}}_N), Q_{\Xi}^N) \leq \tilde{\rho}_N$ . Take  $N$  large enough, set  $\tilde{\rho}_N$  arbitrarily close to  $\epsilon_{N\alpha_N}^p$  and notice that  $\hat{\mathcal{U}}_N(\alpha_N, \epsilon_{N\alpha_N}^p)$  boils down to one single probability measure, the one made up of the  $N\alpha_N$  data points of  $\hat{\mathbb{Q}}_N$  that are the closest to  $\tilde{\Xi}$ . In addition, we have  $\epsilon_{N\alpha_N}^p \rightarrow 0$  with probability one. To see this, take  $K := \lceil N\alpha_N \rceil$  and note that

$$\epsilon_{N\alpha_N}^p \leq \text{dist}(\hat{\xi}_{K:N}, \tilde{\Xi}) \rightarrow \|\hat{\mathbf{z}}_{K:N} - \mathbf{z}^*\| \rightarrow 0$$

almost surely provided that  $\alpha_N \rightarrow 0$  (see [24, Lemmas 2.2 and 2.3]), where  $\hat{\mathbf{z}}_{K:N}$  is the  $\mathbf{z}$ -component of the  $K$ -th nearest neighbor to  $\mathbf{z}^*$  after reordering the data sample  $\{\hat{\xi}_i := (\hat{\mathbf{z}}_i, \hat{\mathbf{y}}_i)\}_{i=1}^N$  in terms of  $\|\hat{\mathbf{z}}_i - \mathbf{z}^*\|$  only.

Therefore, it must hold that  $\mathcal{W}_p(Q_{\Xi}^N, \mathbb{Q}_{\Xi}) \rightarrow 0$  a.s.  $\square$

## B.3 Asymptotic consistency under a nearest neighbors lens

In this section, we show that the asymptotic consistency of our DRO framework for the case  $\mathbb{Q} \ll \lambda^d$  with  $\mathbb{Q}(\tilde{\Xi}) = \alpha = 0$  can also be proved using a nearest-neighbors approach.

If the density of  $\mathbb{Q}$  is sufficiently smooth, it is known that  $\mathbb{Q}_{\Xi}$  can be inferred from information on  $\mathbb{Q}$  within a neighborhood of  $\mathbf{z} = \mathbf{z}^*$ . This essentially means that the portion of mass from the empirical distribution  $\hat{\mathbb{Q}}_N$  that is the closest to  $\tilde{\Xi}$  is statistically

representative of the conditional distribution  $\mathbb{Q}_{\Xi}$ . Inspired by popular data-driven local predictive methods such as  $K$  nearest neighbours and kernel regression, we can solve problem (P) for a series of pairs  $(\alpha_N, \tilde{\rho}_N)$ , both of which tend to zero appropriately as  $N$  increases. Indeed, we will demonstrate that, in doing so, problem  $(P_{(\alpha_N, \tilde{\rho}_N)})$  naturally produces distributionally robustified versions of those popular methods when applied to solve problem (4.1). Next, we formalize these ideas.

**Remark B.5.** *Throughout this section, we will assume that  $\text{dist}(\hat{\xi}_i, \tilde{\Xi}) = \|\hat{\mathbf{z}}_i - \mathbf{z}^*\|$ . This assumption is standard in the technical literature. The geometry of the joint support set  $\Xi$  is expected to have a negligible impact on the asymptotic performance of problem  $(P_{(\alpha_N, \tilde{\rho}_N)})$  (i.e., for large samples), because, under a smoothness condition on  $\mathbb{Q}$  and  $K/N \rightarrow 0$ , it holds that  $\text{dist}(\hat{\xi}_{K:N}, \tilde{\Xi}) \rightarrow \|\hat{\mathbf{z}}_{K:N} - \mathbf{z}^*\| \rightarrow 0$  almost surely (see [24, Lemmas 2.2 and 2.3]), where  $\hat{\mathbf{z}}_{K:N}$  is the  $\mathbf{z}$ -component of the  $K$ -th nearest neighbor to  $\mathbf{z}^*$  after reordering the data sample  $\{\hat{\xi}_i := (\hat{\mathbf{z}}_i, \hat{\mathbf{y}}_i)\}_{i=1}^N$  in terms of  $\|\hat{\mathbf{z}}_i - \mathbf{z}^*\|$  only.*

Here, we show that the solutions of the distributionally robust optimization problem  $(P_{(\alpha_N, \tilde{\rho}_N)})$  converge to the solution of the targeted conditional stochastic program (4.1) as  $N$  increases, for a careful choice of parameters  $\alpha_N$  and  $\tilde{\rho}_N$ . This result is underpinned by the fact that, under that selection of parameters  $\alpha_N$  and  $\tilde{\rho}_N$ , any distribution in  $\hat{\mathcal{U}}_N(\alpha_N, \tilde{\rho}_N)$  converges to the true conditional distribution  $\mathbb{Q}_{\Xi}$ .

**Assumption B.2 (Lipschitz-regularity).** *We assume that there exists an integrable function  $\ell : \mathbb{R}^{d_{\mathbf{y}}} \rightarrow \mathbb{R}_+$  such that for all  $\mathbf{y} \in \mathbb{R}^{d_{\mathbf{y}}}$*

$$|\phi_{\mathbf{y}/\mathbf{z}=\mathbf{z}'}(\mathbf{y}) - \phi_{\mathbf{y}/\mathbf{z}=\mathbf{z}^*}(\mathbf{y})| \leq \ell(\mathbf{y})\|\mathbf{z}' - \mathbf{z}^*\|, \quad \forall \mathbf{z}' \text{ such that } \|\mathbf{z}' - \mathbf{z}^*\| \leq r_0 \quad (\text{B.16})$$

where  $\phi_{\mathbf{y}/\mathbf{z}=\mathbf{z}'}(\cdot)$  stands for the density function of  $\mathbf{y}$  conditional on  $\mathbf{z} = \mathbf{z}'$ .

**Lemma B.1 (Convergence of transported trimmed distributions).** *Suppose that Assumptions 4.3 and B.2 hold. Take  $(\alpha_N, \tilde{\rho}_N)$  such that  $\alpha_N \rightarrow 0$ ,  $\frac{N\alpha_N}{\log(N)} \rightarrow \infty$ , and  $\tilde{\rho}_N \rightarrow 0$  a.s., with  $\tilde{\rho}_N \geq \epsilon_{N\alpha_N}^p$ , where  $\epsilon_{N\alpha_N}$  is the minimum transportation budget as in Definition 4.2. Then, we have that*

$$\mathcal{W}_p(Q_{\Xi}^N, \mathbb{Q}_{\Xi}) \rightarrow 0 \text{ a.s.}$$

where  $Q_{\Xi}^N := \sum_{i=1}^N b_i^N \delta_{(\mathbf{z}^*, \hat{\mathbf{y}}_i)} \in \hat{\mathcal{U}}_N(\alpha_N, \tilde{\rho}_N)$  is the distribution that results from transporting the distribution  $\sum_{i=1}^N b_i^N \delta_{(\hat{\mathbf{z}}_i, \hat{\mathbf{y}}_i)}$  in the trimming set  $\mathcal{R}_{1-\alpha_N}(\hat{\mathbb{Q}}_N)$  onto  $\tilde{\Xi}$ .

**Proof.** Proof Since  $\mathbf{y}$  is bounded, we only need to prove that  $Q_{\Xi}^N$  converges weakly to  $\mathbb{Q}_{\Xi}$ . For this purpose, take a continuous and bounded function  $h$  and let  $m(\mathbf{z}^*) = \mathbb{E}[h(\mathbf{y}) \mid \mathbf{z} = \mathbf{z}^*]$ . We have

$$\left| \sum_{i=1}^N b_i^N h(\widehat{\mathbf{y}}_i) - m(\mathbf{z}^*) \right| \leq \left| \sum_{i=1}^N b_i^N h(\widehat{\mathbf{y}}_i) - \sum_{i=1}^N b_i^N m(\widehat{\mathbf{z}}_i) \right| + \left| \sum_{i=1}^N b_i^N m(\widehat{\mathbf{z}}_i) - m(\mathbf{z}^*) \right|$$

We deal with each of the terms in the inequality above one by one. First, we use [44, Lemma 6] to get

$$\mathbb{P} \left( \left| \sum_{i=1}^N b_i^N h(\widehat{\mathbf{y}}_i) - \sum_{i=1}^N b_i^N m(\widehat{\mathbf{z}}_i) \right| > \varepsilon \mid \widehat{\mathbf{z}}_1, \dots, \widehat{\mathbf{z}}_N \right) \leq 2 \exp \left( \frac{-(N\alpha_N)\varepsilon^2}{4 \|h\|_\infty (2 \|h\|_\infty + \varepsilon)} \right)$$

Given that

$$\mathbb{P} \left( \left| \sum_{i=1}^N b_i^N (h(\widehat{\mathbf{y}}_i) - m(\widehat{\mathbf{z}}_i)) \right| > \varepsilon \right) = \mathbb{E} \left[ \mathbb{P} \left( \left| \sum_{i=1}^N b_i^N (h(\widehat{\mathbf{y}}_i) - m(\widehat{\mathbf{z}}_i)) \right| > \varepsilon \mid \widehat{\mathbf{z}}_1, \dots, \widehat{\mathbf{z}}_N \right) \right]$$

we have

$$\mathbb{P} \left( \left| \sum_{i=1}^N b_i^N h(\widehat{\mathbf{y}}_i) - \sum_{i=1}^N b_i^N m(\widehat{\mathbf{z}}_i) \right| > \varepsilon \right) \leq 2 \exp \left( \frac{-(N\alpha_N)\varepsilon^2}{4 \|h\|_\infty (2 \|h\|_\infty + \varepsilon)} \right) \quad (\text{B.17})$$

Now let  $\ell$  be an integrable function satisfying condition (B.16). Hence, for any  $\mathbf{z}'$  such that  $\|\mathbf{z}' - \mathbf{z}^*\| \leq r_0$

$$|m(\mathbf{z}') - m(\mathbf{z}^*)| \leq \|h\|_\infty \|\ell\|_1 \|\mathbf{z}' - \mathbf{z}^*\| =: L \|\mathbf{z}' - \mathbf{z}^*\| \quad (\text{B.18})$$

In addition,

$$\left| \sum_{i=1}^N b_i^N m(\widehat{\mathbf{z}}_i) - m(\mathbf{z}^*) \right| = \left| \sum_{i=1}^N b_i^N (m(\widehat{\mathbf{z}}_i) - m(\mathbf{z}^*)) \right| \leq \sum_{i=1}^N b_i^N |m(\widehat{\mathbf{z}}_i) - m(\mathbf{z}^*)|$$

Let  $J$  be the number of samples such that their distance from the set  $\widetilde{\Xi}$  is smaller than or equal to  $r_0$ . We can write

$$\begin{aligned} \left| \sum_{i=1}^N b_i^N m(\widehat{\mathbf{z}}_i) - m(\mathbf{z}^*) \right| &\leq \sum_{i=1}^J b_{i:N}^N |m(\widehat{\mathbf{z}}_{i:N}) - m(\mathbf{z}^*)| + \sum_{i=J+1}^N b_{i:N}^N |m(\widehat{\mathbf{z}}_{i:N}) - m(\mathbf{z}^*)| \\ &\leq L \sum_{i=1}^J b_{i:N}^N \|\widehat{\mathbf{z}}_{i:N} - \mathbf{z}^*\| + 2 \|h\|_\infty \sum_{i=J+1}^N b_{i:N}^N \\ &\leq L \mathcal{W}_1 \left( Q_{\widetilde{\Xi}}^N, \sum_{i=1}^N b_i^N \delta_{(\widehat{\mathbf{z}}_i, \widehat{\mathbf{y}}_i)} \right) + 2 \|h\|_\infty \sum_{i=J+1}^N b_{i:N}^N \end{aligned}$$



$$\begin{aligned}
&\leq L \mathcal{W}_p \left( Q_{\Xi}^N, \sum_{i=1}^N b_i^N \delta_{(\hat{\mathbf{z}}_i, \hat{\mathbf{y}}_i)} \right) + 2 \|h\|_{\infty} \sum_{i=J+1}^N b_{i:N}^N \\
&\leq L (\tilde{\rho}_N)^{\frac{1}{p}} + 2 \|h\|_{\infty} \sum_{i=J+1}^N b_{i:N}^N
\end{aligned}$$

Next we upper bound the second term in right-hand side of the last inequality.

$$\begin{aligned}
\sum_{i=J+1}^N b_{i:N}^N &\leq \sup_{0 \leq b_{i:N}^N \leq \frac{1}{N\alpha_N}, \forall i} \left\{ \sum_{i=J+1}^N b_{i:N}^N, \sum_{i=1}^N b_{i:N}^N = 1; \sum_{i=1}^N b_{i:N}^N \|\hat{\mathbf{z}}_{i:N} - \mathbf{z}^*\|^p \leq \tilde{\rho}_N \right\} \\
&= \inf \left\{ \frac{1}{N\alpha_N} \sum_{i=1}^N \mu_{i:N} + \theta + \lambda \tilde{\rho}_N, \mu_{i:N} + \theta + \lambda \|\hat{\mathbf{z}}_{i:N} - \mathbf{z}^*\|^p - \gamma_{i:N} = 0, \forall i = 1, \dots, J; \right. \\
&\quad \left. \mu_{i:N} + \theta + \lambda \|\hat{\mathbf{z}}_{i:N} - \mathbf{z}^*\|^p - \gamma_{i:N} = 1, \forall i = J+1, \dots, N; \lambda \geq 0; \gamma_{i:N}, \mu_{i:N} \geq 0, \forall i \right\}
\end{aligned}$$

It suffices to take a feasible solution. In particular, we consider  $\mu_{i:N} = 0, \forall i, \theta = 0$ , and  $\lambda = 1/r_0^p$ , which renders

$$\sum_{i=J+1}^N b_{i:N}^N \leq \frac{\tilde{\rho}_N}{r_0^p}$$

Hence,

$$\left| \sum_{i=1}^N b_i^N m(\hat{\mathbf{z}}_i) - m(\mathbf{z}^*) \right| \leq L (\tilde{\rho}_N)^{\frac{1}{p}} + \frac{2 \|h\|_{\infty}}{r_0^p} \tilde{\rho}_N$$

Consequently, we essentially need that  $\lim_{N \rightarrow \infty} \tilde{\rho}_N = 0$  with probability one. To show this, as  $\tilde{\rho}_N \geq \underline{\epsilon}_{N\alpha_N}^p$ , we decompose  $\tilde{\rho}_N$  into  $\underline{\epsilon}_{N\alpha_N}^p$  plus  $\Delta\tilde{\rho}_N$  and use  $(\underline{\epsilon}_{N\alpha_N}^p + \Delta\tilde{\rho}_N)^{1/p} \leq \underline{\epsilon}_{N\alpha_N} + (\Delta\tilde{\rho}_N)^{1/p}$  to recast the expression above as

$$\left| \sum_{i=1}^N b_i^N m(\hat{\mathbf{z}}_i) - m(\mathbf{z}^*) \right| \leq L \underline{\epsilon}_{N\alpha_N} + \frac{2 \|h\|_{\infty}}{r_0^p} \underline{\epsilon}_{N\alpha_N}^p + L (\Delta\tilde{\rho}_N)^{\frac{1}{p}} + \frac{2 \|h\|_{\infty}}{r_0^p} \Delta\tilde{\rho}_N$$

Importantly, the budget  $\Delta\tilde{\rho}_N$  is under the decision-maker's control, who simply needs to guarantee that  $\Delta\tilde{\rho}_N \rightarrow 0$  so that the last two terms on the right-hand side of the previous inequality vanishes. Group these two terms into  $a_N(\Delta\tilde{\rho}_N)$ , set  $K := \lceil N\alpha_N \rceil$  and note that  $\underline{\epsilon}_{N\alpha_N} \leq \|\hat{\mathbf{z}}_{K:N} - \mathbf{z}^*\|$ .

Thus, for any arbitrary  $\varepsilon > 0$ ,

$$\begin{aligned}
\mathbb{P} \left( \left| \sum_{i=1}^N b_i^N m(\hat{\mathbf{z}}_i) - m(\mathbf{z}^*) \right| - a_N(\Delta\tilde{\rho}_N) > \varepsilon \right) &\leq \mathbb{P} \left( L \|\hat{\mathbf{z}}_{K:N} - \mathbf{z}^*\| > \frac{\varepsilon}{2} \right) \\
&\quad + \mathbb{P} \left( \frac{2 \|h\|_{\infty}}{r_0^p} \|\hat{\mathbf{z}}_{K:N} - \mathbf{z}^*\|^p > \frac{\varepsilon}{2} \right)
\end{aligned}$$

In turn,

$$\begin{aligned} \mathbb{P}\left(L \|\widehat{\mathbf{z}}_{K:N} - \mathbf{z}^*\| > \frac{\varepsilon}{2}\right) &= \mathbb{P}\left(\|\widehat{\mathbf{z}}_{K:N} - \mathbf{z}^*\| > \frac{\varepsilon}{2L}\right) \\ \mathbb{P}\left(\frac{2\|h\|_\infty}{r_0^p} \|\widehat{\mathbf{z}}_{K:N} - \mathbf{z}^*\|^p > \frac{\varepsilon}{2}\right) &= \mathbb{P}\left(\|\widehat{\mathbf{z}}_{K:N} - \mathbf{z}^*\| > r_0 \left(\frac{\varepsilon}{4\|h\|_\infty}\right)^{\frac{1}{p}}\right) \end{aligned}$$

Furthermore, due to the first point in Assumption 4.3, it holds that

$$\mathbb{P}(\|\widehat{\mathbf{z}}_{K:N} - \mathbf{z}^*\| > \eta) \leq \exp\left(-\frac{\tilde{C}}{8} N \eta^{d_{\mathbf{z}}}\right)$$

for any  $0 < \eta \leq r_0$  and provided that  $\frac{K}{N} \leq \frac{\tilde{C}}{2} \eta^{d_{\mathbf{z}}}$  (see [95, formula (34)], which is an application of the lower-tail of Chernoff's bound).

Therefore, in that case,

$$\begin{aligned} \mathbb{P}\left(\left|\sum_{i=1}^N b_i^N m(\widehat{\mathbf{z}}_i) - m(\mathbf{z}^*)\right| - a_N(\Delta\tilde{\rho}_N) > \varepsilon\right) &\leq \exp\left(-\frac{\tilde{C}}{8} N \left(\frac{\varepsilon}{2L}\right)^{d_{\mathbf{z}}}\right) \\ &\quad + \exp\left(-\frac{\tilde{C}}{8} N r_0^{d_{\mathbf{z}}} \left(\frac{\varepsilon}{4\|h\|_\infty}\right)^{\frac{d_{\mathbf{z}}}{p}}\right) \end{aligned}$$

whenever

$$\frac{K}{N} \leq \min\left\{\frac{\tilde{C}}{2} \left(\frac{\varepsilon}{2L}\right)^{d_{\mathbf{z}}}, \frac{\tilde{C} r_0^{d_{\mathbf{z}}}}{2} \left(\frac{\varepsilon}{4\|h\|_\infty}\right)^{\frac{d_{\mathbf{z}}}{p}}\right\}$$

which we guarantee, for  $N$  large enough, by enforcing  $\alpha_N \rightarrow 0$ .

This way, for any arbitrarily small  $\varepsilon > 0$ , we finally have

$$\begin{aligned} \mathbb{P}\left(\left|\sum_{i=1}^N b_i^N h(\widehat{\mathbf{y}}_i) - m(\mathbf{z}^*)\right| - a_N(\Delta\tilde{\rho}_N) > \varepsilon\right) &\leq 2 \exp\left(\frac{-(N\alpha_N)(\varepsilon/3)^2}{4\|h\|_\infty(2\|h\|_\infty + \varepsilon/3)}\right) \\ &\quad + \exp\left(-\frac{\tilde{C}}{8} N \left(\frac{\varepsilon}{3L}\right)^{d_{\mathbf{z}}}\right) \\ &\quad + \exp\left(-\frac{\tilde{C}}{8} N r_0^{d_{\mathbf{z}}} \left(\frac{\varepsilon}{6\|h\|_\infty}\right)^{\frac{d_{\mathbf{z}}}{p}}\right) \quad (\text{B.19}) \end{aligned}$$

The last two terms on the right-hand side of (B.19) are summable over  $N$ , while the first one is summable if  $\frac{N\alpha_N}{\log(N)} \rightarrow \infty$ . Consequently, the Borel-Cantelli Lemma allows us to conclude that

$$\begin{aligned} \mathbb{P}\left(\lim_{N \rightarrow \infty} \left|\sum_{i=1}^N b_i^N h(\widehat{\mathbf{y}}_i) - m(\mathbf{z}^*)\right| - a_N(\Delta\tilde{\rho}_N) = 0\right) &= \mathbb{P}\left(\lim_{N \rightarrow \infty} \left|\sum_{i=1}^N b_i^N h(\widehat{\mathbf{y}}_i) - m(\mathbf{z}^*)\right| = 0\right) \\ &= 1 \end{aligned}$$

given that  $a_N \rightarrow 0$  when  $\Delta\tilde{\rho}_N \rightarrow 0$ . Thus,  $Q_{\Xi}^N$  converges weakly to  $Q_{\Xi}$  almost surely.  $\square$

The following corollary extends the convergence to any distribution in the proposed ambiguity set (apart from the transported trimmings of the empirical distribution).

**| Corollary B.2 (Convergence of conditional distributions).** *Suppose that the conditions in Lemma B.1 hold. Then, it follows that*

$$\mathcal{W}_p(Q_{\Xi}^N, Q_{\Xi}) \rightarrow 0 \text{ a.s.}$$

where  $Q_{\Xi}^N$  is any distribution from the ambiguity set  $\hat{\mathcal{U}}_N(\alpha_N, \tilde{\rho}_N)$ .

*Proof.* Proof This corollary is an immediate result of the previous lemma. With some abuse of notation, let  $\sum_{i=1}^N b_i^N \delta_{(\hat{\mathbf{z}}_i, \hat{\mathbf{y}}_i)}$  be the distribution in the trimming set  $\mathcal{R}_{1-\alpha_N}(\hat{\mathbb{Q}}_N)$  such that  $\mathcal{W}_p(\mathcal{R}_{1-\alpha_N}(\hat{\mathbb{Q}}_N), Q_{\Xi}^N) = \mathcal{W}_p(\sum_{i=1}^N b_i^N \delta_{(\hat{\mathbf{z}}_i, \hat{\mathbf{y}}_i)}, Q_{\Xi}^N)$ .

By the triangle inequality, we have

$$\mathcal{W}_p(Q_{\Xi}^N, Q_{\Xi}) \leq \mathcal{W}_p\left(Q_{\Xi}^N, \sum_{i=1}^N b_i^N \delta_{(\hat{\mathbf{z}}_i, \hat{\mathbf{y}}_i)}\right) + \mathcal{W}_p\left(\sum_{i=1}^N b_i^N \delta_{(\hat{\mathbf{z}}_i, \hat{\mathbf{y}}_i)}, Q_{\Xi}\right) \quad (\text{B.20})$$

where  $\mathcal{W}_p^p\left(Q_{\Xi}^N, \sum_{i=1}^N b_i^N \delta_{(\hat{\mathbf{z}}_i, \hat{\mathbf{y}}_i)}\right) \leq \tilde{\rho}_N$ , because  $Q_{\Xi}^N \in \hat{\mathcal{U}}_N(\alpha_N, \tilde{\rho}_N)$ . We again use the triangle inequality to upper bound the second term on the right-hand side of (B.20).

$$\mathcal{W}_p\left(\sum_{i=1}^N b_i^N \delta_{(\hat{\mathbf{z}}_i, \hat{\mathbf{y}}_i)}, Q_{\Xi}\right) \leq \mathcal{W}_p\left(\sum_{i=1}^N b_i^N \delta_{(\hat{\mathbf{z}}_i, \hat{\mathbf{y}}_i)}, \sum_{i=1}^N b_i^N \delta_{(\mathbf{z}^*, \hat{\mathbf{y}}_i)}\right) + \mathcal{W}_p\left(\sum_{i=1}^N b_i^N \delta_{(\mathbf{z}^*, \hat{\mathbf{y}}_i)}, Q_{\Xi}\right)$$

where  $\sum_{i=1}^N b_i^N \delta_{(\mathbf{z}^*, \hat{\mathbf{y}}_i)}$  is the distribution with support on  $\Xi$  that is the closest (in  $p$ -Wasserstein distance) to  $\sum_{i=1}^N b_i^N \delta_{(\hat{\mathbf{z}}_i, \hat{\mathbf{y}}_i)}$ . Therefore,

$$\mathcal{W}_p^p\left(\sum_{i=1}^N b_i^N \delta_{(\hat{\mathbf{z}}_i, \hat{\mathbf{y}}_i)}, \sum_{i=1}^N b_i^N \delta_{(\mathbf{z}^*, \hat{\mathbf{y}}_i)}\right) \leq \mathcal{W}_p^p\left(\sum_{i=1}^N b_i^N \delta_{(\hat{\mathbf{z}}_i, \hat{\mathbf{y}}_i)}, Q_{\Xi}^N\right) \leq \tilde{\rho}_N$$

That is,  $\sum_{i=1}^N b_i^N \delta_{(\mathbf{z}^*, \hat{\mathbf{y}}_i)}$  is in  $\hat{\mathcal{U}}_N(\alpha_N, \tilde{\rho}_N)$  and is precisely one of the transported trimmed distributions to which Lemma B.1 refers.

Hence,

$$\mathcal{W}_p(Q_{\Xi}^N, Q_{\Xi}) \leq 2(\tilde{\rho}_N)^{\frac{1}{p}} + \mathcal{W}_p\left(\sum_{i=1}^N b_i^N \delta_{(\mathbf{z}^*, \hat{\mathbf{y}}_i)}, Q_{\Xi}\right)$$

Since both  $\tilde{\rho}_N \rightarrow 0$  and  $\mathcal{W}_p\left(\sum_{i=1}^N b_i^N \delta_{(\mathbf{z}^*, \hat{\mathbf{y}}_i)}, Q_{\Xi}\right) \rightarrow 0$  a.s. by Lemma B.1, the claim of the corollary follows.  $\square$

Finally, the following theorem formally states the asymptotic consistency guarantee of our model.

**Theorem B.4 (Asymptotic consistency).** *Suppose that the assumptions in Corollary B.2 hold. Then, we have*

- (i) *If for any fixed  $\xi \in \tilde{\Xi}$ ,  $f(\cdot, \xi)$  is continuous on  $X$ , and for any fixed value  $\mathbf{x} \in X$ ,  $f(\mathbf{x}, \xi)$  is continuous in  $\xi$  and there is  $L \geq 0$  such that  $|f(\mathbf{x}, \xi)| \leq L(1 + \|\xi\|^p)$  for all  $\mathbf{x} \in X$  and  $\xi \in \tilde{\Xi}$ , then we have that  $\hat{J}_N \rightarrow J^*$  almost surely when  $N$  grows to infinity.*
- (ii) *Let  $X_N, X^*$  be the set of optimal solutions of problems  $(P_{(\alpha_N, \tilde{\rho}_N)})$  and (4.22), respectively. If the assumptions in (i) are satisfied, the feasible set  $X$  is closed and  $X_N, X^*$  are non-empty, then we have that any accumulation point of the sequence  $\{\hat{\mathbf{x}}_N\}_N$  is almost surely an optimal solution of problem (4.22).*

*Proof.* Proof Set  $v_N(\mathbf{x}) = \sup_{Q_{\Xi} \in \hat{\mathcal{U}}_N(\alpha_N, \tilde{\rho}_N)} \mathbb{E}_{Q_{\Xi}}[f(\mathbf{x}, \xi)]$  and  $v(\mathbf{x}) = \mathbb{E}_{Q_{\Xi}}[f(\mathbf{x}, \xi)]$ . Let  $\mathcal{F}$  be the class of random functions defined as follows

$$\begin{aligned} \mathcal{F} := \{ & f(\xi) := f(\mathbf{x}, \xi) \text{ continuous such that } \mathbf{x} \in X \\ & \text{and } \exists L \geq 0 \text{ with } |f(\mathbf{x}, \xi)| \leq L(1 + \|\xi\|^p), \forall \mathbf{x} \in X, \forall \xi \in \tilde{\Xi} \} \end{aligned} \quad (\text{B.21})$$

and let  $\mathcal{D}$  be the pseudometric between two probability measures  $P$  and  $Q$  given by

$$\mathcal{D}(P, Q) := \sup_{f \in \mathcal{F}} |\mathbb{E}_P[f] - \mathbb{E}_Q[f]|$$

For two sets of probability measures  $\mathcal{U}_1$  and  $\mathcal{U}_2$ , define the *excess* of  $\mathcal{U}_1$  over  $\mathcal{U}_2$  as

$$\mathcal{D}(\mathcal{U}_1, \mathcal{U}_2) := \sup_{P \in \mathcal{U}_1} \inf_{Q \in \mathcal{U}_2} \mathcal{D}(P, Q)$$

First, we show that  $v_N(\mathbf{x}) < \infty$  for all  $\mathbf{x} \in X$ . Fix  $\mathbf{x} \in X$  and define

$$\mathcal{V} := \{\mathbb{E}_{Q_{\Xi}}[f(\mathbf{x}, \xi)]\} \text{ and } \mathcal{V}_N := \{\mathbb{E}_{Q_{\Xi}}[f(\mathbf{x}, \xi)] : Q_{\Xi} \in \hat{\mathcal{U}}_N(\alpha_N, \tilde{\rho}_N)\}.$$

The function  $f$  satisfies the following uniform-integrability-type condition for all  $\mathbf{x}$ ,

$$\lim_{t \rightarrow \infty} \sup_{Q_{\Xi} \in \hat{\mathcal{U}}_N(\alpha_N, \tilde{\rho}_N)} \int_{\{\tilde{\Xi}: |f(\mathbf{x}, \xi)| \geq t\}} |f(\mathbf{x}, \xi)| Q_{\Xi}(d\xi) = 0$$

due to the limitation on the maximum growth of  $f$  established in point (i) and the  $p$ -uniform integrability of  $\hat{\mathcal{U}}_N(\alpha_N, \tilde{\rho}_N)$ . Furthermore, the set  $\hat{\mathcal{U}}_N(\alpha_N, \tilde{\rho}_N)$  is also tight. Consequently, using [129, Proposition 1], we deduce that the set  $\mathcal{V}_N$  is compact (and hence bounded). Thus,  $v_N(\mathbf{x}) < \infty$ .

Let  $a_N := \inf_{v \in \mathcal{V}_N} v$ ,  $b_N := \sup_{v \in \mathcal{V}_N} v$  and  $c := \inf_{v \in \mathcal{V}} v = \sup_{v \in \mathcal{V}} v$ . Now, denote the Hausdorff distance between the respective convex hulls of the sets  $\mathcal{V}$  and  $\mathcal{V}_N$  as

$\mathbb{H}(\text{conv}\mathcal{V}, \text{conv}\mathcal{V}_N)$ . We have

$$\mathbb{H}(\text{conv}\mathcal{V}, \text{conv}\mathcal{V}_N) = \mathbb{H}(\mathcal{V}, \text{conv}\mathcal{V}_N) = \max\{|b_N - c|, |c - a_N|\}$$

where

$$\begin{aligned} b_N - c &= \max_{Q_{\Xi} \in \hat{\mathcal{U}}_N(\alpha_N, \tilde{\rho}_N)} \mathbb{E}_{Q_{\Xi}}[f(\mathbf{x}, \boldsymbol{\xi})] - \mathbb{E}_{Q_{\Xi}}[f(\mathbf{x}, \boldsymbol{\xi})] \\ c - a_N &= \mathbb{E}_{Q_{\Xi}}[f(\mathbf{x}, \boldsymbol{\xi})] - \min_{Q_{\Xi} \in \hat{\mathcal{U}}_N(\alpha_N, \tilde{\rho}_N)} \mathbb{E}_{Q_{\Xi}}[f(\mathbf{x}, \boldsymbol{\xi})] \end{aligned}$$

On the other hand, by [75, Proposition 2.1 (c)] and the definition of the Hausdorff distance, the following holds

$$\mathbb{H}(\mathcal{V}, \text{conv}\mathcal{V}_N) \leq \mathbb{H}(\mathcal{V}, \mathcal{V}_N) = \max(\mathbb{D}(\mathcal{V}, \mathcal{V}_N), \mathbb{D}(\mathcal{V}_N, \mathcal{V})) = \mathbb{D}(\mathcal{V}_N, \mathcal{V})$$

where

$$\begin{aligned} \mathbb{D}(\mathcal{V}_N, \mathcal{V}) &= \max_{v' \in \mathcal{V}_N} d(v', \mathcal{V}) = \max_{v' \in \mathcal{V}_N} \min_{v \in \mathcal{V}} |v' - v| \\ &= \max_{Q_{\Xi} \in \hat{\mathcal{U}}_N(\alpha_N, \tilde{\rho}_N)} \left| \mathbb{E}_{Q_{\Xi}}[f(\mathbf{x}, \boldsymbol{\xi})] - \mathbb{E}_{Q_{\Xi}}[f(\mathbf{x}, \boldsymbol{\xi})] \right| \\ &\leq \max_{Q_{\Xi} \in \hat{\mathcal{U}}_N(\alpha_N, \tilde{\rho}_N)} \sup_{\mathbf{x} \in X} \left| \mathbb{E}_{Q_{\Xi}}[f(\mathbf{x}, \boldsymbol{\xi})] - \mathbb{E}_{Q_{\Xi}}[f(\mathbf{x}, \boldsymbol{\xi})] \right| \\ &= \max_{Q_{\Xi} \in \hat{\mathcal{U}}_N(\alpha_N, \tilde{\rho}_N)} \mathcal{D}(Q_{\Xi}, Q_{\Xi}) \\ &= \mathcal{D}(\hat{\mathcal{U}}_N(\alpha_N, \tilde{\rho}_N), Q_{\Xi}) \end{aligned}$$

Note that  $\mathcal{D}(\hat{\mathcal{U}}_N(\alpha_N, \tilde{\rho}_N), Q_{\Xi}) \xrightarrow{N \rightarrow \infty} 0$ , because, for any  $f \in \mathcal{F}$ , we have that  $\mathbb{E}_{Q_{\Xi}}[f] \xrightarrow{N \rightarrow \infty} \mathbb{E}_{Q_{\Xi}}[f]$  under Corollary B.2 and Proposition B.1. Thus,

$$\mathbb{H}(\mathcal{V}, \text{conv}\mathcal{V}_N) \leq \mathbb{H}(\mathcal{V}, \mathcal{V}_N) = \mathbb{D}(\mathcal{V}_N, \mathcal{V}) \leq \mathcal{D}(\hat{\mathcal{U}}_N(\alpha_N, \tilde{\rho}_N), Q_{\Xi})$$

Therefore,

$$|v_N(\mathbf{x}) - v(\mathbf{x})| \leq \mathbb{H}(\mathcal{V}, \text{conv}\mathcal{V}_N) \leq \mathcal{D}(\hat{\mathcal{U}}_N(\alpha_N, \tilde{\rho}_N), Q_{\Xi}) \xrightarrow{N \rightarrow \infty} 0$$

Hence, since the inequality above is independent of the value of  $\mathbf{x}$ , we have  $\lim_{N \rightarrow \infty} \sup_{\mathbf{x} \in X} |v_N(\mathbf{x}) - v(\mathbf{x})| = 0$  a.s.

Now, we show that the functions  $v_N(\mathbf{x})$  and  $v(\mathbf{x})$  are continuous in  $\mathbf{x} \in X$ : Fix an arbitrary  $\mathbf{x} \in X$  and consider a sequence  $(\mathbf{x}_N)_N$  such that  $\mathbf{x}_N \rightarrow \mathbf{x}$  as  $N$  grows to infinity. We want to prove that  $v_N(\mathbf{x}_N) \rightarrow v_N(\mathbf{x})$  and  $v(\mathbf{x}_N) \rightarrow v(\mathbf{x})$ . First, there exist  $Q_{\mathbf{x}_N}, Q_{\mathbf{x}} \in \hat{\mathcal{U}}_N(\alpha_N, \tilde{\rho}_N)$  such that  $v_N(\mathbf{x}_N) = \mathbb{E}_{Q_{\mathbf{x}_N}} f(\mathbf{x}_N, \boldsymbol{\xi})$  and  $v_N(\mathbf{x}) = \mathbb{E}_{Q_{\mathbf{x}}} f(\mathbf{x}, \boldsymbol{\xi})$ . For any  $\varepsilon > 0$ , there exists  $N' > 0$  sufficiently large such that for  $N \geq N'$  the following

holds:

$$\begin{aligned}
|v_N(\mathbf{x}_N) - v_N(\mathbf{x})| &= |\mathbb{E}_{Q_{\mathbf{x}_N}} f(\mathbf{x}_N, \boldsymbol{\xi}) - \mathbb{E}_{Q_{\mathbf{x}}} f(\mathbf{x}, \boldsymbol{\xi})| \\
&\leq |\mathbb{E}_{Q_{\mathbf{x}_N}} f(\mathbf{x}_N, \boldsymbol{\xi}) - \mathbb{E}_{Q_{\mathbf{x}_N}} f(\mathbf{x}, \boldsymbol{\xi})| + |\mathbb{E}_{Q_{\mathbf{x}_N}} f(\mathbf{x}, \boldsymbol{\xi}) - \mathbb{E}_{Q_{\mathbf{x}}} f(\mathbf{x}, \boldsymbol{\xi})| \\
&\leq \varepsilon/2 + \varepsilon/2 = \varepsilon
\end{aligned}$$

since  $|\mathbb{E}_{Q_{\mathbf{x}_N}} f(\mathbf{x}_N, \boldsymbol{\xi}) - \mathbb{E}_{Q_{\mathbf{x}_N}} f(\mathbf{x}, \boldsymbol{\xi})| < \varepsilon/2$  because  $f$  is continuous in  $\mathbf{x}$  and

$$|\mathbb{E}_{Q_{\mathbf{x}_N}} f(\mathbf{x}, \boldsymbol{\xi}) - \mathbb{E}_{Q_{\mathbf{x}}} f(\mathbf{x}, \boldsymbol{\xi})| \leq \mathcal{D}(Q_{\mathbf{x}_N}, Q_{\mathbf{x}}) \leq \mathcal{D}(Q_{\mathbf{x}_N}, Q_{\Xi}) + \mathcal{D}(Q_{\Xi}, Q_{\mathbf{x}}) \xrightarrow{N \rightarrow \infty} 0,$$

because  $\mathcal{D}(\widehat{\mathcal{U}}_N(\alpha_N, \tilde{\rho}_N), Q_{\Xi}) \xrightarrow{N \rightarrow \infty} 0$ . As  $\varepsilon > 0$  is arbitrary, this implies that the function  $v_N(\mathbf{x})$  is continuous in  $\mathbf{x} \in X$ . Similarly, since  $f$  is continuous in  $\mathbf{x}$ , we have that the function  $v(\mathbf{x})$  is continuous in  $\mathbf{x} \in X$ . Finally, as  $v_N(\mathbf{x})$  and  $v(\mathbf{x})$  are continuous in  $\mathbf{x} \in X$  and  $\lim_{N \rightarrow \infty} \sup_{\mathbf{x} \in X} |v_N(\mathbf{x}) - v(\mathbf{x})| = 0$  a.s., we deduce from [140, Lemma 3.4] that  $\widehat{J}_N \rightarrow J^*$  a.s. and the proof of (i) is complete.

The proof of (ii) is given by the application of [94, Lemma 3.8].  $\square$

**Remark B.6.** *The theoretical framework underpinned by Lemma B.1, Corollary B.2 and B.4 leaves the decision maker with considerable freedom to choose the values for  $\alpha_N$  and  $\tilde{\rho}_N$ . In the following two corollaries, we show that our framework naturally produces distributionally robust variants of popular non-parametric regression techniques such as the  $K$ -nearest neighbors and the Nadaraya-Watson kernel regression. This could serve to guide the selection of  $\alpha_N$  and  $\tilde{\rho}_N$ .*

**Corollary B.3 (Distributionally robust  $K$ -nearest neighbors).** *Let  $K_N$  be the number of nearest neighbors, chosen such that  $K_N \rightarrow \infty$ ,  $K_N/N \rightarrow 0$  and  $\frac{K_N}{\log N} \rightarrow \infty$  when the sample size  $N$  grows to infinity. This defines a standard KNN regression method.*

*Take problem  $(P_{(\alpha_N, \tilde{\rho}_N)})$ , set  $\alpha_N := K_N/N$  and compute the minimum transportation budget  $\epsilon_{K_N}$  as in Definition 4.2. Problem  $(P_{(\alpha_N, \tilde{\rho}_N)})$  for any sequence of  $\tilde{\rho}_N$ ,  $N \in \mathbb{N}$ , such that  $\tilde{\rho}_N = \epsilon_{K_N}^p + \Delta \tilde{\rho}_N$  with  $\Delta \tilde{\rho}_N \downarrow 0$  is a distributionally robust variant of that KNN method.*

**Proof.** The proof of this claim directly follows from the fact that all the conditions in Lemma B.1 are satisfied if we choose  $\alpha_N = K_N/N$ . Actually, if we set  $\tilde{\rho}_N = \epsilon_{K_N}^p$ , the ambiguity set consisting of all distributions  $Q_{\Xi}^N$  such that  $Q_{\Xi}^N \in \widehat{\mathcal{U}}_N(\alpha_N, \tilde{\rho}_N)$  is reduced, for each  $N \in \mathbb{N}$ , to the singleton  $Q_{\Xi}^N := \sum_{i=1}^{K_N} \frac{1}{K_N} \delta_{(\mathbf{z}^*, \hat{\mathbf{y}}_{i:N})}$ , where  $\hat{\mathbf{y}}_{i:N}$  represents the  $\mathbf{y}$ -coordinate of the data point in the sample that is the  $i$ -th nearest neighbor. The decision maker can thus use the extra budget  $\Delta \tilde{\rho}_N$  to control the degree of robustness of the KNN solution.  $\square$

**Corollary B.4 (Distributionally robust Nadaraya-Watson kernel regres-**

**sion).** Consider a Nadaraya-Watson (NW) kernel regression method with bandwidth  $h_N$  such that  $h_N \rightarrow 0$  and  $Nh_N^{d_z}/\log(N) \rightarrow \infty$  when  $N$  grows to infinity. Also, assume that the non-negative Kernel  $\mathcal{K}$  of the NW method satisfies that there exist positive numbers  $c_1$ ,  $c_2$  and  $r$  such that  $c_1 \mathbb{I}_{\{\|\mathbf{v}\| \leq r\}} \leq \mathcal{K}(\mathbf{v}) \leq c_2 \mathbb{I}_{\{\|\mathbf{v}\| \leq r\}}$ .

Let  $w_i$ ,  $i = 1, \dots, N$  be the weights given by the NW method to the data points in a certain sample of size  $N$  and let  $w^{\max} := \max_i w_i$ . Compute

$$\tilde{\rho}_N^{\text{NW}} = \sum_{i=1}^N w_i \text{dist} \left( (\hat{\mathbf{z}}_i, \hat{\mathbf{y}}_i), \tilde{\Xi} \right)^p.$$

The choices  $\alpha_N := 1/(Nw^{\max})$  and  $\tilde{\rho}_N := \tilde{\rho}_N^{\text{NW}} + \Delta\tilde{\rho}_N$  with  $\Delta\tilde{\rho}_N \downarrow 0$  produce an asymptotically consistent and distributionally robust Nadaraya-Watson kernel regression method.

*Proof.* Proof To prove this corollary, we will use the following lemma, which appears in [43].

**Lemma B.2 (Lemma 4.1 from [43]).** If  $n$  is a binomial random variable with parameters  $N$  and  $\hat{p}$ , then

$$\sum_{N=1}^{\infty} \mathbb{E}[\exp(-sn)] < \infty, \text{ for all } s > 0$$

whenever  $N\hat{p}/\log N \rightarrow \infty$ .

Define  $A_i$  as the event  $(\|\hat{\mathbf{z}}_i - \mathbf{z}^*\| \leq rh_N)$ . Then,  $n = \sum_{i=1}^N \mathbb{I}_{A_i}$  is a binomial random variable with parameters  $N$  and  $\hat{p} = \mathbb{P}(\|\hat{\mathbf{z}}_i - \mathbf{z}^*\| \leq rh_N)$  that represents the number of samples that are given a weight different from zero by the NW method. By Assumption 4.3, it follows that  $\hat{p} \geq \tilde{C}r^{d_z}h_N^{d_z}$ , when  $rh_N < r_0$ . Furthermore, by the way the weights are constructed in this method and the choice of  $\alpha_N$ , we have that  $\tilde{\rho}_N^{\text{NW}} \geq \epsilon_{N\alpha_N}^p$ , provided that  $n \geq 1$ . In that case, it also holds  $1/n \leq w^{\max} \leq c_2/(c_1 n)$  and thus,  $(c_1 n)/c_2 \leq N\alpha_N \leq n$ . Note that the event  $(n = 0)$  can happen only in a finite number of instances as  $N$  increases. Indeed, for  $N$  sufficiently large,  $\mathbb{P}(n = 0) = (1 - \hat{p})^N \leq \exp(-N\hat{p}) \leq \exp(-N\tilde{C}r^{d_z}h_N^{d_z})$ , which is summable over  $N$ , because  $Nh_N^{d_z}/\log(N) \rightarrow \infty$ . Therefore, in practice, the bandwidth of the NW method could be occasionally augmented in those specific instances so that  $n \geq 1$ , without affecting the convergence of the method.

Thus, we have

$$\tilde{\rho}_N^{\text{NW}} \leq \frac{c_2}{c_1 n} \sum_{i=1}^N \|\hat{\mathbf{z}}_i - \mathbf{z}^*\|^p \mathbb{I}_{A_i} \leq \frac{c_2 r^p h_N^p}{c_1} \rightarrow 0$$

because  $h_N$  tends to 0 as  $N$  grows to infinity.

Now, we need to revisit Equation (B.17), since  $N\alpha_N$  is random here (contingent on the training sample). In particular, we have

$$\begin{aligned} \mathbb{P} \left( \left| \sum_{i=1}^N b_i^N h(\hat{\mathbf{y}}_i) - \sum_{i=1}^N b_i^N m(\hat{\mathbf{z}}_i) \right| > \varepsilon \mid \hat{\mathbf{z}}_1, \dots, \hat{\mathbf{z}}_N \right) &\leq 2 \exp \left( \frac{-(N\alpha_N)\varepsilon^2}{4 \|h\|_\infty (2 \|h\|_\infty + \varepsilon)} \right) \\ &\leq 2 \exp \left( \frac{-(c_1 n/c_2)\varepsilon^2}{4 \|h\|_\infty (2 \|h\|_\infty + \varepsilon)} \right) \end{aligned}$$

for any arbitrary  $\varepsilon > 0$ .

Hence,

$$\mathbb{P} \left( \left| \sum_{i=1}^N b_i^N h(\hat{\mathbf{y}}_i) - \sum_{i=1}^N b_i^N m(\hat{\mathbf{z}}_i) \right| > \varepsilon \right) \leq \mathbb{E} \left[ 2 \exp \left( \frac{-(c_1 n/c_2)\varepsilon^2}{4 \|h\|_\infty (2 \|h\|_\infty + \varepsilon)} \right) \right]$$

The summability with respect to  $N$  of the expectation on the right-hand side of the inequality above is ensured by Lemma B.2, given that, for  $N$  large enough,  $N\hat{p}/\log N \geq \tilde{C}r^{d_{\mathbf{z}}}Nh_N^{d_{\mathbf{z}}}/\log(N) \rightarrow \infty$ . The Borel-Cantelli lemma does the rest to conclude the proof.

While not explicitly required in this proof, it is easy to check that  $\alpha_N \rightarrow 0$  almost surely as well. Note that  $\frac{c_1 n}{c_2 N} \leq \alpha_N \leq \frac{n}{N}$ , with  $\mathbb{E} \left[ \frac{n}{N} \right] = \hat{p} \rightarrow 0$ , since  $h_N \rightarrow 0$ . Using [44, Lemma 6], we get, for any  $\varepsilon > 0$ ,

$$\mathbb{P} \left( \left| \frac{n}{N} - \hat{p} \right| > \varepsilon \right) = \mathbb{P} \left( \left| \sum_{i=1}^N \frac{1}{N} (\mathbb{I}_{A_i} - \hat{p}) \right| > \varepsilon \right) \leq 2 \exp \left( -\frac{N\varepsilon^2}{2(1+\varepsilon)} \right)$$

which is summable with respect to  $N$ . Thus,  $\lim_{N \rightarrow \infty} \frac{n}{N} = \hat{p} = 0$  with probability one (as expected) and consequently,  $\alpha_N \rightarrow 0$  a.s.  $\square$

Similarly as before, the extra budget  $\Delta\tilde{\rho}_N$  can be used by the decision-maker to robustify the NW solution. Nevertheless, in this case, as  $\tilde{\rho}_N^{NW} \geq \underline{\epsilon}_{N\alpha_N}^p$ , the ambiguity set is not necessarily a singleton, meaning that our DRO approach already confers some degree of robustness on the decision vector  $\mathbf{x}$  even if we set  $\tilde{\rho}_N = \tilde{\rho}_N^{NW}$ .

We conclude this section with a corollary that extends Lemma B.1 to the case of unbounded uncertainty  $\mathbf{y}$  under certain conditions. This extension guarantees that the solution to problem  $(P_{(\alpha_N, \tilde{\rho}_N)})$  is asymptotically consistent also for this case.

**Corollary B.5 (Extension of Lemma B.1 to unbounded  $\mathbf{y}$ ).** *Suppose that Assumptions 4.3.1 and B.2 hold. Consider the true data-generating distribution  $\mathbb{Q}$  of the random vector  $\boldsymbol{\xi} := (\mathbf{z}, \mathbf{y})$  with support  $\Xi := \Xi_{\mathbf{z}} \times \mathbb{R}^{d_{\mathbf{y}}}$  and define  $m(\mathbf{z}^*) = \mathbb{E}[\|\mathbf{y}\|^a \mid \mathbf{z} = \mathbf{z}^*]$ , for some  $a \geq p$ .*

*Assume that there exists a constant  $\bar{m} > 0$  such that  $m(\mathbf{z}) < \bar{m}$  for almost all  $\mathbf{z} \in \Xi_{\mathbf{z}}$ ,*



and that there are non-negative numbers  $(\sigma, \nu)$  such that

$$\log \mathbb{E} [\exp \{t(\|\mathbf{y}\|^a - m(\mathbf{z}))\} \mid \mathbf{z} = \mathbf{z}^*] \leq \sigma^2 t^2 / 2, \quad |t| \leq 1/\nu,$$

for almost all  $\mathbf{z}^* \in \Xi_{\mathbf{z}}$ . Then, if the sequence  $(\alpha_N, \tilde{\rho}_N)$ ,  $N \in \mathbb{N}$ , meets the conditions stated in Lemma B.1, we have that the convergence result stated in that lemma, also applies in the following two cases: i)  $a = p$  and function  $\ell : \mathbb{R}^{d_{\mathbf{y}}} \rightarrow \mathbb{R}_+$  in Assumption B.2 is such that  $\int \|\mathbf{y}\|^p \ell(\mathbf{y}) d\mathbf{y} < R < \infty$ ; and ii)  $a > p$ .

*Proof.* Proof Since the weak convergence of distributions is guaranteed by way of Lemma B.1, we just need to prove that  $\int_{\Xi} \|\mathbf{y}\|^p dQ_{\Xi}^N \rightarrow \int_{\Xi} \|\mathbf{y}\|^p dQ_{\Xi}$  (i.e., convergence of the  $p$ -th moment, see Proposition B.1). For this purpose, we will use different strategies in cases i) and ii).

- Case i): Here we follow a similar strategy to that used to prove Lemma B.1.

We have

$$\left| \sum_{i=1}^N b_i^N \|\hat{\mathbf{y}}_i\|^p - m(\mathbf{z}^*) \right| \leq \left| \sum_{i=1}^N b_i^N \|\hat{\mathbf{y}}_i\|^p - \sum_{i=1}^N b_i^N m(\hat{\mathbf{z}}_i) \right| + \left| \sum_{i=1}^N b_i^N m(\hat{\mathbf{z}}_i) - m(\mathbf{z}^*) \right|$$

To upper bound the first term on the right-hand side of the above inequality, we exploit the subexponential character of  $\|\hat{\mathbf{y}}_i\|^p$ ,  $i = 1, \dots, N$  (understood as random variables). To this end, we employ the following technical result, which corresponds to Theorem 2.51 in [14].

**Theorem B.5 (Theorem 2.51 from [14]).** *Let  $Z_1, \dots, Z_n$  be a finite sequence of independent and centered random variables such that, for all  $1 \leq k \leq n$ , the random variable  $Z_k$  satisfies  $\log \mathbb{E}[\exp(tZ_k)] \leq l(t)$  for any  $t \geq 0$ , with  $l(t)$  being a function from  $[0, \infty)$  to  $[0, \infty]$  with a concave derivative such that  $l(0) = l'(0) = 0$ .*

Denote  $S_N = b_1 Z_1 + \dots + b_N Z_N$  for some positive real numbers  $b_1, \dots, b_N$ . For any positive  $\varepsilon$ ,

$$\mathbb{P}(S_N \geq \varepsilon) \leq \exp \left( - \frac{\|b\|_1^2}{\|b\|_2^2} l^* \left( \frac{\varepsilon}{\|b\|_1} \right) \right)$$

where  $l^*$  stands for the convex conjugate of  $l$ .

By assumption, we have

$$\log \mathbb{E} [\exp \{t(\|\mathbf{y}\|^p - m(\mathbf{z}^*))\} \mid \mathbf{z} = \mathbf{z}^*] \leq \frac{\sigma^2 t^2}{2} \text{ for } 0 \leq t \leq 1/\nu \text{ and for almost all } \mathbf{z}^* \in \Xi_{\mathbf{z}}$$

We take then  $l(t) := \frac{\sigma^2 t^2}{2}$ , if  $0 \leq t \leq 1/\nu$ , and  $l(t) := \infty$ , if  $t > 1/\nu$ . Therefore,  $l^*(s) = \frac{s^2}{2\sigma^2}$ , if  $0 < s \leq \sigma^2/\nu$  and  $l^*(s) = \frac{s}{\nu} - \frac{\sigma^2}{2\nu^2}$ , if  $s > \sigma^2/\nu$ .

Thus, for any arbitrary  $\varepsilon > 0$ ,

$$\mathbb{P} \left( \left| \sum_{i=1}^N b_i^N \|\widehat{\mathbf{y}}_i\|^p - \sum_{i=1}^N b_i^N m(\widehat{\mathbf{z}}_i) \right| \geq \varepsilon \mid \widehat{\mathbf{z}}_1, \dots, \widehat{\mathbf{z}}_N \right) \leq 2 \exp \left( -\frac{\|b\|_1^2}{\|b\|_2^2} t^* \left( \frac{\varepsilon}{\|b\|_1} \right) \right)$$

It holds  $\|b\|_1 = 1$ ,  $\|b\|_2^2 \leq 1/N\alpha_N$ , and  $\frac{s}{\nu} - \frac{\sigma^2}{2\nu^2} > \frac{s}{2\nu}$ , if  $s > \sigma^2/\nu$ . Hence,

$$\begin{aligned} \mathbb{P} \left( \left| \sum_{i=1}^N b_i^N \|\widehat{\mathbf{y}}_i\|^p - \sum_{i=1}^N b_i^N m(\widehat{\mathbf{z}}_i) \right| \geq \varepsilon \right) &\leq 2 \exp \left( -\frac{N\alpha_N \varepsilon^2}{2\sigma^2} \right) \mathbb{I}_{(\varepsilon \leq \sigma^2/\nu)} \\ &\quad + 2 \exp \left( -\frac{N\alpha_N \varepsilon}{2\nu} \right) \mathbb{I}_{(\varepsilon > \sigma^2/\nu)} \end{aligned}$$

which is summable with respect to  $N$  because  $\frac{N\alpha_N}{\log(N)} \rightarrow \infty$ .

To deal with the term  $\left| \sum_{i=1}^N b_i^N m(\widehat{\mathbf{z}}_i) - m(\mathbf{z}^*) \right|$ , we first note that

$$\left| \sum_{i=1}^N b_i^N m(\widehat{\mathbf{z}}_i) - m(\mathbf{z}^*) \right| \leq \sum_{i=1}^N b_i^N |m(\widehat{\mathbf{z}}_i) - m(\mathbf{z}^*)|$$

where

$$\begin{aligned} |m(\widehat{\mathbf{z}}_i) - m(\mathbf{z}^*)| &= \left| \int \|\mathbf{y}\|^p \phi_{\mathbf{y}/\mathbf{z}=\widehat{\mathbf{z}}_i}(\mathbf{y}) d\mathbf{y} - \int \|\mathbf{y}\|^p \phi_{\mathbf{y}/\mathbf{z}=\mathbf{z}^*}(\mathbf{y}) d\mathbf{y} \right| \\ &\leq \int \|\mathbf{y}\|^p |(\phi_{\mathbf{y}/\mathbf{z}=\widehat{\mathbf{z}}_i} - \phi_{\mathbf{y}/\mathbf{z}=\mathbf{z}^*})(\mathbf{y})| d\mathbf{y} \\ &\leq \|\widehat{\mathbf{z}}_i - \mathbf{z}^*\| \int \|\mathbf{y}\|^p \ell(y) d\mathbf{y} \\ &\leq R \|\widehat{\mathbf{z}}_i - \mathbf{z}^*\| \end{aligned}$$

for any  $\widehat{\mathbf{z}}_i$  such that  $\|\widehat{\mathbf{z}}_i - \mathbf{z}^*\| \leq r_0$ .

We finish the proof of case i) here, because, from this point on, the process is the same as in Lemma B.1, just replacing  $L$  and  $2\|h\|_\infty$  with  $R$  and  $\bar{m}$ , respectively.

- Case ii): Based on the corollary to [27, Theorem 25.12], it suffices to show that

$$\sup_N \int_{\mathbb{R}^{d_{\mathbf{y}}}} \|\mathbf{y}\|^a dQ_{\Xi}^N < \infty$$

We first compute the integral for a fixed  $N$ .

$$\int_{\mathbb{R}^{d_{\mathbf{y}}}} \|\mathbf{y}\|^a dQ_{\Xi}^N = \sum_{i=1}^N b_i^N \|\widehat{\mathbf{y}}_i\|^a = \sum_{i=1}^N b_i^N (\|\widehat{\mathbf{y}}_i\|^a - m(\widehat{\mathbf{z}}_i)) + \sum_{i=1}^N b_i^N m(\widehat{\mathbf{z}}_i)$$

$$\leq \left| \sum_{i=1}^N b_i^N (\|\hat{\mathbf{y}}_i\|^a - m(\hat{\mathbf{z}}_i)) \right| + \bar{m}$$

By Theorem B.5, we have, for any arbitrary  $\epsilon > 0$ ,

$$\begin{aligned} \mathbb{P} \left( \left| \sum_{i=1}^N b_i^N \|\hat{\mathbf{y}}_i\|^a - \sum_{i=1}^N b_i^N m(\hat{\mathbf{z}}_i) \right| \geq \epsilon \right) &\leq 2 \exp \left( -\frac{N\alpha_N \epsilon^2}{2\sigma^2} \right) \mathbb{I}_{(\epsilon \leq \sigma^2/\nu)} \\ &\quad + 2 \exp \left( -\frac{N\alpha_N \epsilon}{2\nu} \right) \mathbb{I}_{(\epsilon > \sigma^2/\nu)} \end{aligned}$$

which is summable with respect to  $N$ , because  $\frac{N\alpha_N}{\log(N)} \rightarrow \infty$ . Take  $\epsilon := \epsilon_0 > 0$ , there must then exist a sufficiently large  $N_0$  such that

$$\left| \sum_{i=1}^N b_i^N (\|\hat{\mathbf{y}}_i\|^a - m(\hat{\mathbf{z}}_i)) \right| < \epsilon_0$$

for  $N \geq N_0$  with probability one.

Therefore,

$$\int_{\mathbb{R}^{d_{\mathbf{y}}}} \|\mathbf{y}\|^a dQ_{\Xi}^N \leq \epsilon_0 + \bar{m}$$

for large enough  $N \geq N_0$ .

Thus,

$$\sup_N \int_{\mathbb{R}^{d_{\mathbf{y}}}} \|y\|^a dQ_{\Xi}^N \leq \max \left\{ \sup_{N < N_0} \int_{\mathbb{R}^{d_{\mathbf{y}}}} \|y\|^a dQ_{\Xi}^N, \epsilon_0 + \bar{m} \right\} < \infty \text{ a.s.}$$

□

**Remark B.7.** *The proof of Corollary B.5 is considerably simplified if it holds*

$$|m(\mathbf{z}) - m(\mathbf{z}^*)| \leq R \|\mathbf{z} - \mathbf{z}^*\|$$

for almost all  $\mathbf{z} \in \Xi_{\mathbf{z}}$  and some  $R > 0$ . In this case, for instance, we do not need the almost-everywhere boundedness condition on random variable  $m(\mathbf{z})$ .

# Appendix C

## Additional material to Section 4.4

### C.1 Notation used in Section 4.4

The main notation used throughout the Section 4.4 is stated below for quick reference. Other symbols are defined as required.

#### C.1.1 Sets, numbers and indices

$\mathcal{B}$  Set of buses, indexed by  $b$ .

$\mathcal{L}$  Set of lines, indexed by  $\ell$ .

$\mathcal{G}$  Set of generators (dispatchable units), indexed by  $j$ .

$\mathcal{W}$  Set of wind power plants, indexed by  $m$ .

#### C.1.2 Parameters and functions

$\mathbf{f}$  Array of forecasted power outputs [MW].

$\tilde{\mathbf{f}}$  Array of nominal (p.u.) forecasted power outputs.

$\mathbf{L}$  Array of loads [MW].

$\mathbf{g}^{\min}, \mathbf{g}^{\max}$  Array of upper and lower capacity limits of generators [MW].

$\mathbf{Cap}$  Array of line capacities [MW].

$\overline{\mathbf{C}}$  Array of installed capacities of the wind power plants [MW].

$\mathbf{M}^{\mathcal{G}/\mathcal{W}/\mathcal{B}}$  Matrix of DC power transfer distribution factors, which maps nodal power injections to line flows for generators/wind farms/loads.

$\mathbf{c}^D, \mathbf{c}^U$  Array of downward and upward reserve capacity costs [\$/MW].

$C(\cdot)$  Total production cost function, which is given by the sum of  $|\mathcal{G}|$  convex piecewise linear cost functions with  $S_j$  pieces for generator  $j$ , i.e.,

$$C(\tilde{\mathbf{g}}(\boldsymbol{\omega})) := \sum_{j \in \mathcal{G}} \max_{s=1, \dots, S_j} \{m_{js} \tilde{g}_j(\boldsymbol{\omega}) + n_{js}\},$$

where  $m_{js}, n_{js}$  are the slope and the intercept of the  $s$ -th piece for generator  $j$ , respectively [\\$].

### C.1.3 Random variables and uncertain parameters

|   |   |
|---|---|
| $\mathbf{z}$                              | Random vector of features/covariates.   |
| $\boldsymbol{\omega}$                     | Random vector representing the wind power forecast errors of the $ \mathcal{W} $ wind power plants [MW].  |
| $\boldsymbol{\xi}$                        | Random vector representing the pair of features/covariates and the wind power forecast errors of the $ \mathcal{W} $ wind farms, that is, $\boldsymbol{\xi} := (\mathbf{z}, \boldsymbol{\omega})$ .   |
| $W_m$                                     | Actual power output at wind power plant $m \in \mathcal{W}$ in per unit.  |
| $\Xi_{\boldsymbol{\omega}}$               | Support set of the random vector $\boldsymbol{\omega}$ .  |
| $\tilde{\Xi}_{\boldsymbol{\omega}}$       | Support set of the random vector $\boldsymbol{\omega}$ conditional on $\mathbf{z} = \mathbf{f}$ , which is given by the hypercube $\prod_{m \in \mathcal{W}} [-f_m, \bar{C}_m - f_m]$ .   |
| $\Xi$                                     | Support set of the random vector $(\mathbf{z}, \boldsymbol{\omega})$ .  |
| $\tilde{\Xi}$                             | Contextual information, that is, the event $(\mathbf{z} = \mathbf{f}; \boldsymbol{\omega} \in \tilde{\Xi}_{\boldsymbol{\omega}})$ .   |
| $\Omega$                                  | Random variable defined as $\sum_{m \in \mathcal{W}} \omega_m$ , which describes the system-wise aggregate wind power forecast error [MW].  |
| $\tilde{\Xi}_{\Omega}$                    | Contextual information linked to the random vector $(\mathbf{z}, \Omega)$ , that is, the event $(\mathbf{z} = \mathbf{f}; \Omega \in [\underline{\Omega}, \bar{\Omega}])$ , with $[\underline{\Omega}, \bar{\Omega}] = [-\sum_{m \in \mathcal{W}} f_m, \sum_{m \in \mathcal{W}} (\bar{C}_m - f_m)]$ . |
| $\tilde{\mathbf{g}}(\boldsymbol{\omega})$ | Array of power generation outputs of generators (random vector) [MW].   |
| $\tilde{\mathbf{r}}(\boldsymbol{\omega})$ | Array of reserves deployed by generators (random vector) [MW].  |
| $\mathbb{E}_Q$                            | Expectation operator with respect to the probability measure $Q$ .  |
| $\delta_{\boldsymbol{\xi}}$               | Dirac distribution at $\boldsymbol{\xi}$ .  |

**C.1.4 Variables**

|                              |   |
|------------------------------|---|
| $\mathbf{g}$                 | Generators' power dispatch [MW].  |
| $\beta$                      | Array of generators' participation factors.   |
| $\mathbf{r}^D, \mathbf{r}^U$ | Array of downward/upward reserve capacities provided by generators [MW].  |
| $\mathbf{x}$                 | Vector of decision variables, that is, $\mathbf{x} := (\mathbf{g}, \beta, \mathbf{r}^D, \mathbf{r}^U)$ .  |
| $\mathbf{y}$                 | Vector of first-stage decision variables (power dispatch and reserve capacity provision), that is, $\mathbf{y} := (\mathbf{g}, \mathbf{r}^D, \mathbf{r}^U)$ . |

**C.1.5 Other symbols**

|  |  |
|--|--|
| $\mathbf{1}$                                     | Array of ones (of appropriate dimension).  |
| $\mathbf{0}$                                     | Array of zeros (of appropriate dimension).   |
| $ A $  | Cardinal of a set $A$ .  |
| $(x)^+$  | Positive part of $x$ , i.e., $\max\{x, 0\}$ .  |
| $\lfloor x \rfloor$                              | Floor function of $x$ , given by $\max\{m \in \mathbb{Z} / m \leq x\}$ .   |
| $\langle \cdot, \cdot \rangle$                   | Dot product.   |
| $W_1$  | 1-Wasserstein distance.  |
| $\mathcal{P}_1(\Xi), \mathcal{P}_1(\tilde{\Xi})$ | The set of all probability distributions with finite first moment supported on $\Xi, \tilde{\Xi}$ , respectively.  |
| $\mathcal{R}_{1-\alpha}(P)$                      | The set of all $(1 - \alpha)$ -trimmings of the probability distribution $P$ .   |
| $\rho$   | Robustness parameter.  |
| $\epsilon_{N\alpha}$                             | Minimum transportation budget.   |
| $S_B$  | Support function of a set $B \subseteq \mathbb{R}^d$ , defined as $S_B(a) := \sup_{b \in B} \langle a, b \rangle$ .  |
| $Q - \mathbf{CVaR}_\epsilon(\phi(\omega))$       | Conditional Value at Risk at level $\epsilon \in (0, 1)$ of $\phi(\omega)$ under the probability measure $Q$ ; that is, the value $\inf_{\tau \in \mathbb{R}} \{\tau + \frac{1}{\epsilon} \mathbb{E}_Q[(\phi(\omega) - \tau)^+]\}$ . |

## C.2 Proof of Proposition 4.4

The proof of this proposition is a direct application of [50, Theorem 1], after noticing that the inner supremum in constraint (4.40) involves a maximum of linear functions. Variables  $\mathbf{v}_{ik}$  and  $\gamma_{ik}$  in optimization problem (4.42) are auxiliary variables that result from the dualization of the inner supremum that appears in the **CVaR** approximation of the chance constraint system, that is, in (4.39). This dualization is critical to the reformulation of (4.39) as a tractable mathematical program, see [50, Theorem 1].

## C.3 Proof of Proposition 4.5

Based on [50, Theorem 1], the DRO problem defined by (4.45) can be reformulated as follows:

$$\inf_{\lambda \geq 0; \bar{\mu}_i, \forall i \leq N; \theta \in \mathbb{R}} \lambda \rho + \theta + \frac{1}{N\alpha} \sum_{i=1}^N \bar{\mu}_i + \langle \mathbf{c}^D, \mathbf{r}^D \rangle + \langle \mathbf{c}^U, \mathbf{r}^U \rangle \quad (\text{C.1})$$

$$\text{s.t. } \bar{\mu}_i + \theta + \lambda \|\mathbf{z}^* - \hat{\mathbf{z}}_i\| \geq \sup_{\Omega \in [\underline{\Omega}, \bar{\Omega}]} \left( \sum_{j \in \mathcal{G}} \max_{s=1, \dots, S_j} \left\{ m_{js} [g_j - \beta_j \Omega] + n_{js} \right\} - \lambda |\Omega - \hat{\Omega}_i| \right), \quad \forall i \leq N \quad (\text{C.2})$$

$$\bar{\mu}_i \geq 0, \quad \forall i \leq N \quad (\text{C.3})$$

Now, constraint (C.2) is equivalent to the following ones:

$$\bar{\mu}_i + \theta + \lambda \|\mathbf{z}^* - \hat{\mathbf{z}}_i\| \geq t_i, \quad \forall i \leq N \quad (\text{C.4})$$

$$t_i \geq \sup_{\Omega \in [\underline{\Omega}, \bar{\Omega}]} \sum_{j \in \mathcal{G}} \max_{s=1, \dots, S_j} \left\{ m_{js} [g_j - \beta_j \Omega] + n_{js} \right\} - \lambda |\Omega - \hat{\Omega}_i|, \quad \forall i \leq N \quad (\text{C.5})$$

In order to reformulate the supremum on the right-hand side of (C.5), we resort to the following partition of the set  $\{1, \dots, N\}$ :

$$\underline{I} := \{i \in \{1, \dots, N\} : \hat{\Omega}_i < \underline{\Omega}\} \quad (\text{C.6})$$

$$I := \{i \in \{1, \dots, N\} : \hat{\Omega}_i \in [\underline{\Omega}, \bar{\Omega}]\} \quad (\text{C.7})$$

$$\bar{I} := \{i \in \{1, \dots, N\} : \hat{\Omega}_i > \bar{\Omega}\} \quad (\text{C.8})$$

In this way, because of the convexity of the sum of a maximum of affine functions, constraint (C.5) can be replaced by the following set of constraints:

$$t_i \geq \sum_{j \in \mathcal{G}} \max_{s=1, \dots, S_j} \left\{ m_{js} [g_j - \beta_j \underline{\Omega}] + n_{js} \right\} - \lambda(\underline{\Omega} - \hat{\Omega}_i), \quad \forall i \in \underline{I} \quad (\text{C.9})$$

$$t_i \geq \sum_{j \in \mathcal{G}} \max_{s=1, \dots, S_j} \left\{ m_{js} [g_j - \beta_j \overline{\Omega}] + n_{js} \right\} - \lambda(\overline{\Omega} - \hat{\Omega}_i), \quad \forall i \in \underline{I} \quad (\text{C.10})$$

$$t_i \geq \sum_{j \in \mathcal{G}} \max_{s=1, \dots, S_j} \left\{ m_{js} [g_j - \beta_j \underline{\Omega}] + n_{js} \right\} + \lambda(\underline{\Omega} - \hat{\Omega}_i), \quad \forall i \in \bar{I} \quad (\text{C.11})$$

$$t_i \geq \sum_{j \in \mathcal{G}} \max_{s=1, \dots, S_j} \left\{ m_{js} [g_j - \beta_j \overline{\Omega}] + n_{js} \right\} + \lambda(\overline{\Omega} - \hat{\Omega}_i), \quad \forall i \in \bar{I} \quad (\text{C.12})$$

$$t_i \geq \sum_{j \in \mathcal{G}} \max_{s=1, \dots, S_j} \left\{ m_{js} [g_j - \beta_j \overline{\Omega}] + n_{js} \right\} - \lambda(\overline{\Omega} - \hat{\Omega}_i), \quad \forall i \in I \quad (\text{C.13})$$

$$t_i \geq \sum_{j \in \mathcal{G}} \max_{s=1, \dots, S_j} \left\{ m_{js} [g_j - \beta_j \underline{\Omega}] + n_{js} \right\} + \lambda(\underline{\Omega} - \hat{\Omega}_i), \quad \forall i \in I \quad (\text{C.14})$$

$$t_i \geq \sum_{j \in \mathcal{G}} \max_{s=1, \dots, S_j} \left\{ m_{js} [g_j - \beta_j \hat{\Omega}_i] + n_{js} \right\}, \quad \forall i \in I \quad (\text{C.15})$$

Introducing epigraphical auxiliary variables  $\underline{t}_{ij}$ ,  $\bar{t}_{ij}$  and  $\hat{t}_{ij}$ , we finish the proof.  $\square$

## C.4 Real-time re-dispatch problem

This appendix contains the optimization program used to evaluate the out-of-sample performance of a given solution of the chance-constrained DC-OPF problem. Given  $N$ , a data-driven solution  $\mathbf{y}_N := (\mathbf{g}, \mathbf{r}^D, \mathbf{r}^U)_N$  and a realization of the forecast error  $\hat{\omega}_i$ , the operator of the system solves the following deterministic linear program:

$$\min_{\mathbf{r}, \Delta \mathbf{d}, \Delta \omega} C(\mathbf{g}_N + \mathbf{r}) + \langle \mathbf{c}^{\text{shed}}, \Delta \mathbf{d} \rangle + \langle \mathbf{c}^D, \mathbf{r}_N^D \rangle + \langle \mathbf{c}^U, \mathbf{r}_N^U \rangle \quad (\text{C.16})$$

$$\text{s.t. } \mathbf{0} \leq \Delta \mathbf{d} \leq \mathbf{L} \quad (\text{C.17})$$

$$\mathbf{0} \leq \Delta \omega \leq \mathbf{f} + \hat{\omega}_i \quad (\text{C.18})$$

$$-\mathbf{r}_N^D \leq \mathbf{r} \leq \mathbf{r}_N^U \quad (\text{C.19})$$

$$\langle \mathbf{1}, \mathbf{r} \rangle + \langle \mathbf{1}, \Delta \mathbf{d} \rangle + \langle \mathbf{1}, \hat{\omega}_i - \Delta \omega \rangle = 0 \quad (\text{C.20})$$

$$-\text{Cap} \leq \mathbf{M}^G(\mathbf{g}_N + \mathbf{r}) + \mathbf{M}^W(\mathbf{f} + \hat{\omega}_i - \Delta \omega) - \mathbf{M}^B(\mathbf{L} - \Delta \mathbf{d}) \leq \text{Cap} \quad (\text{C.21})$$

where  $\mathbf{r}$ ,  $\Delta \mathbf{d}$  and  $\Delta \omega$  are the deployed reserves, load shedding and wind spillage vector of decision variables; and the parameter  $c_b^{\text{shed}}$  is the load shedding cost at bus  $b$ . The



objective function in (C.16) minimizes the total operational cost of the system, which comprises the electricity generation cost, the load shedding cost and the total cost of up- and down-reserve capacities. The latter is known and constant and thus, does not intervene in the minimization. Constraints (C.17) and (C.18) limit the amount of load involuntarily curtailed and the amount of wind power unused to the actual realizations of the load and the wind power production, respectively. Constraint (C.19) ensures that the deployed reserves are kept within the reserve capacities scheduled in the forward stage. Constraint (C.20) constitutes the real-time power balance equation and, finally, constraints (C.21) enforce the transmission capacity limits.

## C.5 Illustrative example (3-bus system)

In this appendix, we use a small three-node system to illustrate how our DRO framework based on probability trimmings, named DROTRIMM, compares to that proposed in [19], which we refer to as KNNDRO. This other approach is based on a local inference method (specifically, a  $K$ -nearest neighbors), to construct, from the joint data sample, the conditional empirical distribution at which a Wasserstein ball is centered. We also compare DROTRIMM with the DROW approach introduced in Section 4.4.5 of the main text. Actually, DROW becomes equivalent to KNNDRO when taking  $K = N$ .

The topology of the three-bus system has been taken from [100]. It includes three lines connecting buses 1–2, 2–3, and 1–3, three generators located at nodes 1, 2 and 3, and a 200-MW load connected to bus 3. The production cost of the three generators is modeled as a piecewise function consisting of three pieces. Further details on the generators' and network's parameters can be found in Appendix C.6.

We consider one wind power plant placed at bus 2 with an installed capacity of  $\bar{C}_1 = 60$  MW. Its predicted power output is assumed to be  $z^* = f_1 = 30$  MW in this example. We also assume that the system operator has a series of data pairs given by past point forecasts of the power output at the wind farm and their associated forecast error. Figure C.1 shows a heat map of the true bivariate joint distribution of the forecast power output and its error, together with a kernel estimate of the probability density function of the random forecast error  $\omega$  conditional on  $z^* = 30$  MW. The joint support set  $\Xi$  is polyhedral, and recall that the conditional support varies with the value of  $z$  (see the red line in Figure C.1).

The box plots corresponding to the total downward and upward reserve capacity that is procured, the violation probability and the expected cost delivered out of sample by each of the considered CC-DRO OPF models is depicted in Figures C.2 and C.4 as a function of their respective robustness parameter, estimated over 200 independent runs for a fixed sample size  $N = 30$  and  $N = 2000$ , respectively. Similarly, the plots in Figures C.3 and C.5 pertain to the generators' dispatch and their participation factors.

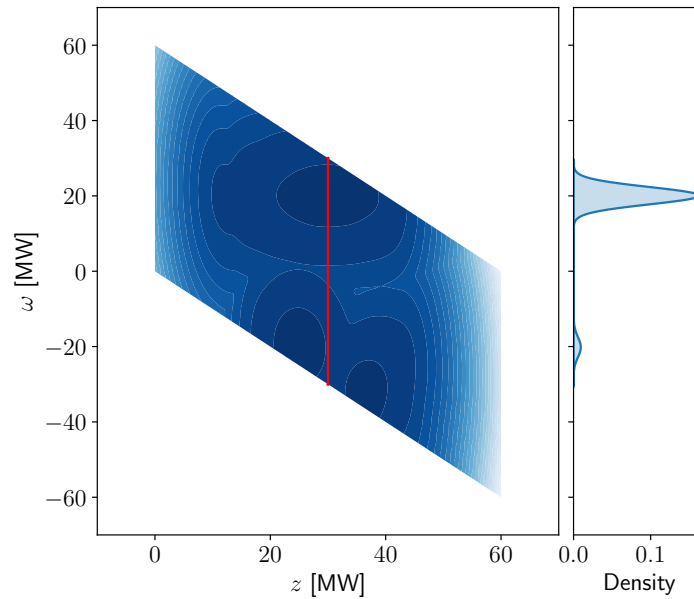


Figure C.1: Heat map of the true joint distribution and kernel estimate of the true conditional density given  $z^* = 30$  MW

The robustness parameter of KNNDRO and DROW corresponds to the radius of the Wasserstein ball these methods use as the ambiguity set. For its part, the robustness parameter for DROTRIMM is to be greater than or equal to the minimum transportation budget (see Definition 4.2). Therefore, what the system operator really needs to tune in DROTRIMM is the budget excess as done in Section 4.2. This is what we represent in the  $x$ -axes of the aforementioned figures for this method.

As in the case study in Section 4.4.5 of the main text, the color-shaded areas have been obtained by joining the minimum and maximum edge cases of the box plots, while the associated bold colored lines link their means. The number of neighbors  $K_N$  we have considered for KNNDRO is given by the logarithmic rule, that is, if  $N$  is the sample size of the joint data, then the number of neighbors is computed as  $K_N = \lfloor N/(\log(N+1)) \rfloor$ . To ensure a fair comparison, we have also taken  $\alpha_N = K_N/N$  for DROTRIMM. Figures C.2–C.5 provide information on the ability of each method to discover good dispatch solutions, that is, scheduling plans for power production and reserve capacity provision that are cost-efficient in expectation while guaranteeing the desired system reliability. We consider two settings, namely, the case of a small sample size ( $N = 30$ ), for which the available joint data is expected to carry little information on the true conditional distribution of the wind power forecast error (Figures C.2 and C.3), and another one where the sample size is notably higher, specifically  $N = 2000$  (Figures C.4 and C.5).

Figure C.2 shows that, for a value of the robustness parameter equal to or greater than  $10^2$ , DROTRIMM recovers the *robust* dispatch, that is, the dispatch that performs the best under the worst-case value of the forecast error, having predicted 30 MW of

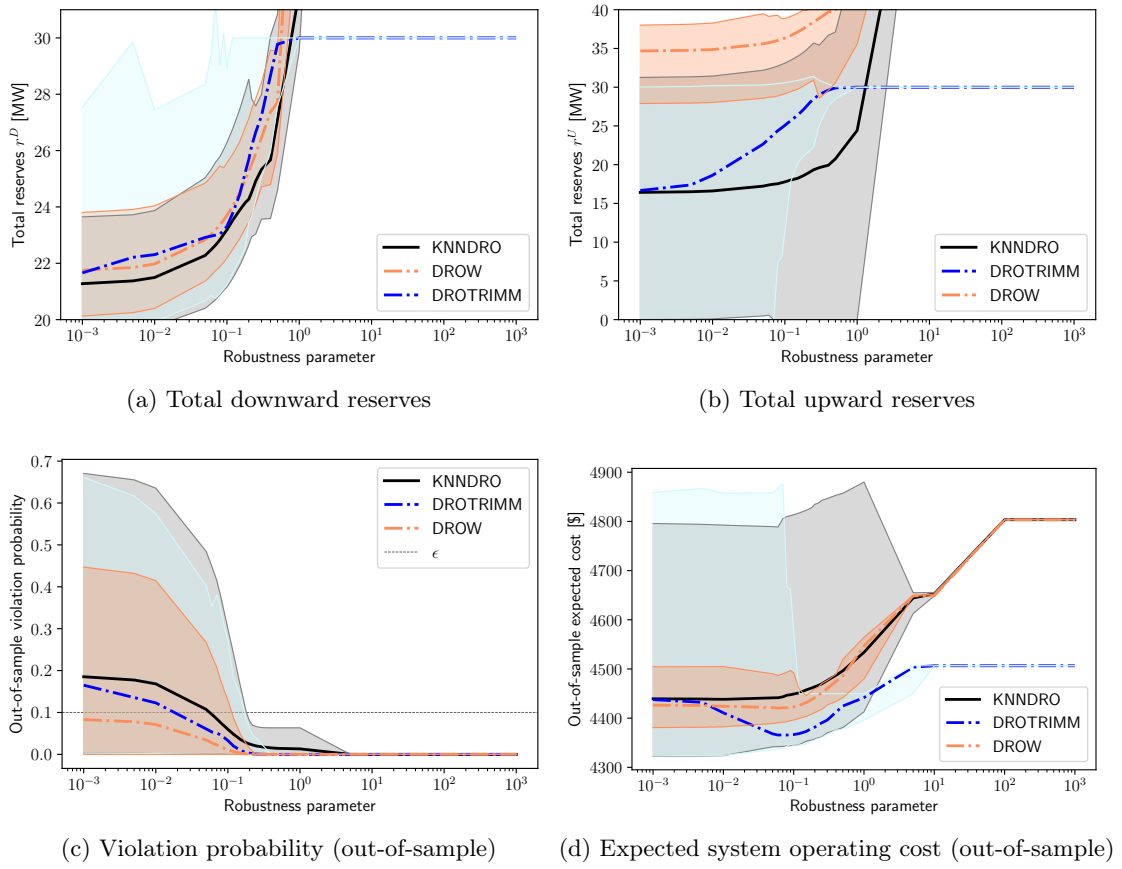


Figure C.2: Three-bus system, sample size  $N = 30$  and  $\epsilon = 0.1$ : Total downward and upward reserves and performance metrics

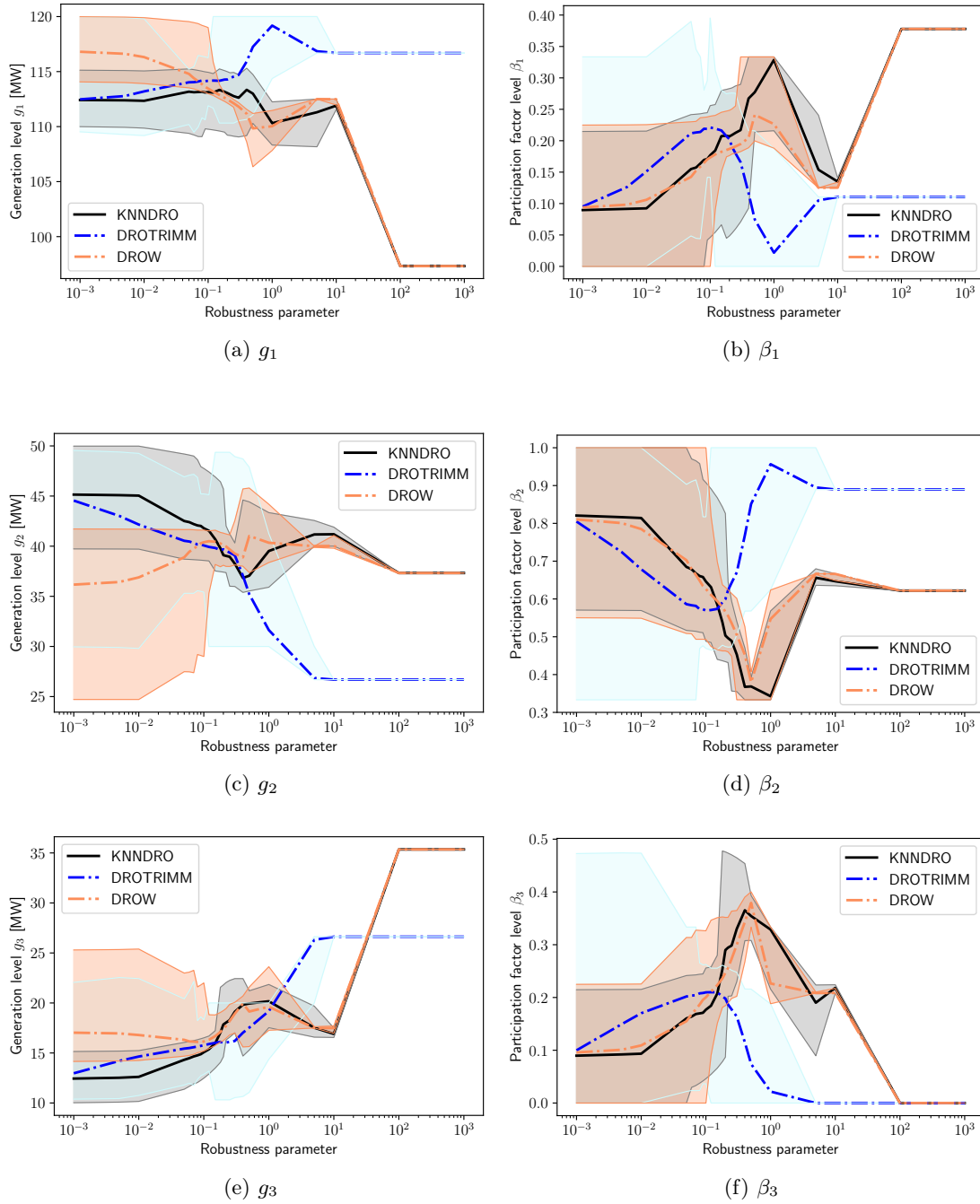


Figure C.3: Three-bus system, sample size  $N = 30$  and  $\epsilon = 0.1$ : Generators' dispatch and participation factors

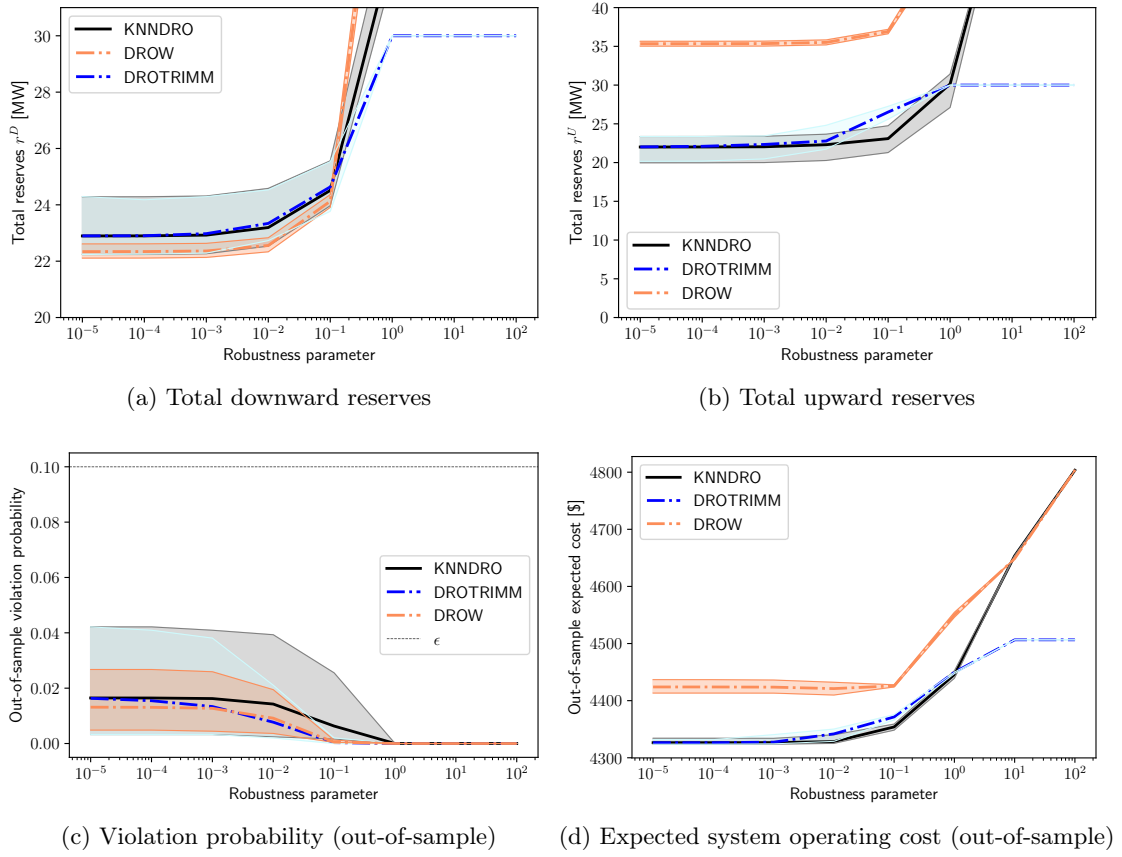


Figure C.4: Three-bus system, sample size  $N = 2000$  and  $\epsilon = 0.1$ : Total downward and upward reserves and performance metrics

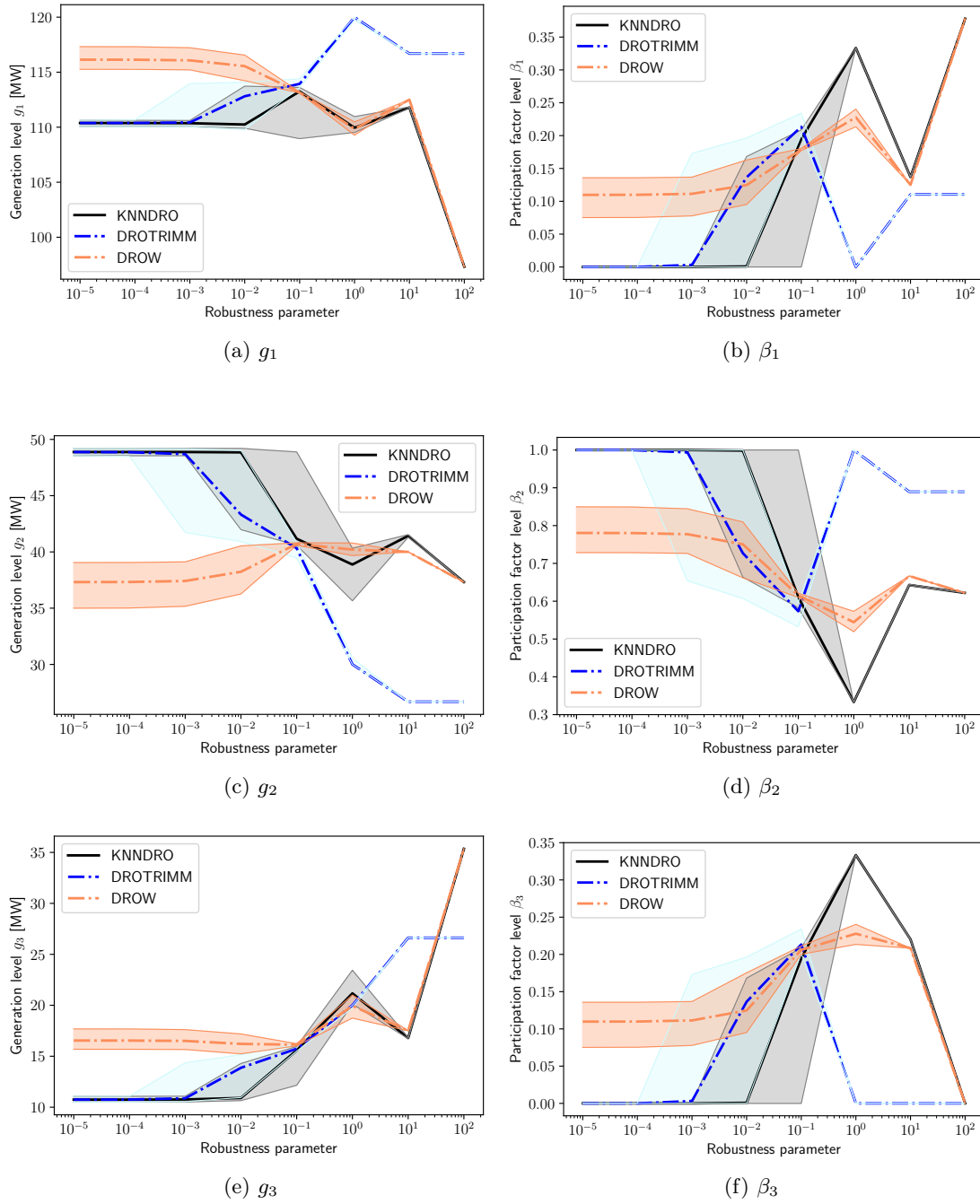


Figure C.5: Three-bus system, sample size  $N = 2000$  and  $\epsilon = 0.1$ : Generators' dispatch and participation factors

wind power production. On the contrary, KNNDRO and DROW recover the *unconditional robust* dispatch needing a substantially higher amount of reserves than the ones required by DROTRIMM. This is so because the probability distributions in the Wasserstein balls that DROW and KNNDRO use are not necessarily supported on the red line shown in Figure C.1. Additionally, note that DROTRIMM achieves a violation probability that is always lower than or equal to that attained by KNNDRO. Moreover, if we consider values of the robustness parameter smaller than  $10^{-1}$  in Figure C.2c, many of the dispatch solutions delivered by DROTRIMM, DROW, and KNNDRO lead to a violation probability higher than the threshold  $\epsilon = 0.1$ , which highlights the value of *distributional robustness* in the OPF problem under uncertainty.

When  $N = 30$ , DROW exhibits a remarkably robust behavior in the sense that the OPF solutions that this method delivers exhibit comparatively lower variance in terms of both expected cost and reliability. However, like DROTRIMM, DROW also needs a value of its robustness parameter around  $10^{-1}$  (or greater) to ensure that almost none of the dispatch solutions it provides violate the reliability threshold  $\epsilon = 0.1$ . In fact, it is within a neighborhood of that value for the robustness parameter where the three methods provide reliable OPF solutions with the best cost performance on average. Nevertheless, DROTRIMM is able to identify dispatch solutions that are about 1% cheaper in expectation than those given by DROW and KNNDRO. In this regard, Figure C.3 reveals that DROTRIMM achieves this percentage point in expected cost savings through a slightly different generators' dispatch and participation factors. Indeed, see how close to one another the colored bold lines in that figure are for a value of the robustness parameter around  $10^{-1}$ . Nonetheless, it is also noteworthy that from  $10^{-1}$  onward, the OPF solutions given by DROTRIMM start to diverge from those provided by DROW and KNDDRO in a noticeable way.

DROTRIMM's solutions are systematically more cost-efficient on average and variance (and even more reliable) than those of KNNDRO. The reason for this is that KNNDRO does not explicitly protect the dispatch against the potential conditional inference error incurred by the local predictive method it relies on.

When the sample size is high enough, e.g.,  $N = 2000$ , Figure C.4c shows that all the methods provide reliable dispatch solutions (i.e., power dispatches that ensure the threshold  $\epsilon = 0.1$ ) for any value of their respective robustness parameter. Moreover, the three methods provide the best OPF solutions in terms of expected cost for low values of this parameter. However, the solutions given by DROW are about 2.3% more expensive, because this method procures far more upward reserve capacity than needed (see Figures C.4b). For this, as can be seen in Figure C.5, DROW provides a dispatch for the generators and values for their participation factors that are starkly different from those given by the other two methods. On the other hand, the performance of DROTRIMM and KNNDRO tends to be similar in a large-sample regime. Indeed,

both are able to disclose dispatches that result in similar system operating costs while guaranteeing the reliability of the system, although DROTRIMM is still comparatively better (see Figures C.4c and C.4d in the range of the robustness parameter between  $10^{-5}$  y  $10^{-3}$ ). This result is hardly surprising because, as the sample data grows in size, the uncertainty intrinsic to the conditional inference diminishes.

Since we have detected in our numerical experiments that DROTRIMM always provide dispatches with a performance similar or superior to those given by KNNDRO, we have not discussed the latter in the case study of Section 4.4.5.

## C.6 Data for the illustrative example (3-bus system)

This appendix compiles data pertaining to the illustrative example that has been presented and discussed in C.5.

| Generator index | Bus index | $c_j^D$ | $c_j^U$ | $g_j^{\min}$ | $g_j^{\max}$ |
|-----------------|-----------|---------|---------|--------------|--------------|
| 1               | 1         | 6       | 3       | 0            | 120          |
| 2               | 2         | 2       | 5       | 0            | 80           |
| 3               | 3         | 4       | 8       | 0            | 100          |

Table C.1: Generators' location, power output limits and reserve capacity costs

| Generator index | Slopes $m_s$ |         |         | Intercepts $n_s$ |         |         |
|-----------------|--------------|---------|---------|------------------|---------|---------|
|                 | piece 1      | piece 2 | piece 3 | piece 1          | piece 2 | piece 3 |
| 1               | 22           | 26      | 30      | 0                | -173    | -493    |
| 2               | 29           | 37      | 45      | 0                | -231    | -658    |
| 3               | 38           | 55      | 71      | 0                | -601    | -1715   |

Table C.2: Slopes ( $m_s$ ) and intercepts ( $n_s$ ) of the generators' piecewise linear cost functions

| index | from bus | to bus | X (reactance p.u.) | Cap (MW) |
|-------|----------|--------|--------------------|----------|
| 1     | 1        | 2      | 0.13               | 100      |
| 2     | 1        | 3      | 0.13               | 100      |
| 3     | 2        | 3      | 0.13               | 100      |

Table C.3: Transmission line parameters



# References

- [1] Agulló Antolín, M.: Trimming methods for model validation and supervised classification in the presence of contamination. Ph.D. thesis, University of Valladolid (2018). URL <http://uvadoc.uva.es/handle/10324/31682>
- [2] Álvarez-Esteban, Del Barrio, E., Cuesta-Albertos, J.A., Matrán, C.: Similarity of samples and trimming. *Bernoulli* **18**(2), 606–634 (2012)
- [3] Álvarez-Esteban, P.C., del Barrio, E., Cuesta-Albertos, J.A., Matrán, C.: Uniqueness and approximate computation of optimal incomplete transportation plans. *Annales de l’Institut Henri Poincaré, Probabilités et Statistiques* **47**(2), 358 – 375 (2011)
- [4] Ambrosio, L., Gigli, N., Savaré, G.: *Gradient Flows: In Metric Spaces and in the Space of Probability Measures*. Springer Science & Business Media (2005)
- [5] Andersson, J., Jörnsten, K., Nonås, S.L., Sandal, L., Ubøe, J.: A maximum entropy approach to the newsvendor problem with partial information. *European Journal of Operational Research* **228**(1), 190–200 (2013)
- [6] Arrigo, A., Kazempour, J., De Grève, Z., Toubéau, J.F., Vallée, F.: *Embedding Dependencies Within Distributionally Robust Optimization of Modern Power Systems* (2021). URL <http://arxiv.org/abs/2104.08101>
- [7] Arrigo, A., Ordoudis, C., Kazempour, J., De Grève, Z., Toubéau, J.F., Vallée, F.: Wasserstein distributionally robust chance-constrained optimization for energy and reserve dispatch: An exact and physically-bounded formulation. *European Journal of Operational Research* **296**(1), 304–322 (2022)
- [8] Balghithi, O.E., Elmachoub, A.N., Grigas, P., Tewari, A.: Generalization bounds in the predict-then-optimize framework (2019). URL <http://arxiv.org/abs/1905.11488>
- [9] Ban, G.Y., Rudin, C.: The big data newsvendor: Practical insights from machine learning. *Operations Research* **67**(1), 90–108 (2019)

- [10] del Barrio, E., Matrán, C.: Rates of convergence for partial mass problems. *Probability Theory and Related Fields* **155**(3-4), 521–542 (2013)
- [11] Bayraksan, G., Love, D.K.: Data-driven stochastic programming using phi-divergences. In: *The Operations Research Revolution*, pp. 1–19. INFORMS (2015)
- [12] Ben-Tal, A., El Ghaoui, L., Nemirovski, A.: *Robust optimization*. Princeton university press (2009)
- [13] Ben-Tal, A., den Hertog, D., De Waegenaere, A., Melenberg, B., Rennen, G.: Robust solutions of optimization problems affected by uncertain probabilities. *Management Science* **59**(2), 341–357 (2013)
- [14] Bercu, B., Delyon, B., Rio, E.: *Concentration Inequalities for Sums and Martingales*. SpringerBriefs in Mathematics. Springer International Publishing, Cham (2015)
- [15] Bertsimas, D., Gupta, V., Kallus, N.: Data-driven robust optimization. *Mathematical Programming* **167**(2), 235–292 (2018)
- [16] Bertsimas, D., Gupta, V., Kallus, N.: Robust sample average approximation. *Mathematical Programming* **171**(1-2), 217–282 (2018)
- [17] Bertsimas, D., Kallus, N.: From predictive to prescriptive analytics. *Management Science* **66**(3), 1025–1044 (2020)
- [18] Bertsimas, D., McCord, C.: Optimization over continuous and multi-dimensional decisions with observational data (2018). URL <http://arxiv.org/abs/1807.04183>
- [19] Bertsimas, D., McCord, C., Sturt, B.: Dynamic optimization with side information. *European Journal of Operational Research* (2022). DOI 10.1016/j.ejor.2022.03.030
- [20] Bertsimas, D., Shtern, S., Sturt, B.: A Data-Driven Approach to Multistage Stochastic Linear Optimization. *Management Science* (2022). DOI 10.1287/mnsc.2022.4352
- [21] Bertsimas, D., Shtern, S., Sturt, B.: Technical Note—Two-Stage Sample Robust Optimization. *Operations Research* **70**(1), 624–640 (2022)
- [22] Bertsimas, D., Van Parys, B.: Bootstrap robust prescriptive analytics (2017). URL <http://arxiv.org/abs/1711.09974>
- [23] Bhattacharya, B.: Testing multinomial parameters under order restrictions. *Communications in Statistics - Theory and Methods* **26**(8), 1839–1865 (1997)

- [24] Biau, G., Devroye, L.: Lectures on the Nearest Neighbor Method. Springer Series in the Data Sciences. Springer International Publishing, Cham (2015)
- [25] Bienstock, D., Chertkov, M., Harnett, S.: Chance-constrained optimal power flow: Risk-aware network control under uncertainty. *SIAM Rev.* **56**(3), 461–495 (2014)
- [26] Billingsley, P. (ed.): Convergence of probability measures. Wiley Series in Probability and Statistics. John Wiley & Sons, Inc., Hoboken, NJ, USA (1999)
- [27] Billingsley, P.: Probability and Measure. Wiley Series in Probability and Statistics. Wiley (2012)
- [28] Blanchet, J., Kang, Y., Murthy, K., Zhang, F.: Data-driven optimal transport cost selection for distributionally robust optimization. In: 2019 Winter Simulation Conference (WSC), pp. 3740–3751 (2019)
- [29] Blanchet, J., Kang, Y., Zhang, F., He, F., Hu, Z.: Doubly robust data-driven distributionally robust optimization. In: Applied Modeling Techniques and Data Analysis 1, pp. 75–90. Wiley (2021)
- [30] Blanchet, J., Murthy, K., Nguyen, V.A.: Statistical analysis of wasserstein distributionally robust estimators. In: Tutorials in Operations Research: Emerging Optimization Methods and Modeling Techniques with Applications, pp. 227–254. INFORMS (2021)
- [31] Bludszuweit, H., Dominguez-Navarro, J., Llobart, A.: Statistical analysis of wind power forecast error. *IEEE Transactions on Power Systems* **23**(3), 983–991 (2008)
- [32] Boyd, S., Boyd, S.P., Vandenberghe, L.: Convex Optimization. Cambridge University Press (2004)
- [33] Cascos, I., López-Díaz, M.: Consistency of the  $\alpha$ -trimming of a probability. Applications to central regions. *Bernoulli* **14**(2), 580–592 (2008)
- [34] Chen, R.: Distributionally robust learning under the Wasserstein metric. Ph.D. thesis (2019). URL <https://open.bu.edu/handle/2144/38236>
- [35] Chen, R., Paschalidis, I.C.: A robust learning approach for regression models based on distributionally robust optimization. *Journal of Machine Learning Research* **19**(13), 1–48 (2018)
- [36] Chen, X., Lin, Q., Xu, G.: Distributionally robust optimization with confidence bands for probability density functions. *INFORMS Journal on Optimization* **4**(1), 65–89 (2022)

- [37] Chen, Z., Kuhn, D., Wiesemann, W.: Data-driven chance constrained programs over Wasserstein balls. *Operations Research* (2022). DOI 10.1287/opre.2022.2330
- [38] Chen, Z., Sim, M., Xiong, P.: Robust stochastic optimization made easy with rsome. *Management Science* **66**(8), 3329–3339 (2020)
- [39] Choi, T.M. (ed.): Handbook of Newsvendor Problems, *International Series in Operations Research & Management Science*, vol. 176. Springer New York, New York, NY (2012)
- [40] Cisneros-Velarde, P., Petersen, A., Oh, S.Y.: Distributionally robust formulation and model selection for the graphical lasso. In: S. Chiappa, R. Calandra (eds.) *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics, Proceedings of Machine Learning Research*, vol. 108, pp. 756–765. PMLR, Online (2020)
- [41] Dangeti, P.: *Statistics for machine learning: Techniques for exploring supervised, unsupervised, and reinforcement learning models with Python and R*. Packt Publishing (2017)
- [42] Delage, E., Ye, Y.: Distributionally robust optimization under moment uncertainty with application to data-driven problems. *Operations Research* **58**(3), 595–612 (2010)
- [43] Devroye, L.: On the almost everywhere convergence of nonparametric regression function estimates. *The Annals of Statistics* **9**(6), 1310–1319 (1981)
- [44] Devroye, L.: Necessary and sufficient conditions for the pointwise convergence of nearest neighbor regression function estimates. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete* **61**(4), 467–481 (1982)
- [45] Diao, S., Sen, S.: Distribution-free algorithms for learning enabled predictive stochastic programming (2020). URL [http://www.optimization-online.org/DB\\_HTML/2020/03/7661.html](http://www.optimization-online.org/DB_HTML/2020/03/7661.html)
- [46] Donti, P., Amos, B., Kolter, J.Z.: Task-based end-to-end model learning in stochastic optimization. *Adv. Neural Inf. Process Syst.* pp. 5484–5494 (2017)
- [47] Duchi, J.C., Glynn, P.W., Namkoong, H.: Statistics of robust optimization: A generalized empirical likelihood approach. *Mathematics of Operations Research* p. moor.2020.1085 (2021)
- [48] Elmachtoub, A.N., Grigas, P.: Smart “predict, then optimize”. *Management Science* (2021). DOI 10.1287/mnsc.2020.3922

- [49] Erdoğan, E., Iyengar, G.: Ambiguous chance constrained problems and robust optimization. *Mathematical Programming* **107**(1-2), 37–61 (2006)
- [50] Esteban-Pérez, A., Morales, J.M.: Distributionally robust stochastic programs with side information based on trimmings - Extended version (2020). URL <http://arxiv.org/abs/2009.10592>
- [51] Esteban-Pérez, A., Morales, J.M.: Distributionally robust optimal power flow with contextual information. *arXiv preprint arXiv:2109.07896* (2021)
- [52] Esteban-Pérez, A., Morales, J.M.: Distributionally robust optimal power flow with contextual information – Codes and data. *GitHub repository* (2021). URL [https://github.com/groupoasys/DRO\\_DCOPF\\_CONTEXTUAL](https://github.com/groupoasys/DRO_DCOPF_CONTEXTUAL)
- [53] Esteban-Pérez, A., Morales, J.M.: Distributionally robust stochastic programs with side information based on trimmings. *Mathematical Programming* (2021). DOI 10.1007/s10107-021-01724-0
- [54] Esteban-Pérez, A., Morales, J.M.: Distributionally robust stochastic programs with side information based on trimmings – Codes and Data. *GitHub repository* (2021). URL [https://github.com/groupoasys/DRO\\_CONDITIONAL\\_TRIMMINGS](https://github.com/groupoasys/DRO_CONDITIONAL_TRIMMINGS)
- [55] Esteban-Pérez, A., Morales, J.M.: Partition-based distributionally robust optimization via optimal transport with order cone constraints. *4OR* (2021). DOI 10.1007/s10288-021-00484-z
- [56] Fabbri, A., Gomez San Roman, T., Rivier Abbad, J., Mendez Quezada, V.: Assessment of the cost associated with wind generation prediction errors in a liberalized electricity market. *IEEE Transactions on Power Systems* **20**(3), 1440–1446 (2005)
- [57] Falk, M., Hüsler, J., Reiss, R.D.: *Laws of Small Numbers: Extremes and Rare Events*. Springer Science & Business Media (2010)
- [58] Farokhi, F.: Why does regularization help with mitigating poisoning attacks? *Neural Processing Letters* **53**(4), 2933–2945 (2021)
- [59] Fournier, N., Guillin, A.: On the rate of convergence in Wasserstein distance of the empirical measure. *Probability Theory and Related Fields* **162**(3-4), 707–738 (2015)
- [60] Gabriel, S.A., Conejo, A.J., Fuller, J.D., Hobbs, B.F., Ruiz, C.: *Complementarity modeling in energy markets*, vol. 180. Springer Science & Business Media (2012)

- [61] Gallego, G., Moon, I.: The distribution free newsboy problem: Review and extensions. *The Journal of the Operational Research Society* **44**(8), 825 (1993)
- [62] Gao, R.: Finite-sample guarantees for Wasserstein distributionally robust optimization: Breaking the curse of dimensionality. *Operations Research* (2022). DOI 10.1287/opre.2022.2326
- [63] Gao, R., Kleywegt, A.J.: Distributionally Robust Stochastic Optimization with Wasserstein Distance (2016). URL <http://arxiv.org/abs/1604.02199>
- [64] Gao, R., Kleywegt, A.J.: Distributionally Robust Stochastic Optimization with Dependence Structure (2017). URL <http://arxiv.org/abs/1701.04200>
- [65] Gibbs, A.L., Su, F.E.: On choosing and bounding probability metrics. *International statistical review* **70**(3), 419–435 (2002)
- [66] Graf, S., Luschgy, H.: Foundations of Quantization for Probability Distributions, *Lecture Notes in Mathematics*, vol. 1730. Springer Berlin Heidelberg, Berlin, Heidelberg (2000)
- [67] Gray, R.M.: Probability, Random Processes, and Ergodic Properties. Springer US, Boston, MA (2009). DOI 10.1007/978-1-4419-1090-5
- [68] Guo, S., Xu, H.: Distributionally robust shortfall risk optimization model and its approximation. *Mathematical Programming* **174**(1), 473–498 (2019)
- [69] Hanasusanto, G.A., Kuhn, D.: Robust data-driven dynamic programming. *Advances in Neural Information Processing Systems* **26**, 827–835 (2013)
- [70] Hanasusanto, G.A., Kuhn, D., Wallace, S.W., Zymmler, S.: Distributionally robust multi-item newsvendor problems with multimodal demand distributions. *Mathematical Programming* **152**(1-2), 1–32 (2015)
- [71] Hanasusanto, G.A., Kuhn, D., Wiesemann, W.: A comment on “computational complexity of stochastic programming problems”. *Mathematical Programming* **159**(1), 557–569 (2016)
- [72] Hanasusanto, G.A., Roitch, V., Kuhn, D., Wiesemann, W.: A distributionally robust perspective on uncertainty quantification and chance constrained programming. *Mathematical Programming* **151**(1), 35–62 (2015)
- [73] Hanasusanto, G.A., Roitch, V., Kuhn, D., Wiesemann, W.: Ambiguous joint chance constraints under mean and dispersion information. *Operations Research* **65**(3), 751–767 (2017)

- [74] Hannah, L., Powell, W., Blei, D.: Nonparametric density estimation for stochastic optimization with an observable state variable. *Advances in Neural Information Processing Systems* **23**, 820–828 (2010)
- [75] Hess, C.: Conditional expectation and martingales of random sets. *Pattern Recognition* **32**(9), 1543–1567 (1999)
- [76] Hong, T., Pinson, P., Fan, S., Zareipour, H., Troccoli, A., Hyndman, R.J.: Probabilistic energy forecasting: Global energy forecasting competition 2014 and beyond. *Int. J. Forecast.* **32**(3), 896–913 (2016)
- [77] Huber, J., Müller, S., Fleischmann, M., Stuckenschmidt, H.: A data-driven newsvendor problem: From data to decision. *European Journal of Operational Research* **278**(3), 904–915 (2019)
- [78] Jabr, R.A.: Distributionally robust CVaR constraints for power flow optimization. *IEEE Transactions on Power Systems* **35**(5), 3764–3773 (2020)
- [79] Ji, R., Lejeune, M.A.: Data-driven optimization of reward-risk ratio measures. *INFORMS Journal on Computing* p. ijoc.2020.1002 (2020)
- [80] Jia, M., Hug, G., Shen, C.: Iterative decomposition of joint chance constraints in OPF. *IEEE Transactions on Power Systems* **36**(5), 4836–4839 (2021)
- [81] Jiang, N., Xie, W.: ALSO-X and ALSO-X+: Better Convex Approximations for Chance Constrained Programs. *Operations Research* (2022). DOI 10.1287/opre.2021.2225
- [82] Kannan, R., Bayraksan, G., Luedtke, J.: Heteroscedasticity-aware residuals-based contextual stochastic optimization. *arXiv preprint arXiv:2101.03139* (2021)
- [83] Kannan, R., Bayraksan, G., Luedtke, J.R.: Data-driven sample average approximation with covariate information. *Optimization Online*. URL: [http://www.optimization-online.org/DB\\_HTML/2020/07/7932.html](http://www.optimization-online.org/DB_HTML/2020/07/7932.html) (2020)
- [84] Kannan, R., Bayraksan, G., Luedtke, J.R.: Residuals-based distributionally robust optimization with covariate information (2020). URL <http://arxiv.org/abs/2012.01088>
- [85] Keith, A.J., Ahner, D.K.: A survey of decision making and optimization under uncertainty. *Annals of Operations Research* **300**(2), 319–353 (2021)
- [86] Koduri, N.: Essays on decision making under uncertainty. Ph.D. thesis, Massachusetts Institute of Technology (2021). URL <https://dspace.mit.edu/handle/1721.1/139032>

- [87] Kuhn, D., Esfahani, P.M., Nguyen, V.A., Shafieezadeh-Abadeh, S.: Wasserstein distributionally robust optimization: Theory and applications in machine learning. In: Operations Research & Management Science in the Age of Analytics, pp. 130–166. INFORMS (2019)
- [88] Li, B., Jiang, R., Mathieu, J.L.: Distributionally robust risk-constrained optimal power flow using moment and unimodality information. In: 2016 IEEE 55th Conf. on Decision and Control (CDC), pp. 2425–2430 (2016)
- [89] Li, B., Jiang, R., Mathieu, J.L.: Ambiguous risk constraints with moment and unimodality information. *Mathematical Programming* **173**(1-2), 151–192 (2019)
- [90] Li, B., Jiang, R., Mathieu, J.L.: Ambiguous risk constraints with moment and unimodality information. *Mathematical Programming* **173**(1-2), 151–192 (2019)
- [91] Li, B., Jiang, R., Mathieu, J.L.: Distributionally robust chance-constrained optimal power flow assuming unimodal distributions with misspecified modes. *IEEE Trans. Control Netw. Syst.* **6**(3), 1223–1234 (2019)
- [92] Liu, J., Chen, Z., Lisser, A., Xu, Z.: Closed-form optimal portfolios of distributionally robust mean-CVaR problems with unknown mean and variance. *Applied Mathematics & Optimization* **79**(3), 671–693 (2019)
- [93] Liu, Q., Wu, J., Xiao, X., Zhang, L.: A note on distributionally robust optimization under moment uncertainty. *Journal of Numerical Mathematics* **26**(3), 141–150 (2018)
- [94] Liu, Y., Xu, H.: Stability analysis of stochastic programs with second order dominance constraints. *Mathematical Programming* **142**(1-2), 435–460 (2013)
- [95] Loubes, J.M., Pelletier, B.: Prediction by quantization of a conditional distribution. *Electronic Journal of Statistics* **11**(1), 2679–2706 (2017)
- [96] Lubin, M., Dvorkin, Y., Backhaus, S.: A robust approach to chance constrained optimal power flow with renewable generation. *IEEE Transactions on Power Systems* **31**(5), 3840–3849 (2016)
- [97] Lucchetti, R.: Convexity and Well-Posed Problems. CMS Books in Mathematics. Springer New York, New York, NY (2006)
- [98] Mehrotra, S., Papp, D.: A cutting surface algorithm for semi-infinite convex programming with an application to moment robust optimization. *SIAM Journal on Optimization* **24**(4), 1670–1697 (2014)



- [99] Mohajerin Esfahani, P., Kuhn, D.: Data-driven distributionally robust optimization using the Wasserstein metric: performance guarantees and tractable reformulations. *Mathematical Programming* **171**(1-2), 115–166 (2018)
- [100] Morales, J.M., Conejo, A.J., Liu, K., Zhong, J.: Pricing electricity in pools with wind producers. *IEEE Transactions on Power Systems* **27**(3), 1366–1376 (2012)
- [101] Muñoz, M.A., Pineda, S., Morales, J.M.: A bilevel framework for decision-making under uncertainty with contextual information. *Omega* **108**, 102575 (2022)
- [102] Nakao, H., Shen, S., Chen, Z.: Network design in scarce data environment using moment-based distributionally robust optimization. *Computers & Operations Research* **88**, 44–57 (2017)
- [103] Namkoong, H., Duchi, J.: Stochastic gradient methods for distributionally robust optimization with f-divergences. *Nips 2016* (2016)
- [104] Németh, A., Németh, S.: Isotonic regression and isotonic projection. *Linear Algebra and its Applications* **494**, 80 – 89 (2016)
- [105] Nemirovski, A., Shapiro, A.: Convex approximations of chance constrained programs. *SIAM Journal on Optimization* **17**(4), 969–996 (2007)
- [106] Nguyen, V.A., Zhang, F., Blanchet, J., Delage, E., Ye, Y.: Distributionally robust local non-parametric conditional estimation (2020). DOI 10.05373. URL <http://arxiv.org/abs/2010.05373>
- [107] Nguyen, V.A., Zhang, F., Blanchet, J., Delage, E., Ye, Y.: Robustifying conditional portfolio decisions via optimal transport (2021). URL <http://arxiv.org/abs/2103.16451>
- [108] Ordoudis, C., Nguyen, V.A., Kuhn, D., Pinson, P.: Energy and reserve dispatch with distributionally robust joint chance constraints. *Operations Research Letters* **49**(3), 291–299 (2021)
- [109] Panaretos, V.M., Zemel, Y.: An Invitation to Statistics in Wasserstein Space. SpringerBriefs in Probability and Mathematical Statistics. Springer International Publishing, Cham (2020)
- [110] Pando, V., San-José, L.A., García-Laguna, J., Sicilia, J.: A newsboy problem with an emergency order under a general backorder rate function. *Omega* **41**(6), 1020–1028 (2013)
- [111] Pando, V., San-José, L.A., García-Laguna, J., Sicilia, J.: Some general properties for the newsboy problem with an extraordinary order. *TOP* **22**(2), 674–693 (2014)

- [112] Pang Ho, C., Hanasusanto, G.A.: On data-driven prescriptive analytics with side information: A regularized Nadaraya-Watson approach (2019). URL [http://www.optimization-online.org/DB\\_HTML/2019/01/7043.html](http://www.optimization-online.org/DB_HTML/2019/01/7043.html)
- [113] Peña-Ordieres, A., Molzahn, D.K., Roald, L.A., Wächter, A.: DC optimal power flow with joint chance constraints. *IEEE Transactions on Power Systems* **36**(1), 147–158 (2020)
- [114] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E.: Scikit-learn: Machine learning in python. *Journal of Machine Learning Research* **12**, 2825–2830 (2011)
- [115] Pichler, A., Xu, H.: Quantitative stability analysis for minimax distributionally robust risk optimization. *Mathematical Programming* **191**(1), 47–77 (2022)
- [116] Rahimian, H., Bayraksan, G., Homem-de Mello, T.: Identifying effective scenarios in distributionally robust stochastic programs with total variation distance. *Mathematical Programming* **173**(1-2), 393–430 (2019)
- [117] Rahimian, H., Mehrotra, S.: Distributionally Robust Optimization: A Review (2019). URL <http://arxiv.org/abs/1908.05659>
- [118] Rockafellar, R.T., Uryasev, S.: Optimization of conditional value-at-risk. *Journal of Risk* **2**, 21–41 (2000)
- [119] Ruiz, C., Conejo, A., García-Bertrand, R.: Some analytical results pertaining to Cournot models for short-term electricity markets. *Electric Power Systems Research* **78**(10), 1672–1678 (2008)
- [120] Santambrogio, F.: Optimal transport for applied mathematicians, vol. 55. Springer (2015)
- [121] Sen, S., Deng, Y.: Learning enabled optimization: Towards a fusion of statistical learning and stochastic programming (2018). URL [http://www.optimization-online.org/DB\\_HTML/2017/03/5904.html](http://www.optimization-online.org/DB_HTML/2017/03/5904.html)
- [122] Shafieezadeh-Abadeh, S., Kuhn, D., Esfahani, P.M.: Regularization via mass transportation. *Journal of Machine Learning Research* **20**(103), 1–68 (2019)
- [123] Shapiro, A.: On duality theory of conic linear problems. In: *Semi-infinite programming*, pp. 135–165. Springer (2001)
- [124] Shapiro, A.: Distributionally robust stochastic programming. *SIAM Journal on Optimization* **27**(4), 2258–2275 (2017)

- [125] Shapiro, A., Dentcheva, D., Ruszczyński, A.: Lectures on stochastic programming: modeling and theory. SIAM (2021)
- [126] Silvapulle, M.J., Sen, P.K.: Constrained Statistical Inference: Inequality, Order, and Shape Restrictions. Wiley Series in Probability and Statistics. John Wiley & Sons, Inc., Hoboken, NJ, USA (2011)
- [127] Sion, M.: On general minimax theorems. Pacific Journal of Mathematics (1958). DOI 1103040253
- [128] Stott, B., Jardim, J., Alsac, O.: DC power flow revisited. IEEE Transactions on Power Systems **24**(3), 1290–1300 (2009)
- [129] Sun, H., Xu, H.: Convergence analysis for distributionally robust optimization and equilibrium problems. Mathematics of Operations Research **41**(2), 377–401 (2016)
- [130] Sutter, T., Van Parys, B.P.G., Kuhn, D.: A General Framework for Optimal Data-Driven Optimization (2020). URL <http://arxiv.org/abs/2010.06606>
- [131] Van Parys, B.P., Esfahani, P.M., Kuhn, D.: From data to decisions: Distributionally robust optimization is optimal. Management Science **67**(6), 3387–3402 (2021)
- [132] Villani, C.: Topics in Optimal Transportation, *Graduate Studies in Mathematics*, vol. 58. American Mathematical Society, Providence, Rhode Island (2003)
- [133] Villani, C.: Optimal Transport: Old and New. Grundlehren der mathematischen Wissenschaften. Springer Berlin Heidelberg (2008)
- [134] Vrakopoulou, M., Margellos, K., Lygeros, J., Andersson, G.: A probabilistic framework for reserve scheduling and  $n - 1$  security assessment of systems with high wind power penetration. IEEE Transactions on Power Systems **28**(4), 3885–3896 (2013)
- [135] Wang, Z., Glynn, P.W., Ye, Y.: Likelihood robust optimization for data-driven problems. Computational Management Science **13**(2), 241–261 (2016)
- [136] Xie, W.: On distributionally robust chance constrained programs with Wasserstein distance. Mathematical Programming **186**(1-2), 115–155 (2021)
- [137] Xie, W., Ahmed, S.: Distributionally robust chance constrained optimal power flow with renewables: A conic reformulation. IEEE Transactions on Power Systems **33**(2), 1860–1867 (2018)

- [138] Xie, W., Ahmed, S.: Bicriteria approximation of chance-constrained covering problems. *Operations Research* **68**(2), 516–533 (2020)
- [139] Xin, L., Goldberg, D.A.: Time (in)consistency of multistage distributionally robust inventory models with moment constraints. *European Journal of Operational Research* **289**(3), 1127–1141 (2021)
- [140] Xu, H., Meng, F.: Convergence analysis of sample average approximation methods for a class of stochastic mathematical programs with equality constraints. *Mathematics of Operations Research* **32**(3), 648–668 (2007)
- [141] Zhang, H., Li, P.: Chance constrained programming for optimal power flow under uncertainty. *IEEE Transactions on Power Systems* **26**(4), 2417–2424 (2011)
- [142] Zhao, C., Guan, Y.: Data-Driven Risk-Averse Two-Stage Stochastic Program with  $\zeta$ -Structure Probability Metrics (2015). URL [http://www.optimization-online.org/DB\\_FILE/2015/07/5014.pdf](http://www.optimization-online.org/DB_FILE/2015/07/5014.pdf)
- [143] Zhen, J., Kuhn, D., Wiesemann, W.: Mathematical foundations of robust and distributionally robust optimization (2021). URL <http://arxiv.org/abs/2105.00760>
- [144] Zhou, A., Yang, M., Wang, M., Zhang, Y.: A linear programming approximation of distributionally robust chance-constrained dispatch with Wasserstein distance. *IEEE Transactions on Power Systems* **35**(5), 3366–3377 (2020)
- [145] Zymler, S., Kuhn, D., Rustem, B.: Distributionally robust joint chance constraints with second-order moment information. *Mathematical Programming* **137**(1-2), 167–198 (2013)