# Models and Algorithms for Safeguarded Data-Driven Decision Making

Aras Selvi

Imperial Business School, Department of Analytics and Operations

Imperial College London

Submitted in partial fulfillment of the requirements of the degree of Doctor of Philosophy at Imperial College London.

| | |
|---|---|
| Doctoral Advisor: | Wolfram Wiesemann, *Imperial College London* |
| Internal Examiner: | Fanyin Zheng, *Imperial College London* |
| External Examiner: | Coralia Cartis, *University of Oxford* |

July 2025

*To my family, Veronik, Artür, Aren*

# Statement of Originality

I certify that this thesis and the research to which it refers are the product of my own work. Any ideas, data, or quotations from the work of others, published or unpublished, are fully acknowledged in accordance with the standard referencing practices of the discipline. Each section of the thesis clearly indicates where the corresponding work has been published or submitted, and lists the names of my co-authors where applicable.

Aras Selvi
July 2025, London, UK

# Copyright Declaration

# Abstract

This thesis focuses on making *smart* decisions. Early works often equated smartness with the ability to do more with data (*e.g.*, more accurate predictions, more profitable prescriptions), leading to ever progressing algorithms without *safeguards*. Safeguards are vital, however, to enforce fairness/ethics, interpretability, privacy, and robustness of decisions. A decision can only be truly smart if it is optimized while adhering to these values. With this perspective, this thesis focuses on the following key themes:

*(i)* **Developing safeguarded data-driven decision making models.** I formulate safeguarded data-driven decision making models to address concerns such as privacy and robustness.

*(ii)* **Deriving tractable algorithms to solve safeguarded problems.** These safeguards often make the underlying computational tasks intractable; therefore, I derive efficient approximation schemes with rigorous performance guarantees for the optimization problems that arise in such settings.

*(iii)* **Analyzing safeguarding within multi-stage systems.** Data-driven decision making involves multiple stages, typically starting with estimating the unseen truth from data and ending with optimizing decisions over the perceived reality. I aim to understand at which stage these safeguards should be imposed for the best outcomes.

The structure of this thesis is as follows.

The first chapter, Chapter I, summarizes my work on optimal ethical decision making. Broadly speaking, this chapter focuses on settings where decision makers face trade-offs between ethical constraints and the optimality of decisions with respect to a given objective. One example, drawn from the domain of privacy as an ethical constraint, is as follows. Modifying a data-driven decision making algorithm to ensure that its outcomes, if publicly released, do not

compromise the privacy of individuals whose data is used introduces a privacy–utility trade-off: the utility gained from data typically decreases relative to a setting with no privacy considerations. As a result, some practical approaches gradually relax privacy requirements to improve decision quality. In contrast, my work focuses on optimizing decisions subject to a prespecified level of privacy. I aim to maximize the utility obtained from data while strictly adhering to a prespecified definition and level of privacy.

Chapter II focuses on another class of safeguards, namely robustness, within the field of machine learning. In this context, I study how to mitigate the optimizer's curse: the degradation of performance in future deployment of machine learning models due to overfitting to historical data. I develop distributionally robust learning methods that account for such statistical errors, with an emphasis on real-world applicability. In particular, I address settings that are common in practice but often overlooked in the literature, such as the presence of mixed (continuous and categorical) features in our datasets and adversarial attacks that may occur in the future, during model deployment. Accordingly, this chapter presents my contributions to robust machine learning under several practical extensions.

Chapter III summarizes my work on fundamental classes of nonconvex problems in robust and stochastic optimization, which often arise as reformulations or subproblems within safeguarded decision making. In several settings throughout this thesis, the resulting optimization problems are NP-hard. I address these challenges by developing efficient approximation algorithms with provable guarantees, leveraging tools from stochastic and robust optimization such as decision rules and convexification techniques. This chapter presents my contributions to this direction, including tractable methods for convex maximization and reformulation techniques that enable stronger relaxations in practice.

The final chapter, Chapter IV, outlines directions that I believe warrant further investigation, including several that I am already pursuing or have initiated collaborations on.

Chapters I–III each consist of two research papers that I submitted to peer-reviewed conferences and journals during my PhD. Summaries of these publications, along with the list of co-authors, are provided at the beginning of each chapter. Each section contains numerical subsections corresponding to the main content, and alphabetical subsections corresponding to appendices, which include proofs and details of numerical experiments. The numbering of environments such as theorems and equations is continuous throughout the thesis and does not reset within sections.

# Acknowledgements

When I started writing my acknowledgments, I intended them to be short and to the point. But as I began reflecting on the many people who shaped my path over these years, the words kept coming. Perhaps it is because I speak a lot in real life, or perhaps because, as a first-generation student, I have relied on the love, encouragement, and support of many individuals in pursuing my PhD. This journey would not have been possible without them, and I want to take this space to express my gratitude. The completion of this PhD is not mine alone; it belongs to all those who walked alongside me.

First and foremost, I want to express my deepest gratitude to Wolfram Wiesemann for being an exceptional advisor. When I first joined Imperial, I had a meeting with my academic grandfather, Berç Rustem, who had advised Wolfram. Berç's only advice to me was: "In the next several years, one of the best things you can do is to understand and adopt the way Wolfram thinks, in his impressively unique way". Following this advice became the most rewarding part of my PhD. The way Wolfram approaches problems, both in academia and beyond, is remarkably simple, free of unnecessary complications, and yet consistently intelligent and elegant. I know that in the years ahead, whenever I face a difficult decision, asking myself "how would Wolfram think?" will be invaluable. Beyond his intellectual traits, I have learned from Wolfram many essential miscellaneous lessons: how to judge whether a brilliant idea merely improves a field by 'epsilon', how to remain calm in frustrating moments and focus on solutions rather than mourning with frustration, and how to treat every individual with empathy and fairness. Wolfram has contributed more than anyone else to my professional and intellectual growth. I will always remember and cherish our Skype (RIP) meetings, walks outside for research meetings in fresh air, and cold-calls whenever one of us had an exciting idea.

The research group Wolfram built has been a family to me. Zhengchao Wang and I worked side by side throughout the PhD program, and he has been my best friend, with a shared dream of buying our local pub, the Queen's Arms, and hundreds of joint bets, many of which I lost.

Esma Koca, Clint Chin Pang Ho, Zhi Chen, and Man-Chung Yue have been older siblings, always looking out for me. I had the chance to work with the brilliant Huikang Liu, whose mathematical proofs and sleeping schedule never ceased to amaze me. I am glad to leave this group in the presence of such a great set of new members: Reza Belbasi, Lingjun Meng, Bowen Li, and Fupeng Sun. I am also excited to be exploring new directions with a great set of co-authors I met through Wolfram: Çağıl Koçyiğit, Bill Tang, and Phebe Vayanos.

I want to extend my heartfelt gratitude to Coralia Cartis and Fanyin Zheng for serving as examiners of my thesis–two academics whom I deeply respect.

I want to thank those who mentored me throughout the years. I did a research master's degree in Operations Research at Tilburg University under the supervision of Dick den Hertog. Dick is the person who introduced me to the world of robust optimization and, more broadly, the one who taught me research-level optimization. At the time of writing, Dick is a professor at the University of Amsterdam, and the building that housed Tilburg's Operations Research department has been demolished. Yet the things Dick taught me in that building will always stay with me. I am lucky to be still mentored by and collaborating with him. Thanks to Dick, I met *the* Aharon (Ronny) Ben-Tal, whom I am proud to call a mentor and collaborator. I learn a new concept in optimization from him nearly every day, and his passion and love for research and for solving difficult optimization problems have been truly inspirational. Visiting Ronny, working with him side by side, and pausing research for burgers at the nearest mall have been a remarkable part of my PhD. I am grateful to Juan Vera for playing a pivotal role in the beginning of my research career. I am also grateful to my undergraduate mentors, Burcu Balçık, Enis Kayış, Erhun Kundakcıoğlu, and İhsan Yanıkoğlu, for convincing me to take a PhD journey.

At Imperial Business School, I had a great cohort with whom we survived an extensive set of coursework, the COVID-19 pandemic, and the academic job market. I relied on the friendship of Louiza Bartzoka, Barbara Duffek, Narges Mohammadi, Win Monroe, Hanzhao Wang, and Zhengchao Wang. My PhD journey was enriched by friends from other cohorts, including Beste Akbaş, David Licher, Gülce Opuz, Yanwei Sun, Moritz Wiedemann, and Niloofar Zamani Foroushani. At the School, we were also fortunate to be visited by Martina Luzzi, Haolin Ruan, Zhiyu Zeng, and Haiyuan Zhu, from whom I learned many things, ranging from how to make Calabrian-style 'nduja pasta to how to model Markov decision processes in C++. I also had the chance to learn a lot from faculty members at the School, notably Gah-Yi Ban, Martin Haugh, Xiaocheng Li, and Heikki Peura.

I want to thank the PhD program committee, who were the behind-the-scenes heroes of my

PhD. There was not a single problem that could not be solved under the aegis of Nicola Hann and Alex Michaelides. I am grateful to Laura Barkaway, Owen Howells Ottin, Pamela Monteith, and Kathleen Ellis for making every administrative step as smooth as it could possibly be.

I am thankful to the Computational Optimisation Group (COG) for welcoming me as an affiliate, despite being an outsider to the Department of Computing. COG is home to many brilliant minds, and working alongside them was truly humbling. In particular, thanks to COG, I met Daniel Lengyel, known for being the fastest Python coder, with an impressively wide knowledge of mathematics. I have been lucky enough to be flatmates with him and Kathinka Kriens Nakstad, who together formed what we fondly called the North Acton Family. Thank you, Daniel and Kathinka, for patiently listening to all my stories, and apologies that I was never able to hear my alarms in the mornings. Thanks to COG, I met Anastasia Borovykh, whose friendship I deeply value, and who is now conquering the world of LLMs after mastering partial differential equations. I have been lucky to call Calvin Tsay my most productive friend, Alexander Thebelt my most humorous friend, Onur Demiray a lifelong friend and upcoming collaborator, and Jose Pablo Folch my go-to person for festivals. I am grateful to Ruth Misener and Berç Rustem for always being available to share their wisdom with me.

Living in London allowed me to interact, work closely with, and be friends with many brilliant PhD researchers. I would like to thank Cem Aydın for a deep friendship and for the many hours of heated discussions on abstract concepts we both care about. Staying in London will be all the more meaningful because we will continue to share those moments together. I am proud to be friends with Alp Süngü, who writes striking papers and is just as engaging in conversation as he is on the page. I cherish my friendship with Arturo Estrada Rodriguez, Naireet Ghosh, Sıdıka Tunç Candoğan, and Baizhi Song. Among the academics I met in London, I would like to thank Mirko Kremer, Sven Mikolon, and Yuri Mishina for all the beers and schnapps we shared, and I look forward to continuing the tradition. I am grateful to the 'academeeks', Ryan Cory-Wright and Jean Pauphilet, for both our culinary trips and the endless advice I received from them. During my time as a guest teacher in the LSE Mathematics Department, I enjoyed working with and learning from Gregory Sorkin and Neil Olver. I thank them both for five great years of affiliation with LSE. Outside academia, I always felt the support of my dear friends in London: Ekin Alkan, Begüm Aydın, Florian Baronyai, Melis Birgen, Defne Calbay, Emre Göynügür, Victoria Graham, Anna Konovalenko, Milena Lefèvre, Lockie Michalski, Zoltan Nagy, Melis Okur, and Oleksandr Vertyporokh.

I thank the Alan Turing Institute for awarding me a 9-month fellowship, during which I met brilliant PhD students from across the UK, including Berker Banar, Jacob Bradley, Mauricio Figueroa, Anna George, and Vincent Jeanselme. Thanks to Turing, I met the online learning guru Stephen Pasteris, whom I am excited to be collaborating with. During my internship at JP Morgan AI Research, I was lucky enough to work in the same group with Mohsen Ghassemi, Eleonora Kreacic, and Vamsi Potluru, as well as be colleagues and friends with Ben Batten, Mattia Villani, Tom Walshe, and Sarah Zheng.

During my academic journey, I had the opportunity to meet, and even be friends with, some of my academic heroes. Knowing Amine Bennouna is itself a great achievement of my PhD. Not only is Amine extremely smart and working on an exciting research agenda, but he also shows a great level of academic maturity, is incredibly kind, and is a wonderful person to exchange ideas with. Amine, thank you for being by my side on my hardest days. I was lucky enough to work closely with Daniel Kuhn's research group at EPFL. I thank Daniel for being a great academic co-grandfather, and his student, Mengmeng Li, for being an excellent partner in organizing joint seminars and minisymposia. Meeting Anthony Karahalios and being exposed to his infectious energy and positivity has been a remarkable experience. Sina Aghaei, Liviu Aolaritei, Kyle Hunt, Laixi Shi, and Luaho Zhang, thank you all for making international conferences so fun.

In addition to my generous PhD funding from Imperial Business School, I would like to sincerely thank the Sargent Centre, G-Research, the NeurIPS Foundation, SIAM, and INFORMS for their support of my research through travel grants and additional funding over the years.

Perhaps the most unusual and yet rewarding experience was being on the academic job market. It is one of the few things in life that is nearly impossible to navigate without strong support from people both inside and outside academia, regardless of their seniority or focus area. I am grateful to my 'job market consulting team'. To Nazlı Sönmez, for following my job market updates live and offering thoughtful advice on both my applications and the best restaurants in any part of the world. To Lily Xu, for adopting me at the start of the process and staying on as a friend. To the Uzunoğulları family, Merve, Mert, and Rico, for hosting me in Seattle and for energizing me during the INFORMS Annual Meeting 2024, as well as for being wonderful friends over many years. To Shuvomoy Das Gupta and Amine Bennouna, for the hours-long phone calls regardless of time zones. To Canberk Üçel, for the constant positivity throughout the job market. To Barbara Duffek and Laci Duffek, for offering the wisest suggestions. To Ryan Cory-Wright, for being the fastest thinker I know and for always generously sharing his thought process whenever I had a question. To Büşra Karamancı, for matching my nocturnal hours and providing support from far away. To Ruslan Momot, for always giving the most direct and constructive advice. To Erman Acar, for supporting my academic growth since my time in the Netherlands. And to Yufei Qiu, for patiently providing feedback on each of my application materials. I have been humbled to be in the same job market with Bo Lin, Jerry Anunrojwong, and Akshit Kumar, who I am now happy to call friends. Finally, I am grateful to my recommenders, Wolfram Wiesemann, Dick den Hertog, Aharon Ben-Tal, as well as to Kamalini Ramdas and Peter Allen who wrote letters to recommend my teaching.

I am deeply honored to be joining the UCL School of Management as an Assistant Professor in the Operations and Technology group–a group that I have long respected and whose work I have deeply admired. Thank you all for accepting me as a colleague, and making me already feel at home at UCL: Yi-Chun Akchen, Kenan Arifoğlu, Bilal Gökpınar, Rouba Ibrahim, Ersin Körpeoğlu, Yuri Resende Fonseca, Siddharth Prakash Singh, Onesun Steve Yoo, Dongyuan Zhan, and Zhenyu Zhang. I am very much looking forward to all the work we will do together at UCL. I am also thankful to Bartolomeo Stellato for hosting me for a postdoctoral research position in the ORFE department at Princeton University. I have been a fan of Bartolomeo's work for a long time and have always looked forward to discussing my ideas with him at conferences. I am excited for the opportunity to work under his supervision in the coming year.

Throughout my PhD, I always leaned on my friends from home. I am grateful to Roy Mıgırdiçoğlu and Zorbey Seymen for being my oldest friends and for always being there for me. I also feel incredibly lucky to have great friends in Istanbul: Cem Bozkır, Mert Avşar, Göktuğ Gür, Mustafa Kurdoğlu, Muzaffer Turgut, Dikris Bahadır, Baturay Ofluoğlu, Onurcan Aktürk, Süheyl Reyhani, and Ilayda Türker–thank you all for always finding the time and space for our friendship despite the long distances.

Most importantly, I owe everything to my family, and I dedicate this thesis to them. They are the reason I have always had a sense of home, no matter where I am. I have two homes. The first is London, where I have my job, community, and friends. The second is my home in Istanbul, where my family is, and where I have always received unconditional love and support. Without the latter, it would have been impossible to build the former. I am grateful to my parents, Veronik and Artür, for their endless support and for investing as much effort in my PhD as I did. I am grateful to my brother, Aren, who is the most important person to me, and who makes me proud every day with his brilliant business mind. I thank my grandparents, Jan Selvi and Takuhi Abkar, for always rooting for me. My grandmother, Sona Selvi, who passed away during my PhD, will always be deeply in my thoughts. She was especially proud that I was pursuing a PhD, and I carried that pride with me throughout this journey.

# Contents

# List of Figures

# List of Tables

# Chapter I

# Optimal Ethical Decision Making

During my PhD, I worked on decision making subject to ethical constraints. The research theme is making optimal decisions under pre-specified definitions of privacy (for individuals whose private information appears in the data used for data-driven decision making), fairness (as a result of our decisions, for groups of individuals characterized by distinct sensitive attributes), or interpretability (of the models used to make decisions). While I have ongoing work on all of these themes, this chapter focuses on privacy. In Chapter IV, I discuss my works on fairness and interpretability that follow a similar approach.

The first section of this chapter, Section 1, is based on the following work (Selvi et al. 2025):

**Aras Selvi, Huikang Liu, Wolfram Wiesemann.** (2025). Differential privacy via distributionally robust optimization. *Forthcoming in* **Operations Research**.

- 2024 George Nicholson Student Paper Competition (Second Place)
- 2023 Imperial Best PhD Paper in Operations (First Place)
- 2023 INFORMS Optimization Society Student Paper Prize (Honorable Mention)

The second section, Section 2, is based on the following work:

**Huikang Liu, Aras Selvi, Wolfram Wiesemann.** (2025). Mixtures of Gaussians in approximate differential privacy. *Under Review.*

# 1 Differential Privacy via Distributionally Robust Optimization

## Abstract

In recent years, differential privacy has emerged as the *de facto* standard for sharing statistics of datasets while limiting the disclosure of private information about the involved individuals. This is achieved by randomly perturbing the statistics to be published, which in turn leads to a privacy-accuracy trade-off: larger perturbations provide stronger privacy guarantees, but they result in less accurate statistics that offer lower utility to the recipients. Of particular interest are therefore optimal mechanisms that provide the highest accuracy for a pre-selected level of privacy. To date, work in this area has focused on specifying families of perturbations *a priori* and subsequently proving their asymptotic and/or best-in-class optimality.

In this work, we develop a class of mechanisms that enjoy non-asymptotic and unconditional optimality guarantees. To this end, we formulate the mechanism design problem as an infinite-dimensional distributionally robust optimization problem. We show that the problem affords a strong dual, and we exploit this duality to develop converging hierarchies of finite-dimensional upper and lower bounding problems. Our upper (primal) bounds correspond to implementable perturbations whose suboptimality can be bounded by our lower (dual) bounds. Both bounding problems can be solved within seconds via cutting plane techniques that exploit the inherent problem structure. Our numerical experiments demonstrate that our perturbations can outperform the previously best results from the literature on artificial as well as standard benchmark problems.

## 1.1 Introduction

When organizations collect personal data about individuals, it is their responsibility to protect that data when they share information about it with third parties. Data anonymization, which aims to alter the data so that individuals are no longer identifiable, is often insufficient in that regard as it is prone to, among others, reconstruction attacks (Dinur and Nissim 2003) and de-identification attacks (Sweeney 1997, Heffetz and Ligett 2014). To address this issue, manifold definitions of data privacy have been proposed, including $k$-map (Sweeney 2001, §4.3), $k$-anonymity (Sweeney 2002), $\ell$-diversity (Machanavajjhala et al. 2007) and $\delta$-presence (Nergiz et al. 2007); see also the review of Desfontaines (2020, §2.1). Among those, *differential privacy* (DP), first proposed by Dwork et al. (2006b), has arguably received the most attention among researchers and practitioners.

DP considers databases $D \in \mathcal{D}$, where $\mathcal{D} := \mathbb{U}^n$ denotes the set of databases with $n$ individuals (or rows), each of which stems from a data universe $\mathbb{U}$ that characterizes the admissible attribute vectors (*e.g.*, the possible values of the predictors and the output for a supervised learner). For any $\varepsilon, \delta \geq 0$, a randomized algorithm $\mathcal{A}$ mapping databases $D \in \mathcal{D}$ to random outputs $\omega \in \Omega$ is $(\varepsilon, \delta)$-*differentially private* if

$$\mathbb{P}[\mathcal{A}(D) \in A] \leq e^{\varepsilon} \cdot \mathbb{P}[\mathcal{A}(D') \in A] + \delta \qquad \forall (D, D') \in \mathcal{N},\ \forall A \in \mathcal{F},$$

where $(\Omega, \mathcal{F}, \mathbb{P})$ is a probability space and

$$\mathcal{N} = \{(D, D') \in \mathcal{D} \times \mathcal{D} \ : \ D' = (D_{-k}, d) \text{ for some } k = 1, \ldots, n \text{ and } d \in \mathbb{U}\}$$

denotes the (symmetric) set of neighboring databases $(D, D')$, where $D'$ emerges from $D$ by replacing its $k$-th element with any $d \in \mathbb{U}$ (Dwork et al. 2006a,b). Intuitively, under DP with $\delta = 0$, an adversary cannot confidently estimate any single row of the database $D$ from a single sample $\omega \sim \mathcal{A}(D)$ even if she knows all other rows of $D$ and the implementation of $\mathcal{A}$ (Dwork 2011). In the general case where $\delta > 0$, $(\varepsilon, \delta)$-DP is a sufficient (but not necessary) condition for the aforementioned $(\varepsilon, 0)$-DP to hold with a probability of at least $1 - \delta$ (Dinur and Nissim 2003, Meiser 2018, Canonne et al. 2020). Viewed through a Bayesian lens, $(\varepsilon, \delta)$-DP allows the adversary to update her prior on $D$ by at most an amount that is bounded by a function of $\varepsilon$ and $\delta$ upon seeing a realization of $\mathcal{A}(D)$, see Vadhan (2017, §1.6).

Compared to other notions of privacy, DP enjoys several desirable features. The composition theorem (Dwork and Roth 2014, Theorem 3.16), for example, implies that sharing $k$ different statistics, where statistic $i$ has been generated by a $(\varepsilon_i, \delta_i)$-DP mechanism, $i = 1, \ldots, k$, satisfies $(\sum_{i=1}^{k} \varepsilon_i, \sum_{i=1}^{k} \delta_i)$-DP. Likewise, the post processing property asserts that any analysis derived from the output of a differentially private mechanism remains differentially private with the same privacy guarantees (Dwork and Roth 2014, Proposition 2.1). Due to these and other features, DP has found manifold recent applications in statistics and machine learning (Chaudhuri and Monteleoni 2008, Friedman and Schuster 2010, Chaudhuri et al. 2011, Abadi et al. 2016, Cai and Kou 2019), optimization (Mangasarian 2011, Hsu et al. 2014, Han et al. 2016, Hsu et al. 2016), mechanism design (McSherry and Talwar 2007) and revenue management (Chen et al. 2022, 2023, Lei et al. 2024). DP has also been widely applied in industry, ranging from emoji recommender systems that learn from user behavior (Apple Differential Privacy Team 2017), databases that publish user interactions on Facebook (Messing et al. 2020) and COVID-19 vaccination search insights at Google (Bavadekar et al. 2021) to insights from LinkedIn's Economic Graph (Rogers et al. 2020), the U.S. Broadband Coverage dataset posted by Microsoft (Pereira et al. 2021) and the earnings distribution published by the U.S. Census Bureau (Foote et al. 2019).

In this work, we study algorithms that perturb the output of a scalar *query function* $f : \mathcal{D} \mapsto \mathbb{R}$ so as to guarantee $(\varepsilon, \delta)$-DP. The query function $f$ could be a simple statistical query, such as the average, the median or a quantile of a real-valued attribute across all rows, a count of the rows satisfying a user-specified condition, or it could be part of a machine learning model that is trained on the database. We illustrate our setting with the following motivating example.

**Example 1.** *The popular Kaggle* salary dataset[1] *reports the monthly salaries (in thousands of Indian Rupees; INR) of* $6,704$ *individuals whose features include education level and job title. Consider the query function* $f$ *that computes the average salary of PhD graduates working in research. Across the* $194$ *PhD researchers in the database, the salary varies between* $120k$ *INR and* $190k$ *INR, and the average salary amounts to* $165.65k$ *INR. Returning this average would violate DP, however, since an adversary with knowledge of the salaries of the first* $193$ *PhD researchers could readily compute the salary of the* $194th$ *PhD researcher from the query result.*

The Laplace mechanism (Dwork et al. 2006b) achieves $(\varepsilon, 0)$-DP, also referred to as *pure* differential privacy, by returning $f(D) + \tilde{X}$, where the random variable $\tilde{X}$ follows a zero-mean Laplace distribution with scale parameter $\Delta f / \varepsilon$ and $\Delta f := \sup_{(D,D') \in \mathcal{N}} |f(D) - f(D')|$ denoting the sensitivity of the query $f$.

**Example 1** (cont'd)**.** *Assume that the salaries of PhD researchers vary between* $120k$ *INR and* $190k$ *INR, which are the minimum and maximum PhD researcher salaries recorded in the* salary dataset*. Thus, the sensitivity of the average PhD researcher salary query is* $\Delta f = \frac{190k - 120k}{194}$ *INR* $\approxeq 0.36k$ *INR, which is attained by any two databases* $(D, D') \in \mathcal{N}$ *that differ in the salary of one PhD researcher, with one database recording a salary of* $120k$ *INR and the other reporting a salary of* $190k$ *INR. To achieve* $(1, 0)$*-DP, the Laplace mechanism returns as average salary the sum of* $165.65k$ *INR and the realization of a zero-mean Laplace distribution with scale parameter* $\Delta f / \varepsilon \approxeq 0.36k$ *INR. Assume that this sum amounts to* $165.5k$ *INR. Equipped with her knowledge of the salaries of the first* $193$ *PhD researchers, the parameters* $\Delta f$ *and* $\varepsilon$ *of the Laplace mechanism as well as the query response of* $165.5k$ *INR, the adversary could try to deduce the salary of the* $194th$ *PhD researcher via maximum likelihood estimation. The likelihood of the* $194th$ *PhD researcher's salary being* $s$*, given the above information, is*

[1] URL: https://www.kaggle.com/datasets/mohithsairamreddy/salary-data/data

*$L(s) \approxeq p(165.5k \,|\, [31{,}966.10k+s]/194, 0.36k)$, where $p(\cdot \,|\, \mu, b)$ is the density function of a Laplace distribution with location parameter $\mu$ and scale parameter $b$ and $31{,}966.10k$ INR is the sum of the first $193$ PhD researcher salaries (that are known to the adversary). The maximum likelihood estimator is $s^\star = 140.9k$ INR, which is exactly the value that makes the average across the first $193$ PhD researcher salaries and the unknown last salary equal to the observed query output of $165.5k$ INR. One readily computes that $L(s^\star) \approxeq 1.389$ while $\min\{L(s) : s \in [120k, 190k]\} \approxeq 0.688$. The ratio between the maximum and the minimum likelihood is $1.389/0.688 \approxeq \exp(0.7)$, and DP guarantees that this ratio is bounded from above by $\exp(\varepsilon)$. In other words, even if the adversary knew all but one of the PhD researcher salaries, she could not confidently estimate the unknown salary from a single observation of the query output.*

Pure differential privacy bounds the probability ratio of outputs within any measurable set $A \in \mathcal{F}$, no matter how small the involved probabilities are. This significantly restricts the design of admissible algorithms $\mathcal{A}$, particularly in their tail behavior, and it has spurred research into other DP notions that relax the privacy requirement for unlikely events (Desfontaines and Pejó 2020). The most prominent notion is $(\varepsilon, \delta)$-DP, also known as *approximate* differential privacy. The Gaussian mechanism (Dwork and Roth 2014, Appendix A), for example, achieves $(\varepsilon, \delta)$-DP for any $\varepsilon, \delta \in (0, 1)$ by returning $f(D) + \tilde{Y}$, where the random variable $\tilde{Y}$ follows a zero-mean Gaussian distribution with variance $2\ln(1.25/\delta)(\Delta f/\varepsilon)^2$. In Example 1, the $(1, 0)$-DP Laplace mechanism adds a noise with standard deviation 510.28 INR to the query result, whereas the $(1, 0.2)$-DP Gaussian mechanism—despite satisfying a relaxed notion of DP—increases the standard deviation of the noise term to 689.21 INR.

The Laplace and Gaussian mechanisms are *data independent additive noise mechanisms* as their additive noises $\tilde{X}$ and $\tilde{Y}$ do not depend on the database $D$. In contrast, the noise of a *data dependent mechanism* may depend on the database $D$. One of the earliest data dependent mechanisms is the exponential mechanism (McSherry and Talwar 2007), which achieves $(\varepsilon, 0)$-DP for query functions with nominal outputs. Instead of adding data independent noise to the query output, the exponential mechanism ensures that the output is in the range of nominal outputs by randomly selecting one of finitely many outputs according to some score function. For query functions with real-valued outputs, smooth sensitivity mechanisms (Nissim et al. 2007) achieve $(\varepsilon, \delta)$-DP at a higher accuracy than their data independent counterparts by adding noise whose variance is smaller for databases in neighborhoods with a low variation of the query outputs.

Algorithms with stronger privacy guarantees tend to offer less utility from data (Alvim et al. 2011). This is clearly seen for the Laplace and Gaussian mechanisms, whose variances increase with smaller values of $\varepsilon$ and $\delta$. It is therefore natural to study whether those mechanisms are optimal, that is, whether they minimize a pre-specified loss function among the respective classes of $(\varepsilon, 0)$-DP and $(\varepsilon, \delta)$-DP algorithms. In particular, since the $(1, 0.2)$-DP Gaussian mechanism results in a larger standard deviation than the $(1, 0)$-DP Laplace mechanism in Example 1, we conclude that Gaussian mechanisms are not optimal for the $\ell_2$-loss function in general. The early work on optimal mechanisms has focused on specific query functions (such as count queries), and is reviewed, among others, by Geng and Viswanath (2014, §1) and Sommer (2021, §4.6). Among the first papers that investigate optimal mechanisms for generic query functions is the work of Soria-Comas and Domingo-Ferrer (2013), who show that for large classes of $(\varepsilon, 0)$-DP data independent additive noise mechanisms and loss functions, a necessary optimality condition is that no probability mass in the distribution of the random noise can be moved towards zero without violating the privacy guarantee. The Laplace mechanism violates this condition and is thus not optimal, whereas the condition is satisfied by piecewise constant 'staircase distributions' that move the probability mass of the Laplace distribution closer to zero. Geng and Viswanath (2014) show that for classes of $\ell_1$- and $\ell_2$-loss functions, the Laplace mechanism is asymptotically optimal as $\varepsilon \to 0$, but that it can be significantly suboptimal for larger values of $\varepsilon$. They also propose an $(\varepsilon, 0)$-DP data independent additive noise mechanism based on piecewise constant staircase distributions that is optimal across all $(\varepsilon, 0)$-DP algorithms for a large class of symmetric and increasing loss functions. While determining the optimal staircase distribution generally requires the tuning of a single parameter, closed-form characterizations are provided for the special cases of $\ell_1$- and $\ell_2$-loss functions.

In contrast, the design of optimal mechanisms for $\delta \neq 0$ is much less well understood. For the special case where $\varepsilon = 0$ (also known as additive differential privacy), Geng et al. (2019) show that sampling the additive noise from the product of a uniform and a Bernoulli random variable is optimal for symmetric and increasing loss functions among the class of $(0, \delta)$-DP data independent additive noise mechanisms with decreasing noise distributions. Closed-form characterizations of the optimal distributions are provided for $\ell_p$-loss functions, whereas the general case requires the tuning of a single parameter. Balle and Wang (2018) show that the parameter choice of the aforementioned Gaussian mechanism is suboptimal. They propose the analytic Gaussian mechanism, which is optimal among the family of Gaussian mechanisms, by numerically computing the smallest variance that satisfies $(\varepsilon, \delta)$-DP. For the general class of $(\varepsilon, \delta)$-DP data independent additive noise mechanisms, Geng and Viswanath (2016) show that the suboptimality of uniform and discretized Laplace distributions can be bounded by a multiplicative constant for integer-valued queries under $\ell_1$- and $\ell_2$-loss functions when $\varepsilon \to 0$ and $\delta \to 0$ simultaneously. Tighter suboptimality bounds have been derived by Geng et al. (2020) for real-valued queries when the data independent additive noise is governed by a truncated Laplace distribution. In their analysis, the authors decompose the support of the distribution into a 'body' that achieves $(\varepsilon, 0)$-DP and a 'tail' that breaches privacy but is limited to a probability mass of $\delta$. The resulting mechanisms are $(\varepsilon, \delta)$-DP, and they are asymptotically optimal under $\ell_1$- and $\ell_2$-loss functions when $\varepsilon \to 0$ and $\delta \to 0$ simultaneously. The authors show that the truncated Laplace mechanism outperforms the aforementioned analytic Gaussian mechanism under $\ell_1$- and $\ell_2$-loss functions. To our best knowledge, the truncated Laplace mechanism provides the strongest optimality guarantees among the currently known $(\varepsilon, \delta)$-DP mechanisms for generic queries. In Example 1, the noise added by the $(1, 0.2)$-DP analytic Gaussian mechanism has a standard deviation of 300.96 INR, whereas the truncated Laplace mechanism reduces the standard deviation to 273.48 INR.

The definition of optimality is more involved for data dependent mechanisms since their expected losses vary with the database. Minimizing the worst-case expected loss across all potential databases $D \in \mathcal{D}$ (which the *minimax optimality* criterion attempts to achieve) is overly conservative since it implies that data independent mechanisms remain optimal under mild conditions (Geng and Viswanath 2014). Instead, instance specific optimality criteria have been proposed that compare the expected loss for each database with a lower bound that is tailored to the database. Local minimax optimality (Asi and Duchi 2020a,b), for example, requires that a mechanism's expected loss for any database $D \in \mathcal{D}$ is within a constant factor of the expected loss of any other DP mechanism for at least one database in a neighborhood of $D$. Local minimax optimality recovers the earlier notion of minimax optimality when the neighborhood contains all databases $D \in \mathcal{D}$. Asi and Duchi (2020a,b) show that under the expected $\ell_1$-loss, the inverse sensitivity mechanism first proposed by Johnson and Shmatikov (2013) satisfies local minimax optimality for various query functions, and that it outperforms the Laplace and the smooth Laplace mechanisms under mild conditions.

The aforementioned contributions to the design of optimal $(\varepsilon, \delta)$-DP mechanisms have in common that they limit their attention *a priori* to specific classes of mechanisms (such as Gaussian; standard, truncated or discretized Laplace; staircase; uniform-Bernoulli product or uniform distributions) and subsequently prove either optimality among the mechanisms in their respective classes or asymptotic optimality among larger families of mechanisms as $\varepsilon \to 0$ and $\delta \to 0$ simultaneously. In this work, we propose to formulate and solve the optimal $(\varepsilon, \delta)$-

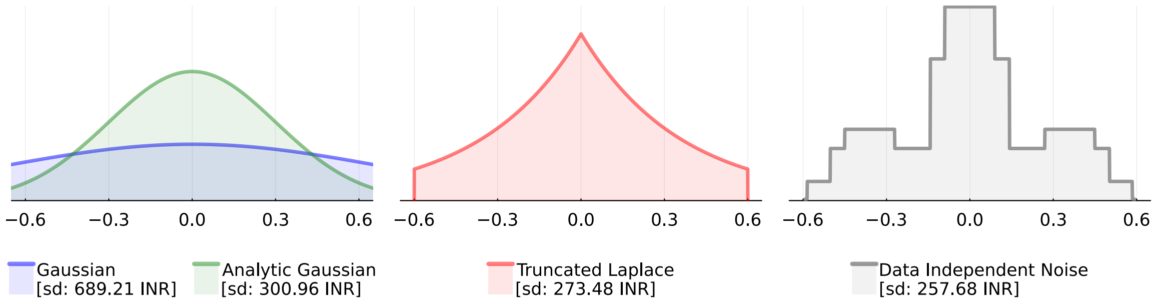


Figure 1: *Different noise distributions that guarantee* $(1, 0.2)$*-DP in Example 1.*

DP mechanism design problem as an infinite-dimensional distributionally robust optimization (DRO) problem (Delage and Ye 2010, Wiesemann et al. 2014, Kuhn et al. 2019). To this end, we minimize an expected loss function over all noise distributions, subject to the satisfaction of the DP constraints. In contrast to much of the existing literature, our formulation caters for generic loss functions (including asymmetric ones such as the pinball loss), and it can restrict the noise via support constraints. We show that our formulation affords a strong dual, and we develop hierarchies of finite-dimensional conservative approximations to the primal and dual formulations to derive converging upper and lower bounds on the optimal expected loss. Our upper bounds correspond to implementable perturbations whose optimality gaps can be certified by the lower bounds. Our bounding problems can be solved efficiently via cutting plane techniques that leverage the inherent problem structure. Our numerical experiments show that our optimal mechanisms can outperform the previously best results from the literature on artificial as well as two standard machine learning benchmark problems.

**Example 1** (cont'd)**.** *To guarantee* $(1, 0.2)$*-DP, our data independent additive noise algorithm from Section 1.2 adds a noise with standard deviation* 257.68 *INR. As Figure 1 demonstrates, our noise distribution does not appear to admit a simple analytical characterization.*

The contributions of this work may be summarized as follows.

*(i)* We formulate the data independent additive noise problem as a DRO problem. Our formulation is flexible enough to cater for a large range of loss functions, and it extends to various problem variants such as the data dependent and the instance optimal problem.

*(ii)* We show that our primal and dual formulations can be bounded from above and below by converging hierarchies of large-scale linear programs that can be solved efficiently via tailored cutting plane techniques.

*(iii)* In contrast to the existing optimality results, which are either restricted to specific mechanisms or hold asymptotically, our formulation affords optimality guarantees that are non-asymptotic and that apply to any choice of $\varepsilon$ and $\delta$. Our numerical results showcase the advantages of our approach on a range of artificial as well as benchmark problems.

In our view, the optimization perspective on DP put forward in this work opens up several opportunities for future research. On one hand, our proposed hierarchy of primal and dual bounds appears to extend to other existing and new definitions of DP, it may allow for the development of approximation schemes for $(\varepsilon, \delta)$-DP with either *a priori* or *a posteriori* optimality guarantees, and it may generalize to multi-dimensional queries. On the other hand, the DP mechanism design problem gives rise to novel classes of DRO problems that have not been studied previously and that may find applications elsewhere. Most importantly, we believe that the design of optimal DP mechanisms should be viewed and addressed as a DRO problem, and the DRO community has developed a rich arsenal of techniques that can be leveraged beneficially to contribute with novel insights and algorithms.

The remainder of this work unfolds as follows. Section 1.2 formulates the data independent additive noise problem as an infinite-dimensional DRO problem, it shows that the problem affords a strong dual, and it develops a hierarchy of converging finite-dimensional upper and lower bounding problems. Building upon this analysis, Section 1.3 studies the data dependent additive noise problem, which optimizes over an uncountable family of noise distributions. Section 1.4 develops a cutting plane algorithm to solve the bounding problems of Sections 1.2 and 1.3. We report numerical experiments in Section 1.5 and offer concluding remarks in Section 1.6, respectively. All proofs as well as some additional numerical results are relegated to the e-companion. Finally, the GitHub repository accompanying this work contains the sourcecodes of all algorithms that we implemented as part of this work, the datasets used in our numerical experiments as well as additional numerical experiments and extended discussions of the related literature.[2]

**Notation.** Bold lower case letters denote vectors, while scalars are assigned standard lower case letters. The sets $\{1, \ldots, N\}$ and $\{-N, \ldots, N\}$ are abbreviated by $[N]$ and $[\pm N]$, respectively. For a set $\mathcal{S}$ and a scalar $a \in \mathbb{R}$, we let $\mathcal{S} + a = \{s + a \ : \ s \in \mathcal{S}\}$ denote the Minkowski sum of the sets $\mathcal{S}$ and $\{a\}$; similarly, $\mathcal{S} - a$ abbreviates $\mathcal{S} + (-a)$. For a measurable interval $I \subseteq \mathbb{R}^n$, $|I| \in \mathbb{R} \cup \{+\infty\}$ denotes the Lebesgue measure of $I$. Unless otherwise stated, measures of real sets are defined on the corresponding Borel $\sigma$-algebras. The sets of sign-unrestricted and

[2] https://github.com/selvi-aras/DP-via-DRO

non-negative measures that are defined on the power-set $\sigma$-algebra of some set $A$ are denoted by $\mathcal{M}(A)$ and $\mathcal{M}_+(A)$, respectively. Finally, $\mathbb{1}[\mathcal{E}]$ is the indicator function taking value 1 if the condition $\mathcal{E}$ is satisfied and 0 otherwise.

## 1.2 Data Independent Noise Optimization

We start with data independent additive noise mechanisms that perturb the query output $f(D)$ of a database $D \in \mathcal{D}$ by adding a random noise $\tilde{X}$ independent of $D$ so as to minimize an expected loss while satisfying $(\varepsilon, \delta)$-DP. We formalize this problem in Section 1.2.1, and Section 1.2.2 derives a converging hierarchy of finite-dimensional upper and lower bounding problems.

### 1.2.1 The Data Independent Optimization Problem

We study the problem

$$\begin{array}{ll} \underset{\gamma}{\text{minimize}} & \displaystyle\int_{x\in\mathbb{R}} c(x)\, \mathrm{d}\gamma(x) \\ \text{subject to} & \gamma \in \mathcal{P}_0 \\ & \displaystyle\int_{x\in\mathbb{R}} \mathbb{1}[f(D)+x \in A]\, \mathrm{d}\gamma(x) \leq e^{\varepsilon} \cdot \int_{x\in\mathbb{R}} \mathbb{1}[f(D')+x \in A]\, \mathrm{d}\gamma(x) + \delta \\ & \hfill \forall (D,D') \in \mathcal{N},\ \forall A \in \mathcal{F}, \end{array} \tag{1}$$

where $(\mathbb{R}, \mathcal{F})$ is a measurable space with the Borel $\sigma$-algebra $\mathcal{F}$ on $\mathbb{R}$ and the set $\mathcal{P}_0$ of probability measures supported on $\mathbb{R}$. Problem (1) selects a probability measure $\gamma$ governing the random noise $\tilde{X}$ so as to minimize the expected value of the Borel loss function $c : \mathbb{R} \mapsto \mathbb{R}_+$, subject to $(\varepsilon, \delta)$-DP for $\varepsilon, \delta > 0$. Note in particular that the integrals on both sides of the DP constraint evaluate the probabilities $\mathbb{P}[\mathcal{A}(D) \in A]$ and $\mathbb{P}[\mathcal{A}(D') \in A]$ for the randomized query outputs $\mathcal{A}(D) = f(D) + \tilde{X}$ and $\mathcal{A}(D') = f(D') + \tilde{X}$ with $\tilde{X} \sim \gamma$ as per our definition of $(\varepsilon, \delta)$-DP from the previous section. We assume that $c$ satisfies the following regularity conditions.

**Assumption 1** (Loss Function)**.** *The loss function $c : \mathbb{R} \mapsto \mathbb{R}_+$ satisfies the following properties:*

*(a)* Continuity. *$c$ is continuous on $\mathbb{R}$.*

*(b)* Unboundedness. *For any $r \in \mathbb{R}$ we have $c(x) \geq r$ for $|x|$ sufficiently large.*

Assumption *(a)* enables us to construct discrete approximations to problem (1) and its dual that converge as we refine their granularity. Assumption *(b)* allows us to restrict these approximations to a bounded support of the involved measures without incurring an unbounded loss. Loss functions typically used in the literature, such as the noise amplitude ($\ell_1$-loss with $c(x) = |x|$) and the noise power ($\ell_2$-loss with $c(x) = x^2$), satisfy Assumption 1.

Recall that $\Delta f := \sup\{f(D') - f(D) \ : \ (D, D') \in \mathcal{N}\}$ is the global sensitivity of the query $f$ over the set of neighboring databases $\mathcal{N}$. We assume that $f$ is surjective in the following sense.

**Assumption 2** (Query Function). *For each $\varphi \in [-\Delta f, \Delta f]$, we have $f(D') - f(D) = \varphi$ for some $(D, D') \in \mathcal{N}$.*

Assumption 2 is standard (Geng and Viswanath 2014, Geng et al. 2020), and it is satisfied by common descriptive statistics including the mean, median, minimum/maximum and standard deviation, as well as several popular machine learning algorithms (*cf.* Section 1.5), if the data is numeric. Our theory continues to apply if Assumption 2 is violated, but our reformulation of problem (1) will be conservative as it guarantees DP over the entire range $\varphi \in [-\Delta f, \Delta f]$, as opposed to the subset of query output differences that can actually be observed over $\mathcal{N}$.

**Observation 1.** *Under Assumption 2, the* data independent noise optimization problem *is*

$$\begin{array}{ll} \underset{\gamma}{\text{minimize}} & \displaystyle\int_{x \in \mathbb{R}} c(x) \, \mathrm{d}\gamma(x) \\ \text{subject to} & \gamma \in \mathcal{P}_0 \\ & \displaystyle\int_{x \in \mathbb{R}} \mathbb{1}[x \in A] \, \mathrm{d}\gamma(x) \leq e^{\varepsilon} \cdot \int_{x \in \mathbb{R}} \mathbb{1}[x + \varphi \in A] \, \mathrm{d}\gamma(x) + \delta \quad \forall (\varphi, A) \in \mathcal{E}, \end{array} \tag{P}$$

*where $\mathcal{E} := [-\Delta f, \Delta f] \times \mathcal{F}$.*

Problem P has uncountably many decision variables and constraints and thus appears to be challenging to solve. Despite the linearity of the problem, standard tools from finite-dimensional linear programming, such as strong duality, are typically not available without further assumptions (Anderson and Nash 1987). Problem P is feasible since its constraints are satisfied, for example, by Laplace (Dwork et al. 2006b) and Gaussian measures (Dwork and Roth 2014, Theorem A.1). Convexity of the feasible region implies that mixtures of such measures are also feasible.

Problem P can be interpreted as an uncertainty quantification problem from the distributionally robust optimization literature (Owhadi et al. 2013, Han et al. 2015, Hanasusanto et al. 2015). Under this view, the constraints of P correspond to an uncountable number of moment conditions that define an ambiguity set from which nature selects a distribution $\gamma$ that minimizes the expected profit of the decision maker's action. Problem P differs from the uncertainty quantification problems typically studied in the literature in both the number and the structure of these moment constraints. The constraints of P can also be interpreted as robust constraints that have to be satisfied for all realizations $(\varphi, A) \in \mathcal{E}$ of the 'uncertain parameters' $\varphi$ and $A$ (Ben-Tal et al. 2009, Bertsimas and den Hertog 2022). In contrast to the standard robust optimization literature, however, the uncertain parameter $A$ in our problem is infinite-dimensional.

Problem P is also reminiscent of continuous linear programs (Anderson and Nash 1987), which comprise uncountably many decision variables and constraints as well. Owing to their continuous-time control heritage, however, the constraints in continuous linear programs are indexed by a single bounded real scalar $x \in [0, T]$, whereas our constraint indices additionally involve the set of all Borel sets $\mathcal{F}$. Moreover, the decision variable of a continuous linear program has a bounded support $[0, T]$ and is assumed to admit a density, whereas our decision variable $\gamma$ has an unbounded support $\mathbb{R}$ and may not admit a density. Both of these additional complications imply that we cannot directly use the theory of continuous linear programming and instead have to derive bounding problems and prove their convergence from first principles.

### 1.2.2 A Hierarchy of Converging Bounding Problems

To obtain a tractable upper bound on problem P, we first introduce a restriction of P that replaces the generic measure $\gamma$ with the piecewise constant function

$$\gamma(A) = \sum_{i \in \mathbb{Z}} p(i) \cdot \frac{|A \cap I_i(\beta)|}{\beta} \quad \forall A \in \mathcal{F}, \tag{2}$$

where $p : \mathbb{Z} \mapsto \mathbb{R}_+$ satisfies $\sum_{i \in \mathbb{Z}} p(i) = 1$, and $\{I_i(\beta)\}_{i \in \mathbb{Z}}$ partitions $\mathbb{R}$ into disjoint intervals $I_i(\beta) := [i \cdot \beta, (i+1) \cdot \beta)$, $i \in \mathbb{Z}$, of some pre-selected length $\beta > 0$. To simplify the exposition, we assume that $\Delta f$ is divisible by $\beta$. Under restriction (2), most constraints in P become redundant.

**Lemma 1.** *Under restriction* (2)*,* P *has the same optimal value as*

$$\begin{array}{ll} \underset{p}{\text{minimize}} & \displaystyle\sum_{i \in \mathbb{Z}} c_i(\beta) \cdot p(i) \\ \text{subject to} & \displaystyle p : \mathbb{Z} \mapsto \mathbb{R}_+, \ \sum_{i \in \mathbb{Z}} p(i) = 1 \\ & \displaystyle\sum_{i \in \mathbb{Z}} \mathbb{1}[I_i(\beta) \subseteq A] \cdot p(i) \leq e^{\varepsilon} \cdot \sum_{i \in \mathbb{Z}} \mathbb{1}[I_i(\beta) + \varphi \subseteq A] \cdot p(i) + \delta \quad \forall (\varphi, A) \in \mathcal{E}(\beta), \end{array} \tag{P($\beta$)}$$

*where* $c_i(\beta) := \beta^{-1} \cdot \int_{x \in I_i(\beta)} c(x) \mathrm{d}x$ *and* $\mathcal{E}(\beta) := \mathscr{B}(\beta) \times \mathcal{F}(\beta)$ *with* $\mathscr{B}(\beta) := \{-\Delta f, -\Delta f + \beta, \ldots, \Delta f\}$ *and* $\mathcal{F}(\beta) := \left\{ \bigcup_{i \in \mathcal{I}} I_i(\beta) \ : \ \mathcal{I} \subseteq \mathbb{Z} \right\}$*.*

In contrast to P, problem P($\beta$) has countably many decision variables. By Cantor's theorem, however, it still comprises uncountably many constraints since $\mathcal{F}(\beta)$ is indexed by the power set of infinitely many intervals $\{I_i(\beta)\}_{i\in\mathbb{Z}}$. To bound $\mathrm{P}(\beta)$ from above by a finite-dimensional linear optimization problem, we constrain $\mathrm{P}(\beta)$ further by restricting the discrete probability measure $p$ to a bounded support. Formally, we impose that there is $L \in \mathbb{N}$ such that

$$p(i) = 0 \quad \forall i \in \mathbb{Z} \setminus [\pm L], \tag{3}$$

that is, we bound the overall support to $2L+1$ intervals $I_i(\beta)$ centered at 0. Restriction (3) allows us to remove from $\mathrm{P}(\beta)$ any privacy constraint that relates to intervals $I_i(\beta)$ and $I_j(\beta)$ whose indices $i$ and $j$ both lie outside the support $[\pm L]$.

**Proposition 1.** *With the additional constraint* (3)*,* $\mathrm{P}(\beta)$ *has the same optimal value as*

$$\begin{array}{ll} \underset{p}{\text{minimize}} & \displaystyle\sum_{i\in[\pm L]} c_i(\beta)\cdot p(i) \\ \text{subject to} & p : [\pm L] \mapsto \mathbb{R}_+, \ \displaystyle\sum_{i\in[\pm L]} p(i) = 1 \\ & \displaystyle\sum_{i\in[\pm L]} \mathbb{1}[I_i(\beta)\subseteq A]\cdot p(i) \leq e^{\varepsilon}\cdot \sum_{i\in[\pm L]} \mathbb{1}[I_i(\beta)+\varphi\subseteq A]\cdot p(i) + \delta \quad \forall(\varphi, A)\in\mathcal{E}(L,\beta), \end{array} \tag{P(L,\beta)}$$

*where* $\mathcal{E}(L,\beta) := \mathscr{B}(\beta)\times\mathcal{F}(L,\beta)$ *with* $\mathcal{F}(L,\beta) := \left\{\bigcup_{i\in\mathcal{L}} I_i(\beta) \ : \ \mathcal{L}\subseteq[\pm L]\right\}$.

Problem $\mathrm{P}(L,\beta)$ constitutes a large-scale but finite-dimensional linear optimization problem that will serve as a building block to our cutting plane algorithm in Section 1.4. We emphasize that the finiteness of the problem is solely due to our measure discretization (2) as well as the restriction (3) to a bounded support, that is, we did not impose any additional assumptions on the structure of the worst-case events $A$ to arrive at a finite-dimensional model. In terms of optimal values, we have the relationship $\mathrm{P}(L,\beta) \geq \mathrm{P}(\beta) \geq \mathrm{P}$ for all support sizes $L\in\mathbb{N}$ and interval lengths $\beta$.

We next derive a lower bound on $\mathrm{P}$. To this end, we employ a strategy widely adopted in distributionally robust optimization and first propose a dual to problem $\mathrm{P}$:

$$\begin{array}{ll} \underset{\theta,\psi}{\text{maximize}} & \theta - \delta\displaystyle\int_{(\varphi,A)\in\mathcal{E}} \mathrm{d}\psi(\varphi,A) \\ \text{subject to} & \theta\in\mathbb{R}, \ \psi\in\mathcal{M}_+(\mathcal{E}) \\ & \theta \leq c(x) + \displaystyle\int_{(\varphi,A)\in\mathcal{E}} \mathbb{1}[x\in A]\mathrm{d}\psi(\varphi,A) - e^{\varepsilon}\cdot\int_{(\varphi,A)\in\mathcal{E}} \mathbb{1}[x+\varphi\in A]\mathrm{d}\psi(\varphi,A) \\ & \hfill \forall x\in\mathbb{R}. \end{array} \tag{D}$$

The integrals in this problem are well-defined due to the domain of $\psi$ specified in the first constraint. Problem D affords a natural interpretation: suppose that $\delta = 0$ and replace in the objective function the epigraphical variable $\theta$ with

$$\inf_{x\in\mathbb{R}}\ c(x) + \int_{(\varphi,A)\in\mathcal{E}} \mathbb{1}[x \in A]\mathrm{d}\psi(\varphi, A) - e^{\varepsilon} \cdot \int_{(\varphi,A)\in\mathcal{E}} \mathbb{1}[x + \varphi \in A]\mathrm{d}\psi(\varphi, A).$$

Problem D then determines a conic combination of database-event pairs $(\varphi, A)$ that maximizes the sum of noise-related costs $c(x)$ and cumulative DP shortfall (*i.e.*, the cumulative violation of all DP constraints) under the most benign realization $x$ of the random noise $\tilde{X}$.

We can readily establish weak duality between the problems P and D.

**Proposition 2** (Weak Duality)**.** *For any $\gamma$ feasible in* P *and $(\theta, \psi)$ feasible in* D*, we have*

$$\int_{x\in\mathbb{R}} c(x)\,\mathrm{d}\gamma(x) \geq \theta - \delta \int_{(\varphi,A)\in\mathcal{E}} \mathrm{d}\psi(\varphi, A).$$

It is tempting to conclude that strong duality should hold as well between P and D due the linearity of both problems. However, strong duality does not typically hold in infinite-dimensional optimization without further assumptions (Anderson and Nash 1987). We therefore defer the discussion of strong duality between P and D to the end of this section.

Similar to problem P, problem D appears challenging to solve since it comprises uncountably many decision variables and constraints. To construct a tractable lower bound on D, we first remove all variables $\psi(\varphi, A)$ indexed by $(\varphi, A) \in \mathcal{E} \setminus \mathcal{E}(\beta)$, that is, we impose that

$$\int_{(\varphi,A)\in\mathcal{E}\setminus\mathcal{E}(\beta)} \mathrm{d}\psi(\varphi, A) = 0. \tag{4}$$

Restriction (4) can be understood as the dual pendant to our discretization (2); in fact, the dual variables unaffected by (4) correspond precisely to the constraints in problem P($\beta$). Under restriction (4), most constraints of D become redundant.

**Lemma 2.** *With the additional constraint* (4)*,* D *has the same optimal value as*

$$\begin{aligned}
\underset{\theta,\psi}{\text{maximize}} \quad & \theta - \delta \cdot \int_{(\varphi,A)\in\mathcal{E}(\beta)} \mathrm{d}\psi(\varphi, A) \\
\text{subject to} \quad & \theta \in \mathbb{R},\ \psi \in \mathcal{M}_+(\mathcal{E}(\beta)) \\
& \theta \leq \underline{c}_i(\beta) + \int_{(\varphi,A)\in\mathcal{E}(\beta)} \mathbb{1}[I_i(\beta) \subseteq A]\mathrm{d}\psi(\varphi, A) - e^{\varepsilon} \cdot \int_{(\varphi,A)\in\mathcal{E}(\beta)} \mathbb{1}[I_i(\beta) + \varphi \subseteq A]\mathrm{d}\psi(\varphi, A) \\
& \hfill \forall i \in \mathbb{Z},
\end{aligned} \tag{D($\beta$)}$$

*where* $\underline{c}_i(\beta) := \inf\{c(x) : x \in I_i(\beta)\},\ i \in \mathbb{Z}$*.*

In contrast to problem D, which comprises uncountably many constraints, problem D($\beta$) has countably many constraints. However, the problem still contains infinitely many constraints as well as uncountably many variables. To bound D($\beta$) from below by a finite-dimensional linear optimization problem, we set $\psi(\mathcal{E}(\beta) \setminus \mathcal{E}(L, \beta)) = 0$ for some $L \in \mathbb{N}$, that is, we impose that

$$\int_{(\varphi,A)\in\mathcal{E}(\beta)\setminus\mathcal{E}(L,\beta)} \mathrm{d}\psi(\varphi, A) = 0. \tag{5}$$

Restriction (5) is the dual pendant to our support constraint (3). It removes variables associated with events that contain intervals sufficiently far away from 0 since $(\varphi, A) \in \mathcal{E}(\beta)\setminus\mathcal{E}(L, \beta)$ implies that $A \not\subseteq \bigcup_{i\in[\pm L]} I_i(\beta)$.

**Proposition 3.** *With the additional constraint* (5)*,* D($\beta$) *has the same optimal value as*

$$\begin{array}{lll} \underset{\theta,\psi}{\text{maximize}} & \theta - \delta \cdot \displaystyle\sum_{(\varphi,A)\in\mathcal{E}(L,\beta)} \psi(\varphi, A) & \\ \text{subject to} & \theta \in \mathbb{R},\ \psi : \mathcal{E}(L, \beta) \mapsto \mathbb{R}_+ & (\mathrm{D}(L, \beta)) \\ & \theta \le \underline{c}_i(\beta) + \displaystyle\sum_{(\varphi,A)\in\mathcal{E}(L,\beta)} \mathbb{1}[I_i(\beta) \subseteq A] \cdot \psi(\varphi, A) - e^{\varepsilon} \cdot \displaystyle\sum_{(\varphi,A)\in\mathcal{E}(L,\beta)} \mathbb{1}[I_i(\beta) + \varphi \subseteq A] \cdot \psi(\varphi, A) & \\ & \qquad \forall i \in [\pm(L + \Delta f/\beta)]. & \end{array}$$

Similar to P($L, \beta$), problem D($L, \beta$) constitutes a large-scale but finite-dimensional linear optimization problem that will serve as a building block to our cutting plane algorithm in Section 1.4. In terms of optimal values, we have the relationship $\mathrm{D}(L, \beta) \le \mathrm{D}(\beta) \le \mathrm{D}$ for all support sizes $L \in \mathbb{N}$ and interval lengths $\beta$. In particular, P and D are sandwiched by the finite-dimensional linear optimization problems P($L, \beta$) and D($L, \beta$).

We close this section with an analysis of the convergence of the finite-dimensional linear optimization problems P($L, \beta$) and D($L, \beta$). To this end, recall that by our earlier assumption, $\beta$ divides $\Delta f$, which allows us to equivalently represent $\beta$ as $\Delta f / k$ for some $k \in \mathbb{N}$.

**Theorem 1.** *For any* $\xi > 0$*, there is* $\Lambda' \in \mathbb{N}$ *and* $k' \in \mathbb{N}$ *such that*

$$\mathrm{P}(\Lambda \cdot k, \Delta f/k) - \mathrm{D}(\Lambda \cdot k, \Delta f/k) \le \xi \qquad \forall \Lambda \ge \Lambda',\ \forall k \ge k'.$$

Intuitively, Theorem 1 states that both the discretization granularity $\Delta f/k$ needs to shrink *and* the support $[-\Lambda \cdot \Delta f, \Lambda \cdot (\Lambda + 1/k) \cdot \Delta f)$ of the noise distribution needs to grow for the

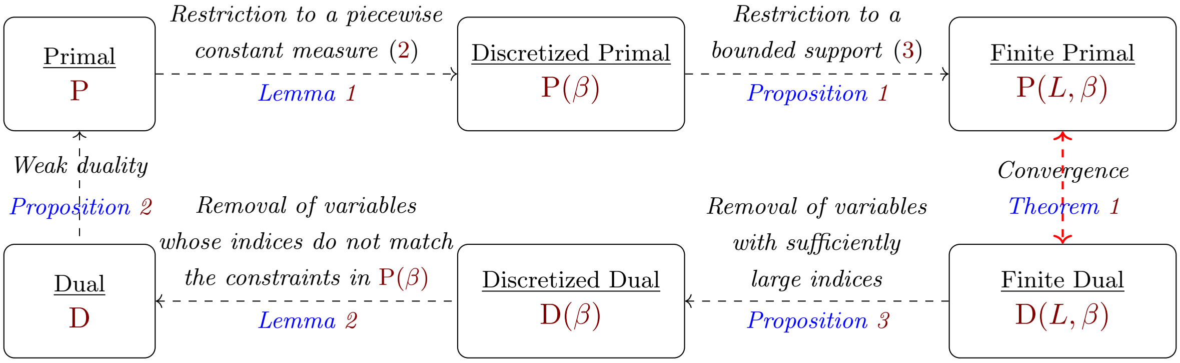

Primal
P
Restriction to a piecewise constant measure (2)
Lemma 1
Discretized Primal
P($\beta$)
Restriction to a bounded support (3)
Proposition 1
Finite Primal
P($L, \beta$)
Weak duality
Proposition 2
Convergence
Theorem 1
Removal of variables whose indices do not match the constraints in P($\beta$)
Lemma 2
Removal of variables with sufficiently large indices
Proposition 3
Dual
D
Discretized Dual
D($\beta$)
Finite Dual
D($L, \beta$)

Figure 2: *Summary of the results in Section 1.2. Directed arrows $x \dashrightarrow y$ indicate upper bound relationships $x \leq y$, whereas the double arrow confirms the convergence of optimal values as L increases and $\beta$ decreases.*

primal and dual approximations to converge. In particular, keeping the support fixed (which amounts to fixing $\Lambda$ in Theorem 1) and merely increasing $k$ is *not* sufficient for convergence as the dual approximation provides a lower bound for *all* noise distributions (of potentially unbounded support), as opposed to only the noise distributions that share the same support as the primal approximation. We elaborate further on this in our GitHub supplement, where we also derive conditions on $\Delta$ and $k$ that ensure feasibility of the upper bound P($L, \beta$). Note that Theorem 1 also implies strong duality of the two infinite-dimensional problems P and D.

We close this section by showing that the conditions of Assumption 1 are in a sense minimal requirements to guarantee the correctness of Theorem 1.

**Proposition 4.** *There are loss functions that satisfy one condition of Assumption 1 (but not both) and for which* P($L, \beta$) *and* D($L, \beta$) *do not converge for any* $L \in \mathbb{N}$ *and* $\beta > 0$.

Figure 2 summarizes the key results of this section.

### 1.3 Data Dependent Noise Optimization

We next study data dependent noise mechanisms whose additive perturbation $\tilde{X}(f(D))$ of the query output $f(D)$ of a database $D \in \mathcal{D}$ may depend on $f(D)$. Section 1.3.1 formalizes this problem, and Section 1.3.2 develops finite-dimensional upper and lower bounding problems.

### 1.3.1 The Data Dependent Optimization Problem

We study the problem

$$\begin{aligned}
\underset{\gamma}{\text{minimize}} \quad & \int_{\phi \in \Phi} w(\phi) \cdot \left[ \int_{x \in \mathbb{R}} c(x) \, \mathrm{d}\gamma(x \mid \phi) \right] \mathrm{d}\phi \\
\text{subject to} \quad & \gamma \in \Gamma \\
& \int_{x \in \mathbb{R}} \mathbb{1}[f(D) + x \in A] \, \mathrm{d}\gamma(x \mid f(D)) \leq e^{\varepsilon} \cdot \int_{x \in \mathbb{R}} \mathbb{1}[f(D') + x \in A] \, \mathrm{d}\gamma(x \mid f(D')) + \delta \\
& \hspace{20em} \forall (D, D') \in \mathcal{N}, \ \forall A \in \mathcal{F},
\end{aligned} \tag{6}$$

where $\Phi := \{f(D) : D \in \mathcal{D}\}$ is the set of possible query outputs and $(\mathbb{R}, \mathcal{F})$ is again a measurable space with the Borel $\sigma$-algebra $\mathcal{F}$ on $\mathbb{R}$. Problem (6) selects a family $\{\gamma(\cdot \mid \phi)\}_{\phi \in \Phi}$ of conditional probability measures governing the random noise $\tilde{X}(\cdot)$ so as to minimize an iterated expectation of the Borel loss function $c : \mathbb{R} \mapsto \mathbb{R}_+$ satisfying Assumption 1, subject to satisfaction of $(\varepsilon, \delta)$-DP. The inner expectation in the objective function evaluates the expected loss for a specific query output $\phi \in \Phi$, whereas the outer expectation weighs different query outputs according to the continuous probability density function $w : \Phi \mapsto \mathbb{R}_+$. The domain of $\gamma$ is now defined as

$$\Gamma := \left\{ \gamma \ : \ \begin{bmatrix} \gamma(\cdot \mid \phi) \in \mathcal{P}_0, & \phi \in \Phi \\ \phi \mapsto \gamma(A \mid \phi) \text{ measurable}, & A \in \mathcal{F} \end{bmatrix} \right\},$$

where $\mathcal{P}_0$ is again the set of probability measures supported on $\mathbb{R}$. The domain $\Gamma$ restricts $\gamma$ to the set of Markov kernels with continuous state $\phi \in \Phi$. The first condition ensures that the inner expectation in the objective function is well-defined, while the second condition ensures that this expectation is measurable in the outer expectation.

To ensure privacy, the weighting $w$ must not reveal anything about the true database (which is ensured, for example, by choosing the uniform distribution over $\Phi$) or it must itself be kept private. In practice, the application domain often guides the choice of $w$; we discuss this further in Section 1.5. Similar to Assumption 2, we impose the following surjectivity requirement on $f$.

**Assumption 3** (Query Function). *$\Phi$ is a bounded interval, and for each $D \in \mathcal{D}$ and $\varphi \in [-\Delta f, \Delta f] \cap (\Phi - f(D))$, we have $f(D') - f(D) = \varphi$ for some $(D, D') \in \mathcal{N}$.*

Assumption 3 reduces to Assumption 2 if we set $\Phi = \mathbb{R}$. To simplify the exposition, however, we assume that $\Phi$ is bounded; otherwise, additional steps would have to be taken to restrict the states $\phi \in \Phi$ of $\gamma$ to bounded intervals without incurring a potentially unbounded loss.

**Observation 2.** *Under Assumption 3, the* data dependent noise optimization problem *is*

$$\begin{array}{ll}\underset{\gamma}{\text{minimize}} & \displaystyle\int_{\phi\in\Phi} w(\phi)\cdot\left[\int_{x\in\mathbb{R}} c(x)\,\mathrm{d}\gamma(x\mid\phi)\right]\mathrm{d}\phi \\ \text{subject to} & \gamma\in\Gamma \\ & \displaystyle\int_{x\in\mathbb{R}} \mathbb{1}[x\in A]\,\mathrm{d}\gamma(x\mid\phi)\le e^{\varepsilon}\cdot\int_{x\in\mathbb{R}}\mathbb{1}[x+\varphi\in A]\,\mathrm{d}\gamma(x\mid\phi+\varphi)+\delta \\ & \qquad\qquad\qquad\qquad\qquad\qquad \forall\phi\in\Phi,\ \forall(\varphi,A)\in\mathcal{E}'(\phi),\end{array}\tag{P$'$}$$

*where* $\mathcal{E}'(\phi):=[[-\Delta f,\Delta f]\cap(\Phi-\phi)]\times\mathcal{F}$ *for* $\phi\in\Phi$.

Problem P′ is not a generalization of problem P from Section 1.2 *per se*, but P would be recovered from P′ if we introduced the additional requirement that all $\gamma(\cdot|\phi)$, $\phi\in\Phi$, in P′ must coincide. This argument implies that problem P′ is guaranteed to be feasible. Note that P′ does not decompose into separate problems for $\phi\in\Phi$ since the DP constraint couples the conditional measures of neighbouring databases. Similar to P, problem P′ contains uncountably many decision variables and constraints and thus appears challenging to solve. The data dependent noise optimization problem P′ is more involved than its data independent counterpart P, however, since it optimizes over an uncountable family of noise distributions. In the following, we will re-use the insights of Section 1.2 to reduce the variables and constraints relating to each individual noise distribution, which allows us to focus on the new challenge of uncountably many noise distributions that is unique to the data dependent setting.

### 1.3.2 A Hierarchy of Converging Bounding Problems

To obtain a tractable upper bound on the data *independent* noise optimization problem, Section 1.2.2 bounds problem P from above by finite-dimensional approximations that live on a partition $\{I_i(\beta)\}_{i\in\mathbb{Z}}$ of the possible noise realizations into disjoint intervals $I_i(\beta)=[i\cdot\beta,(i+1)\cdot\beta)$ of length $\beta>0$. In this section, we retain our earlier assumption that $\Delta f$ is divisible by $\beta$, and we additionally stipulate that $\Phi=\bigcup_{k\in[K]}\Phi_k(\beta)$ with $\Phi_k(\beta):=I_{t+k}(\beta)$ for some $t\in\mathbb{Z}$ and $K\in\mathbb{N}$. This will allow us to partition the set of possible query outputs $\Phi$ in the same way, using a single granularity parameter $\beta$.

To bound problem P′ from above, we first restrict the uncountable family $\{\gamma(\cdot|\phi)\}_{\phi\in\Phi}$ of probability measures in P′ to a finite subset that is piecewise constant on the intervals $\Phi_k(\beta)$:

$$\gamma(\cdot\mid\phi)=\gamma(\cdot\mid\phi')\quad\forall k\in[K],\ \forall\phi,\phi'\in\Phi_k(\beta).\tag{7a}$$

Under restriction (7a), P$'$ optimizes over finitely many probability measures $\gamma_k$, $k \in [K]$, but it still involves uncountably many decision variables and constraints. To further simplify the problem, we restrict each probability measure in P$'$ to a piecewise constant function via

$$\gamma_k(A) = \sum_{i \in \mathbb{Z}} p_k(i) \cdot \frac{|A \cap I_i(\beta)|}{\beta} \quad \forall k \in [K], \ \forall A \in \mathcal{F} \tag{7b}$$

for a family of probability measures $\{p_k : \mathbb{Z} \mapsto \mathbb{R}_+\}_{k \in [K]}$ satisfying $\sum_{i \in \mathbb{Z}} p_k(i) = 1$ for all $k \in [K]$. We also restrict each probability measure $\gamma_k$ to a bounded support by imposing that

$$p_k(i) = 0 \quad \forall k \in [K], \ \forall i \in \mathbb{Z} \setminus [\pm L] \tag{7c}$$

for some $L \in \mathbb{N}$. The restrictions (7b) and (7c) are akin to (2) and (3) from Section 1.2, respectively.

**Proposition 5.** *With the additional constraints* (7), P$'$ *has the same optimal value as*

$$\begin{array}{ll} \underset{p}{\text{minimize}} & \beta \cdot \displaystyle\sum_{k \in [K]} w_k(\beta) \cdot \Big[ \sum_{i \in [\pm L]} c_i(\beta) \cdot p_k(i) \Big] \\ \text{subject to} & p_k : [\pm L] \mapsto \mathbb{R}_+, \ \displaystyle\sum_{i \in [\pm L]} p_k(i) = 1, \ k \in [K] \\ & \displaystyle\sum_{i \in [\pm L]} \mathbb{1}[I_i(\beta) \subseteq A] \cdot p_k(i) \leq e^{\varepsilon} \cdot \sum_{i \in [\pm L]} \mathbb{1}[I_i(\beta) + \varphi \subseteq A] \cdot p_m(i) + \delta \\ & \qquad\qquad\qquad\qquad \forall k, m \in [K], \ \forall (\varphi, A) \in \mathcal{E}'_{km}(L, \beta) \end{array} \tag{P$'(L,\beta)$}$$

*where* $w_k(\beta) := \beta^{-1} \cdot \int_{\phi \in \Phi_k(\beta)} w(\phi) \, \mathrm{d}\phi$ *and* $\mathcal{E}'_{km}(L, \beta) := [\mathscr{B}(\beta) \cap \{(m - k - 1) \cdot \beta, \ (m - k) \cdot \beta, \ (m - k + 1) \cdot \beta\}] \times \mathcal{F}(L, \beta)$ *with* $\mathscr{B}(\beta)$ *and* $\mathcal{F}(L, \beta)$ *defined as in Section* 1.2.

Problem P$'(L, \beta)$ constitutes a large-scale but finite-dimensional linear optimization problem that will serve as a building block to our cutting plane algorithm in Section 1.4. In terms of optimal values, we have the relationship P$'(L, \beta) \geq$ P$'$ for all $L \in \mathbb{N}$ and $\beta > 0$.

To obtain a lower bound on P$'$, we first propose the dual problem

$$\begin{array}{ll} \underset{\theta, \psi}{\text{maximize}} & \displaystyle\int_{\phi \in \Phi} \left[ \theta(\phi) - \delta \cdot \int_{(\varphi, A) \in \mathcal{E}'(\phi)} \mathrm{d}\psi(\varphi, A \mid \phi) \right] \mathrm{d}\phi \\ \text{subject to} & \theta : \Phi \mapsto \mathbb{R} \text{ measurable}, \ \psi \in \Psi \\ & \theta(\phi) \leq \displaystyle\int_{(\varphi, A) \in \mathcal{E}'(\phi)} \mathbb{1}[x \in A] \, \mathrm{d}\psi(\varphi, A \mid \phi) - e^{\varepsilon} \cdot \int_{(-\varphi, A) \in \mathcal{E}'(\phi)} \mathbb{1}[x + \varphi \in A] \, \mathrm{d}\psi(\varphi, A \mid \phi - \varphi) \\ & \qquad\qquad\qquad\qquad + c(x) \cdot w(\phi) \quad \forall \phi \in \Phi, \ \forall x \in \mathbb{R}, \end{array} \tag{D$'$}$$

where the dual measure is defined over the set

$$\Psi := \left\{ \psi \;:\; \begin{bmatrix} \psi(\cdot \mid \phi) \in \mathcal{M}_+(\mathcal{E}'(\phi)),\ \phi \in \Phi \\ \exists \psi_0 \in \mathcal{M}(\mathcal{E}) \;:\; \psi(\cdot \mid \phi) \ll \psi_0,\ \phi \in \Phi \text{ with } (\varphi, A, \phi) \mapsto \dfrac{\mathrm{d}\psi(\varphi, A \mid \phi)}{\mathrm{d}\psi_0(\varphi, A)} \text{ measurable} \end{bmatrix} \right\}.$$

Here, the first condition resembles the domain that we imposed on $\psi$ in problem D from Section 1.2.2. The second condition is new, and it ensures that the integral on the right-hand side of the DP constraint in D$'$, which varies with $\varphi$, is well-defined.

We first establish weak duality between the problems P$'$ and D$'$.

**Proposition 6** (Weak Duality). *For any $\gamma$ feasible in* P$'$ *and $(\theta, \psi)$ feasible in* D$'$*, we have*

$$\int_{\phi \in \Phi} w(\phi) \cdot \left[ \int_{x \in \mathbb{R}} c(x) \, \mathrm{d}\gamma(x \mid \phi) \right] \mathrm{d}\phi \geq \int_{\phi \in \Phi} \left[ \theta(\phi) - \delta \cdot \int_{(\varphi, A) \in \mathcal{E}'(\phi)} \mathrm{d}\psi(\varphi, A \mid \phi) \right] \mathrm{d}\phi.$$

As in the previous section, we defer the discussion of strong duality between P$'$ and D$'$ to the end of this section. To construct a tractable lower bound on D$'$, we impose that

$$\psi(\cdot \mid \phi) = \psi(\cdot \mid \phi') \text{ and } \theta(\phi) = \theta(\phi') \qquad \forall k \in [K],\ \forall \phi, \phi' \in \Phi_k(\beta), \tag{8a}$$

that is, $\psi(\cdot|\phi)$ and $\theta(\phi)$ have to be piecewise constant over the intervals $\Phi_k(\beta)$, and we remove all variables $\psi(\varphi, A|\phi)$ indexed by $(\varphi, A) \in \mathcal{E}'(\phi) \setminus \mathcal{E}(L, \beta)$, that is, we impose that

$$\int_{(\varphi, A) \in \mathcal{E}'(\phi) \setminus \mathcal{E}(L, \beta)} \mathrm{d}\psi(\varphi, A \mid \phi) = 0 \qquad \forall \phi \in \Phi, \tag{8b}$$

which implies that $\psi(\cdot|\phi)$ vanishes for all $\phi \in \Phi$ on $(\varphi, A)$ with $\varphi$ not divisible by $\beta$ or $A$ outside $\mathcal{F}(L, \beta)$. Constraint (8b) is reminiscent of the constraints (4) and (5) from Section 1.2, and it can be interpreted as the dual pendant of the restriction (7c).

**Proposition 7.** *With the additional constraints (8), $\mathrm{D}'$ has the same optimal value as*

$$\begin{array}{ll} \underset{\theta,\psi}{\text{maximize}} & \beta \cdot \left[ \sum_{k \in [K]} \theta_k - \delta \cdot \sum_{k \in [K]} \sum_{(\varphi, A) \in \mathcal{E}'_k(L,\beta)} \psi_k(\varphi, A) \right] \\ \text{subject to} & \boldsymbol{\theta} \in \mathbb{R}^K, \; \psi_k : \mathcal{E}'_k(L,\beta) \mapsto \mathbb{R}_+, \; k \in [K] \\ & \theta_k \leq \sum_{(\varphi,A) \in \mathcal{E}'_k(L,\beta)} \mathbb{1}[I_i(\beta) \subseteq A] \cdot \psi_k(\varphi, A) - e^{\varepsilon} \cdot \sum_{(-\varphi,A) \in \mathcal{E}'_k(L,\beta)} \mathbb{1}[I_i(\beta) + \varphi \subseteq A] \cdot \psi_{k-\varphi/\beta}(\varphi, A) \\ & \qquad + \underline{c}_i(\beta) \cdot \underline{w}_k(\beta) \quad \forall k \in [K], \; \forall i \in [\pm(L + \Delta f/\beta)], \end{array} \tag{$\mathrm{D}'(L,\beta)$}$$

*where* $\mathcal{E}'_k(L,\beta) := [\mathscr{B}(\beta) \cap (\Phi - \underline{\Phi}_k(\beta))] \times \mathcal{F}(L,\beta)$ *for* $\underline{\Phi}_k(\beta) := \inf\{\phi : \phi \in \Phi_k(\beta)\}$, $\underline{c}_i(\beta) := \inf\{c(x) : x \in I_i(\beta)\}$ *and* $\underline{w}_k(\beta) := \inf\{w(\phi) : \phi \in \Phi_k(\beta)\}$, $k \in [K]$ *and* $i \in \mathbb{Z}$.

The large-scale but finite-dimensional linear optimization problem $\mathrm{D}'(L,\beta)$ will serve as a building block to our cutting plane algorithm in Section 1.4. In terms of optimal values, we have the relationship $\mathrm{D}'(L,\beta) \leq \mathrm{D}'$ for all $L \in \mathbb{N}$ and $\beta > 0$. Similar to Section 1.2, $\mathrm{P}'$ and $\mathrm{D}'$ are sandwiched by the finite-dimensional linear optimization problems $\mathrm{P}'(L,\beta)$ and $\mathrm{D}'(L,\beta)$.

We close this section with an analysis of the convergence of the finite-dimensional linear optimization problems $\mathrm{P}'(L,\beta)$ and $\mathrm{D}'(L,\beta)$. Similarly to Section 1.2, recall that by our earlier assumption, $\beta$ divides $\Delta f$, which allows us to equivalently represent $\beta$ as $\Delta f / k$ for some $k \in \mathbb{N}$.

**Theorem 2.** *For any $\xi > 0$, there is $\Lambda' \in \mathbb{N}$ and $k' \in \mathbb{N}$ such that*

$$\mathrm{P}'(\Lambda \cdot k, \Delta f/k) - \mathrm{D}'(\Lambda \cdot k, \Delta f/k) \leq \xi \qquad \forall \Lambda \geq \Lambda', \; \forall k \geq k'.$$

Intuitively, Theorem 2 has a similar interpretation as Theorem 1, that is, both the discretization granularity $\Delta f/k$ needs to shrink *and* the support $[-\Lambda \cdot \Delta f, (\Lambda + 1/k) \cdot \Delta f)$ of the noise distributions needs to grow for the primal and dual approximations to converge. Here, we additionally observe that shrinking the granularity also results in a larger number of distributions to be optimized over. Similar to Theorem 1 in the data independent setting, Theorem 2 implies strong duality of the two infinite-dimensional problems $\mathrm{P}'$ and $\mathrm{D}'$.

## 1.4 Iterative Solution of the Bounding Problems

The bounding problems of Sections 1.2 and 1.3 employ a uniform partitioning of the noise distribution $\gamma$. Motivated by our numerical experiments, which indicate that (near-)optimal noise distributions tend to combine steep peaks around 0 with gradually declining tails, Section 1.4.1

extends our bounding problems to non-uniform partitions of $\gamma$. Non-uniform partitions allow us to compute noise distributions with similar expected losses in shorter computation times.

Unfortunately, even under a non-uniform partitioning the bounding problems cannot be solved monolithically with an off-the-shelf solver due to their exponential scaling in the problem parameters. Instead, Section 1.4.2 proposes a cutting plane technique that solves those bounding problems iteratively through an increasingly accurate sequence of relaxations. At the heart of our cutting plane technique is the identification of the constraints that our incumbent solutions violate with the largest margins. While a naïve search for these constraints would require an exponential effort, our algorithm scales polynomially in the size of the problem description.

### 1.4.1 Upper and Lower Bounding Problems with Non-Uniform Partitions

Recall that our bounding problems $\mathrm{P}(L,\beta)$ and $\mathrm{D}(L,\beta)$ from Section 1.2 partition the support of the noise distribution $\gamma$ into $2L+1$ uniform intervals $I_i(\beta) = [i\cdot\beta, (i+1)\cdot\beta)$, $i \in [\pm L]$. Let $\boldsymbol{\pi} \in \{-L, \ldots, L+1\}^{N+1}$ be an index vector satisfying

$$-L = \pi_1 \;<\; \pi_2 \;<\; \ldots \;<\; \pi_N \;<\; \pi_{N+1} = L+1,$$

and let $\Pi_j(\beta) = [\pi_j \cdot \beta, \pi_{j+1}\cdot\beta)$, $j \in [N]$, denote the $j$-th interval induced by the consecutive elements of $\boldsymbol{\pi}\cdot\beta$. Consider a variant of our upper bounding problem $\mathrm{P}(L,\beta)$ that enforces equality of all decision variables $p(i)$ and $p(i')$, $i,i' \in [\pm L]$, that satisfy $\pi_j \leq i, i' < \pi_{j+1}$ for some $j \in [N]$. The revised upper bounding problem is equivalent to

$$\begin{array}{ll} \underset{p}{\text{minimize}} & \displaystyle\sum_{j\in[N]} c_j(\boldsymbol{\pi},\beta)\cdot p(j) \\ \text{subject to} & p : [N] \mapsto \mathbb{R}_+, \;\; \displaystyle\sum_{j\in[N]} p(j) = 1 \\ & \displaystyle\sum_{j\in[N]} p(j)\cdot\frac{|A\cap\Pi_j(\beta)|}{|\Pi_j(\beta)|} \leq e^{\varepsilon}\cdot\sum_{j\in[N]} p(j)\cdot\frac{|(A-\varphi)\cap\Pi_j(\beta)|}{|\Pi_j(\beta)|} + \delta \quad \forall(\varphi,A)\in\mathcal{E}(L,\beta), \end{array} \tag{P($\boldsymbol{\pi},\beta$)}$$

where $c_j(\boldsymbol{\pi},\beta) := |\Pi_j(\beta)|^{-1}\cdot\int_{x\in\Pi_j(\beta)} c(x)\,\mathrm{d}x$. The new upper bound $\mathrm{P}(\boldsymbol{\pi},\beta)$ closely resembles the previous bound $\mathrm{P}(L,\beta)$, except that the objective coefficients $c_j$ and the intervals $\Pi_j(\beta)$ in the DP constraints now reflect the new partitioning of $\gamma$.

In a similar fashion, we propose the revised lower bound

$$\begin{aligned} \underset{p}{\text{minimize}} \quad & \sum_{j \in \mathfrak{N}} \underline{c}_j(\boldsymbol{\pi}, \beta) \cdot p(j) \\ \text{subject to} \quad & p : \mathfrak{N} \mapsto \mathbb{R}_+, \ \sum_{j \in \mathfrak{N}} p(j) = 1 \\ & \sum_{j \in \mathfrak{N}} p(j) \cdot \frac{|A \cap \Pi_j(\beta)|}{|\Pi_j(\beta)|} \leq e^{\varepsilon} \cdot \sum_{j \in \mathfrak{N}} p(j) \cdot \frac{|(A - \varphi) \cap \Pi_j(\beta)|}{|\Pi_j(\beta)|} + \delta \quad \forall (\varphi, A) \in \mathcal{E}(L, \beta) \end{aligned} \tag{D($\boldsymbol{\pi}, \beta$)}$$

where $\underline{c}_j(\boldsymbol{\pi}, \beta) := \inf\{c(x) : x \in \Pi_j(\beta)\}$ and the index set $\mathfrak{N} = \{-\frac{\Delta f}{\beta} + 1, \ldots, N + \frac{\Delta f}{\beta}\}$ emerges from the previous index set $[N]$ by padding it at both ends with $\Delta f / \beta$ additional elements whose associated interval indices are set to $\pi_{1-t} = \pi_1 - t$ and $\pi_{N+1+t} = \pi_{N+1} + t$, $t = 1, \ldots, \Delta f / \beta$, with the intervals $\Pi_j(\beta)$ extended to $j \in \mathfrak{N} \setminus [N]$ in the obvious way. Again, the revised lower bound closely resembles the previous bound D($L, \beta$), with minor changes in the objective coefficients $\underline{c}_j$ and the intervals $\Pi_j(\beta)$. The revised bounding problems enjoy convergence properties akin to those from Section 1.2.

**Corollary 1.** *For any $\beta > 0$ and $\boldsymbol{\pi}$ satisfying $-L = \pi_1 < \ldots < \pi_{N+1} = L + 1$ for some $L \in \mathbb{N}$, we have* P($\boldsymbol{\pi}, \beta$) $\geq$ P $=$ D $\geq$ D($\boldsymbol{\pi}, \beta$). *Moreover, for any $\xi > 0$ there is $\Lambda' \in \mathbb{N}$ and $k' \in \mathbb{N}$ such that* P($\boldsymbol{\pi}, \beta$) $-$ D($\boldsymbol{\pi}, \beta$) $\leq \xi$ *for any $\boldsymbol{\pi}$ whose induced partition $\{\Pi_j(\beta)\}_{j \in [N]}$ is a refinement of a uniform partition $\{I_i(\Delta f / k)\}_{i \in [\pm \Lambda \cdot k]}$ with $\Lambda \geq \Lambda'$ and $k \geq k'$.*

Appendix 1.C.3 presents analogous bounding problems for the data dependent case.

### 1.4.2 Cutting Plane Algorithm

Although the revised upper bounding problem P($\boldsymbol{\pi}, \beta$) only contains $N$ decision variables, it remains challenging to solve monolithically since it comprises $\mathcal{O}(2^L \cdot \Delta f / \beta)$ DP constraints. To address this issue, Algorithm 1 solves a sequence of relaxations of P($\boldsymbol{\pi}, \beta$) that only involve those constraints that are active at incumbent solutions. In particular, every iteration of Algorithm 1 determines a constraint $(\varphi^\star, A^\star)$ with maximum *privacy shortfall*, which is the quantity that the DP constraints in P($\boldsymbol{\pi}, \beta$) require to be non-positive:

$$V(\varphi, A) = \sum_{j \in [N]} p(j) \cdot \frac{|A \cap \Pi_j(\beta)|}{|\Pi_j(\beta)|} - e^{\varepsilon} \cdot \sum_{j \in [N]} p(j) \cdot \frac{|A \cap (\Pi_j(\beta) + \varphi)|}{|\Pi_j(\beta)|} - \delta \quad \text{for } (\varphi, A) \in \mathcal{E}(L, \beta). \tag{9}$$

```
Algorithm 1: Cutting plane algorithm for problem $\mathrm{P}(\boldsymbol{\pi}, \beta)$
  input  : $\boldsymbol{\pi}$, $\beta$, $\Delta f$
  output: optimal solution $p^\star$ to problem $\mathrm{P}(\boldsymbol{\pi}, \beta)$
  Initialize $\mathcal{S} = \emptyset$;
  do
      Let $p^\star$ be an optimal solution to the relaxation of $\mathrm{P}(\boldsymbol{\pi}, \beta)$ that only contains the
        privacy constraints indexed by $(\varphi, A) \in \mathcal{S}$.
      Find a constraint $(\varphi^\star, A^\star)$ with maximum privacy shortfall under the incumbent
        solution $p^\star$.
      if constraint $(\varphi^\star, A^\star)$ has positive privacy shortfall then update $\mathcal{S} = \mathcal{S} \cup (\varphi^\star, A^\star)$.
  while $\mathcal{S}$ has been updated;
  return $p^\star$.
```

Since $\mathrm{P}(\boldsymbol{\pi}, \beta)$ contains finitely many constraints, one readily recognizes that Algorithm 1 determines an optimal solution to $\mathrm{P}(\boldsymbol{\pi}, \beta)$ in finitely many iterations.

**Observation 3.** *Algorithm 1 terminates after a finite number of iterations with an optimal solution $p^\star$ to problem $\mathrm{P}(\boldsymbol{\pi}, \beta)$.*

A key step in Algorithm 1 is the identification of a constraint $(\varphi^\star, A^\star) \in \mathcal{E}(L, \beta)$ with maximum privacy shortfall. A naïve implementation of this step would require the evaluation of $\mathcal{O}(2^L \cdot \Delta f / \beta)$ privacy shortfalls in time $\mathcal{O}(N)$ each. Instead, we employ Algorithm 2 to identify a constraint $(\varphi^\star, A^\star)$ with maximum privacy shortfall in polynomial time.

**Proposition 8.** *For a fixed solution $p$, Algorithm 2 can be implemented so as to return a constraint of $\mathrm{P}(\boldsymbol{\pi}, \beta)$ with maximum privacy shortfall in time $\mathcal{O}(N^3)$.*

To illustrate the intuition behind Algorithm 2 and Proposition 8, fix any $\varphi \in \mathscr{B}(\beta)$ in problem $\mathrm{P}(\boldsymbol{\pi}, \beta)$. To construct the DP constraint $(\varphi, A) \in \mathcal{E}(L, \beta)$ with maximum privacy shortfall across all $A \in \mathcal{F}(L, \beta)$, we need to decide for each interval $I_i(\beta) = [i \cdot \beta, (i+1) \cdot \beta)$, $i \in [\pm L]$, whether or not to include the interval $I_i(\beta)$ in $A$. To this end, we first observe that we can include all intervals $I_i(\beta)$ for which $I_i(\beta) \cap (\Pi_{j'}(\beta) + \varphi) = \emptyset$ for all $j' \in [N]$ since the inclusion of those intervals in $A$ cannot decrease $V(\varphi, A)$. For the intervals $I_i(\beta)$ that satisfy both $I_i(\beta) \subseteq \Pi_j(\beta)$ and $I_i(\beta) \subseteq (\Pi_{j'}(\beta) + \varphi)$ for some $j, j' \in [N]$, on the other hand, we compare the magnitude of the positive coefficient $p(j)$ with that of the negative coefficient $-e^\varepsilon \cdot p(j')$ in $V(\varphi, A)$ to decide whether $I_i(\beta)$ should be included in $A$. Algorithm 2 and Proposition 8 refine this idea by *(i)* iterating only over the intervals $\Pi_j(\beta)$, $j \in [N]$, as opposed to the larger set of intervals $I_i(\beta)$, $i \in [\pm L]$; *(ii)* identifying the relevant interval pairs $(j, j') \in [N]^2$ in linear time

**Algorithm 2:** *Identification of a constraint in* $\mathrm{P}(\boldsymbol{\pi}, \beta)$ *with maximum privacy shortfall*

```
input  : $\boldsymbol{\pi}$, $\beta$, $p$, $\Delta f$
output: constraint $(\varphi^\star, A^\star)$ with maximum privacy shortfall $V(\varphi^\star, A^\star)$
Initialize $V^\star = 0$;
for $\varphi \in \big\{(\pi_j - \pi_{j'}) \cdot \beta \ : \ (\pi_j - \pi_{j'}) \cdot \beta \in [-\Delta f, \Delta f]$ and $j, j' \in [N]\big\} \cup \{-\Delta f, \Delta f\}$ do
    Initialize $A = \emptyset$ and $V = 0$;
    for $j = 1, \ldots, N$ do
        Let $A_j = \Pi_j(\beta) \setminus [-L \cdot \beta + \varphi, (L+1) \cdot \beta + \varphi)$ and update
            $$A = A \cup A_j, \quad V = V + |A_j| \cdot \frac{p(j)}{|\Pi_j(\beta)|}.$$
        for $j' = 1, \ldots, N$ do
            if $p(j)/|\Pi_j(\beta)| > e^\varepsilon \cdot p(j')/|\Pi_{j'}(\beta)|$ then
                Let $A_{jj'} = \Pi_j(\beta) \cap (\Pi_{j'}(\beta) + \varphi)$ and update
                    $$A = A \cup A_{jj'}, \quad V = V + |A_{jj'}| \cdot \left[\frac{p(j)}{|\Pi_j(\beta)|} - e^\varepsilon \cdot \frac{p(j')}{|\Pi_{j'}(\beta)|}\right].$$
            end
        end
    end
    if $V > V^\star$ then
        Update $\varphi^\star = \varphi$, $A^\star = A$ and $V^\star = V$.
    end
end
return $(\varphi^\star, A^\star)$ and $V^\star(\varphi, A) = V^\star - \delta$.
```

$\mathcal{O}(N)$ as opposed to quadratic time $\mathcal{O}(N^2)$; and *(iii)* restricting the search over $\varphi \in \mathscr{B}(\beta)$ with cardinality $\mathcal{O}(\Delta f/\beta)$ to the smaller set in the outer for-loop with cardinality $\mathcal{O}(N^2)$.

As discussed in Section 1.2.1, the constraints of $\mathrm{P}(\boldsymbol{\pi}, \beta)$ can be interpreted as semi-infinite robust optimization constraints. Under this lens, Algorithm 1 follows the tradition of cutting plane schemes from the robust optimization literature (*cf.* Bienstock and Özbay 2008, Mutapcic and Boyd 2009, Bertsimas et al. 2016a and Pätzold and Schöbel 2020). The key technical contribution of this section is to identify for a fixed decision $p$ the maximally violated constraints $(\varphi^\star, A^\star) \in \mathcal{E}(L, \beta)$ without inspecting exponentially many events $A \in \mathcal{F}(L, \beta)$.

This section focused on the cutting plane algorithm for the upper bounding problem $\mathrm{P}(\boldsymbol{\pi}, \beta)$ in the data independent setting. Algorithms 1 and 2 immediately extend to the lower bounding problem $\mathrm{D}(\boldsymbol{\pi}, \beta)$ if we replace the objective coefficients $c_j$ with $\underline{c}_j$ and extend the domain of the decisions $p$ from $[N]$ to $\mathfrak{N}$. Both algorithms also readily extend to the data dependent setting (*cf.* Section 1.3), where a constraint with maximum privacy shortfall can be identified in time $\mathcal{O}(K^2 \cdot N)$. For the sake of brevity, we relegate the details of that algorithm variant to the GitHub repository accompanying the journal version of this work.

## 1.5 Numerical Experiments

Our numerical experiments are split into two parts. The first part compares the privacy-accuracy trade-off of our optimization-based approach with popular DP mechanisms from the literature, and it examines the runtime of our cutting plane algorithm as well as the convergence of our bounding problems. Since this part focuses on the quality of our bounds as well as their computation times, we use synthetic instances that give us complete control over all parameters. The second part of our experiments investigates whether the improved accuracy on synthetic instances carries over to a better in-sample and out-of-sample performance in machine learning problems involving standard benchmark instances. To this end, we study differentially private variants of the naïve Bayes classifier and a proximal coordinate descent method for logistic regression.

Our optimization algorithms are implemented in C++ and use the GUROBI 9.5.2 LP solver. The machine learning algorithms from the second part are implemented in Julia, and all data is processed using Python 3. All experiments are conducted on Intel Xeon 2.66GHz cluster processors with 16GB memory in single-core and single-thread mode. All sourcecodes and datasets, together with more detailed descriptions of our numerical experiments, are available open-source on the GitHub repository accompanying this work.

### 1.5.1 Synthetic Experiments

In our first experiment, we compare the privacy-accuracy trade-off of our optimization-based DP scheme with that of popular benchmark mechanisms from the literature. To this end, we consider the data independent noise optimization problem and select 100 combinations of $\varepsilon \in [0.005, 5]$ and $\delta \in [0.005, 0.75]$. We intentionally chose conservative combinations of $\varepsilon$ and $\delta$ (*cf.* Table 1 of Zhao et al. (2019)); larger values of $\varepsilon$ yield results that are more favorable to our algorithm. We solve our upper and lower bounding problems $\mathrm{P}(\boldsymbol{\pi}, \beta)$ and $\mathrm{D}(\boldsymbol{\pi}, \beta)$ with $\boldsymbol{\pi}$ and $\beta$ set appropriately so that their relative optimality gaps, measured as $100\% \cdot (\mathrm{P}(\boldsymbol{\pi}, \beta) - \mathrm{D}(\boldsymbol{\pi}, \beta)) / \mathrm{D}(\boldsymbol{\pi}, \beta)$, are strictly less than 1% (with a median gap of 0.28% across our experiments). We then use the midpoint $O := (\mathrm{P}(\boldsymbol{\pi}, \beta) - \mathrm{D}(\boldsymbol{\pi}, \beta))/2$ of both bounds as a substitute to the optimal privacy-accuracy trade-

| | | $\delta$ | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | **0.005** | **0.010** | **0.020** | **0.050** | **0.100** | **0.200** | **0.250** | **0.300** | **0.500** | **0.750** |
| $\varepsilon$ | **0.005** | 1.87% | 1.21% | 1.20% | 6.10% | 18.53% | 3.55% | 49.59% | 23.18% | 49.96% | 33.34% |
| | **0.010** | 2.89% | 1.42% | 8.00% | 2.32% | 17.08% | 3.08% | 49.18% | 23.03% | 49.92% | 33.35% |
| | **0.020** | 2.84% | 2.83% | 0.56% | 15.09% | 14.19% | 67.44% | 48.39% | 22.74% | 49.83% | 33.37% |
| | **0.050** | 1.85% | 2.11% | 2.57% | 18.36% | 6.22% | 66.03% | 46.15% | 21.95% | 49.58% | 33.42% |
| | **0.100** | 4.78% | 4.18% | 8.68% | 19.37% | 33.32% | 63.85% | 42.87% | 20.85% | 49.17% | 33.50% |
| | **0.200** | 10.43% | 9.60% | 8.90% | 17.29% | 19.82% | 60.17% | 37.70% | 19.36% | 48.33% | 33.63% |
| | **0.500** | 23.10% | 21.41% | 25.45% | 16.36% | 38.70% | 42.56% | 29.52% | 18.48% | 45.85% | 33.79% |
| | **1.000** | 40.11% | 39.73% | 40.23% | 40.41% | 28.62% | 32.53% | 26.63% | 21.43% | 41.80% | 33.36% |
| | **2.000** | 33.96% | 34.18% | 33.45% | 31.63% | 32.46% | 28.70% | 27.12% | 25.67% | 34.35% | 30.51% |
| | **5.000** | 19.32% | 19.31% | 19.29% | 19.23% | 19.15% | 19.01% | 18.94% | 18.88% | 18.65% | 19.06% |

Table 1: *Suboptimality of the best performing benchmark mechanisms on synthetic data independent instances with $\Delta f = 1$, $\ell_1$-loss and various combinations of $\varepsilon$ and $\delta$.*

off, and we measure the suboptimality of different benchmark mechanisms from the literature: the analytic Gaussian (Balle and Wang 2018) and the truncated Laplace (Geng et al. 2020) mechanisms as upper bounds and the 'near optimal lower bound' of Geng and Viswanath (2016, Thm 8) and Geng et al. (2020) as lower bound. Table 1 records the optimality gaps of the best performing upper and lower bounds $B_{\text{UB}}$ and $B_{\text{LB}}$ from the literature, which consistently turn out to be the truncated Laplace mechanism and the lower bound of Geng et al. (2020). The optimality gaps are reported as $100\% \cdot [(B_{\text{UB}} - B_{\text{LB}})/\max\{O, 1\}]$. A breakdown into separate suboptimalities incurred by the upper and lower bounds is presented in Appendix 1.D. The table shows that the suboptimality of the benchmark approaches increases with $\varepsilon$ and $\delta$, and the optimality gaps are significant in most of the considered privacy regimes.[3] We note that if we replace $\max\{O, 1\}$ with $O$ in the denominator of the optimality gap formula, then the gaps in Table 1 increase to more than 700% for $(\varepsilon, \delta) = (5, 0.75)$. We observe qualitatively similar results as in Table 1 also for the $\ell_2$-loss; the corresponding table is relegated to the e-companion.

Our second experiment investigates the runtime and the convergence of our optimization-

[3] The vigilant reader will observe that the suboptimality is not entirely monotone in $\varepsilon$ and $\delta$. This is due to two factors: the computation of the 'near optimal lower bound' involves a non-monotonic parameter rounding, and we approximate the optimal privacy-accuracy trade-off with the midpoint $O$ of our upper and lower bounds.

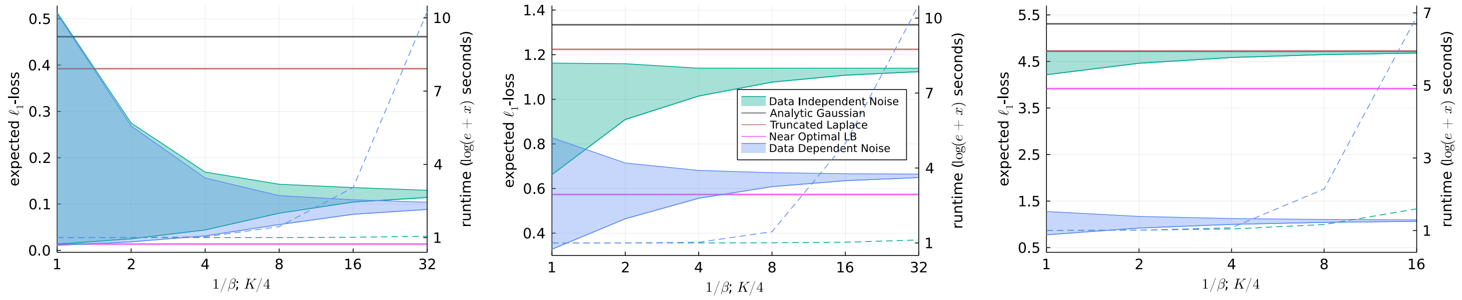


Figure 3: *Comparison of our optimization-based DP schemes with benchmark approaches from the literature on instances with* $\Delta f = 2$*,* $\ell_1$*-loss and* $|\Phi| = 4$ *in low-privacy (*$\varepsilon = 5$ *and* $\delta = 0.25$*; left), medium-privacy (*$\varepsilon = 1$ *and* $\delta = 0.2$*; middle) and high-privacy (*$\varepsilon = 0.2$ *and* $\delta = 0.05$*; right) regimes. All computation times are median values over* 10 *repetitions.*

based upper and lower bounds in the data independent and data dependent settings. To this end, we consider three privacy regimes: a low-privacy setting with $(\varepsilon, \delta) = (5, 0.25)$, a medium-privacy setting with $(\varepsilon, \delta) = (1, 0.2)$, and a high-privacy setting with $(\varepsilon, \delta) = (0.2, 0.05)$. As in the previous experiment, we compare our optimization-based DP schemes with the analytic Gaussian and the truncated Laplace mechanisms as upper bounds and the 'near optimal lower bound' as lower bound. In our DP schemes, we match the support of the truncated Laplace distribution (with the support bounds rounded to nearest integer values) and compute a hierarchy of refined upper and lower bounds by selecting $\beta \in \{1, 1/2, \ldots, 1/32\}$ and—in the data dependent case—$K \in \{4, 8, \ldots, 128\}$. The results are presented in Figure 3. The figure confirms that the truncated Laplace mechanism is asymptotically optimal among all data independent DP schemes in high-privacy settings. However, the figure also reveals that the truncated Laplace mechanism can be substantially outperformed by our optimization-based data independent noise mechanism in low- and medium-privacy regimes, while it is dominated by our optimization-based data dependent noise mechanism in high-privacy regimes. For low-privacy settings, the difference between our data independent and dependent mechanisms is negligible, but it becomes substantial in medium- and high-privacy regimes, where our data dependent mechanisms significantly outperform the data independent ones. The figure also reveals the computational price to be paid for optimal noise distributions. While the data independent problems were all solved within 2.2 secs, the data dependent problems are more challenging: across all instances, it took up to 5.82 secs (980.44 secs) to reduce the gap between our upper and lower bounds to 10% (5%).

Our final synthetic experiment analyzes the shapes of the noise distributions $\gamma$ obtained by

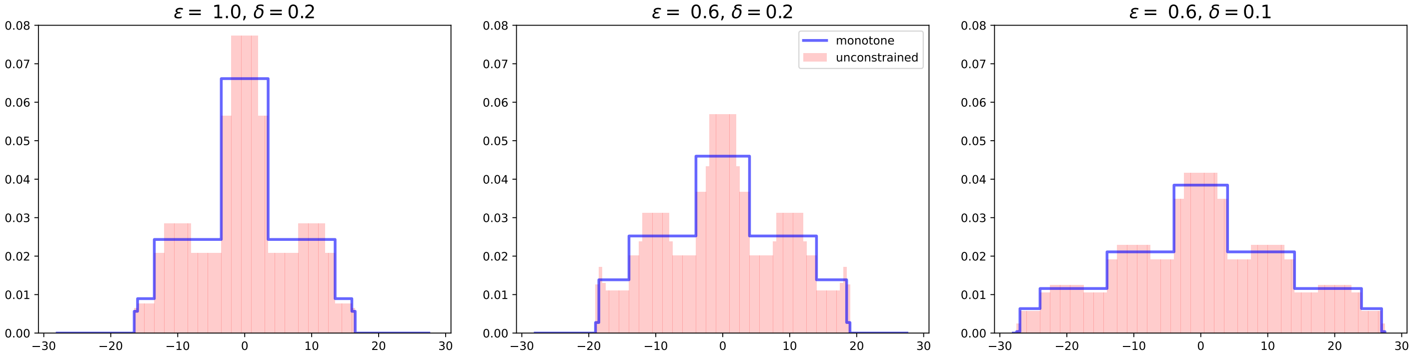


Figure 4: *Optimization-based noise distributions for synthetic data independent instances with* $\Delta f = 10$, $\beta = 0.5$, $\ell_1$*-loss and various combinations of* $\varepsilon$ *and* $\delta$*. The unconstrained distributions are shown in red shading, whereas the best monotone distributions are shown as blue lines.*

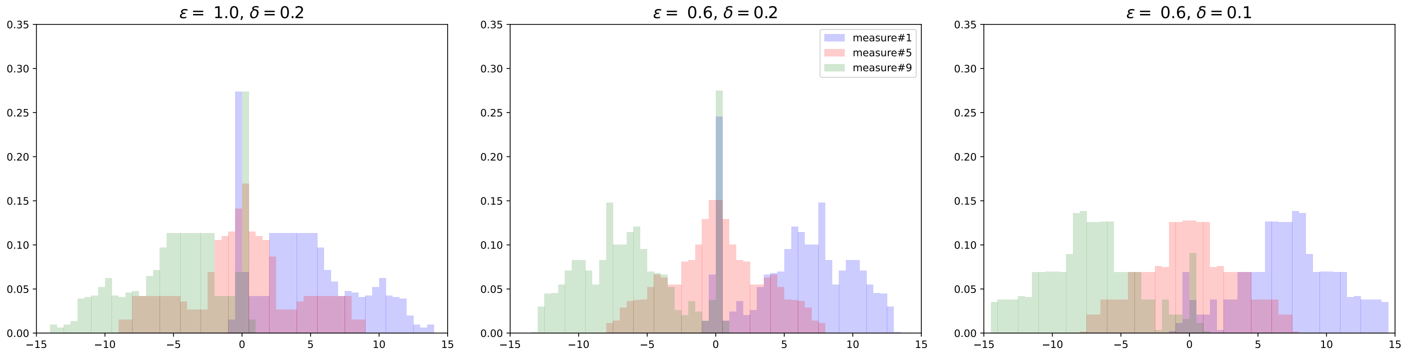


Figure 5: *Optimization-based noise distributions for synthetic data dependent instances with* $\Delta f = 10$, $\beta = 0.5$, $\ell_1$*-loss and various combinations of* $\varepsilon$ *and* $\delta$*. The set* $\Phi = [0, 18)$ *of query outputs has been partitioned into* 9 *intervals of equal length, resulting in* 9 *noise distributions.*

our data independent and data dependent noise optimization problems. For clarity of exposition, we present results for uniform partitions $\{I_i(\beta)\}_{i \in [\pm L]}$ of the noise realizations as well as the range $\Phi$ of query outputs. We emphasize, however, that better results can normally be obtained by non-uniform partitions that combine finer discretizations around 0 with coarser discretizations of the tails. Figure 4 visualizes optimization-based data independent noise distributions for different privacy regimes. We make multiple observations. Firstly, the optimal noise distribution may not be monotone. Indeed, we verified in separate experiments that the lower bounds of the best monotone noise distributions can strictly exceed the upper bounds of the best non-monotone noise distributions (details are available in the GitHub repository). This is noteworthy as all of the benchmark DP mechanisms employ monotone noise distributions, and monotonicity assumptions are commonly made in the literature without scrutinizing their impact on optimality. Secondly, the shapes of the optimization-based noise distributions are non-trivial, and they appear to depend on the problem parameters in a non-trivial manner. Finally, we note that the shapes of the optimization-based noise distributions differ with the loss function; in particular, asymmetric loss functions (such as the pinball loss) result in asymmetric noise distributions (details are relegated to the GitHub repository). We take our last two observations as an indication of the inherent complexity of optimal noise distributions, which emphasizes the need for optimization-based approaches as opposed to closed-form solutions. Figure 5 reports optimization-based data dependent noise distributions for different privacy regimes. In addition to the previous remarks, which continue to apply to the data dependent setting, we additionally observe that the noise distributions corresponding to different intervals of the query output range $\Phi$ differ in non-trivial ways. Again, this confirms our belief that the superior privacy-accuracy trade-offs achieved by optimization-based noise distributions are unlikely to be matched by DP mechanisms relying on closed-form expressions or the tuning of a small number of hyperparameters.

### 1.5.2 Differentially Private Naïve Bayes Classification

Given a dataset $(\boldsymbol{x}^i, y^i)_{i=1}^n$ with feature vectors $\boldsymbol{x}^i$ comprising numerical features $x_v^i$, $v \in \mathcal{V}_{\text{num}}$, and/or categorical features $x_v^i$, $v \in \mathcal{V}_{\text{cat}}$, as well as categorical outputs $y^i \in \mathcal{C}$, the naïve Bayes classifier employs the class-conditional independence assumption to predict the output $c^\star$ corresponding to the feature values $\boldsymbol{x} = \boldsymbol{\chi}$ of a new sample as the label $c \in \mathcal{C}$ that maximizes

$$\mathbb{P}[y = c \mid \boldsymbol{x}] = \frac{\mathbb{P}[y = c] \cdot \prod_{v \in \mathcal{V}_{\text{num}} \cup \mathcal{V}_{\text{cat}}} \mathbb{P}[x_v = \chi_v \mid y = c]}{\mathbb{P}[\boldsymbol{x}]}.$$

The naïve Bayes classifier replaces the unknown probabilities $\mathbb{P}[y = c]$ and $\mathbb{P}[x_v = \chi_v \,|\, y = c]$, $v \in \mathcal{V}_{\text{cat}}$, with their empirical frequencies in the dataset $(\boldsymbol{x}^i, y^i)_{i=1}^n$, and it makes a normality assumption to replace $\mathbb{P}[x_v = \chi_v \,|\, y = c]$, $v \in \mathcal{V}_{\text{num}}$, with an empirical density in the dataset $(\boldsymbol{x}^i, y^i)_{i=1}^n$ using the empirical means $\mu_{\{x_v | y=c\}}$ and standard deviations $\sigma_{\{x_v | y=c\}}$. We refer to Hastie et al. (2009) for a detailed description of the naïve Bayes classifier.

We follow Vaidya et al. (2013) and Lopuhaä-Zwakenberg et al. (2021) to construct a differentially private naïve Bayes classifier. Assuming that the number of training samples $n$ is public knowledge, the only data-related information used by our classifier is the number $n_{\{y=c\}}$ of training samples with label $c \in \mathcal{C}$, the number $n_{\{x_v = \chi_v \wedge y=c\}}$ of training samples with label

| UCI dataset descriptions | | | | | In-sample errors | | | | | Out-of-sample errors | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Dataset | $n$ | $\lvert\mathcal{V}_{\text{num}}\rvert$ | $\lvert\mathcal{V}_{\text{cat}}\rvert$ | $\lvert\mathcal{C}\rvert$ | GN | TLN | OPT | NB | *Imp* | GN | TLN | OPT | NB | *Imp* |
| post-operative | 86 | 1 | 7 | 2 | 36.94% (33.41%) | 32.42% | **31.43%** (29.61%) | 25.65% | *14.62%* | 41.28% (39.61%) | 39.07% | **38.30%** (37.03%) | 35.12% | *19.53%* |
| adult | 45,222 | 5 | 8 | 2 | 38.78% (21.98%) | 21.73% | **20.49%** (18.84%) | 17.37% | *28.25%* | 38.82% (22.04%) | 21.80% | **20.56%** (18.91%) | 17.45% | *28.28%* |
| breast-cancer | 683 | 0 | 9 | 2 | 2.45% (2.23%) | 2.20% | **2.17%** (2.14%) | 2.12% | *36.80%* | 3.82% (3.57%) | 3.54% | **3.51%** (3.48%) | 3.46% | *36.75%* |
| contraceptive | 1,473 | 2 | 7 | 3 | 58.84% (52.84%) | 52.38% | **51.32%** (50.46%) | 49.14% | *32.64%* | 59.53% (54.22%) | 53.76% | **52.89%** (52.19%) | 51.08% | *32.55%* |
| dermatology | 366 | 2 | 32 | 6 | 1.68% (1.03%) | 0.91% | **0.90%** (0.60%) | 0.49% | *1.07%* | 35.96% (35.61%) | 35.53% | **35.52%** (35.28%) | 35.25% | *1.33%* |
| cylinder-bands | 539 | 19 | 14 | 2 | 40.95% (35.63%) | 34.69% | **33.98%** (31.46%) | 23.43% | *6.35%* | 41.83% (37.20%) | 36.42% | **35.81%** (33.70%) | 26.89% | *6.40%* |
| annealing | 898 | 6 | 18 | 5 | 13.65% (7.93%) | 7.47% | **7.39%** (7.31%) | 7.32% | *51.56%* | 14.23% (8.38%) | 7.89% | **7.80%** (7.69%) | 7.70% | *44.41%* |
| spect | 160 | 0 | 22 | 2 | 26.46% (25.99%) | 25.89% | **25.78%** (25.77%) | 25.67% | *47.48%* | 29.34% (28.77%) | 28.66% | **28.59%** (28.57%) | 28.50% | *43.74%* |
| bank | 45,211 | 4 | 12 | 2 | 13.98% (12.40%) | 12.35% | **12.32%** (12.24%) | 12.13% | *14.20%* | 14.05% (12.48%) | 12.43% | **12.40%** (12.33%) | 12.22% | *14.26%* |
| abalone | 4,177 | 7 | 1 | 2 | 39.25% (31.85%) | 31.36% | **30.64%** (28.75%) | 26.46% | *14.77%* | 39.42% (32.14%) | 31.66% | **30.95%** (29.09%) | 26.86% | *14.76%* |
| spambase | 4,601 | 57 | 0 | 2 | 39.40% (37.09%) | 35.50% | **34.99%** (31.56%) | 18.09% | *2.91%* | 39.26% (36.97%) | 35.41% | **34.91%** (31.54%) | 18.26% | *2.90%* |
| ecoli | 336 | 5 | 2 | 2 | 48.09% (31.40%) | 27.00% | **26.51%** (18.76%) | 3.97% | *2.15%* | 48.33% (31.81%) | 27.44% | **26.95%** (19.28%) | 4.52% | *2.12%* |
| absent | 740 | 12 | 8 | 2 | 45.95% (30.22%) | 28.24% | **28.01%** (26.56%) | 24.56% | *2.60%* | 47.14% (33.99%) | 32.26% | **32.03%** (30.73%) | 29.18% | *2.59%* |

Table 2: *In-sample and out-of-sample errors of our optimization-based differentially private naïve Bayes classifier as well as several DP mechanisms from the literature on UCI datasets. Bold printing highlights the smallest errors obtained across all data independent DP mechanisms.*

$c \in \mathcal{C}$ whose categorical feature $v \in \mathcal{V}_{\text{cat}}$ attains value $\chi_v$, as well as the conditional empirical means $\mu_{\{x_v|y=c\}}$ and standard deviations $\sigma_{\{x_v|y=c\}}$ of the numerical features $v \in \mathcal{V}_{\text{num}}$. The post processing property (Dwork and Roth 2014, Proposition 2.1) then allows us to design a differentially private naïve Bayes classifier by perturbing these statistics according to their individual sensitivities and using the perturbed statistics to classify new samples. Since $n_{\{y=c\}}$ and $n_{\{x_v=\chi_v \wedge y=c\}}$, $v \in \mathcal{V}_{\text{cat}}$, are simple counting queries, they can change by at most 1 among any two neighboring datasets. For the numerical features $v \in \mathcal{V}_{\text{num}}$, we assume given upper and lower bounds $u_v$ and $l_v$ for the feature values. In this case, the value of $x_v$ can differ by at most $u_v - l_v$ for any two neighboring datasets, which implies that the sensitivity of $\psi_{\{x_v|y=c\}}$ and $\sigma_{\{x_v|y=c\}}$ is bounded from above by $(u_v - l_v)/n_{\{y=c\}}$ and $(u_v - l_v)/\sqrt{n_{\{y=c\}}}$, respectively. We note that $\mu_{\{x_v|y=c\}}$ and $\sigma_{\{x_v|y=c\}}$ satisfy our surjectivity assumption (*cf.* Assumption 2 in Section 1.2 and Assumption 3 in Section 1.3), whereas the assumption is violated by the count queries $n_{\{y=c\}}$ and $n_{\{x_v=\chi_v \wedge y=c\}}$. Thus, our optimization-based approaches provide feasible but potentially overly conservative noise distributions.

Table 2 presents the in-sample and out-of-sample errors of our optimization-based differentially private naïve Bayes classifier using the $\ell_1$-loss in a data independent (“OPT”) as well as data dependent (“(OPT)”) setting on the most popular UCI classification datasets (Dua and Graff 2017). The table also compares our results with diffentially private naïve Bayes classifiers employing a Gaussian noise (“GN”), an analytic Gaussian noise (“(GN)”) and a truncated Laplace noise (“TLN”), as well as the classical (non-private) naïve Bayes classifier (“NB”). We fix $(\varepsilon, \delta) = (1, 0.1)$ in all DP mechanisms. The reported errors are mean errors over 100 random splits of the datasets into training sets (80% of the data) and test sets (20% of the data) as well as, for each split, 1,000 simulations of all differentially private naïve Bayes implementations. The column 'Imp' records the percentage of the gap between NB and the best method from the literature (which turns out to be TLN) that is closed by OPT. The table reveals that our optimization-based data independent and data dependent noise distributions consistently outperform the considered competitors. To see whether this outperformance is statistically significant, we computed the p-values of a t-test with a null hypothesis that the second best approach is as good as OPT. The t-test averages the 1,000 simulated errors for each training set-test set split and considers the differences of the 100 averages corresponding to different training set-test set splits (Salzberg 1997). In all experiments, the p-values are less than $10^{-7}$, except for dermatology where the p-value is $10^{-1}$. Further details on the experimental setting as well as the applied the t-tests are relegated to the GitHub repository.

### 1.5.3 Differentially Private Proximal Coordinate Descent

Given a dataset $(\boldsymbol{x}^i, y^i)_{i=1}^n$ with feature vectors $\boldsymbol{x}^i \in \mathbb{R}^d$ comprising numerical and/or categorical features as well as binary outputs $y^i \in \{-1, +1\}$, the $\ell_1$-regularized logistic regression assumes that $\mathbb{P}[y \,|\, \boldsymbol{x} = \boldsymbol{\chi}] = [1 + \exp(-y \cdot \boldsymbol{h}^{0\top} \boldsymbol{\chi})]^{-1}$ for some unknown hyperplane $\boldsymbol{h}^0 \in \mathbb{R}^d$, and it determines a hyperplane $\boldsymbol{h}^\star \in \mathbb{R}^d$ that minimizes the empirical logistic loss

$$\frac{1}{n} \cdot \sum_{i=1}^{n} \log(1 + \exp(-y^i \cdot \boldsymbol{h}^\top \boldsymbol{x}^i)) + \lambda \cdot ||\boldsymbol{h}||_1,$$

where $\lambda > 0$ is a hyperparameter. Subsequently, the output of a new sample with feature values $\boldsymbol{x} = \boldsymbol{\chi}$ is predicted to be the label $y \in \{-1, +1\}$ that maximizes $[1 + \exp(-y \cdot \boldsymbol{h}^\top \boldsymbol{\chi})]^{-1}$.

To solve the logistic regression problem, proximal coordinate descent (Friedman et al. 2010, Richtárik and Takáč 2014) starts at a randomly selected initial solution $\boldsymbol{h}^0$ and conducts $t = 1, \ldots, T$ iterations, $T \in \mathbb{N}$, each of which applies the proximal operator to a random subset $i_1^t, \ldots, i_K^t \in [d]$ of the components via

$$h_l^{t,k} = \text{prox}_{\lambda|\cdot|}\left( h_l^{t,k-1} - \frac{1}{n} \cdot \left[ \sum_{i=1}^{n} \frac{\exp(-y^i \cdot \boldsymbol{h}^{t,k\top} \boldsymbol{x}^i)}{1 + \exp(-y^i \cdot \boldsymbol{h}^{t,k\top} \boldsymbol{x}^i)} \cdot (-y^i \cdot x_l^i) \right] \right) \qquad (10)$$

if $l = i_k^t$, and $h_l^{t,k} = h_l^{t,k-1}$ otherwise, for all $k = 1, \ldots, K$. Here, we fix $\boldsymbol{h}^{t,0} = \boldsymbol{h}^{t-1}$, and $\text{prox}_{\lambda|\cdot|}(\cdot)$

| UCI dataset descriptions | | | In-sample errors | | | | | Out-of-sample errors | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Dataset | $n$ | $d$ | GN | TLN | OPT | PCD | *Imp* | GN | TLN | OPT | PCD | *Imp* |
| post-operative | 86 | 14 | 40.23% (34.96%) | 33.75% | **33.37%** (30.67%) | 27.31% | *5.94%* | 45.67% (42.99%) | 42.08% | **41.86%** (41.22%) | 35.56% | *3.38%* |
| adult | 45,222 | 57 | 19.77% (19.77%) | 19.75% | **19.73%** (19.67%) | 19.75% | *573.31%* | 19.79% (19.79%) | 19.77% | **19.75%** (19.69%) | 19.77% | *640.17%* |
| breast-cancer | 683 | 26 | 4.60% (4.36%) | 4.36% | **4.34%** (3.93%) | 4.31% | *28.05%* | 4.75% (4.52%) | 4.51% | **4.50%** (4.09%) | 4.46% | *22.25%* |
| contraceptive | 1,473 | 18 | 38.52% (38.30%) | 38.29% | **38.27%** (37.39%) | 38.21% | *19.28%* | 39.90% (39.71%) | 39.70% | **39.69%** (38.93%) | 39.64% | *20.17%* |
| dermatology | 366 | 98 | 18.09% (14.56%) | 14.23% | **14.14%** (7.95%) | 13.14% | *8.07%* | 21.38% (18.21%) | 17.93% | **17.85%** (11.96%) | 16.95% | *8.17%* |
| cylinder-bands | 539 | 63 | 30.27% (28.74%) | 28.65% | **28.59%** (25.30%) | 28.20% | *12.96%* | 32.65% (31.36%) | 31.29% | **31.24%** (28.56%) | 30.90% | *14.65%* |
| annealing | 898 | 42 | 16.70% (16.27%) | 16.30% | **16.29%** (14.74%) | 16.20% | *6.16%* | 17.52% (17.10%) | 17.13% | **17.13%** (15.61%) | 17.03% | *0.83%* |
| spect | 160 | 23 | 28.89% (23.62%) | 22.74% | **22.57%** (18.60%) | 19.61% | *5.23%* | 31.47% (27.19%) | 26.37% | **26.24%** (23.58%) | 23.71% | *4.84%* |
| bank | 45,211 | 44 | 12.20% (12.20%) | 12.20% | **12.20%** (11.69%) | 12.22% | *25.43%* | 12.21% (12.21%) | 12.21% | **12.21%** (11.71%) | 12.23% | *20.31%* |
| abalone | 4,177 | 10 | 27.45% (27.44%) | 27.44% | **27.43%** (27.34%) | 27.43% | *87.76%* | 27.53% (27.52%) | 27.52% | **27.51%** (27.43%) | 27.51% | *79.42%* |
| spambase | 4,601 | 58 | 39.25% (39.26%) | 39.23% | **39.21%** (38.79%) | 39.26% | *65.85%* | 39.60% (39.60%) | 39.57% | **39.55%** (39.13%) | 39.61% | *41.07%* |
| ecoli | 336 | 8 | 9.38% (7.29%) | 7.09% | **7.01%** (6.31%) | 6.52% | *14.81%* | 9.88% (7.73%) | 7.53% | **7.45%** (6.70%) | 6.96% | *13.67%* |
| absent | 740 | 70 | 33.77% (32.88%) | 32.78% | **32.74%** (29.45%) | 32.54% | *17.16%* | 35.63% (34.83%) | 34.74% | **34.68%** (31.95%) | 34.52% | *26.34%* |
| colon-cancer | 62 | 2,000 | 18.72% (10.85%) | 9.20% | **8.62%** (0.03%) | 0.00% | *6.34%* | 32.09% (31.63%) | 31.62% | **31.54%** (30.41%) | 30.67% | *7.77%* |

Table 3: *In-sample and out-of-sample errors of our optimization-based differentially private logistic classifier as well as several DP mechanisms from the literature on UCI datasets. Bold printing highlights the lowest errors obtained across all data independent DP mechanisms.*

denotes the proximal operator (Parikh et al. 2014) associated with the $\ell_1$-regularization:

$$\operatorname{prox}_{\lambda|\cdot|}(w_j) := \underset{v \in \mathbb{R}}{\arg\min} \left\{ \frac{1}{2} \cdot (w_j - v)^2 + \lambda \cdot |v| \right\} = \begin{cases} w_j - \lambda & \text{if } w_j \geq \lambda \\ w_j + \lambda & \text{if } w_j \leq -\lambda \\ 0 & \text{if } |w_j| \leq \lambda \end{cases}$$

Upon completion of the $K$ applications of the proximal operator in iteration $t$, we set $\boldsymbol{h}^t = (1/K) \cdot \sum_{k=1}^{K} \boldsymbol{h}^{t,k}$ and continue with iteration $t+1$. The algorithm terminates with $\boldsymbol{h}^T$ as an approximately optimal solution to the regularized logistic regression problem.

We follow Mangold et al. (2022) to construct a differentially private proximal coordinate descent method for the logistic regression problem. The only data-related information used by our algorithm is contained in the proximal updates (10). Assuming that the feature vectors $\boldsymbol{x}^i$ are normalized so that $\|\boldsymbol{x}^i\|_\infty \leq 1,\ i \in [n]$, we have

$$\sum_{i=1}^{n} \underbrace{\frac{\exp(-y^i \cdot \boldsymbol{h}^{t,k\top} \boldsymbol{x}^i)}{1 + \exp(-y^i \cdot \boldsymbol{h}^{t,k\top} \boldsymbol{x}^i)}}_{\in (0,1)} \cdot \underbrace{(-y^i \cdot x_l^i)}_{\in [-1,+1]} \in (-n, n),$$

and thus the sensitivity of this summation, which is determined by the maximum change achievable by modifying a single training sample $i \in [n]$, is 2. The post processing property (Dwork and Roth 2014, Proposition 2.1) then allows us to design a differentially private proximal coordinate descent method by perturbing the sum inside the proximal updates (10) accordingly.

Table 3 presents the in-sample and out-of-sample errors of our optimization-based differentially private proximal coordinate descent algorithm in a data independent ("OPT") as well as data dependent ("(OPT)") setting for $T = 100$ iterations and $K = \lceil d/4 \rceil$ proximal updates per iteration, regularization parameter $\lambda = 10^{-8}$ and $(\varepsilon, \delta) = (1, 0.1)$. While our data independent algorithm minimizes the expected $\ell_1$-loss, our data dependent algorithm performs much better under a handcrafted loss function that resembles the $\ell_1$-loss in the vicinity of the origin but has steep slopes of -1,000 and 1,000 for large negative and positive values away from the origin, respectively. The large slopes penalize switching the sign of gradient, which tends to slow down convergence (recall from Figure 5 that noise distributions associated with negative values tend to have large probability mass on the positive side and vice versa). We use the same datasets as in Section 1.5.2, but we *(i)* convert non-binary output labels into binary ones via binning (if the output is ordinal) or distinguishing the majority class from all other classes (if the output is nominal) and *(ii)* apply one-hot encoding for the nominal input features. Additionally, as the proximal coordinate descent method is commonly used for datasets where $d \gg n$, we also include the colon-cancer dataset that is available in LIBSVM (Chang and Lin 2011). As in the previous section, we compare our optimization-based algorithms with differentially private proximal coordinate descent methods employing a Gaussian noise ("GN"), an analytic Gaussian noise ("(GN)") and a truncated Laplace noise ("TLN"), as well as the classical (non-private) proximal coordinate descent scheme ("PCD"). As before, the reported errors are mean errors over 100 random splits of the datasets into training sets (80% of the data) and test sets (20% of the data) as well as, for each split, 1,000 simulations of all differentially private algorithm implementations. The column 'Imp' records the percentage of the gap between PCD and the best method from the literature that is closed by OPT. As in the experiment from the previous section, our optimization-based data independent and data dependent noise distributions consistently outperform the considered competitors. The p-values of a t-test similar to the previous section were always smaller than $10^{-7}$, except for the annealing, breast-cancer and contraceptive datasets, where the $p$-values are $10^{-4}$, and except for the bank dataset, where there is no significance. Further details on the experimental setting can be found in the GitHub repository. Interestingly, we observe that on all datasets except for post-operative, our data dependent differentially private classifier (OPT) outperforms the non-private classifier PCD. This result is due to our handcrafted loss function (see above), and it does not hold for $\ell_1$- and $\ell_2$-losses. Effects of this types have been observed previously in gradient-based learning algorithms (see, *e.g.*, Neelakantan et al. 2015), and they further highlight the benefits of an optimization-based approach towards DP, which readily caters for non-standard loss functions that can be tuned towards the task at hand.

## 1.6 Conclusions

With the widespread adoption of analytics, privacy concerns have witnessed a remarkable resurgence in the public discourse. While DP has established itself as a predominant privacy paradigm in both academic research and industrial practice, the existing DP mechanisms almost exclusively focus on privacy to the detriment of accuracy. The few methodological studies on the privacy-accuracy trade-off focus on asymptotic performance and/or restricted classes of mechanisms.

In our view, the privacy-accuracy trade-off is most naturally studied through the lens of optimization theory, which gives rise to infinite-dimensional DP mechanism design problems that are similar to—but at the same time notedly distinct from—continuous linear programs and distributionally robust optimization problems. We developed a hierarchy of converging upper and lower bounds on these problems that result in DP mechanisms with rigorous privacy guarantees and deterministic bounds on their accuracy. Our numerical results demonstrate that our mechanisms can achieve significant improvements on both synthetic and real-world problems.

A key advantage of an optimization-based DP approach is its versatility. Our upper and lower bounds can be readily extended to incorporate monotonicity and symmetry constraints as well as a bounded range for the query output in the case of data dependent mechanisms. We can account for different loss functions, and multiple loss functions can be combined in a multi-objective framework. Our approach also allows us to incorporate tighter bounds on the probability of distinguishing events $A \in \mathcal{F}$, $\mathbb{P}[\mathcal{A}(D) \in A] > 0$ and $\mathbb{P}[\mathcal{A}(D') \in A] = 0$ for some $(D, D') \in \mathcal{N}$, that enable an adversary to exclude certain databases altogether (Dwork and Rothblum 2016, Remark 1.3).

We regard this work as a first step towards optimization-based DP, and it opens up several avenues for future research. Firstly, while our approach generalizes to multi-dimensional query functions $f$ via the composition theorem, better results can be expected if one directly optimizes over multi-dimensional noise distributions. Since a naïve implementation of our algorithms would scale exponentially in that dimension, this may necessitate the development of further approximations. Our GitHub supplement reports on some initial numerical experiments for the multi-dimensional case as well as a potential avenues to alleviate the curse of dimensionality. Secondly, our work measures accuracy through a simple loss function. In many interesting practical applications, the query output may be the solution to an optimization problem, and in those cases accuracy may be best measured in terms of the expected performance of the perturbed output (*e.g.*, the expected total discounted reward of the perturbed policy in the context of differentially private Markov decision processes). Thirdly, DP as currently defined in the literature is tailored to the traditional noise distributions such as the Laplace and the Gaussian mechanisms. The versatility of an optimization-based view on DP allows us to explore other, potentially more general notions of differential privacy as well. Finally, it would be instructive to further study the connections between DP and robust optimization. Interesting avenues for further exploration include the development of uncertainty set-based mechanisms for DP that may avoid the curse of dimensionality (Bertsimas and Sim 2004), as well as applying robust satisficing techniques (Long et al. 2023) to offer group privacy guarantees. We refer to GitHub supplement for a more detailed discussion of those topics.

## 1.A Proofs of Section 1.2

### 1.A.1 Proof of Observation 1

Assumption 2 allows us to replace the DP constraint in (1) with

$$\begin{aligned}
&\int_{x\in\mathbb{R}} \mathbb{1}[x \in A]\,\mathrm{d}\gamma(x) \leq e^{\varepsilon} \cdot \int_{x\in\mathbb{R}} \mathbb{1}[f(D') - f(D) + x \in A]\,\mathrm{d}\gamma(x) + \delta \quad \forall (D, D') \in \mathcal{N},\ \forall A \in \mathcal{F} \\
\iff &\int_{x\in\mathbb{R}} \mathbb{1}[x \in A]\,\mathrm{d}\gamma(x) \leq e^{\varepsilon} \cdot \int_{x\in\mathbb{R}} \mathbb{1}[x + \varphi \in A]\,\mathrm{d}\gamma(x) + \delta \qquad\qquad\qquad \forall (\varphi, A) \in \mathcal{E}.
\end{aligned}$$

Here, the first line holds since $\{A : A \in \mathcal{F}\} = \{A + f(D) : A \in \mathcal{F}\}$ for any $D \in \mathcal{D}$, whereas the second line is due to Assumption 2. Replacing the latter representation of the DP constraints with those in (1) gives P and thus concludes the observation. □

### 1.A.2 Proof of Lemma 1

The proof of Lemma 1 relies on three auxiliary results, which we state and prove first. Lemma 3 shows that under restriction (2), problem P can be expressed entirely in terms of the decision variables $p$. The resulting problem coincides with P($\beta$) in terms of the decision variables, but it still comprises a larger set of constraints. Lemma 4 identifies for each query output difference $\varphi \in [-\Delta f, \Delta f]$ a constraint $(\varphi, A^{\star}(\varphi))$ that weakly dominates all constraints $(\varphi, A)$, $A \in \mathcal{F}$, and Lemma 5 identifies a set of values for $\varphi$ such that the associated constraints $(\varphi, A^{\star}(\varphi))$ weakly dominate all constraints of the problem. Lemma 1 then combines these results to prove the equivalence between the problems P and P($\beta$) under restriction (2).

**Lemma 3.** *Under restriction* (2), P *has the same optimal value as*

$$\begin{aligned}
\underset{p}{\text{minimize}} \quad & \sum_{i\in\mathbb{Z}} c_i(\beta)\cdot p(i) \\
\text{subject to} \quad & p:\mathbb{Z}\mapsto\mathbb{R}_+,\ \sum_{i\in\mathbb{Z}} p(i)=1 \\
& \sum_{i\in\mathbb{Z}} p(i)\cdot\frac{|A\cap I_i(\beta)|}{\beta} \le e^{\varepsilon}\cdot\sum_{i\in\mathbb{Z}} p(i)\cdot\frac{|(A-\varphi)\cap I_i(\beta)|}{\beta}+\delta \quad \forall(\varphi,A)\in\mathcal{E}.
\end{aligned} \tag{11}$$

*Proof.* We use restriction (2) to replace $\gamma$ in problem P with the new decision variables $p$. To this end, observe first that under restriction (2), $\gamma$ affords a density function via

$$\begin{aligned}
\gamma(A)=\sum_{i\in\mathbb{Z}} p(i)\cdot\frac{|A\cap I_i(\beta)|}{\beta} &= \sum_{i\in\mathbb{Z}} p(i)\cdot\int_{x\in\mathbb{R}}\frac{\mathbb{1}[x\in A]\cdot\mathbb{1}[x\in I_i(\beta)]}{\beta}\,\mathrm{d}x \\
&= \int_{x\in\mathbb{R}}\mathbb{1}[x\in A]\cdot\left(\sum_{i\in\mathbb{Z}} p(i)\cdot\frac{\mathbb{1}[x\in I_i(\beta)]}{\beta}\right)\mathrm{d}x,
\end{aligned}$$

$A\in\mathcal{F}$, where the last step holds by Fubini's theorem since $\gamma(A)\in[0,1]$. This derivation shows that $\sum_{i\in\mathbb{Z}} p(i)\cdot\mathbb{1}[x\in I_i(\beta)]/\beta$ is the density function of $\gamma$. Using this density function, the objective function of P can be rewritten as

$$\begin{aligned}
\int_{x\in\mathbb{R}} c(x)\,\mathrm{d}\gamma(x)=\int_{x\in\mathbb{R}} c(x)\cdot\left(\sum_{i\in\mathbb{Z}} p(i)\cdot\frac{\mathbb{1}[x\in I_i(\beta)]}{\beta}\right)\mathrm{d}x &= \sum_{i\in\mathbb{Z}}\frac{p(i)}{\beta}\cdot\int_{x\in\mathbb{R}} c(x)\cdot\mathbb{1}[x\in I_i(\beta)]\,\mathrm{d}x \\
&= \sum_{i\in\mathbb{Z}} c_i(\beta)\cdot p(i),
\end{aligned}$$

where the second equality holds by Fubini's theorem, and the last step substitutes $c_i(\beta)=\beta^{-1}\cdot\int_{x\in I_i(\beta)} c(x)\mathrm{d}x$ for all $i\in\mathbb{Z}$. The final expression coincides with the objective of problem (11).

In view of the DP constraints in problem P, we note that

$$\int_{x\in\mathbb{R}}\mathbb{1}[x\in A]\,\mathrm{d}\gamma(x)=\gamma(A)=\sum_{i\in\mathbb{Z}} p(i)\cdot\frac{|A\cap I_i(\beta)|}{\beta}$$

as well as

$$\int_{x\in\mathbb{R}}\mathbb{1}[x+\varphi\in A]\,\mathrm{d}\gamma(x)=\gamma(A-\varphi)=\sum_{i\in\mathbb{Z}} p(i)\cdot\frac{|(A-\varphi)\cap I_i(\beta)|}{\beta},$$

and the right-hand side expressions coincide with the expressions on either side of the DP constraints of problem (11). This concludes the proof. □

To reduce the number of DP constraints in problem (11), we first characterize the tightest privacy constraint $(\varphi, A) \in \mathcal{E}$ in (11) for a fixed decision $p$ and a fixed query output difference $\varphi \in [-\Delta f, \Delta f]$. To this end, we define the privacy shortfall as

$$V(\varphi, A) = \sum_{i \in \mathbb{Z}} p(i) \cdot |A \cap I_i(\beta)| - e^{\varepsilon} \cdot \sum_{i \in \mathbb{Z}} p(i) \cdot |(A - \varphi) \cap I_i(\beta)|.$$

Note that $V(\varphi, A)$ coincides with the slack of the DP constraint $(\varphi, A)$ in problem (11), shifted by $-\delta$ and scaled by $\beta$.[4] In particular, maximizers $(\varphi, A) \in \mathcal{E}$ of the privacy shortfall $V$ correspond to the tightest constraints in problem (11).

**Observation 4.** *The privacy shortfall is linear over partitions of $A$, that is, we have $V(\varphi, A) = \sum_{\ell} V(\varphi, A_\ell)$ for any partition $\{A_\ell\}_\ell$ of $A$.*

We next show that for any fixed $\varphi \in [-\Delta f, \Delta f]$, the largest privacy shortfall $\sup\{V(\varphi, A) : A \in \mathcal{F}\}$ is attained by a worst-case event $A^\star(\varphi) \in \mathcal{F}$ of a simple structure.

**Definition 1.** *For any $i \in \mathbb{Z}$, let $j := \lceil i - \varphi/\beta \rceil$ be the unique integer satisfying $\varphi + j \cdot \beta \in I_i(\beta)$, and define $I_i^1(\varphi, \beta)$ and $I_i^2(\varphi, \beta)$ as the following partition of $I_i(\beta)$:*

*(i)* $I_i^1(\varphi, \beta) := I_i(\beta) \cap (I_{j-1}(\beta) + \varphi) = [i \cdot \beta, \varphi + j \cdot \beta)$

*(ii)* $I_i^2(\varphi, \beta) := I_i(\beta) \setminus I_i^1(\varphi, \beta) = I_i(\beta) \cap (I_j(\beta) + \varphi) = [\varphi + j \cdot \beta, (i+1) \cdot \beta)$.

For any $i \in \mathbb{Z}$, Definition 1 implies that $I_i^1(\varphi, \beta) \cap I_{i'}(\beta) = \emptyset$ for all $i' \neq i$. Moreover, since $I_i^1(\varphi, \beta) \subseteq I_{j-1}(\beta) + \varphi$ it also follows that $(I_i^1(\varphi, \beta) - \varphi) \subseteq I_{j-1}(\beta)$ and therefore $(I_i^1(\varphi, \beta) - \varphi) \cap I_{j'}(\beta) = \emptyset$ for all $j' \neq j - 1$. Applying a similar reasoning also to $I_i^2(\varphi, \beta)$ allows us to simplify the expressions for the privacy shortfall over subsets of $I_i^1(\varphi, \beta)$ and $I_i^2(\varphi, \beta)$.

**Observation 5.** *For any $i \in \mathbb{Z}$ and $\varphi \in [-\Delta f, \Delta f]$ we have*

$$V(\varphi, A) = \begin{cases} |A| \cdot (p(i) - e^{\varepsilon} \cdot p(j-1)) & \text{for all } A \subseteq I_i^1(\varphi, \beta), \\ |A| \cdot (p(i) - e^{\varepsilon} \cdot p(j)) & \text{for all } A \subseteq I_i^2(\varphi, \beta), \end{cases}$$

*where $j = \lceil i - \varphi/\beta \rceil$.*

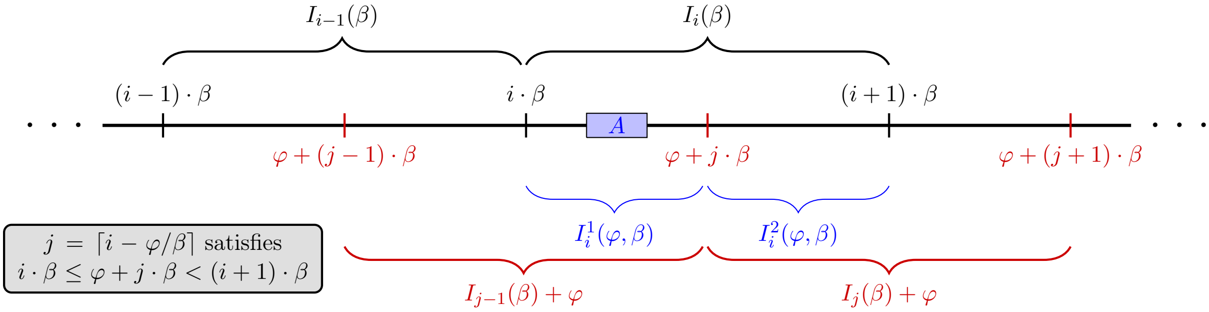


Figure 6: *The interval $I_i^1(\varphi,\beta)$ satisfies $I_i^1(\varphi,\beta) \subseteq I_i(\beta)$ as well as $I_i^1(\varphi,\beta) \subseteq I_{j-1}(\beta)+\varphi$. Therefore, for any $A \subseteq I_i^1(\varphi,\beta)$ we have $|A \cap I_i(\beta)| = |A|$ and $|A \cap I_{i'}(\beta)| = 0$ for all $i' \neq i$; similarly, $|(A-\varphi)\cap I_{j-1}(\beta)| = |A|$ and $|(A-\varphi)\cap I_{j'-1}(\beta)| = 0$ for all $j' \neq j$. This shows the first case in Observation 5; the second case can be verified analogously.*

Figure 6 illustrates the intuition underlying Observation 5. We now show that there is always a worst-case event $A^\star(\varphi)$ that constitutes a union of intervals $I_i^1(\varphi,\beta)$ and $I_i^2(\varphi,\beta)$, $i \in \mathbb{Z}$.

**Lemma 4.** *For any $\varphi \in [-\Delta f, \Delta f]$, there is an event*

$$A^\star(\varphi) = \bigcup_{i\in\mathcal{I}_1} I_i^1(\varphi,\beta) \cup \bigcup_{i\in\mathcal{I}_2} I_i^2(\varphi,\beta) \in \mathcal{F} \qquad \textit{for some } \mathcal{I}_1,\mathcal{I}_2 \subseteq \mathbb{Z} \tag{12}$$

*that attains the largest privacy shortfall* $\sup\{V(\varphi,A) : A \in \mathcal{F}\}$.

*Proof.* By Definition 1, $I_i^1(\varphi,\beta)$ and $I_i^2(\varphi,\beta)$ partition $I_i(\beta)$ for any $i \in \mathbb{Z}$, and thus $\{I_i^1(\varphi,\beta) \cup I_i^2(\varphi,\beta)\}_{i\in\mathbb{Z}}$ partitions $\mathbb{R}$. Observation 4 and the sub-additivity of the supremum operator then imply that

$$\begin{aligned}\sup_{A\in\mathcal{F}}\{V(\varphi,A)\} &= \sup_{A\in\mathcal{F}}\left\{\sum_{i\in\mathbb{Z}} V(\varphi, A\cap I_i^1(\varphi,\beta)) + \sum_{i\in\mathbb{Z}} V(\varphi, A\cap I_i^2(\varphi,\beta))\right\}\\ &\leq \sum_{i\in\mathbb{Z}} \sup_{A\subseteq I_i^1(\varphi,\beta)} \{V(\varphi,A)\} + \sum_{i\in\mathbb{Z}} \sup_{A\subseteq I_i^2(\varphi,\beta)} \{V(\varphi,A)\}.\end{aligned}$$

We will show that each supremum on the right-hand side of the inequality is attained and then construct $A^\star(\varphi) = \bigcup_{i\in\mathbb{Z}} A_i^1(\varphi) \cup \bigcup_{i\in\mathbb{Z}} A_i^2(\varphi)$, where $A_i^1(\varphi), A_i^2(\varphi) \in \mathcal{F}$ are defined as

$$A_i^1(\varphi) \in \operatorname*{arg\,max}_{A\subseteq I_i^1(\varphi,\beta)} \{V(\varphi,A)\} \quad \text{and} \quad A_i^2(\varphi) \in \operatorname*{arg\,max}_{A\subseteq I_i^2(\varphi,\beta)} \{V(\varphi,A)\}.$$

[4] Our definition here differs slightly from a later definition of privacy shortfall in Section 1.4.2. The context will always make it clear which definition is being used.

The statement then follows from the fact that $A^\star(\varphi) \in \mathcal{F}$ by construction.

In view of $A_i^1(\varphi)$, Observation 5 implies that $V(\varphi, A)$ is maximized by

$$A_i^1(\varphi) = \begin{cases} I_i^1(\varphi, \beta) & \text{if } p(i) - e^\varepsilon \cdot p(j-1) > 0, \\ \emptyset & \text{otherwise.} \end{cases}$$

Applying a similar reasoning to $A_i^2(\varphi)$, we observe that

$$A_i^2(\varphi) = \begin{cases} I_i^2(\varphi, \beta) & \text{if } p(i) - e^\varepsilon \cdot p(j) > 0, \\ \emptyset & \text{otherwise.} \end{cases}$$

The statement of the lemma thus follows. □

The next result shows that $V(\varphi, A^\star(\varphi))$ is maximized by $\varphi^\star = k \cdot \beta$ for some $k \in \mathbb{Z}$.

**Lemma 5.** *The function $\varphi \mapsto V(\varphi, A^\star(\varphi))$ is affine over each interval $[k \cdot \beta, (k+1) \cdot \beta]$, $k \in \mathbb{Z}$.*

*Proof.* For any $k \in \mathbb{Z}$, the construction of $A^\star(\varphi)$ in the proof of Lemma 4 implies that the sets $\mathcal{I}_1$ and $\mathcal{I}_2$ in the statement of the lemma coincide for all $\varphi \in [k \cdot \beta, (k+1) \cdot \beta) = I_k(\beta)$. Therefore, for all $\varphi \in I_k(\beta)$ we have

$$\begin{aligned} V(\varphi, A^\star(\varphi)) &= \sum_{i \in \mathcal{I}_1} V(\varphi, I_i^1(\varphi, \beta)) + \sum_{i \in \mathcal{I}_2} V(\varphi, I_i^2(\varphi, \beta)) \\ &= \sum_{i \in \mathcal{I}_1} |I_i^1(\varphi, \beta)| \cdot (p(i) - e^\varepsilon \cdot p(j-1)) + \sum_{i \in \mathcal{I}_2} |I_i^2(\varphi, \beta)| \cdot (p(i) - e^\varepsilon \cdot p(j)) \\ &= (\varphi \bmod \beta) \cdot \sum_{i \in \mathcal{I}_1} (p(i) - e^\varepsilon \cdot p(j-1)) + (\beta - (\varphi \bmod \beta)) \cdot \sum_{i \in \mathcal{I}_2} (p(i) - e^\varepsilon \cdot p(j)) \\ &= (\varphi \bmod \beta) \cdot \left[ \sum_{i \in \mathcal{I}_1} (p(i) - e^\varepsilon \cdot p(j-1)) - \sum_{i \in \mathcal{I}_2} (p(i) - e^\varepsilon \cdot p(j)) \right] + \beta \cdot \sum_{i \in \mathcal{I}_2} (p(i) - e^\varepsilon \cdot p(j)), \end{aligned} \tag{13}$$

where $j = \lceil i - \varphi/\beta \rceil$ as specified by Definition 1. Here, the first equality follows from Lemma 4 and Observation 4, the second equality is due to Observation 5, the third equality holds since

$$|I_i^1(\varphi, \beta)| = \varphi + \lceil i - \varphi/\beta \rceil \cdot \beta - i \cdot \beta = \varphi - \beta \cdot \lfloor \varphi/\beta \rfloor = (\varphi \bmod \beta)$$

and $|I_i^2(\varphi, \beta)| = \beta - |I_i^1(\varphi, \beta)| = \beta - (\varphi \bmod \beta)$. In the final expression, all terms except for

($\varphi$ mod $\beta$) are independent of $\varphi$, and $\varphi \mapsto (\varphi \bmod \beta)$ is affine over $\varphi \in I_k(\beta)$. We thus conclude that $\varphi \mapsto V(\varphi, A^\star(\varphi))$ is affine over $\varphi \in I_k(\beta)$.

To conclude the proof, we show that the result holds for the closure of $I_k(\beta)$, that is, $\varphi \mapsto V(\varphi, A^\star(\varphi))$ is not discontinuous at $\overline{\varphi} = (k+1) \cdot \beta$. In other words, we show that

$$\lim_{\varphi \to \overline{\varphi}} V(\varphi, A^\star(\varphi)) = V(\overline{\varphi}, A^\star(\overline{\varphi})).$$

To this end, we first note that $(\overline{\varphi} \bmod \beta) = 0$ as well as $j = \lceil i - \overline{\varphi}/\beta \rceil = i - k - 1$, and hence (13) implies that

$$\begin{aligned} V(\overline{\varphi}, A^\star(\overline{\varphi})) = \beta \cdot \sum_{i \in \mathcal{I}_2} (p(i) - e^\varepsilon \cdot p(j)) &= \beta \cdot \sum_{i \in \mathcal{I}_2} (p(i) - e^\varepsilon \cdot p(i-k-1)) \\ &= \beta \cdot \sum_{i \in \mathbb{Z}} \max\{p(i) - e^\varepsilon \cdot p(i-k-1), 0\}. \end{aligned}$$

Since we also have $j = \lceil i - \varphi/\beta \rceil = i - k$ for all $k \cdot \beta \leq \varphi < (k+1) \cdot \beta$, it follows that

$$\begin{aligned} \lim_{\varphi \to \overline{\varphi}} V(\varphi, A^\star(\varphi)) =& \lim_{\varphi \to \overline{\varphi}} \Big( (\varphi \bmod \beta) \cdot \Big[ \sum_{i \in \mathcal{I}_1} (p(i) - e^\varepsilon \cdot p(j-1)) - \sum_{i \in \mathcal{I}_2} (p(i) - e^\varepsilon \cdot p(j)) \Big] + \beta \cdot \sum_{i \in \mathcal{I}_2} (p(i) - e^\varepsilon \cdot p(j)) \Big) \\ =& \Big[ \sum_{i \in \mathcal{I}_1} (p(i) - e^\varepsilon \cdot p(j-1)) - \sum_{i \in \mathcal{I}_2} (p(i) - e^\varepsilon \cdot p(j)) \Big] \cdot \lim_{\varphi \to \overline{\varphi}} \{(\varphi \bmod \beta)\} + \beta \cdot \sum_{i \in \mathcal{I}_2} (p(i) - e^\varepsilon \cdot p(j)) \\ =& \beta \cdot \sum_{i \in \mathcal{I}_1} (p(i) - e^\varepsilon \cdot p(j-1)) = \beta \cdot \sum_{i \in \mathcal{I}_1} (p(i) - e^\varepsilon \cdot p(i-k-1)) \\ =& \beta \cdot \sum_{i \in \mathbb{Z}} \max\{p(i) - e^\varepsilon \cdot p(i-k-1), 0\} = V(\overline{\varphi}, A^\star(\overline{\varphi})), \end{aligned}$$

where the first equality follows from (13), the second equality exploits the linearity of limits, the third equality holds since $\lim_{\varphi \to \overline{\varphi}} (\varphi \bmod \beta) = \beta$, and the final equalities follow from substituting $j = i - k$ and using the construction of $\mathcal{I}_1$, which includes all indices $i \in \mathbb{Z}$ for which the incremental privacy shortfall $p(i) - e^\varepsilon \cdot p(i-k-1)$ is positive. This shows that $\varphi \mapsto V(\varphi, A^\star(\varphi))$ is not discontinuous at $\overline{\varphi} = (k+1) \cdot \beta$ and therefore concludes the proof. □

We can now prove Lemma 1 by showing that problem (11) in the statement of Lemma 3 has the same optimal value as problem P($\beta$).

**Proof of Lemma 1.** First notice that the DP constraints of P($\beta$) can be written as

$$\sum_{i \in \mathbb{Z}} p(i) \cdot \frac{|A \cap I_i(\beta)|}{\beta} \leq e^{\varepsilon} \cdot \sum_{i \in \mathbb{Z}} p(i) \cdot \frac{|(A - \varphi) \cap I_i(\beta)|}{\beta} + \delta \quad \forall (\varphi, A) \in \mathcal{E}(\beta)$$

since for any $(\varphi, A) \in \mathcal{E}(\beta)$ we have $|A \cap I_i(\beta)|/\beta = \mathbb{1}[I_i(\beta) \subseteq A]$ as well as $|(A - \varphi) \cap I_i(\beta)|/\beta = \mathbb{1}[I_i(\beta) + \varphi \subseteq A]$ by definition. This shows that P($\beta$) is a relaxation of problem (11) since $\mathcal{E}(\beta) \subset \mathcal{E}$. Hence, if P($\beta$) is infeasible, then so is problem (11), and the result follows. To complete the proof, we show that any $p$ feasible in P($\beta$) is also feasible in problem (11).

Fix any feasible solution $p$ to P($\beta$), and assume to the contrary that $p$ violates a DP constraint $(\varphi, A) \in \mathcal{E}$ in problem (11). In that case, Lemmas 4 and 5 imply that there is a constraint $(\varphi^\star, A^\star(\varphi^\star)) \in \mathcal{E}(\beta)$ with a weakly higher privacy shortfall than $(\varphi, A)$. Indeed, Lemma 5 and our earlier assumption that $\Delta f$ is divisible by $\beta$ imply that $\varphi^\star$ can without loss of generality be chosen such that $\varphi^\star \in [-\Delta f, \Delta f] \cap \{k \cdot \beta\}_{k \in \mathbb{Z}} = \mathscr{B}(\beta)$. Since such $\varphi^\star$ is a multiple of $\beta$, we have $I_i^1(\varphi, \beta) = \emptyset$ and $I_i^2(\varphi, \beta) = I_i(\beta)$ for all $i \in \mathbb{Z}$. Hence, Lemma 4 implies that $A^\star(\varphi^\star)$ can be chosen such that $A^\star(\varphi^\star) = \bigcup_{i \in \mathcal{I}_2} I_i(\beta)$ for some $\mathcal{I}_2 \subseteq \mathbb{Z}$, which in turn implies that $A^\star(\varphi) \in \mathcal{F}(\beta)$. Thus, there must be a violated DP constraint $(\varphi^\star, A^\star(\varphi^\star))$ such that $\varphi^\star \in \mathscr{B}(\beta)$ and $A^\star(\varphi) \in \mathcal{F}(\beta)$, that is, $(\varphi^\star, A^\star(\varphi^\star)) \in \mathcal{E}(\beta)$. This contradicts our earlier assumption that $p$ is feasible in P($\beta$), and thus the result follows. □

### 1.A.3 Proof of Proposition 1

Appending the additional constraint (3) to problem P($\beta$) yields

$$\begin{aligned}
\underset{p}{\text{minimize}} \quad & \sum_{i \in [\pm L]} c_i(\beta) \cdot p(i) \\
\text{subject to} \quad & p : [\pm L] \mapsto \mathbb{R}_+, \; \sum_{i \in [\pm L]} p(i) = 1 \\
& \sum_{i \in [\pm L]} \mathbb{1}[I_i(\beta) \subseteq A] \cdot p(i) \leq e^{\varepsilon} \cdot \sum_{i \in [\pm L]} \mathbb{1}[I_i(\beta) + \varphi \subseteq A] \cdot p(i) + \delta \quad \forall (\varphi, A) \in \mathcal{E}(\beta).
\end{aligned}$$

The result follows if we show that any DP constraint $(\varphi, A) \in \mathcal{E}(\beta) \setminus \mathcal{E}(L, \beta)$ is redundant in the above problem. To this end, fix any $p$ that satisfies the first constraint. For any $(\varphi, A) \in \mathcal{E}(\beta)$, define $A_L := A \cap [-L \cdot \beta, (L+1) \cdot \beta)$ so that $(\varphi, A_L) \in \mathcal{E}(L, \beta)$. We show that if $p$ satisfies the

DP constraint $(\varphi, A_L)$, then it also satisfies the DP constraint $(\varphi, A)$. Indeed, we observe that

$$\begin{aligned}
&\sum_{i\in[\pm L]} \mathbb{1}[I_i(\beta) \subseteq A] \cdot p(i) - e^{\varepsilon} \cdot \sum_{i\in[\pm L]} \mathbb{1}[(I_i(\beta) + \varphi) \subseteq A] \cdot p(i) \\
&= \sum_{i\in[\pm L]} \mathbb{1}[I_i(\beta) \subseteq A_L] \cdot p(i) - e^{\varepsilon} \cdot \sum_{i\in[\pm L]} \mathbb{1}[(I_i(\beta) + \varphi) \subseteq A] \cdot p(i) \\
&\leq \sum_{i\in[\pm L]} \mathbb{1}[I_i(\beta) \subseteq A_L] \cdot p(i) - e^{\varepsilon} \cdot \sum_{i\in[\pm L]} \mathbb{1}[(I_i(\beta) + \varphi) \subseteq A_L] \cdot p(i),
\end{aligned}$$

where the equality follows from

$$\sum_{i\in[\pm L]} \mathbb{1}[I_i(\beta) \subseteq A] \cdot p(i) = \sum_{i\in[\pm L]} \mathbb{1}[I_i(\beta) \subseteq A_L] \cdot p(i) + \sum_{i\in[\pm L]} \underbrace{\mathbb{1}[I_i(\beta) \subseteq A \setminus A_L] \cdot p(i)}_{=0}$$

which holds since no $i \in [\pm L]$ can satisfy $I_i(\beta) \subseteq A \setminus A_L$. The inequality in the third row, on the other hand, follows from $A_L \subseteq A$. Thus, the DP constraints $\mathcal{E}(\beta) \setminus \mathcal{E}(L, \beta)$ are redundant since they are weakly dominated by the DP constraints $(\varphi, A_L) \in \mathcal{E}(L, \beta)$. □

### 1.A.4 Proof of Proposition 2

As $(\theta, \psi) \in \mathbb{R} \times \mathcal{M}_+(\mathcal{E})$ is feasible in D and $\delta > 0$, we have that $\int_{(\varphi,A)\in\mathcal{E}} \mathrm{d}\psi(\varphi, A) < \infty$ from the objective function of D, which shows that $\mathcal{E}$ is $\sigma$-finite with measure $\psi$. Moreover, $\gamma$ is a probability measure on $\mathbb{R}$, and hence it is also $\sigma$-finite. We now observe that

$$\begin{aligned}
&\int_{x\in\mathbb{R}} c(x)\,\mathrm{d}\gamma(x) \\
\geq &\int_{x\in\mathbb{R}} \left[\theta - \int_{(\varphi,A)\in\mathcal{E}} \mathbb{1}[x \in A]\,\mathrm{d}\psi(\varphi, A) + e^{\varepsilon} \cdot \int_{(\varphi,A)\in\mathcal{E}} \mathbb{1}[x + \varphi \in A]\,\mathrm{d}\psi(\varphi, A)\right] \mathrm{d}\gamma(x) \\
= &\theta - \int_{(\varphi,A)\in\mathcal{E}} \left[\int_{x\in\mathbb{R}} \mathbb{1}[x \in A]\,\mathrm{d}\gamma(x) - e^{\varepsilon} \cdot \int_{x\in\mathbb{R}} \mathbb{1}[x + \varphi \in A]\,\mathrm{d}\gamma(x)\right] \mathrm{d}\psi(\varphi, A) \\
\geq &\theta - \int_{(\varphi,A)\in\mathcal{E}} \delta\,\mathrm{d}\psi(\varphi, A).
\end{aligned}$$

Here, the first inequality follows from the constraints of problem D. The equality follows from Fubini's theorem, which is applicable since the indicator functions are integrable on $\mathbb{R} \times \mathcal{E}$ with the associated product measure and the fact that $\int_{x\in\mathbb{R}} \mathrm{d}\gamma(x) = 1$. The second inequality follows from the constraints of problem P as well as the fact that $\psi$ is a non-negative measure. □

### 1.A.5 Proof of Lemma 2

We first show that under the additional constraint (4), the DP constraints in D reduce to

$$\theta \leq c(x) + \int_{(\varphi,A)\in\mathcal{E}(\beta)} \mathbb{1}[I_i(\beta) \subseteq A]\mathrm{d}\psi(\varphi, A) - e^{\varepsilon} \cdot \int_{(\varphi,A)\in\mathcal{E}(\beta)} \mathbb{1}[I_i(\beta) + \varphi \subseteq A]\mathrm{d}\psi(\varphi, A) \qquad \forall i \in \mathbb{Z},\ \forall x \in I_i(\beta). \tag{14}$$

We then argue that for every $i \in \mathbb{Z}$, all constraints (14) indexed by $(i, x)$, $x \in I_i(\beta)$, are simultaneously satisfied if and only if the DP constraint indexed by $i$ is satisfied in D($\beta$). The result then follows since both the decision variables and the objective function in D coincide with their counterparts in D($\beta$), restricted to the elements $(\varphi, A) \in \mathcal{E}(\beta)$ as stipulated by (4).

In view of the first step, fix any $x \in \mathbb{R}$, and select $i \in \mathbb{Z}$ such that $x \in I_i(\beta)$. Under the additional constraint (4), the first integral in the DP constraint of D indexed by $x$ reduces to

$$\begin{aligned}\int_{(\varphi,A)\in\mathcal{E}} \mathbb{1}[x \in A]\mathrm{d}\psi(\varphi, A) &= \int_{(\varphi,A)\in\mathcal{E}(\beta)} \mathbb{1}[x \in A]\mathrm{d}\psi(\varphi, A) + \underbrace{\int_{(\varphi,A)\in\mathcal{E}\setminus\mathcal{E}(\beta)} \mathbb{1}[x \in A]\mathrm{d}\psi(\varphi, A)}_{=0} \\ &= \int_{(\varphi,A)\in\mathcal{E}(\beta)} \mathbb{1}[I_i(\beta) \subseteq A]\mathrm{d}\psi(\varphi, A).\end{aligned}$$

Here, the last integral in the first row vanishes due to (4), whereas the second equality holds since for all $(\varphi, A) \in \mathcal{E}(\beta)$, the requirement that $A \in \mathcal{F}(\beta)$ implies that $x \in A$ only if $I_i(\beta) \subseteq A$. Note that the integral in the second row above coincides with the first integral in (14) indexed by $(i, x)$. A similar argument shows that under the additional constraint (4), the second integral in the DP constraint of D indexed by $x$ reduces to the second integral in (14) indexed by $(i, x)$. In summary, under the additional constraint (4) the DP constraints in D indeed reduce to (14).

As for the second step, note that for any fixed $i \in \mathbb{Z}$, the constraints (14) indexed by $(i, x)$, $x \in I_i(\beta)$, only differ in their additive terms $c(x)$. Thus, for any fixed $i \in \mathbb{Z}$, all constraints (14) indexed by $(i, x)$, $x \in I_i(\beta)$, are satisfied if and only if they are satisfied for the smallest value $c(x)$, $x \in I_i(\beta)$, which is precisely what the DP constraint in D($\beta$) indexed by $i$ stipulates. □

### 1.A.6 Proof of Proposition 3

First observe that the additional constraint (5) reduces the uncountable set $\mathcal{E}(\beta)$ in the definition of the decision variables as well as the objective function and the constraints of D($\beta$) to the finite subset $\mathcal{E}(L, \beta)$, which allows us to replace the measure $\psi \in \mathcal{M}_+(\mathcal{E}(\beta))$ in D($\beta$) with the discrete map $\psi : \mathcal{E}(L,\beta) \mapsto \mathbb{R}_+$ in D$(L,\beta)$ as well as replace all integrals in D$(\beta)$ with sums in D$(L,\beta)$.

The result now follows if we show that under the additional constraint (5), all DP constraints in D$(\beta)$ indexed by $i \in \mathbb{Z} \setminus [\pm(L + \Delta f/\beta)]$ are weakly dominated by DP constraints indexed by $i \in [\pm(L + \Delta f/\beta)]$. Indeed, observe that the DP constraints indexed by $i \in \mathbb{Z} \setminus [\pm(L + \Delta f/\beta)]$ simplify to

$$\theta \leq \underline{c}_i(\beta) \quad \forall i \in \mathbb{Z} \setminus [\pm(L + \Delta f/\beta)] \tag{15a}$$

since $\mathbb{1}[I_i(\beta) \subseteq A] = \mathbb{1}[I_i(\beta) + \varphi \subseteq A] = 0$ for all $(\varphi, A) \in \mathcal{E}(L,\beta)$ whenever $i \in \mathbb{Z} \setminus [\pm(L+\Delta f/\beta)]$. In contrast, the constraints indexed by $i \in [\pm(L + \Delta f/\beta)] \setminus [\pm L]$ simplify to

$$\theta \leq -e^{\varepsilon} \cdot \sum_{(\varphi,A)\in\mathcal{E}(L,\beta)} \mathbb{1}[I_i(\beta) + \varphi \subseteq A] \cdot \psi(\varphi, A) + \underline{c}_i(\beta) \tag{15b}$$

since $\mathbb{1}[I_i(\beta) \subseteq A] = 0$ for all $(\varphi, A) \in \mathcal{E}(L,\beta)$ whenever $i \in [\pm(L + \Delta f/\beta)] \setminus [\pm L]$. Note that $\underline{c}_i(\beta)$ inherits monotonicity from $c_i(\beta)$, that is, we have $\underline{c}_i(\beta) \leq \underline{c}_{i+1}(\beta)$ for all $i \geq 0$ as well as $\underline{c}_i(\beta) \leq \underline{c}_{i-1}(\beta)$ for all $i \leq 0$. This property, along with the non-negativity of $\psi$, shows that the constraints (15a) are implied by constraints (15b), and the result thus follows. □

### 1.A.7 Proof of Theorem 1

The proof of Theorem 1 relies on the feasibility and monotonicity of the upper bounding problems P$(L,\beta)$, which they inherit from the upper bounding problems P$(\beta)$. We prove these results first in Sections 1.A.7.1 and 1.A.7.2, and we subsequently prove Theorem 1 in Section 1.A.7.3.

#### 1.A.7.1 Monotonicity and Feasibility of P$(\beta)$

The upper bound P$(\beta)$ employs a discretization that is parametrized by $\beta$. We first show that the optimal value of this problem is monotonically non-decreasing in $\beta$ in the following sense.

**Lemma 6.** *For any $\varepsilon > 0$, $\delta > 0$ and $\beta > 0$, the optimal value of problem* P$(\beta)$ *satisfies* P$(\beta) \geq$ P$(\beta/k)$ *for all $k \in \mathbb{N}$.*

*Proof.* The result trivially holds if P$(\beta)$ is infeasible. Assume therefore that P$(\beta)$ is feasible and fix an arbitrary feasible solution $p'$ in P($\beta$). For any $k \in \mathbb{N}$, problem P($\beta/k$) can be written as

$$\begin{array}{ll} \underset{p}{\text{minimize}} & \displaystyle\sum_{i\in\mathbb{Z}}\sum_{l\in[k]} c_{il}(\beta)\cdot p(i,l) \\ \text{subject to} & \displaystyle p:\mathbb{Z}\times[k]\mapsto\mathbb{R}_+,\ \sum_{i\in\mathbb{Z}}\sum_{l\in[k]} p(i,l)=1 \\ & \displaystyle\sum_{i\in\mathbb{Z}}\sum_{l\in[k]} \mathbb{1}[I_{il}(\beta)\subseteq A]\cdot p(i,l) \leq e^{\varepsilon}\cdot\sum_{i\in\mathbb{Z}}\sum_{l\in[k]} \mathbb{1}[I_{il}(\beta)+\varphi\subseteq A]\cdot p(i,l)+\delta \\ & \hfill \forall(\varphi,A)\in\mathcal{E}(\beta/k), \end{array} \tag{P($\beta/k$)}$$

where $I_{il}(\beta) := [(i+(l-1)/k)\cdot\beta, (i+l/k)\cdot\beta)$ and $c_{il}(\beta) := (\beta/k)^{-1}\cdot\int_{x\in I_{il}(\beta)} c(x)\,\mathrm{d}x$. One readily verifies that $p''(i,l) = p'(i)/k$, $i\in\mathbb{Z}$ and $l\in[k]$, is feasible in problem P($\beta/k$) and attains the same objective value as $p'$ in P($\beta$). The statement thus follows. □

We next show that problem P($\beta$) is feasible.

**Lemma 7.** *For any $\delta > 0$, there is $M \in \mathbb{R}$ such that P($\Delta f/k$) $\leq M$ for all $\varepsilon > 0$ and $k \in \mathbb{N}$.*

*Proof.* Fix $\delta > 0$ and denote by P$^0$($\Delta f$) the variant of P($\Delta f$) that replaces $\varepsilon$ with 0. We show that there exists $M\in\mathbb{R}$ such that P$^0$($\Delta f$) $\leq M$. The statement then follows since for any $\varepsilon > 0$ and $k\in\mathbb{N}$, we have P$^0$($\Delta f$) $\geq$ P($\Delta f$) $\geq$ P($\Delta f/k$), where the first inequality is direct and the second inequality is due to Lemma 6.

Consider the following solution to P$^0$($\Delta f$):

$$p(i) = \begin{cases} \dfrac{1}{2\lceil 1/(2\delta)\rceil} & \text{if } i\in\{-\lceil 1/(2\delta)\rceil,\ldots,\lceil 1/(2\delta)\rceil-1\}, \\ 0 & \text{otherwise} \end{cases} \qquad \forall i\in\mathbb{Z}$$

To confirm that $p$ is feasible in P$^0$($\Delta f$), first note that by construction, $p$ is a valid probability distribution. To see that $p$ also satisfies the DP constraints, note that in problem P$^0$($\Delta f$), these constraints simplify to

$$\sum_{i\in\mathbb{Z}}\mathbb{1}[I_i(\Delta f)\subseteq A]\cdot p(i) - \sum_{i\in\mathbb{Z}}\mathbb{1}[I_i(\Delta f)+\varphi\subseteq A]\cdot p(i)\leq\delta \quad \forall\varphi\in\{-\Delta f,0,\Delta f\},\ A\in\mathcal{F}(\Delta f).$$

Since the constraints associated with $\varphi = 0$ are vacuously satisfied, it is sufficient to investigate the cases where $\varphi = \pm\Delta f$. Consider the constraints associated with $\varphi = \Delta f$:

$$\sum_{i\in\mathbb{Z}}(\mathbb{1}[I_i(\Delta f)\subseteq A] - \mathbb{1}[I_{i+1}(\Delta f)\subseteq A])\cdot p(i)\leq\delta \qquad \forall A\in\mathcal{F}(\Delta f)$$

$$\iff \quad \sup_{A \in \mathcal{F}(\Delta f)} \left[ \sum_{i \in \mathbb{Z}} (\mathbb{1}[I_i(\Delta f) \subseteq A] - \mathbb{1}[I_{i+1}(\Delta f) \subseteq A]) \cdot p(i) \right] \leq \delta$$

The supremum in the second row is attained, among others, by the worst-case event $A^\star = I_{\lceil 1/(2\delta) \rceil - 1}(\Delta f) \in \mathcal{F}(\Delta f)$. Indeed, one readily observes that $i = \lceil 1/(2\delta) \rceil - 1$ is the only index for which $\mathbb{1}[I_i(\Delta f) \subseteq A] = 1$ and $\mathbb{1}[I_{i+1}(\Delta f) \subseteq A] = 0$. For $A = A^\star$, however, the DP constraint reduces to $p(\lceil 1/(2\delta) \rceil - 1) = 1/(2\lceil 1/(2\delta) \rceil) \leq \delta$, which is satisfied by construction. We thus conclude that $p$ satisfies all DP constraints of $\mathrm{P}^0(\Delta f)$ associated with $\varphi = \Delta f$ and $A \in \mathcal{F}(\Delta f)$. An analogous argument for $\varphi = -\Delta f$ shows that $p$ indeed satisfies all DP constraints of $\mathrm{P}^0(\Delta f)$.

The solution $p$ attains a finite objective value in $\mathrm{P}^0(\Delta f)$, finally, since

$$\sum_{i \in \mathbb{Z}} c_i(\Delta f) \cdot p(i) = \frac{1}{2\lceil 1/(2\delta) \rceil} \cdot \sum_{i = -\lceil 1/(2\delta) \rceil}^{\lceil 1/(2\delta) \rceil - 1} c_i(\Delta f) =: M < \infty.$$

Since $\mathrm{P}^0(\Delta f)$ is a minimization problem, $\mathrm{P}^0(\Delta f) \leq M$ thus follows. □

Lemma 7 implies that $\mathrm{P}(\Delta f / k)$ is feasible for any fixed $\varepsilon, \delta > 0$ and $k \in \mathbb{N}$.

#### 1.A.7.2 Monotonicity and Feasibility of $\mathrm{P}(L, \beta)$

We first show that problem $\mathrm{P}(L, \beta)$ is monotonically non-increasing in $L$ and monotonically non-decreasing in $\beta$ in the following sense.

**Lemma 8.** *For any $\varepsilon > 0$, $\delta > 0$, $L' \in \mathbb{N}$ and $\beta > 0$, the optimal value of problem $\mathrm{P}(L', \beta)$ satisfies $\mathrm{P}(L', \beta) \geq \mathrm{P}(L, \beta/k)$ for all $k \in \mathbb{N}$ and $L \geq L' \cdot k + k - 1$.*

*Proof.* The result trivially holds if $\mathrm{P}(L', \beta)$ is infeasible. We thus assume that $\mathrm{P}(L', \beta)$ is feasible, and we fix any $k \in \mathbb{N}$ as well as $L = L' \cdot k + k - 1$. We proceed in two steps. We first derive an upper bound $\mathrm{P}'(L, \beta/k)$ to $\mathrm{P}(L, \beta/k)$ in which the noise distribution has the same support as in $\mathrm{P}(L', \beta)$. We then show that $\mathrm{P}(L', \beta)$ bounds $\mathrm{P}'(L, \beta/k)$ from above. The result then follows from the fact that $\mathrm{P}(L, \beta/k)$ is monotonically non-increasing in $L$ for any fixed $\beta$ and $k$.

In view of the first step, we construct the upper bound $\mathrm{P}'(L, \beta/k)$ to problem $\mathrm{P}(L, \beta/k)$ by adding to $\mathrm{P}(L, \beta/k)$ the constraint that $p(i) = 0$ for $i = -(L' \cdot k + k - 1), \ldots, -(L' \cdot k + 1)$, that is, we remove the first $k - 1$ elements from the domain of $p$. This ensures that despite its finer interval granularity of $\beta/k$, the support of the noise distribution in problem $\mathrm{P}'(L, \beta/k)$ is the same as in the more coarsely discretized problem $\mathrm{P}(L', \beta)$, namely $[-L' \cdot \beta, (L' + 1) \cdot \beta)$.

As for the second step, note that the upper bound $\mathrm{P}'(L, \beta/k)$ can be formulated as

$$\begin{aligned}
\underset{p}{\text{minimize}} \quad & \sum_{i\in\mathbb{Z}}\sum_{l\in[k]} c_{il}(\beta)\cdot p(i,l) \\
\text{subject to} \quad & p : [\pm L'] \times [k] \mapsto \mathbb{R}_+, \ \sum_{i\in\mathbb{Z}}\sum_{l\in[k]} p(i,l) = 1 \\
& \sum_{i\in[\pm L']}\sum_{l\in[k]} \mathbb{1}[I_{il}(\beta) \subseteq A]\cdot p(i,l) \le e^{\varepsilon}\cdot \sum_{i\in[\pm L']}\sum_{l\in[k]} \mathbb{1}[I_{il}(\beta) + \varphi \subseteq A]\cdot p(i,l) + \delta \\
& \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \forall(\varphi, A) \in \mathcal{E}(\beta/k),
\end{aligned}$$

where $I_{il}(\beta) = [(i + (l-1)/k)\cdot\beta, (i + l/k)\cdot\beta)$ and $c_{il}(\beta) = (\beta/k)^{-1}\cdot\int_{x\in I_{il}(\beta)} c(x)\,\mathrm{d}x$. Fix any feasible solution $p'$ in problem $\mathrm{P}(L', \beta)$ . One readily observes that the solution $p''(i,l) = p'(i)/k$, $i \in [\pm L']$ and $l \in [k]$, is feasible in $\mathrm{P}'(L, \beta/k)$ and attains the same objective value as $p'$ in $\mathrm{P}(L', \beta)$. We thus conclude that $\mathrm{P}(L', \beta)$ bounds $\mathrm{P}(L, \beta/k)$ from above, as desired. □

We next bound the maximum constraint violation of a solution $p'$ in $\mathrm{P}(\beta)$ that is obtained by truncating any feasible solution $p$ in $\mathrm{P}(\beta)$ to a bounded domain. This will later enable us to determine values of $L$ that ensure the feasibility of $\mathrm{P}(L, \beta)$ for any fixed $\beta$.

**Lemma 9.** *Let $p$ be an arbitrary feasible solution to problem $\mathrm{P}(\beta)$ and fix $L \in \mathbb{N}$ such that $\sum_{i\in[\pm L]} p(i) \ge 1 - \tau$ for some $\tau > 0$. Construct another candidate solution $p'$ to $\mathrm{P}(\beta)$ where $p'(i) = 0$ for all $i \in \mathbb{Z}\backslash[\pm L]$, $p'(L) = \sum_{i\ge L} p(i)$ as well as $p'(-L) = \sum_{i\le -L} p(i)$, and $p'(i) = p(i)$ otherwise. Then $p'$ violates the DP constraints of problem $\mathrm{P}(\beta)$ by at most $(1 + e^{\varepsilon})\cdot\tau$, that is,*

$$\sup_{(\varphi,A)\in\mathcal{E}(\beta)} \left\{ \sum_{i\in\mathbb{Z}} \mathbb{1}[I_i(\beta) \subseteq A]\cdot p'(i) - e^{\varepsilon}\cdot\sum_{i\in\mathbb{Z}} \mathbb{1}[(I_i(\beta) + \varphi) \subseteq A]\cdot p'(i) - \delta \right\} \le (1 + e^{\varepsilon})\cdot\tau.$$

Note that the constant $L$ in the statement of Lemma 9 exists since for any probability measure $\gamma \in \mathcal{P}_0$ and any $\tau > 0$, there is $L' \in \mathbb{N}$ such that $\gamma([-L, L]) \ge 1 - \tau$ for all $L \ge L'$.

**Proof of Lemma 9.** Since $p$ is feasible in problem $\mathrm{P}(\beta)$, it satisfies

$$\sum_{i\in\mathbb{Z}} \mathbb{1}[I_i(\beta) \subseteq A]\cdot p(i) \le e^{\varepsilon}\cdot\sum_{i\in\mathbb{Z}} \mathbb{1}[(I_i(\beta) + \varphi) \subseteq A]\cdot p(i) + \delta \quad \forall(\varphi, A) \in \mathcal{E}(\beta).$$

On the other hand, for any $(\varphi, A) \in \mathcal{E}(\beta)$, the constructed solution $p'$ satisfies

$$\sum_{i\in\mathbb{Z}} \mathbb{1}[I_i(\beta) \subseteq A]\cdot p'(i) - e^{\varepsilon}\cdot\sum_{i\in\mathbb{Z}} \mathbb{1}[(I_i(\beta) + \varphi) \subseteq A]\cdot p'(i) - \delta$$

$$= \sum_{i\in\mathbb{Z}} \mathbb{1}[I_i(\beta) \subseteq A] \cdot [p(i) + (p'(i) - p(i))] - e^{\varepsilon} \cdot \sum_{i\in\mathbb{Z}} \mathbb{1}[(I_i(\beta) + \varphi) \subseteq A] \cdot [p(i) + (p'(i) - p(i))] - \delta$$

$$= \underbrace{\sum_{i\in\mathbb{Z}} \mathbb{1}[I_i(\beta) \subseteq A] \cdot p(i) - e^{\varepsilon} \cdot \sum_{i\in\mathbb{Z}} \mathbb{1}[(I_i(\beta) + \varphi) \subseteq A] \cdot p(i) - \delta +}_{\leq 0 \text{ as } p \text{ is feasible in } (\mathrm{P}(\beta))}$$

$$\underbrace{\sum_{i\in\mathbb{Z}} \mathbb{1}[I_i(\beta) \subseteq A] \cdot (p'(i) - p(i))}_{\leq\tau} - e^{\varepsilon} \cdot \underbrace{\sum_{i\in\mathbb{Z}} \mathbb{1}[(I_i(\beta) + \varphi) \subseteq A] \cdot (p'(i) - p(i))}_{\geq -\tau} \leq (1 + e^{\varepsilon}) \cdot \tau,$$

which implies the statement. □

We can now prove the feasibility of $\mathrm{P}(L, \beta)$.

**Lemma 10.** *For any $\varepsilon > 0$, $\delta > 0$ and $\beta > 0$, there exists $L' \in \mathbb{N}$ such that problem $\mathrm{P}(L, \beta)$ is feasible for all $L \geq L'$.*

*Proof.* Denote by $\mathrm{P}_{\delta/2}(\beta)$ and $\mathrm{P}_{\delta/2}(L, \beta)$ the variants of $\mathrm{P}(\beta)$ and $\mathrm{P}(L, \beta)$ that replace $\delta$ with $\delta/2$, respectively. Fix any feasible solution $p$ in problem $\mathrm{P}_{\delta/2}(\beta)$, whose existence is guaranteed by Lemma 7, and choose $L \in \mathbb{N}$ large enough so that $\sum_{i\in[\pm L]} p(i) \geq 1 - (\delta/2)/(1 + e^{\varepsilon})$. Lemma 9 allows us to construct a solution $p'$ from $p$ that violates the DP constraints of $\mathrm{P}_{\delta/2}(L, \beta)$ by at most $\delta/2$. By construction, $p'$ is thus feasible in $\mathrm{P}(L, \beta)$, and the statement follows. □

### 1.A.7.3 Proof of Theorem 1

To show our convergence result, we define the following auxiliary problem:

$$\begin{array}{ll} \underset{p}{\text{minimize}} & \displaystyle\sum_{i\in[\pm(L+\Delta f/\beta)]} c_i(\beta) \cdot p(i) \\ \text{subject to} & p : [\pm(L + \Delta f/\beta)] \mapsto \mathbb{R}_+, \quad \displaystyle\sum_{i\in[\pm(L+\Delta f/\beta)]} p(i) = 1 \\ & \displaystyle\sum_{i\in[\pm(L+\Delta f/\beta)]} \mathbb{1}[I_i(\beta) \subseteq A] \cdot p(i) \leq e^{\varepsilon} \cdot \sum_{i\in[\pm(L+\Delta f/\beta)]} \mathbb{1}[I_i(\beta) + \varphi \subseteq A] \cdot p(i) + \delta \\ & \hfill \forall(\varphi, A) \in \mathcal{E}(L, \beta). \end{array} \tag{M(L,\beta)}$$

In the following, we will show that *(i)* problem $\mathrm{M}(L, \beta)$ differs from $\mathrm{P}(L, \beta)$ only in the domain of the decision variable $p$; *(ii)* problem $\mathrm{M}(L, \beta)$ differs from the strong dual of $\mathrm{D}(L, \beta)$ only in the objective coefficients; and *(iii)* the relationship $\mathrm{P}(L, \beta) \geq \mathrm{M}(L, \beta) \geq \mathrm{D}(L, \beta)$ holds for all

$L \in \mathbb{N}$ and $\beta > 0$. Thus, instead of analyzing the convergence of $\mathrm{P}(L, \beta)$ and $\mathrm{D}(L, \beta)$ directly, we can analyze separately the convergence of $\mathrm{P}(L, \beta)$ and $\mathrm{M}(L, \beta)$ (*cf.* Lemma 12) as well as of $\mathrm{M}(L, \beta)$ and the strong dual of $\mathrm{D}(L, \beta)$ (*cf.* Lemma 13).

Since $\mathrm{M}(L, \beta)$ coincides with $\mathrm{P}(L, \beta)$ except for the additional decision variables $p(i)$, $i \in [\pm(L + \Delta f/\beta)] \setminus [\pm L]$, we have $\mathrm{P}(L, \beta) \geq \mathrm{M}(L, \beta)$. To show convergence of both problems (*cf.* Lemma 12), we need to ensure that these additional decision variables take sufficiently small values in optimal solutions to $\mathrm{M}(L, \beta)$. This is guaranteed by the next result.

**Lemma 11.** *For any $\varepsilon > 0$, $\delta > 0$ and $\tau > 0$, there exists $L' \in \mathbb{N}$ such that for all $k \in \mathbb{N}$ and $L \geq L' \cdot k + k - 1$, any optimal solution $p^\star$ to $\mathrm{M}(L, \Delta f/k)$ satisfies* $\sum\limits_{i \in [\pm(L+k)] \setminus [\pm L]} p^\star(i) < \tau$.

*Proof.* Fix $\varepsilon > 0$, $\delta > 0$ and $\tau > 0$. By Lemma 10, there is $L_1 \in \mathbb{N}$ such that problem $\mathrm{P}(L, \Delta f)$ is feasible for all $L \geq L_1$. Select $L_2 \in \mathbb{N}$ large enough such that

$$c(x) > \frac{\mathrm{P}(L_1, \Delta f)}{\tau} \quad \forall x \in \mathbb{R} : \ |x| \geq L_2 \cdot \Delta f; \tag{16}$$

such values exist due to Assumption 1 *(b)*. We claim that the statement of the lemma holds for $L' = \max\{L_1, L_2\}$. To see this, fix any $k \in \mathbb{N}$ and $L \geq L' \cdot k + k - 1$.

Take any optimal solution $p^\star : [\pm(L + k)] \mapsto \mathbb{R}_+$ to problem $M(L, \Delta f/k)$ and assume to the contrary that $\sum_{i \in [\pm(L+k)] \setminus [\pm L]} p^\star(i) \geq \tau$. We then observe that

$$\begin{aligned} \sum_{i \in [\pm(L+k)]} c_i(\Delta f/k) \cdot p^\star(i) &= \sum_{i \in [\pm L]} c_i(\Delta f/k) \cdot p^\star(i) + \sum_{i \in [\pm(L+k)] \setminus [\pm L]} c_i(\Delta f/k) \cdot p^\star(i) \\ &> \sum_{i \in [\pm(L+k)] \setminus [\pm L]} \frac{\mathrm{P}(L_1, \Delta f)}{\tau} \cdot p^\star(i) \\ &\geq \tau \cdot \frac{\mathrm{P}(L_1, \Delta f)}{\tau} = \mathrm{P}(L_1, \Delta f) \geq \mathrm{P}(L, \Delta f/k) \geq \mathrm{M}(L, \Delta f/k). \end{aligned} \tag{17}$$

Here, the first inequality holds since $c_i(\Delta f/k) \geq 0$ for all $i \in \mathbb{Z}$ as well as

$$c_i(\Delta f/k) \ = \ \left(\frac{\Delta f}{k}\right)^{-1} \cdot \int_{x \in I_i(\Delta f/k)} c(x) \mathrm{d}x \ > \ \left(\frac{\Delta f}{k}\right)^{-1} \cdot \int_{x \in I_i(\Delta f/k)} \frac{\mathrm{P}(L_1, \Delta f)}{\tau} \mathrm{d}x = \frac{\mathrm{P}(L_1, \Delta f)}{\tau}$$

for all $i \in [\pm(L + k)] \setminus [\pm L]$. The second inequality in (17) follows from our earlier assumption that $\sum_{i \in [\pm(L+k)] \setminus [\pm L]} p^\star(i) \geq \tau$. The third inequality in (17) is due to Lemma 8 and the fact that $L \geq L_1 \cdot k + k - 1$, and the last inequality in (17) holds by construction of problem $\mathrm{M}(L, \Delta f/k)$. We thus conclude that $p^\star$ cannot be optimal in problem $M(L, \Delta f/k)$, which yields the desired contradiction. □

Lemma 11 allows us to prove the convergence between problems $\mathrm{M}(L, \beta)$ and $\mathrm{P}(L, \beta)$.

**Lemma 12.** *For any $\varepsilon > 0$, $\delta > 0$ and $\xi > 0$, there exists $L' \in \mathbb{N}$ such that $\mathrm{P}(L, \Delta f/k) - \mathrm{M}(L, \Delta f/k) \leq \xi$ for all $k \in \mathbb{N}$ and all $L \geq L' \cdot k + k - 1$.*

Intuitively, Lemma 12 shows that $\mathrm{P}(L, \beta) - \mathrm{M}(L, \beta) \to 0$ for any $\beta > 0$ as long as $L$ grows sufficiently quickly relative to $1/\beta$. Recall that the size of the support of the noise distribution $\gamma$ is $L \cdot \beta$. Thus, $\mathrm{P}(L, \beta) - \mathrm{M}(L, \beta) \to 0$ for any $\beta > 0$ as long as the support of $\gamma$ grows large.

**Proof of Lemma 12.** Fix $\varepsilon > 0$, $\delta > 0$ and $\xi > 0$, select any $\alpha \in (0, \delta)$, set $\hat{\delta} = \delta - \alpha$ and denote by $\mathrm{P}_{\hat{\delta}}(L, \Delta f/k)$ the variant of $\mathrm{P}(L, \Delta f/k)$ that replaces $\delta$ with $\hat{\delta}$. We invoke Lemma 10 to select $L_0 \in \mathbb{N}$ so that $\mathrm{P}_{\hat{\delta}}(L, \Delta f)$ is feasible for all $L \geq L_0$, and we denote by $M$ the optimal value of $\mathrm{P}_{\hat{\delta}}(L_0, \Delta f)$. We next invoke Lemma 8 to conclude that $\mathrm{P}_{\hat{\delta}}(L, \Delta f/k)$ remains feasible with an optimal value bounded from above by $M$ for all $k \in \mathbb{N}$ and all $L \geq L_0 \cdot k + k - 1$.

Set $\tau = \xi \cdot \alpha / M$. The remainder of the proof shows the statement in four steps. Step 1 constructs a solution $p_\tau$ to problem $\mathrm{P}(L, \Delta f/k)$ whose expected loss is bounded from above by the optimal value of $\mathrm{M}(L, \Delta f/k)$, but that may violate the DP constraints in $\mathrm{P}(L, \Delta f/k)$ by up to $\tau$. Step 2 then constructs a convex combination $p^\star$ of $p_\tau$ and $p_{\hat{\delta}}$, where the latter is an optimal solution to problem $\mathrm{P}_{\hat{\delta}}(L, \Delta f/k)$. Step 3 shows that the convex combination $p^\star$ is feasible in $\mathrm{P}(L, \Delta f/k)$, and Step 4 shows that the expected loss of $p^\star$ in $\mathrm{P}(L, \Delta f/k)$ is bounded from above by $\mathrm{M}(L, \Delta f/k) + \xi$, as desired.

In view of Step 1, note that problem $\mathrm{M}(L, \Delta f/k)$ is feasible by construction, and it is bounded since the objective coefficients are non-negative. Lemma 11 then ensures the existence of $L_1 \in \mathbb{N}$ such that any optimal solution to $\mathrm{M}(L, \Delta f/k)$, $L \geq L_1 \cdot k + k - 1$, places a probability of strictly less than $\tau/(1 + e^\varepsilon)$ outside the index range $[\pm L]$. Select $L_2 \in \mathbb{N}$ large enough such that

$$c(x) \geq \max_{x' : |x'| \leq \Delta f} c(x') \quad \forall x \in \mathbb{R} : \ |x| \geq L_2 \cdot \Delta f; \tag{18}$$

such values exist due to Assumption 1 *(b)*. We claim that $L' = \max\{L_0, L_1, L_2\}$ satisfies the statement of this lemma. Take any $k \in \mathbb{N}$ and any $L \geq L' \cdot k + k - 1$, and denote by $p_\mathrm{M}$ an optimal solution to problem $\mathrm{M}(L, \Delta f/k)$, which satisfies $\sum_{i \in [\pm(L+k)] \setminus [\pm L]} p_\mathrm{M}(i) < \tau/(1 + e^\varepsilon)$ by the selection of $L'$. Construct a new truncated solution $p_\tau$ via

$$p_\tau(i) = \begin{cases} 0 & \text{if } i \in [\pm(L+k)] \setminus [\pm L] \\ p_{\mathrm{M}}(0) + \sum_{i'>L} p_{\mathrm{M}}(i') + \sum_{i'<-L} p_{\mathrm{M}}(i') & \text{if } i = 0 \\ p_{\mathrm{M}}(i) & \text{otherwise} \end{cases} \qquad \forall i \in [\pm(L+k)].$$

We then have

$$\begin{aligned} \sum_{i \in [\pm L]} c_i(\Delta f/k) \cdot p_\tau(i) &= \sum_{i \in [\pm L]} c_i(\Delta f/k) \cdot p_{\mathrm{M}}(i) + c_0(\Delta f/k) \cdot \left( \sum_{i'>L} p_{\mathrm{M}}(i') + \sum_{i'<-L} p_{\mathrm{M}}(i') \right) \\ &\leq \sum_{i \in [\pm(L+k)]} c_i(\Delta f/k) \cdot p_{\mathrm{M}}(i), \end{aligned} \tag{19}$$

where the inequality holds because (18) implies

$$c_0(\Delta f/k) = (\Delta f/k)^{-1} \int_{x \in I_0(\Delta f/k)} c(x) \leq (\Delta f/k)^{-1} \int_{x \in I_0(\Delta f/k)} \max_{x' : |x'| \leq \Delta f} c(x') = \max_{x' : |x'| \leq \Delta f} c(x') \leq c_i(\Delta f/k)$$

for all $i$ satisfying $|i| > L$. This shows that the truncated solution $p_\tau$ achieves a weakly smaller objective value in problem $\mathrm{P}(L, \Delta f/k)$ than the optimal value of $\mathrm{M}(L, \Delta f/k)$. However, a similar reasoning as in the proof of Lemma 9 shows that $p_\tau$ can violate the DP constraints in $\mathrm{P}(L, \Delta f/k)$ by up to $\tau$.

As for Step 2, we define $p_{\hat{\delta}}$ as an optimal solution to problem $\mathrm{P}_{\hat{\delta}}(L, \Delta f/k)$, which exists as $\mathrm{P}_{\hat{\delta}}(L, \Delta f/k)$ is feasible by the selection of $L$. We then construct the solution $p^\star$ via

$$p^\star = \lambda \cdot p_\tau + (1 - \lambda) \cdot p_{\hat{\delta}} \quad \text{for } \lambda = \frac{\alpha}{\tau + \alpha}.$$

The next two steps will show that $p^\star$ is feasible in problem $\mathrm{P}(L, \Delta f/k)$ and that its expected loss is bounded from above by $\mathrm{M}(L, \Delta f/k) + \xi$.

In view of Step 3, first notice that $p^\star(i) = 0$ for all $i \in [\pm(L+k)] \setminus [\pm L]$ since $p^\star$ is a convex combination of two solutions, neither of which places positive probability on indices $i \in [\pm(L+k)] \setminus [\pm L]$. Consider now any DP constraint $(\varphi, A)$ in problem $\mathrm{P}(L, \Delta f/k)$. We observe that

$$\sum_{i \in [\pm L]} \mathbb{1}[I_i(\Delta f/k) \subseteq A] \cdot p^\star(i) - e^\varepsilon \cdot \sum_{i \in [\pm L]} \mathbb{1}[I_i(\Delta f/k) + \varphi \subseteq A] \cdot p^\star(i)$$

$$
\begin{aligned}
&= \lambda \cdot \left( \sum_{i \in [\pm L]} \mathbb{1}[I_i(\Delta f/k) \subseteq A] \cdot p_\tau(i) - e^\varepsilon \cdot \sum_{i \in [\pm L]} \mathbb{1}[I_i(\Delta f/k) + \varphi \subseteq A] \cdot p_\tau(i) \right) + \\
&\qquad (1-\lambda) \cdot \left( \sum_{i \in [\pm L]} \mathbb{1}[I_i(\Delta f/k) \subseteq A] \cdot p_{\hat{\delta}}(i) - e^\varepsilon \cdot \sum_{i \in [\pm L]} \mathbb{1}[I_i(\Delta f/k) + \varphi \subseteq A] \cdot p_{\hat{\delta}}(i) \right) \\
&\leq \lambda \cdot (\delta + \tau) + (1-\lambda) \cdot \hat{\delta} \\
&= \lambda \cdot (\delta + \tau) + (1-\lambda) \cdot (\delta - \alpha) = \lambda \cdot (\tau + \alpha) + \delta - \alpha = \delta,
\end{aligned}
$$

where the first equality uses the definition of $p^\star$, the inequality holds since $p_\tau$ violates the DP constraints in problem $\mathrm{P}(L, \beta)$ by up to $\tau$ and $p_{\hat{\delta}}$ is feasible in $\mathrm{P}_{\hat{\delta}}(L, \Delta f/k)$, and the final equalities follow from the definition of $\hat{\delta}$, rearranging terms and from the definition of $\lambda$, respectively. We thus conclude that $p^\star$ satisfies the DP constraint $(\varphi, A)$ in problem $\mathrm{P}(L, \Delta f/k)$, and since the choice of $(\varphi, A)$ was arbitrary, $p^\star$ must indeed be feasible in $\mathrm{P}(L, \Delta f/k)$.

As for Step 4, first note that the solution $p^\star$ achieves an objective value of $\sum_{i \in [\pm L]} c_i(\Delta f/k) \cdot p^\star(i)$ in $\mathrm{P}(L, \Delta f/k)$, which itself satisfies the following due to (19):

$$
\begin{aligned}
\sum_{i \in [\pm L]} c_i(\Delta f/k) \cdot p^\star(i) &= \lambda \sum_{i \in [\pm L]} c_i(\Delta f/k) \cdot p_\tau(i) + (1-\lambda) \cdot \sum_{i \in [\pm L]} c_i(\Delta f/k) \cdot p_{\hat{\delta}}(i) \\
&\leq \lambda \sum_{i \in [\pm (L+k)]} c_i(\Delta f/k) \cdot p_{\mathrm{M}}(i) + (1-\lambda) \cdot \sum_{i \in [\pm L]} c_i(\Delta f/k) \cdot p_{\hat{\delta}}(i). \qquad (20)
\end{aligned}
$$

We can use (20) to bound the difference $\mathrm{P}(L, \Delta f/k) - \mathrm{M}(L, \Delta f/k)$ as follows:

$$
\begin{aligned}
\mathrm{P}(L, \Delta f/k) - \mathrm{M}(L, \Delta f/k) &\leq \sum_{i \in [\pm L]} c_i(\Delta f/k) \cdot p^\star(i) - \sum_{i \in [\pm (L+k)]} c_i(\Delta f/k) \cdot p_{\mathrm{M}}(i) \\
&\leq (1-\lambda) \left( \sum_{i \in [\pm L]} c_i(\Delta f/k) \cdot p_{\hat{\delta}}(i) - \sum_{i \in [\pm (L+k)]} c_i(\Delta f/k) \cdot p_{\mathrm{M}}(i) \right) \\
&\leq (1-\lambda) M = \frac{\tau}{\tau + \alpha} \cdot M = \xi.
\end{aligned}
$$

Here, the first inequality holds since $p^\star$ is feasible in $\mathrm{P}(L, \Delta f/k)$ and $p_{\mathrm{M}}$ is optimal in $\mathrm{M}(L, \Delta f/k)$. The second inequality employs (20). The third inequality bounds the objective value of $p_{\hat{\delta}}$ from above by $M$ and the objective value of $p_{\mathrm{M}}$ from below by 0, respectively. The two identities, finally, follow from substituting back the definitions of $\lambda$ and $\tau$, respectively. □

We next prove the convergence between problems $\mathrm{M}(L, \beta)$ and $\mathrm{D}(L, \beta)$.

**Lemma 13.** *For any $\varepsilon > 0$, $\delta > 0$, $\xi > 0$ and $\Lambda \in \mathbb{N}$, there exists $k' \in \mathbb{N}$ such that $\mathrm{M}(\Lambda \cdot k, \Delta f/k) - \mathrm{D}(\Lambda \cdot k, \Delta f/k) \leq \xi$ for all $k \geq k'$.*

Intuitively, Lemma 13 shows that $\mathrm{M}(L, \beta) - \mathrm{D}(L, \beta) \to 0$ for any fixed size of the support of the noise distribution $\gamma$ as long as the discretization granularity $\beta$ vanishes to zero.

**Proof of Lemma 13.** Fix $\varepsilon > 0$, $\delta > 0$, $\xi > 0$ and $\Lambda \in \mathbb{N}$. We will show that there is $k' \in \mathbb{N}$ such that $\mathrm{M}(\Lambda \cdot k, \Delta f/k) - \overline{D}(\Lambda \cdot k, \Delta f/k) \leq \xi$ for all $k \geq k'$, where $\overline{D}(\Lambda \cdot k, \Delta f/k)$ denotes the dual to $\mathrm{D}(\Lambda \cdot k, \Delta f/k)$. Indeed, strong duality holds between $\mathrm{D}(\Lambda \cdot k, \Delta f/k)$ and $\overline{D}(\Lambda \cdot k, \Delta f/k)$ since $(\theta, \psi)$ with $\theta = \min\{\underline{c}_i(\Delta f/k) : i \in [\pm(\Lambda \cdot k + k)]\}$ and $\psi(\varphi, A) = 0$ for all $\mathcal{E}(\Lambda \cdot k, \Delta f/k)$ is feasible in $\mathrm{D}, \Delta f/k)$. The dual problem $\overline{D}(\Lambda \cdot k, \Delta f/k)$ can be formulated as

$$
\begin{aligned}
\underset{p}{\text{minimize}} \quad & \sum_{i \in [\pm(\Lambda \cdot k + k)]} \underline{c}_i(\Delta f/k) \cdot p(i) \\
\text{subject to} \quad & p : [\pm(\Lambda \cdot k + k)] \mapsto \mathbb{R}_+, \quad \sum_{i \in [\pm(\Lambda \cdot k + k)]} p(i) = 1 \\
& \sum_{i \in [\pm(\Lambda \cdot k + k)]} \mathbb{1}[I_i(\Delta f/k) \subseteq A] \cdot p(i) \leq e^{\varepsilon} \cdot \sum_{i \in [\pm(\Lambda \cdot k + k)]} \mathbb{1}[I_i(\Delta f/k) + \varphi \subseteq A] \cdot p(i) + \delta \\
& \hspace{16em} \forall (\varphi, A) \in \mathcal{E}(\Lambda \cdot k, \Delta f/k).
\end{aligned}
\tag{$\overline{\mathrm{D}}(\Lambda \cdot k, \Delta f/k)$}
$$

Note that $\overline{\mathrm{D}}(\Lambda \cdot k, \Delta f/k)$ and $\mathrm{M}(\Lambda \cdot k, \Delta f/k)$ only differ in their objective coefficients $\underline{c}_i(\Delta f/k)$ and $c_i(\Delta f/k)$, respectively. We now show that the difference between those two coefficient sets can be made arbitrarily small, uniformly across all $i \in [\pm(\Lambda \cdot k + k)]$, by increasing $k$. Indeed, the loss function $c$ is continuous by Assumption 1 *(a)*, and it is therefore uniformly continuous over the (closure of the) finite interval $\bigcup\{I_i(\Delta f/k) : i \in [\pm(\Lambda \cdot k + k)]\}$ by the Heine-Cantor theorem. (Note in particular that for any $k$, this interval is contained in the set $[-(\Lambda + 1) \cdot \Delta f, (\Lambda + 2) \cdot \Delta f)$ that is independent of $k$, which justifies our use of the Heine-Cantor theorem.) For the selected $\xi > 0$, we can thus find $k' \in \mathbb{N}$ such that for all $k \geq k'$, we have

$$
\begin{aligned}
c_i(\Delta f/k) - \underline{c}_i(\Delta f/k) \; &= \; \left(\frac{\Delta f}{k}\right)^{-1} \cdot \int_{x \in I_i(\Delta f/k)} c(x) \, \mathrm{d}x - \inf_{x \in I_i(\Delta f/k)} c(x) \\
&\leq \; \sup_{x \in I_i(\Delta f/k)} c(x) - \inf_{x \in I_i(\Delta f/k)} c(x) \; \leq \; \xi,
\end{aligned}
$$

uniformly across all $i \in [\pm(\Lambda \cdot k + k)]$. Here, the identity replaces $c_i(\Delta f/k)$ and $\underline{c}_i(\Delta f/k)$ with their respective definitions, the first inequality exploits that $c(x) \leq \sup_{x \in I_i(\Delta f/k)} c(x)$ for all

$x \in I_i(\Delta f/k)$, and the last inequality makes use of the uniform continuity of $c$.

To bound $\mathrm{M}(\Lambda \cdot k, \Delta f/k) - \overline{\mathrm{D}}(\Lambda \cdot k, \Delta f/k)$, take any optimal solution $p^\star$ to problem $\overline{\mathrm{D}}(\Lambda \cdot k, \Delta f/k)$ and notice that $p^\star$ is feasible in $\mathrm{M}(\Lambda \cdot k, \Delta f/k)$ since both problems only differ in their objective coefficients. We thus have

$$\mathrm{M}(\Lambda \cdot k, \Delta f/k) - \overline{\mathrm{D}}(\Lambda \cdot k, \Delta f/k) \;\leq\; \sum_{i \in [\pm(\Lambda \cdot k + k)]} [c_i(\Delta f/k) - \underline{c}_i(\Delta f/k)] \cdot p^\star(i) \;\leq\; \xi,$$

where the first inequality holds since $p^\star$ is optimal in $\overline{\mathrm{D}}(\Lambda \cdot k, \Delta f/k)$ but feasible (and possibly not optimal) in $\mathrm{M}(\Lambda \cdot k, \Delta f/k)$, and the second inequality is due to our earlier uniform bound on $c_i(\Delta f/k) - \underline{c}_i(\Delta f/k)$ and the fact that $p^\star$ is a probability distribution. □

**Proof of Theorem 1.** Fix $\xi > 0$ as well as any $\xi_1, \xi_2 > 0$ satisfying $\xi_1 + \xi_2 = \xi$. Since

$$\mathrm{P}(\Lambda \cdot k, \Delta f/k) - \mathrm{D}(\Lambda \cdot k, \Delta f/k) = [\mathrm{P}(\Lambda \cdot k, \Delta f/k) - \mathrm{M}(\Lambda \cdot k, \Delta f/k)] + [\mathrm{M}(\Lambda \cdot k, \Delta f/k) - \mathrm{D}(\Lambda \cdot k, \Delta f/k)]\,,$$

for any $\Lambda \in \mathbb{N}$ and $k \in \mathbb{N}$, it suffices to show that there is $\Lambda' \in \mathbb{N}$ and $k' \in \mathbb{N}$ such that

$$\mathrm{P}(\Lambda \cdot k, \Delta f/k) - \mathrm{M}(\Lambda \cdot k, \Delta f/k) \leq \xi_1 \quad \text{and} \quad \mathrm{M}(\Lambda \cdot k, \Delta f/k) - \mathrm{D}(\Lambda \cdot k, \Delta f/k) \leq \xi_2 \tag{21}$$

simultaneously hold for all $\Lambda \geq \Lambda'$ and $k \geq k'$.

In view of the first inequality in (21), we invoke Lemma 12 to select $L' \in \mathbb{N}$ such that $\mathrm{P}(L, \Delta f/k) - \mathrm{M}(L, \Delta f/k) \leq \xi_1$ for all $k \in \mathbb{N}$ and all $L \geq L' \cdot k + k - 1$. Fix $\Lambda' = L' + 1$ and notice that $\Lambda' \cdot k = (L' + 1) \cdot k \geq L' \cdot k + k - 1$. For our choice of $\Lambda'$, we thus have $\mathrm{P}(\Lambda' \cdot k, \Delta f/k) - \mathrm{M}(\Lambda' \cdot k, \Delta f/k) \leq \xi_1$ for all $k \in \mathbb{N}$. As for the second inequality in (21), we invoke Lemma 13 to select $k' \in \mathbb{N}$ such that $\mathrm{M}(\Lambda' \cdot k, \Delta f/k) - \mathrm{D}(\Lambda' \cdot k, \Delta f/k) \leq \xi_2$ for all $k \geq k'$. So far, we have shown that there exists $\Lambda' \in \mathbb{N}$ and $k' \in \mathbb{N}$ such that $\mathrm{P}(\Lambda' \cdot k, \Delta f/k) - \mathrm{D}(\Lambda' \cdot k, \Delta f/k) \leq \xi$ holds for all $k \geq k'$. Since we have $\mathrm{P}(\Lambda' \cdot k, \Delta f/k) \geq \mathrm{P}(\Lambda \cdot k, \Delta f/k)$ and $\mathrm{D}(\Lambda' \cdot k, \Delta f/k) \leq \mathrm{D}(\Lambda \cdot k, \Delta f/k)$ for all $\Lambda \geq \Lambda'$, we can conclude the proof. □

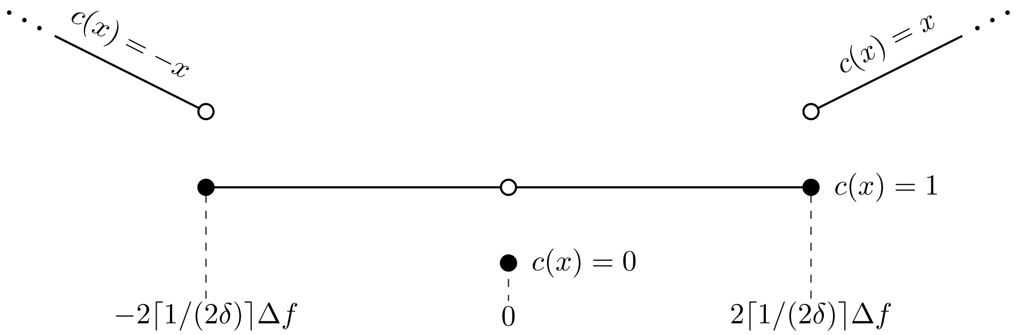


Figure 7: *Visualization of the loss function c in Section 1.A.8.1.*

### 1.A.8 Proof of Proposition 4

In the following, we present two counterexamples that jointly prove the statement of Proposition 4. Both examples rely on the following strong dual of D$(L, \beta)$:

$$\begin{aligned}
&\underset{p}{\text{minimize}} && \sum_{i \in [\pm(L+\Delta f/\beta)]} \underline{c}_i(\beta) \cdot p(i) \\
&\text{subject to} && p : [\pm(L + \Delta f/\beta)] \mapsto \mathbb{R}_+, \quad \sum_{i \in [\pm(L+\Delta f/\beta)]} p(i) = 1 \\
& && \sum_{i \in [\pm(L+\Delta f/\beta)]} \mathbb{1}[I_i(\beta) \subseteq A] \cdot p(i) \leq e^{\varepsilon} \cdot \sum_{i \in [\pm(L+\Delta f/\beta)]} \mathbb{1}[I_i(\beta) + \varphi \subseteq A] \cdot p(i) + \delta \\
& && \hspace{20em} \forall (\varphi, A) \in \mathcal{E}(L, \beta)
\end{aligned} \tag{22}$$

This dual has been derived earlier in the proof of Lemma 13; the version above is adapted slightly to match the notation of our counterexamples.

#### 1.A.8.1 Example 1: Violation of Assumption 1 *(a)*

Fix any $\Delta f, \varepsilon, \delta > 0$ and consider the following loss function $c$:

$$c(x) = \begin{cases} 0 & \text{if } x = 0, \\ 1 & \text{if } x \neq 0 \text{ and } |x| \leq 2 \cdot \lceil 1/(2\delta) \rceil \cdot \Delta f, \\ 1 + |x| & \text{otherwise.} \end{cases}$$

Figure 7 shows that $c$ satisfies Assumption 1 *(b)* since for any $r \in \mathbb{R}$ we have $c(x) \geq r$ for all $x$ satisfying $|x| \geq \max\{2\lceil 1/(2\delta)\rceil \Delta f, r\}$. However, $c$ violates the continuity condition of Assumption 1 at $x = 0$ and $x = \pm 2\lceil 1/(2\delta)\rceil \Delta f$. In the following, we will show that for this loss function, P$(L, \beta)$ and D$(L, \beta)$ differ by at least $\delta/2$ for all $L \in \mathbb{N}$ and $\beta > 0$. We do so in two steps: We first show that P$(\beta) \geq 1$, and we subsequently argue that D$(L, \beta) \leq 1 - \delta/2$. The statement then follows since P$(L, \beta) \geq$ P$(\beta)$.

To see that P$(\beta) \geq 1$, we note that the objective coefficients in P$(\beta)$ satisfy

$$c_i(\beta) = \beta^{-1} \cdot \int_{x \in I_i(\beta)} c(x) \, \mathrm{d}x \geq 1 \quad \forall i \in \mathbb{Z},$$

where the inequality holds since $c(x) \geq 1$ almost everywhere (except at $x = 0$). The claim then follows from the fact that the objective function in P$(\beta)$ constitutes a convex combination of those objective coefficients.

To show that D$(L, \beta) \leq 1 - \delta/2$, we can without loss of generality assume that $L \geq 2 \cdot \lceil 1/(2\delta)\rceil \cdot k$ for $k \in \mathbb{N}$ satisfying $\beta = \Delta f / k$. Indeed, the result then follows for all smaller values of $L$ due to the monotonicity of D$(L, \beta)$ in $L$. To see that D$(L, \beta) \leq 1 - \delta/2$, we construct a feasible solution $p$ to the strong dual (22) of D$(L, \beta)$ that attains the objective value $1 - \delta/2$. To this end, define $\mathcal{I}^k := \{-2 \cdot \lceil 1/(2\delta)\rceil \cdot k, \ldots, 2 \cdot \lceil 1/(2\delta)\rceil \cdot k - 1\}$ and set

$$p(i) \;=\; \begin{cases} \delta/2 & \text{if } i = 0, \\ \dfrac{1 - \delta/2}{4\lceil 1/(2\delta)\rceil k - 1} & \text{if } i \in \mathcal{I}^k \setminus \{0\}, \\ 0 & \text{if } i \in [\pm(L + \Delta f/\beta)] \setminus \mathcal{I}^k \end{cases} \qquad \forall i \in \mathbb{Z}.$$

Note that $p$ indeed constitutes a probability distribution over $i \in \mathcal{Z}$, and that $p(i) = 0$ for all $i \notin [\pm L]$ since $L \geq 2 \cdot \lceil 1/(2\delta)\rceil \cdot k$. Moreover, we have

$$\sum_{i \in [\pm(L+\Delta f/\beta)]} \underline{c}_i(\beta) \cdot p(i) = \sum_{i \in \mathcal{I}^k} \underline{c}_i(\beta) \cdot p(i) = \underline{c}_0(\beta) \cdot p(0) + \sum_{i \in \mathcal{I}^k \setminus \{0\}} p(i) = 1 - \delta/2,$$

where the last identity follows from the fact that $\underline{c}_0(\beta) = \inf\{c(x) : x \in I_0(\beta)\} = 0$ since $0 \in I_0(\beta)$ as well as $\underline{c}_i(\beta) = 1$ for all $i \in \mathcal{I}^k \setminus \{0\}$ since $|x| \leq 2 \cdot \lceil 1/(2\delta)\rceil \cdot \Delta f$ for all $x \in I_i(\beta)$ and $i \in \mathcal{I}^k \setminus \{0\}$. To see that $p$ satisfies the constraints in (22), fix any $(\varphi, A) \in \mathcal{E}(L, \beta)$ and observe that

$$\sum_{i \in [\pm(L+\Delta f/\beta)]} \mathbb{1}[I_i(\beta) \subseteq A] \cdot p(i) - e^{\varepsilon} \cdot \sum_{i \in [\pm(L+\Delta f/\beta)]} \mathbb{1}[I_i(\beta) + \varphi \subseteq A] \cdot p(i)$$

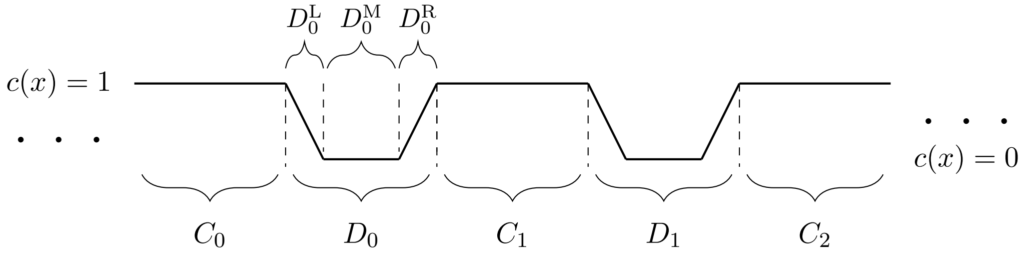


Figure 8: *Visualization of the loss function c in Section 1.A.8.2.*

$$
\begin{aligned}
&\leq \sum_{i \in [\pm(L+\Delta f/\beta)]} (\mathbb{1}[I_i(\beta) \subseteq A] - \mathbb{1}[I_i(\beta) + \varphi \subseteq A]) \cdot p(i) \\
&\leq \sum_{i \in [\pm L]} (\mathbb{1}[I_i(\beta) \subseteq A] - \mathbb{1}[I_i(\beta) + \varphi \subseteq A]) \cdot p(i) \\
&\leq \sum_{i \in [\pm L]} \mathbb{1}[I_i(\beta) \subseteq A \wedge I_i(\beta) + \varphi \not\subseteq A] \cdot p(i) \\
&\leq \sum_{i \in [\pm L]} \mathbb{1}[I_i(\beta) + \varphi \not\subseteq A] \cdot p(i) \\
&\leq \sum_{i \in [\pm L]} \mathbb{1}[I_i(\beta) + \varphi \not\subseteq \bigcup_{i' \in [\pm L]} I_{i'}(\beta)] \cdot p(i) \;\leq\; \max_{i_1 < \ldots < i_k} \{p(i_1) + \ldots + p(i_k)\} \;\leq\; \delta.
\end{aligned}
$$

Here, the first inequality holds since $\varepsilon \geq 0$, and the second inequality uses the fact that $\mathbb{1}[I_i(\beta) \subseteq A] - \mathbb{1}[I_i(\beta) + \varphi \subseteq A] \leq 0$ for all $i \in [\pm(L+k)] \setminus [\pm L]$. The third inequality holds because $\mathbb{1}[I_i(\beta) \subseteq A] - \mathbb{1}[I_i(\beta) + \varphi \subseteq A] = 1$ whenever $I_i(\beta) \subseteq$ and $I_i(\beta) + \varphi \not\subseteq A$, whereas $\mathbb{1}[I_i(\beta) \subseteq A] - \mathbb{1}[I_i(\beta) + \varphi \subseteq A] \in \{-1, 0\}$ otherwise. The fourth inequality disregards one of operands in the conjunction inside the indicator terms. The fifth inequality exploits the fact that $A \subseteq \bigcup_{i' \in [\pm L]} I_{i'}(\beta)$ for any $A \in \mathcal{F}(L, \beta)$. The sixth inequality holds since there are at most $k$ indices $i \in [\pm L]$ satisfying $I_i(\beta) + \varphi \not\subseteq \bigcup_{i' \in [\pm L]} I_{i'}(\beta)$ since $\varphi \in \{-\Delta f, -\Delta f + \beta, \ldots, \Delta f\}$ for $\beta = \Delta f / k$. The last inequality, finally, holds by construction of $p$, whose non-zero values are $\delta/2$ (at $i = 0$) and $(1 - \delta/2)/(4\lceil 1/(2\delta) \rceil k - 1)$ (at $i \in \mathcal{I}^k \setminus \{0\}$), and the fact that $\delta/2 + (k-1) \cdot (1 - \delta/2)/(4\lceil 1/(2\delta) \rceil k - 1) = \delta$.

#### 1.A.8.2 Example 2: Violation of Assumption 1 *(b)*

Fix any $\Delta f, \varepsilon, \delta > 0$ and consider the following loss function $c$:

$$c(x) \;=\; \begin{cases} 1 & \text{if } x \in C_i \text{ for some } i \in \mathbb{Z}, \\ 5 + 8i - \dfrac{8x}{\Delta f} & \text{if } x \in D_i^{\mathrm{L}} \text{ for some } i \in \mathbb{Z}, \\ 0 & \text{if } x \in D_i^{\mathrm{M}} \text{ for some } i \in \mathbb{Z}, \\ \dfrac{8x}{\Delta f} - 8i - 7 & \text{if } x \in D_i^{\mathrm{R}} \text{ for some } i \in \mathbb{Z}. \end{cases}$$

Here, $C_i := [i \cdot \Delta f, (i + \frac{1}{2}) \cdot \Delta f)$, and $D_i := [(i + \frac{1}{2}) \cdot \Delta f, (i + 1) \cdot \Delta f),\ i \in \mathbb{Z}$, are intervals that partition $\mathbb{R}$, and

$$\begin{aligned} D_i^{\mathrm{L}} &= \Big[\big(i + \frac{1}{2}\big) \cdot \Delta f,\ \big(i + \frac{5}{8}\big) \cdot \Delta f\Big) &&\text{with } |D_i^{\mathrm{L}}| = \frac{\Delta f}{8}, \\ D_i^{\mathrm{M}} &= \Big[\big(i + \frac{5}{8}\big) \cdot \Delta f,\ \big(i + \frac{7}{8}\big) \cdot \Delta f\Big) &&\text{with } |D_i^{\mathrm{M}}| = \frac{\Delta f}{4}, \\ D_i^{\mathrm{R}} &= \Big[\big(i + \frac{7}{8}\big) \cdot \Delta f,\ \big(i + 1\big) \cdot \Delta f\Big) &&\text{with } |D_i^{\mathrm{R}}| = \frac{\Delta f}{8} \end{aligned}$$

further partition each interval $D_i$ into a left (superscript 'L'), middle (superscript 'M') and right (superscript 'R') sub-interval, respectively. Figure 8 shows that $c$ satisfies Assumption 1 *(a)* since $c$ is piecewise linear without discontinuities. However, $c$ violates the unboundedness condition of Assumption 1 since its image is contained in the interval $[0, 1]$. In the following, we will show that for this loss function, $\mathrm{P}(L, \beta)$ and $\mathrm{D}(L, \beta)$ differ by at least $(1 - \delta)/(1 + e^{\varepsilon})$ for all $L \in \mathbb{N}$ and $\beta > 0$. We do so in two steps: We first show that $\mathrm{P} \geq (1 - \delta)/(1 + e^{\varepsilon})$, and we subsequently argue that $\mathrm{D}(L, \beta) \leq 0$. The statement then follows since $\mathrm{P}(L, \beta) \geq \mathrm{P}$.

To see that $\mathrm{P} \geq (1-\delta)/(1+e^{\varepsilon})$, define $C = \bigcup_{i \in \mathbb{Z}} C_i$ and $D = \bigcup_{i \in \mathbb{Z}} D_i$ such that $D - (\Delta f/2) = C$. The DP constraint of $\mathrm{P}$ indexed by $(\varphi, A) \in \mathcal{E}$ with $\varphi = \Delta f/2$ and $A = D$ implies that

$$\int_{x \in \mathbb{R}} \mathbb{1}[x \in D] \, \mathrm{d}\gamma(x) \leq e^{\varepsilon} \cdot \int_{x \in \mathbb{R}} \mathbb{1}[x \in D - (\Delta f/2)] \, \mathrm{d}\gamma(x) + \delta = e^{\varepsilon} \cdot \int_{x \in \mathbb{R}} \mathbb{1}[x \in C] \, \mathrm{d}\gamma(x) + \delta.$$

Let us refer to $\int_{x \in \mathbb{R}} \mathbb{1}[x \in C] \, \mathrm{d}\gamma(x)$ as $\gamma(C)$ and $\int_{x \in \mathbb{R}} \mathbb{1}[x \in D] \, \mathrm{d}\gamma(x)$ as $\gamma(D)$, respectively, so that the above inequality reads as $\gamma(D) \leq e^{\varepsilon} \cdot \gamma(C) + \delta$. Since $C$ and $D$ partition $\mathbb{R}$ and $\gamma(\mathbb{R}) = 1$, we have $\gamma(D) = 1 - \gamma(C)$ and thus $\gamma(C) \geq (1 - \delta)/(1 + e^{\varepsilon})$. Since $c(x) = 1$ for $x \in C$ and $c(x) \geq 0$

for all $x \in \mathbb{R}$, any feasible solution to P must satisfy

$$\int_{x \in \mathbb{R}} c(x) \, \mathrm{d}\gamma(x) = \int_{x \in C} c(x) \, \mathrm{d}\gamma(x) + \int_{x \in D} c(x) \, \mathrm{d}\gamma(x) \geq \int_{x \in C} \mathrm{d}\gamma(x) = \frac{1-\delta}{1+e^{\varepsilon}},$$

which shows that indeed $\mathrm{P} \geq (1-\delta)/(1+e^{\varepsilon})$.

To show that $\mathrm{D}(L, \beta) \leq 0$ for any $L \in \mathbb{N}$ and $\beta > 0$, we construct a feasible solution $p$ to the strong dual (22) of $\mathrm{D}(L, \beta)$ that attains the objective value 0. We do so in two steps: We first show that there is an index $i^\star \in \mathbb{Z}$ with $\underline{c}_{i^\star}(\beta) = 0$, and we subsequently argue that the solution $p : [\pm(L + \Delta f/\beta)] \mapsto \mathbb{R}_+$ with $p(i) = \mathbb{1}[i = i^\star]$ is feasible in (22). In view of the first step, note that by construction of $c$, any interval of length $\Delta f$ has a non-empty intersection with some sub-interval $D_i^{\mathrm{M}}$, $i \in \mathbb{Z}$. Since the union $\bigcup_{i=L+1}^{L+\Delta f/\beta} I_i(\beta)$ is an interval of length $\Delta f$, there must thus exist an index $i^\star \in \{L+1, \ldots, L + \Delta f/\beta\}$ such that $I_{i^\star}(\beta) \cap D_i^{\mathrm{M}} \neq \emptyset$ for some $i \in \mathbb{Z}$. We then have $\underline{c}_{i^\star}(\beta) = \inf_{x \in I_{i^\star}(\beta)} \{c(x)\} = 0$ since $c(x) = 0$ for $x \in D_i^{\mathrm{M}}$. As for the second step, we note that for any $(\varphi, A) \in \mathcal{E}(L, \beta)$, the left-hand side of the constraint in (22) satisfies

$$\sum_{i \in [\pm(L + \Delta f/\beta)]} \mathbb{1}[I_i(\beta) \subseteq A] \cdot p(i) = \mathbb{1}[I_{i^\star}(\beta) \subseteq A] \; = \; 0,$$

where the first equality holds by construction of $p$ and the second equality uses the fact that $I_{i^\star}(\beta) \not\subseteq A$ since $i^\star \in \{L+1, \ldots, L + \Delta f/\beta\}$ whereas $A \subseteq \bigcup_{i \in [\pm L]} I_i(\beta)$ because $A \in \mathcal{F}(L, \beta)$. Since the right-hand side of the constraint in (22) is non-negative by construction, we thus conclude that $p$ is feasible in (22).

## 1.B Proofs of Section 1.3

### 1.B.1 Proof of Observation 2

Assumption 3 allows us to replace the DP constraint in (6) with

$$\begin{aligned}
& \int_{x \in \mathbb{R}} \mathbb{1}[x \in A] \, \mathrm{d}\gamma(x \mid f(D)) \leq e^{\varepsilon} \cdot \int_{x \in \mathbb{R}} \mathbb{1}[f(D') - f(D) + x \in A] \, \mathrm{d}\gamma(x \mid f(D')) + \delta \\
& \hspace{20em} \forall (D, D') \in \mathcal{N}, \; \forall A \in \mathcal{F} \\
\iff & \int_{x \in \mathbb{R}} \mathbb{1}[x \in A] \, \mathrm{d}\gamma(x \mid \phi) \leq e^{\varepsilon} \cdot \int_{x \in \mathbb{R}} \mathbb{1}[x + \varphi \in A] \, \mathrm{d}\gamma(x \mid \phi + \varphi) + \delta \\
& \hspace{20em} \forall \phi \in \Phi, \; \forall (\varphi, A) \in \mathcal{E}'(\phi).
\end{aligned}$$

Here, the first line holds since $\{A : A \in \mathcal{F}\} = \{A + f(D) : A \in \mathcal{F}\}$ for any $D \in \mathcal{D}$, whereas the second line is due to Assumption 3. Replacing the latter representation of the DP constraints with those in (6) gives P′ and thus concludes the observation. □

### 1.B.2 Proof of Proposition 5

The proof of Proposition 5 relies on three auxiliary lemmas, each of which is devoted to the inclusion of one of the restrictions (7a), (7b) and (7c) into problem P′. We state and prove these auxiliary lemmas first.

**Lemma 14.** *With the additional constraint* (7a)*,* P′ *has the same optimal value as*

$$\begin{array}{ll} \underset{\gamma}{\text{minimize}} & \beta \cdot \sum_{k \in [K]} w_k(\beta) \cdot \int_{x \in \mathbb{R}} c(x) \, \mathrm{d}\gamma_k(x) \\ \text{subject to} & \gamma_k \in \mathcal{P}_0, \;\; k \in [K] \\ & \int_{x \in \mathbb{R}} \mathbb{1}[x \in A] \, \mathrm{d}\gamma_k(x) \leq e^{\varepsilon} \cdot \int_{x \in \mathbb{R}} \mathbb{1}[x + \varphi \in A] \, \mathrm{d}\gamma_m(x) + \delta \\ & \hspace{8em} \forall k, m \in [K], \;\; \forall (\varphi, A) \in \mathcal{E}''_{km}(\beta), \end{array} \tag{23}$$

*where* $\mathcal{E}''_{km}(\beta) := [[-\Delta f, \Delta f] \cap (\Phi_m(\beta) - \Phi_k(\beta))] \times \mathcal{F}$ *and* $w_k(\beta) := \beta^{-1} \cdot \int_{\phi \in \Phi_k(\beta)} w(\phi) \, \mathrm{d}\phi$*.*

*Proof.* We use restriction (7a) to replace the uncountable family of measures $\{\gamma(\cdot|\phi)\}_{\phi \in \Phi}$ in problem P′ with the finite set of measures $\{\gamma_k\}_{k \in [K]}$. Note that in this case, the requirement $\gamma \in \Gamma$ simplifies to $\gamma_k \in \mathcal{P}_0$ for all $k \in [K]$.

We can now equivalently reformulate the objective function of problem P′ as

$$\begin{aligned} \int_{\phi \in \Phi} w(\phi) \cdot \left[ \int_{x \in \mathbb{R}} c(x) \, \mathrm{d}\gamma(x \mid \phi) \right] \mathrm{d}\phi &= \sum_{k \in [K]} \int_{\phi \in \Phi_k(\beta)} w(\phi) \cdot \Big[ \int_{x \in \mathbb{R}} c(x) \, \mathrm{d} \underbrace{\gamma(x \mid \phi)}_{= \gamma_k(x)} \Big] \, \mathrm{d}\phi \\ &= \sum_{k \in [K]} \underbrace{\Big[ \int_{\phi \in \Phi_k(\beta)} w(\phi) \, \mathrm{d}\phi \Big]}_{= \beta \cdot w_k(\beta)} \cdot \Big[ \int_{x \in \mathbb{R}} c(x) \, \mathrm{d}\gamma_k(x) \Big], \end{aligned}$$

which coincides with the objective function of (23).

Next, fix any DP constraint $(\phi, \varphi, A)$ in P′, and note that $(\phi, \phi + \varphi) \in \Phi_k(\beta) \times \Phi_m(\beta)$ for a unique pair $(k, m) \in [K]$. In terms of our new measures $\{\gamma_k\}_{k \in [K]}$, the constraint becomes

$$\int_{x \in \mathbb{R}} \mathbb{1}[x \in A] \, \mathrm{d}\gamma_k(x) \leq e^{\varepsilon} \cdot \int_{x \in \mathbb{R}} \mathbb{1}[x + \varphi \in A] \, \mathrm{d}\gamma_m(x) + \delta,$$

and this constraint is indeed included in (23) since $\varphi \in [-\Delta f, \Delta f] \cap (\Phi_m(\beta) - \Phi_k(\beta))$. Similarly, one readily verifies that all DP constraints in (23) have corresponding constraints in problem P′. We thus conclude that P′ and (23) are indeed equivalent under restriction (7a), as desired. □

In contrast to P′, problem (23) comprises finitely many probability measures $\{\gamma_k\}_{k\in[K]}$. The next result shows that the restriction (7b) allows us to equivalently represent each measure $\gamma_k$ by countably many decision variables.

**Lemma 15.** *With the additional constraint* (7b)*, problem* (23) *has the same optimal value as*

$$\begin{array}{ll}\underset{p}{\text{minimize}} & \beta \cdot \sum\limits_{k\in[K]} w_k(\beta) \cdot \sum\limits_{i\in\mathbb{Z}} c(i)\cdot p_k(i) \\ \text{subject to} & p_k : \mathbb{Z} \mapsto \mathbb{R}_+,\ \sum\limits_{i\in\mathbb{Z}} p_k(i) = 1,\ \ k\in[K] \\ & \sum\limits_{i\in\mathbb{Z}} \mathbb{1}[I_i(\beta)\subseteq A]\cdot p_k(i) \leq e^{\varepsilon}\cdot \sum\limits_{i\in\mathbb{Z}} \mathbb{1}[I_i(\beta)+\varphi\subseteq A]\cdot p_m(i) + \delta \\ & \qquad\qquad\qquad\qquad \forall k,m\in[K],\ \forall(\varphi,A)\in\mathcal{E}'_{km}(\beta),\end{array} \tag{P′(β)}$$

*where* $\mathcal{E}'_{km}(\beta) := [\mathscr{B}(\beta)\cap\{(m-k-1)\cdot\beta,\ (m-k)\cdot\beta,\ (m-k+1)\cdot\beta\}]\times\mathcal{F}(\beta)$*.*

*Proof.* Similar arguments as in the proof of Lemma 3 allow us to replace each measure $\gamma_k$ with a countable set of decision variables $p_k : \mathbb{Z}\mapsto\mathbb{R}_+$, $k\in[K]$. The resulting reformulation of problem (23) under restriction (7b) reads as follows.

$$\begin{array}{ll}\underset{p}{\text{minimize}} & \beta \cdot \sum\limits_{k\in[K]} w_k(\beta) \cdot \sum\limits_{i\in\mathbb{Z}} c(i)\cdot p_k(i) \\ \text{subject to} & \sum\limits_{i\in\mathbb{Z}} p_k(i) = 1,\ p_k : \mathbb{Z} \mapsto \mathbb{R}_+,\ \ k\in[K] \\ & \sum\limits_{i\in\mathbb{Z}} p_k(i)\cdot\dfrac{|A\cap I_i(\beta)|}{\beta} \leq e^{\varepsilon}\cdot \sum\limits_{i\in\mathbb{Z}} p_m(i)\cdot\dfrac{|(A-\varphi)\cap I_i(\beta)|}{\beta} + \delta \\ & \qquad\qquad\qquad\qquad \forall k,m\in[K],\ \forall(\varphi,A)\in\mathcal{E}''_{km}(\beta)\end{array}$$

In contrast to P′(β), this problem still employs the larger constraint set $(\varphi,A)\in\mathcal{E}''_{km}(\beta)$. Applying similar arguments as in the proof of Lemma 4 shows that the constraints $(\varphi,A)$ satisfying $A\in\mathcal{F}\setminus\mathcal{F}(\beta)$ are weakly dominated by the constraints $(\varphi,A')$ satisfying $A'\in\mathcal{F}(\beta)$, and arguments similar to those in the proof of Lemma 5 show that the constraints $(\varphi,A)$ satisfying $A\in\mathcal{F}(\beta)$ and $\varphi\in[-\Delta f,\Delta f]\cap(\Phi_m(\beta)-\Phi_k(\beta))$ are weakly dominated by the constraints $(\varphi',A)$ satisfying $\varphi'\in\mathscr{B}(\beta)\cap\text{cl}(\Phi_m(\beta)-\Phi_k(\beta))$ whenever $[-\Delta f,\Delta f]\cap(\Phi_m(\beta)-\Phi_k(\beta))$ is nonempty.

We can thus replace in the above optimization problem the constraints $(\varphi, A) \in \mathcal{E}''_{km}(\beta)$ with the smaller set of constraints $(\varphi, A) \in \mathcal{E}'_{km}(\beta)$. In that case, however, the identities

$$\frac{|A \cap I_i(\beta)|}{\beta} = \mathbb{1}[I_i(\beta) \subseteq A] \quad \text{and} \quad \frac{|(A-\varphi) \cap I_i(\beta)|}{\beta} = \mathbb{1}[I_i(\beta) + \varphi \subseteq A]$$

hold for all $i \in \mathbb{Z}$ (*cf.* the proof of Lemma 1), which concludes the proof. □

Problem $\mathrm{P}'(\beta)$ still comprises a countably infinite number of decision variables and uncountably many constraints. The next result shows that the restriction (7c) allows us to reformulate $\mathrm{P}'(\beta)$ as a finite-dimensional linear program.

**Lemma 16.** *With the additional constraint* (7c), $\mathrm{P}'(\beta)$ *has the same optimal value as the finite-dimensional linear program* $\mathrm{P}'(L, \beta)$.

*Proof.* Under restriction (7c), we can reduce each countable set of decisions $p_k : \mathbb{Z} \mapsto \mathbb{R}_+$ to a finite set $p_k : [\pm L] \mapsto \mathbb{R}_+$, $k \in [K]$. Moreover, similar arguments as in the proof of Proposition 1 allow us to show that in the resulting problem, each constraint $(\varphi, A) \in \mathcal{E}'_{km}(\beta) \setminus \mathcal{E}'_{km}(L, \beta)$ is weakly dominated by the a constraint $(\varphi, A_L) \in \mathcal{E}'_{km}(L, \beta)$. This concludes the proof. □

**Proof of Proposition 5.** The proof directly follows from Lemmas 14, 15 and 16. □

### 1.B.3 Proof of Proposition 6

Fix any $\gamma$ feasible in $\mathrm{P}'$ and any $(\theta, \psi)$ feasible in $\mathrm{D}'$. We then have

$$\begin{aligned}
&\int_{\phi \in \Phi} w(\phi) \cdot \left[ \int_{x \in \mathbb{R}} c(x) \, \mathrm{d}\gamma(x \mid \phi) \right] \mathrm{d}\phi \\
&= \int_{\phi \in \Phi} \int_{x \in \mathbb{R}} \Big[ w(\phi) \cdot c(x) \Big] \, \mathrm{d}\gamma(x \mid \phi) \, \mathrm{d}\phi \\
&\geq \int_{\phi \in \Phi} \int_{x \in \mathbb{R}} \Bigg[ \theta(\phi) - \int_{(\varphi, A) \in \mathcal{E}'(\phi)} \mathbb{1}[x \in A] \, \mathrm{d}\psi(\varphi, A \mid \phi) + \\
&\qquad\qquad e^{\varepsilon} \cdot \int_{(-\varphi, A) \in \mathcal{E}'(\phi)} \mathbb{1}[x + \varphi \in A] \, \mathrm{d}\psi(\varphi, A \mid \phi - \varphi) \Bigg] \, \mathrm{d}\gamma(x \mid \phi) \, \mathrm{d}\phi \\
&= \underbrace{\int_{\phi \in \Phi} \int_{x \in \mathbb{R}} \theta(\phi) \, \mathrm{d}\gamma(x \mid \phi) \, \mathrm{d}\phi}_{(i)} - \underbrace{\int_{\phi \in \Phi} \int_{x \in \mathbb{R}} \int_{(\varphi, A) \in \mathcal{E}'(\phi)} \mathbb{1}[x \in A] \, \mathrm{d}\psi(\varphi, A \mid \phi) \, \mathrm{d}\gamma(x \mid \phi) \, \mathrm{d}\phi}_{(ii)}
\end{aligned}$$

$$+ e^{\varepsilon} \cdot \underbrace{\int_{\phi \in \Phi} \int_{x \in \mathbb{R}} \int_{(-\varphi, A) \in \mathcal{E}'(\phi)} \mathbb{1}[x + \varphi \in A] \, \mathrm{d}\psi(\varphi, A \mid \phi - \varphi) \, \mathrm{d}\gamma(x \mid \phi) \, \mathrm{d}\phi,}_{\textit{(iii)}}$$

where the inequality follows from the constraints in D′ and the fact that $\gamma(\cdot|\phi)$ is a non-negative measure, and the final equality is due to the linearity of the integration operator.

The above term *(i)* simplifies to

$$\int_{\phi \in \Phi} \int_{x \in \mathbb{R}} \theta(\phi) \, \mathrm{d}\gamma(x \mid \phi) \, \mathrm{d}\phi = \int_{\phi \in \Phi} \theta(\phi) \left[ \int_{x \in \mathbb{R}} \mathrm{d}\gamma(x \mid \phi) \right] \mathrm{d}\phi = \int_{\phi \in \Phi} \theta(\phi) \, \mathrm{d}\phi,$$

where we used the fact that $\gamma(\cdot|\phi)$ is a probability measure for every $\phi \in \Phi$.

The above term *(ii)* can be reformulated as

$$\begin{aligned} &\int_{\phi \in \Phi} \int_{x \in \mathbb{R}} \int_{(\varphi, A) \in \mathcal{E}'(\phi)} \mathbb{1}[x \in A] \, \mathrm{d}\psi(\varphi, A \mid \phi) \, \mathrm{d}\gamma(x \mid \phi) \, \mathrm{d}\phi \\ = &\int_{\phi \in \Phi} \int_{(\varphi, A) \in \mathcal{E}'(\phi)} \left[ \int_{x \in \mathbb{R}} \mathbb{1}[x \in A] \, \mathrm{d}\gamma(x \mid \phi) \right] \mathrm{d}\psi(\varphi, A \mid \phi) \, \mathrm{d}\phi, \end{aligned}$$

where we used Fubini's theorem (whose applicability follows from similar arguments as in the proof of Proposition 2) to change the order of integration.

Finally, the above term *(iii)* can be rewritten as

$$\begin{aligned} &\int_{\phi \in \Phi} \int_{x \in \mathbb{R}} \int_{(-\varphi, A) \in \mathcal{E}'(\phi)} \mathbb{1}[x + \varphi \in A] \, \mathrm{d}\psi(\varphi, A \mid \phi - \varphi) \, \mathrm{d}\gamma(x \mid \phi) \, \mathrm{d}\phi \\ = &\int_{\phi \in \Phi} \int_{(-\varphi, A) \in \mathcal{E}'(\phi)} \left[ \int_{x \in \mathbb{R}} \mathbb{1}[x + \varphi \in A] \, \mathrm{d}\gamma(x \mid \phi) \right] \mathrm{d}\psi(\varphi, A \mid \phi - \varphi) \, \mathrm{d}\phi \\ = &\int_{\phi \in \Phi} \int_{(\varphi, A) \in \mathcal{E}'(\phi)} \left[ \int_{x \in \mathbb{R}} \mathbb{1}[x + \varphi \in A] \, \mathrm{d}\gamma(x \mid \phi + \varphi) \right] \mathrm{d}\psi(\varphi, A \mid \phi) \, \mathrm{d}\phi. \end{aligned}$$

where the first equality follows from Fubini's theorem. The second equality is due to a change of variables. Specifically, we use the definition of $\mathcal{E}'(\phi)$ to rewrite the region of integration as

$$\left\{(\phi, \varphi, A) \ : \ \phi \in \Phi, (-\varphi, A) \in \mathcal{E}'(\phi)\right\} = \left\{(\phi, \varphi, A) \in \Phi \times [-\Delta f, \Delta f] \times \mathcal{F} \ : \phi - \varphi \in \Phi\right\}.$$

Introducing a new variable $\phi' = \phi - \varphi$, we observe that $\varphi + \phi' = \phi \in \Phi$. Hence the region of integration region can be expressed as

$$\left\{(\phi', \varphi, A) \in \Phi \times [-\Delta f, \Delta f] \times \mathcal{F} \ : \varphi \in \Phi - \phi'\right\} = \left\{(\phi', \varphi, A) \ : \ \phi' \in \Phi, (\varphi, A) \in \mathcal{E}'(\phi')\right\}$$

if we replace $\phi$ with $\phi' + \varphi$ in the integrals. The second equality now holds if we relabel $\phi'$ as $\phi$.

Replacing the terms *(i)–(iii)* with their equivalent expressions derived above, we obtain

$$\begin{aligned}
&\int_{\phi\in\Phi} w(\phi)\cdot\left[\int_{x\in\mathbb{R}} c(x)\,\mathrm{d}\gamma(x\mid\phi)\right]\mathrm{d}\phi \\
\geq &\int_{\phi\in\Phi}\theta(\phi)\,\mathrm{d}\phi \;-\int_{\phi\in\Phi}\int_{(\varphi,A)\in\mathcal{E}'(\phi)}\Bigg[\int_{x\in\mathbb{R}}\mathbb{1}[x\in A]\,\mathrm{d}\gamma(x\mid\phi) - \\
&\qquad\qquad e^{\varepsilon}\cdot\int_{x\in\mathbb{R}}\mathbb{1}[x+\varphi\in A]\,\mathrm{d}\gamma(x\mid\phi+\varphi)\Bigg]\,\mathrm{d}\psi(\varphi,A\mid\phi)\,\mathrm{d}\phi \\
\geq &\int_{\phi\in\Phi}\left[\theta(\phi)-\delta\cdot\int_{(\varphi,A)\in\mathcal{E}'(\phi)}\mathrm{d}\psi(\varphi,A\mid\phi)\right]\mathrm{d}\phi,
\end{aligned}$$

where the final inequality is due to the constraints in P′ and the fact that $\psi(\cdot|\phi)$ is a non-negative measure. The last expression coincides with the objective function of D′, as desired. □

### 1.B.4 Proof of Proposition 7

Our proof proceeds in two steps. We first use restriction (8a) to reduce the number of decision variables in D′, and we subsequently use restriction (8b) to remove the integrals as well as reduce the number of constraints in D′. This will yield the formulation in the statement of Proposition 7.

In view of the first step, we use restriction (8a) to replace the measure $\theta : \Phi \mapsto \mathbb{R}$ with a vector $\boldsymbol{\theta} \in \mathbb{R}^K$ and $\psi$ with a finite family of unconditional measures $\psi_k \in \mathcal{M}_+(\mathcal{E}'(\phi))$, $k \in [K]$. Under those substitutions, problem D′ simplifies to the following formulation.

$$\begin{aligned}
\underset{\boldsymbol{\theta},\psi}{\text{maximize}}\quad & \beta\cdot\sum_{k\in[K]}\theta_k - \delta\cdot\sum_{k\in[K]}\int_{\phi\in\Phi_k(\beta)}\int_{(\varphi,A)\in\mathcal{E}'(\phi)}\mathrm{d}\psi_k(\varphi,A)\,\mathrm{d}\phi \\
\text{subject to}\quad & \boldsymbol{\theta}\in\mathbb{R}^K,\;\; \psi_k\in\mathcal{M}_+(\mathcal{E}'(\phi)),\;\; k\in[K] \\
& \theta_k \leq \int_{(\varphi,A)\in\mathcal{E}'(\phi)}\mathbb{1}[x\in A]\,\mathrm{d}\psi_k(\varphi,A) - \\
&\qquad e^{\varepsilon}\cdot\sum_{m\in[K]}\int_{(-\varphi,A)\in\mathcal{E}'(\phi)}\mathbb{1}[x+\varphi\in A]\cdot\mathbb{1}[\phi-\varphi\in\Phi_m(\beta)]\,\mathrm{d}\psi_m(\varphi,A) \\
&\qquad\qquad\qquad + c(x)\cdot w(\phi)\quad \forall k\in[K],\;\forall\phi\in\Phi_k(\beta),\;\forall x\in\mathbb{R}
\end{aligned}$$

Here, our reformulation uses the fact that $\{\Phi_k(\beta)\}_{k\in[K]}$ partitions $\Phi$, and the reformulated first term of the objective function additionally exploits that $|\Phi_k(\beta)| = \beta$ for all $k \in [K]$.

As for the second step, we note that under restriction (8b), the second expression in the objective function simplifies to

$$\sum_{k\in[K]} \int_{\phi\in\Phi_k(\beta)} \int_{(\varphi,A)\in\mathcal{E}'(\phi)} \mathrm{d}\psi_k(\varphi,A)\,\mathrm{d}\phi = \sum_{k\in[K]} \int_{\phi\in\Phi_k(\beta)} \int_{(\varphi,A)\in\mathcal{E}'_k(L,\beta)} \mathrm{d}\psi_k(\varphi,A)\,\mathrm{d}\phi$$
$$= \beta\cdot\sum_{k\in[K]} \int_{(\varphi,A)\in\mathcal{E}'_k(L,\beta)} \mathrm{d}\psi_k(\varphi,A),$$

where the first equality is due to restriction (8b) and the fact that $\mathcal{E}'(\phi)\cap\mathcal{E}(L,\beta)=\mathcal{E}'_k(L,\beta)$, and the second equality follows from taking the inner integral outside (as it is not parameterized by $\phi$) and from $|\Phi_k(\beta)|=\beta$. The resulting objective function coincides with that of problem $\mathrm{D}'(L,\beta)$. For any fixed $k\in[K]$, $\phi\in\Phi_k(\beta)$ and $x\in\mathbb{R}$, the first integral in the constraints simplifies to

$$\int_{(\varphi,A)\in\mathcal{E}'(\phi)} \mathbb{1}[x\in A]\,\mathrm{d}\psi_k(\varphi,A) = \int_{(\varphi,A)\in\mathcal{E}'_k(L,\beta)} \mathbb{1}[x\in A]\,\mathrm{d}\psi_k(\varphi,A).$$

Likewise, the second integral simplifies to

$$\sum_{m\in[K]} \int_{(-\varphi,A)\in\mathcal{E}'(\phi)} \mathbb{1}[x+\varphi\in A]\cdot\mathbb{1}[\phi-\varphi\in\Phi_m(\beta)]\,\mathrm{d}\psi_m(\varphi,A)$$
$$= \sum_{m\in[K]} \int_{(-\varphi,A)\in\mathcal{E}'_k(L,\beta)} \mathbb{1}[x+\varphi\in A]\cdot\mathbb{1}[\phi-\varphi\in\Phi_m(\beta)]\,\mathrm{d}\psi_m(\varphi,A)$$
$$= \int_{(-\varphi,A)\in\mathcal{E}'_k(L,\beta)} \mathbb{1}[x+\varphi\in A]\,\mathrm{d}\psi_{k-\varphi/\beta}(\varphi,A),$$

where the first equality is due to restriction (8b) and the second equality exploits the fact that for $\phi\in\Phi_k(\beta)$ and $-\varphi\in\mathscr{B}(\beta)$, we have $\phi-\varphi\in\Phi_m(\beta)$ if and only if $m=k-\varphi/\beta$. In summary, the constraints simplify to

$$\theta_k \le \int_{(\varphi,A)\in\mathcal{E}'_k(L,\beta)} \mathbb{1}[x\in A]\,\mathrm{d}\psi_k(\varphi,A) - e^{\varepsilon}\cdot\int_{(-\varphi,A)\in\mathcal{E}'_k(L,\beta)} \mathbb{1}[x+\varphi\in A]\,\mathrm{d}\psi_{k-\varphi/\beta}(\varphi,A)$$
$$+c(x)\cdot w(\phi)\quad \forall k\in[K],\ \forall\phi\in\Phi_k(\beta),\ \forall x\in\mathbb{R}.$$

Note that the index $\phi\in\Phi_k(\beta)$ only affects the last term in this constraint, and the constraint is thus equivalent to

$$\theta_k \le \int_{(\varphi,A)\in\mathcal{E}'_k(L,\beta)} \mathbb{1}[x\in A]\,\mathrm{d}\psi_k(\varphi,A) - e^{\varepsilon}\cdot\int_{(-\varphi,A)\in\mathcal{E}'_k(L,\beta)} \mathbb{1}[x+\varphi\in A]\,\mathrm{d}\psi_{k-\varphi/\beta}(\varphi,A)$$
$$+c(x)\cdot \underline{w}_k(\beta)\quad \forall k\in[K],\ \forall x\in\mathbb{R}.$$

We can equivalently express the index $x \in \mathbb{R}$ in this constraint with the double index $(i, x) \in \mathbb{Z} \times I_i(\beta)$, and arguments similar to those in the proof of Lemma 2 show that the constraint subsequently simplifies to

$$\theta_k \leq \int_{(\varphi,A)\in\mathcal{E}'_k(L,\beta)} \mathbb{1}[I_i(\beta) \subseteq A] \, \mathrm{d}\psi_k(\varphi, A) - e^{\varepsilon} \cdot \int_{(-\varphi,A)\in\mathcal{E}'_k(L,\beta)} \mathbb{1}[I_i(\beta) + \varphi \subseteq A] \, \mathrm{d}\psi_{k-\varphi/\beta}(\varphi, A)$$
$$+ \underline{c}_i(\beta) \cdot \underline{w}_k(\beta) \quad \forall k \in [K], \; \forall i \in \mathbb{Z}.$$

Similar arguments as in the proof of Proposition 3 allow us to further restrict the index $i \in \mathbb{Z}$ in the above constraint to $i \in [\pm(L + \Delta f/\beta)]$. The restriction (8b) also allows us to replace the measures $\psi_k \in \mathcal{M}_+(\mathcal{E}'(\phi))$ with discrete measures $\psi_k : \mathcal{E}'_k(L, \beta) \mapsto \mathbb{R}_+$, $k \in [K]$, and replace the integrals in the objective function and the constraints with sums. This results in the formulation $\mathrm{D}'(L, \beta)$ and thus concludes the proof. □

### 1.B.5 Proof of Theorem 2

We employ the same strategy as in the proof of Theorem 1. We define the auxiliary problem

$$\begin{aligned}
\underset{p}{\text{minimize}} \quad & (\Delta f/\ell) \cdot \sum_{k\in[K]} w_k(\Delta f/\ell) \cdot \Big[ \sum_{i\in[\pm(\Lambda\cdot\ell+\ell)]} c_i(\Delta f/\ell) \cdot p_k(i) \Big] \\
\text{subject to} \quad & p_k : [\pm(\Lambda \cdot \ell + \ell)] \mapsto \mathbb{R}_+, \quad \sum_{i\in[\pm(\Lambda\cdot\ell+\ell)]} p_k(i) = 1, \;\; k \in [K] \\
& \sum_{i\in[\pm(\Lambda\cdot\ell+\ell)]} \mathbb{1}[I_i(\Delta f/\ell) \subseteq A] \cdot p_k(i) \leq e^{\varepsilon} \cdot \sum_{i\in[\pm(\Lambda\cdot\ell+\ell)]} \mathbb{1}[I_i(\Delta f/\ell) + \varphi \subseteq A] \cdot p_m(i) + \delta \\
& \qquad \forall k, m \in [K], \;\; \forall (\varphi, A) \in \mathcal{E}'_{km}(\Lambda \cdot \ell, \Delta f/\ell),
\end{aligned} \qquad (\mathrm{M}'(\Lambda \cdot \ell, \Delta f/\ell))$$

and we show that the optimal values of $\mathrm{P}'(\Lambda \cdot \ell, \Delta f/\ell)$ and $\mathrm{D}'(\Lambda \cdot \ell, \Delta f/\ell)$ converge to that of $\mathrm{M}'(\Lambda \cdot \ell, \Delta f/\ell)$ when $\ell$ increases (which, in return, refines the granularity $\Delta f/\ell$) and $\Lambda$ increases (which, in return, increases the support $[-\Lambda \cdot \Delta f, \; (\Lambda + 1/\ell) \cdot \Delta f))$. Note that the number $K$ of intervals in $\Phi$ depends on $\ell$ due to the partitioning $\Phi = \bigcup_{k\in[K]} \Phi_k(\Delta f/\ell)$.

To see that the optimal value of $\mathrm{P}'(\Lambda \cdot \ell, \Delta f/\ell)$ converges to that of $\mathrm{M}'(\Lambda \cdot \ell, \Delta f/\ell)$, we first note that $\mathrm{M}'(\Lambda \cdot \ell, \Delta f/\ell)$ differs from $\mathrm{P}'(\Lambda \cdot \ell, \Delta f/\ell)$ only in the existence of the additional decision variables $p_k(i)$, $i \in [\pm(\Lambda \cdot \ell + \ell)] \backslash [\pm \Lambda \cdot \ell]$, which also implies that $\mathrm{P}'(\Lambda \cdot \ell, \Delta f/\ell) \geq \mathrm{M}'(\Lambda \cdot \ell, \Delta f/\ell)$. Using similar arguments as in the proof of Lemma 11, we can show that for any $\varepsilon > 0$, $\delta > 0$ and $\tau > 0$, there is $\Lambda' \in \mathbb{N}$ such that any optimal solution $p^\star$ to $\mathrm{M}'(\Lambda \cdot \ell, \Delta f/\ell)$ satisfies

$\sum_{i \in [\pm(\Lambda \cdot \ell + \ell)] \setminus [\pm \Lambda \cdot \ell]} p_k^\star(i) < \tau$ for all $k \in [K]$, $\ell \in \mathbb{N}$ and $\Lambda \geq \Lambda'$. Similar arguments as in the proof of Lemma 12 then allow us to show that there is $\Lambda' \in \mathbb{N}$ such that $\mathrm{P}'(\Lambda' \cdot \ell, \Delta f/\ell) - \mathrm{M}'(\Lambda' \cdot \ell, \Delta f/\ell) \leq \xi$ for all $\ell \in \mathbb{N}$. For the remainder of the proof, we fix such a value of $\Lambda'$.

To see that the optimal value of $\mathrm{D}'(\Lambda' \cdot \ell, \Delta f/\ell)$ converges to that of $\mathrm{M}'(\Lambda' \cdot \ell, \Delta f/\ell)$ , on the other hand, we note that $\mathrm{M}'(\Lambda' \cdot \ell, \Delta f/\ell)$ differs from the strong dual of $\mathrm{D}'(\Lambda' \cdot \ell, \Delta f/\ell)$ essentially only in the objective coefficients, which change from $\underline{w}_k(\Delta f/\ell)$ and $\underline{c}_i(\Delta f/\ell)$ in the strong dual of $\mathrm{D}'(\Lambda' \cdot \ell, \Delta f/\ell)$ to $w_k(\Delta f/\ell)$ and $c_i(\Delta f/\ell)$ in $\mathrm{M}'(\Lambda' \cdot \ell, \Delta f/\ell)$, respectively. Similar arguments as in the proof of Lemma 13 show that for any $\varepsilon > 0$, $\delta > 0$ and $\xi > 0$, there exists $\ell' \in \mathbb{N}$ such that $\mathrm{M}'(\Lambda' \cdot \ell, \Delta f/\ell) - \mathrm{D}'(\Lambda' \cdot \ell, \Delta f/\ell) \leq \xi$ for all $\ell \geq \ell'$. Here, the uniform continuity of $c$ ensures the convergence of the terms $c_i(\Delta f/\ell)$ and $\underline{c}_i(\Delta f/\ell)$, while our earlier assumption that $w$ is a continuous probability density function ensures the convergence of the terms $w_k(\Delta f/\ell)$ and $\underline{w}_k(\Delta f/\ell)$. Moreover, since $\{w_k(\Delta f/\ell)\}_\ell$ and $\{\underline{w}_k(\Delta f/\ell)\}_\ell$ are non-negative and sum up to at most 1, the overall objective functions of both problem converge despite the growing number $K$ of subsets of $\Phi$.

So far, we have shown that there exists $\Lambda' \in \mathbb{N}$ and $\ell' \in \mathbb{N}$ such that $\mathrm{P}'(\Lambda' \cdot \ell, \Delta f/\ell) - \mathrm{D}'(\Lambda' \cdot \ell, \Delta f/\ell) \leq \xi$ holds for all $\ell \geq \ell'$. Since we have $\mathrm{P}'(\Lambda' \cdot \ell, \Delta f/\ell) \geq \mathrm{P}'(\Lambda \cdot \ell, \Delta f/\ell)$ and $\mathrm{D}'(\Lambda' \cdot \ell, \Delta f/\ell) \leq \mathrm{D}'(\Lambda \cdot \ell, \Delta f/\ell)$ for all $\Lambda \geq \Lambda'$, we can conclude the proof. Further details of this proof are relegated to the GitHub supplement.

## 1.C Proofs of Section 1.4

### 1.C.1 Proof of Corollary 1

$\mathrm{P}(\boldsymbol{\pi}, \beta)$ and $\mathrm{D}(\boldsymbol{\pi}, \beta)$ sandwich $\mathrm{P}$ and $\mathrm{D}$ from above and below since $\mathrm{P}(\boldsymbol{\pi}, \beta)$ and $\mathrm{D}(\boldsymbol{\pi}, \beta)$ constitute restrictions of the earlier problems $\mathrm{P}(L, \beta)$ and $\mathrm{D}(L, \beta)$ that satisfy the same inequality (*cf.* Lemmas 1 and 2 as well as Propositions 1 and 3).

In view of the second part of the statement, recall from Theorem 1 that there is $\Lambda' \in \mathbb{N}$ and $k' \in \mathbb{N}$ such that $\mathrm{P}(\Lambda \cdot k, \Delta f/k) - \mathrm{D}(\Lambda \cdot k, \Delta f/k) \leq \xi$ holds for all $\Lambda \geq \Lambda'$ and $k \geq k'$. Moreover, we have $\mathrm{P}(\boldsymbol{\pi}, \beta) \leq \mathrm{P}(\Lambda \cdot k, \Delta f/k)$ if $\{\Pi_j(\beta)\}_{j \in [N]}$ is a refinement of $\{I_i(\Delta f/k)\}_{i \in [\pm \Lambda \cdot k]}$. Indeed, $\mathrm{P}(\Lambda \cdot k, \Delta f/k)$ is equivalent to a variant of $\mathrm{P}(\boldsymbol{\pi}, \beta)$ that includes the additional constraints $p(j) = p(j')$ for all $j, j' \in [N]$ satisfying $\Pi_j(\beta), \Pi_{j'}(\beta) \subseteq I_i(\Delta f/k)$ for some $i \in [\pm \Lambda \cdot k]$. A similar argument shows that $\mathrm{D}(\boldsymbol{\pi}, \beta) \geq \mathrm{D}(\Lambda \cdot k, \Delta f/k)$, which concludes the proof.

### 1.C.2 Proof of Proposition 8

We prove Proposition 8 via three auxiliary results. Lemma 17 proves that each inner loop over $j$ in Algorithm 2 determines a worst-case event $A \in \arg\max\{V(\varphi, A) \, : \, A \in \mathcal{F}(L, \beta)\}$ for the query output difference $\varphi \in \mathscr{B}(\beta)$ fixed by the outer loop. Subsequently, Lemma 18 proves that each outer loop over $\varphi$ determines a maximally violated constraint $(\varphi, A)$, which concludes the correctness of Algorithm 2. Finally, Lemma 19 shows that Algorithm 2 can be implemented such that it determines a maximally violated constraint $(\varphi, A)$ in time $\mathcal{O}(N^3)$.

**Lemma 17.** *For any $\varphi \in \mathscr{B}(\beta)$ fixed by the outer loop of Algorithm 2, the event $A$ constructed in the inner loop satisfies $A \in \arg\max\{V(\varphi, A) \, : \, A \in \mathcal{F}(L, \beta)\}$.*

*Proof.* Fix an arbitrary $\varphi \in \mathscr{B}(\beta)$ and recall that for any $A \in \mathcal{F}(L, \beta)$, the privacy shortfall $V(\varphi, A)$ can be expressed as

$$
\begin{aligned}
&\sum_{j \in [N]} p(j) \cdot \frac{|A \cap \Pi_j(\beta)|}{|\Pi_j(\beta)|} - e^{\varepsilon} \cdot \sum_{j \in [N]} p(j) \cdot \frac{|A \cap (\Pi_j(\beta) + \varphi)|}{|\Pi_j(\beta)|} \\
&= \sum_{i \in [\pm L]} \mathbb{1}[I_i(\beta) \subseteq A] \cdot \left[ \sum_{j \in [N]} p(j) \cdot \frac{|I_i(\beta) \cap \Pi_j(\beta)|}{|\Pi_j(\beta)|} - e^{\varepsilon} \cdot \sum_{j \in [N]} p(j) \cdot \frac{|(I_i(\beta) - \varphi) \cap \Pi_j(\beta)|}{|\Pi_j(\beta)|} \right] \\
&= \sum_{i \in [\pm L]} \mathbb{1}[I_i(\beta) \subseteq A] \cdot \beta \cdot \left[ \sum_{j \in [N]} p(j) \cdot \frac{\mathbb{1}[I_i(\beta) \subseteq \Pi_j(\beta)]}{|\Pi_j(\beta)|} - e^{\varepsilon} \cdot \sum_{j \in [N]} p(j) \cdot \frac{\mathbb{1}[I_i(\beta) \subseteq (\Pi_j(\beta) + \varphi)]}{|\Pi_j(\beta)|} \right],
\end{aligned}
$$

where we disregard the constant $-\delta$ since it does not affect the relative order of privacy shortfalls across the constraints $(\varphi, A)$. Here, the first identity exploits that $A = \bigcup_{i \in \mathcal{L}} I_i(\beta)$ for some $\mathcal{L} \subseteq [\pm L]$. The second identity holds since $|I_i(\beta)| = \beta$ and $I_i(\beta)$ is either entirely contained in or intersection free with $\Pi_j(\beta)$ and $\Pi_j(\beta) + \varphi$, $i \in [\pm L]$ and $j \in [N]$. Thus, $I_i(\beta)$ must be contained in the worst-case event $A$ whenever

$$
\sum_{j \in [N]} p(j) \cdot \frac{\mathbb{1}[I_i(\beta) \subseteq \Pi_j(\beta)]}{|\Pi_j(\beta)|} - e^{\varepsilon} \cdot \sum_{j' \in [N]} p(j') \cdot \frac{\mathbb{1}[I_i(\beta) \subseteq \Pi_{j'}(\beta) + \varphi]}{|\Pi_{j'}(\beta)|} \tag{24}
$$

is strictly positive; $I_i(\beta)$ can be (but does not have to be) included in $A$ if the above quantity is zero; and it must not be contained in $A$ if the above quantity is negative. In the remainder, we fix $i \in [\pm L]$ and distinguish between two cases: *(i)* there is no $j' \in [N]$ satisfying $I_i(\beta) \subseteq \Pi_{j'}(\beta) + \varphi$; and *(ii)* there is $j' \in [N]$ satisfying $I_i(\beta) \subseteq \Pi_{j'}(\beta) + \varphi$.

In case *(i)*, we can include $I_i(\beta)$ in the worst-case event $A$ since the second term in (24) vanishes, whereas the first term is always non-zero by construction. Note that the events $A_j$, $j \in [N]$, constructed in the first part of the inner loop of Algorithm 2 comprise precisely all intervals $I_i(\beta)$ falling under case *(i)*.

In case *(ii)*, the existence of $j' \in [N]$ satisfying $I_i(\beta) \subseteq \Pi_{j'}(\beta) + \varphi$ implies that (24) equals to $p(j)/|\Pi_j(\beta)| - e^{\varepsilon} \cdot p(j')/|\Pi_{j'}(\beta)|$ for some $j \in [N]$, and $I_i(\beta)$ should be included in the worst-case event if this quantity is positive. Note that the events $A_{jj'}$, $j, j' \in [N]$, constructed in the second part of the inner loop of Algorithm 2 comprise precisely all intervals $I_i(\beta)$ falling under case *(ii)* that satisfy $p(j)/|\Pi_j(\beta)| - e^{\varepsilon} \cdot p(j')/|\Pi_{j'}(\beta)| > 0$. □

**Lemma 18.** *Algorithm 2 returns a constraint* $(\varphi, A)$ *with maximum privacy shortfall.*

*Proof.* Recall that the DP constraints of $\mathrm{P}(\boldsymbol{\pi}, \beta)$ are indexed by $(\varphi, A) \in \mathcal{E}(L, \beta) = \mathscr{B}(\beta) \times \mathcal{F}(L, \beta)$ and that Lemma 17 proved that for any fixed $\varphi \in \mathscr{B}(\beta)$, the inner loop of Algorithm 2 constructs a worst-case event $A$ associated with $\varphi$. We show in this proof that it is sufficient to consider the values $\varphi \in \mathscr{B}(\beta, \boldsymbol{\pi}) \subseteq \mathscr{B}(\beta)$, where

$$\mathscr{B}(\beta, \boldsymbol{\pi}) := \{(\pi_j - \pi_{j'}) \cdot \beta \ : \ (\pi_j - \pi_{j'}) \cdot \beta \in [-\Delta f, \Delta f] \text{ and } j, j' \in [N]\} \cup \{-\Delta f, \Delta f\},$$

which is what the outer loop of Algorithm 2 does.

Our earlier arguments of this section have shown that for any $\varphi \in \mathscr{B}(\beta)$, the maximum privacy shortfall $\max\{V(\varphi, A) \ : \ A \in \mathcal{F}(L, \beta)\}$ satisfies

$$\sum_{j \in [N]} |A_j(\varphi)| \cdot \frac{p(j)}{|\Pi_j(\beta)|} + \sum_{j,j' \in [N]} |A_{jj'}(\varphi)| \cdot \left[ \frac{p(j)}{|\Pi_j(\beta)|} - e^{\varepsilon} \cdot \frac{p(j')}{|\Pi_{j'}(\beta)|} \right]^+ - \delta,$$

where $[x]^+ = \max\{0, x\}$ and the only quantities varying with $\varphi$ are

$$A_j(\varphi) = \Pi_j(\beta) \setminus [-L \cdot \beta + \varphi, (L+1) \cdot \beta + \varphi) \quad \text{and} \quad A_{jj'}(\varphi) = \Pi_j(\beta) \cap (\Pi_{j'}(\beta) + \varphi), \ \ j, j' \in [N].$$

One readily verifies that both $|A_j(\varphi)|$ and $|A_{jj'}(\varphi)|$, $j, j' \in [N]$, are affine between any two consecutive points in $\mathscr{B}(\beta, \boldsymbol{\pi})$. In other words, the maximum privacy shortfall is piecewise affine with breakpoints $\mathscr{B}(\beta, \boldsymbol{\pi})$ or a subset thereof, which implies that its maximum must be attained at one of the points $\varphi \in \mathscr{B}(\beta, \boldsymbol{\pi})$. This concludes the proof. □

**Lemma 19.** *Algorithm 2 can be implemented such that it terminates in time* $\mathcal{O}(N^3)$.

*Proof.* Since all individual steps in Algorithm 2 take constant time, the runtime of the algorithm is determined by the numbers of iterations in the outer and inner loops. In the naïve implementation of the main text, both loops comprise $\mathcal{O}(N^2)$ iterations, and thus the overall complexity of that implementation is $\mathcal{O}(N^4)$. We show in this proof that the inner loop can be implemented such that it comprises $\mathcal{O}(N)$ iterations only, which implies the statement of the lemma.

Note that the inner loop in Algorithm 2 constructs all events $A_j$, $j \in [N]$, in time $\mathcal{O}(N)$, and thus we only need to consider the construction of the events $A_{jj'}$, $j, j' \in [N]$. Instead of the naïve implementation from the main text, which probes all pairs of subsets $(j, j') \in [N]^2$, we consider the following variant of the Bentley-Ottmann algorithm used to identify crossings in a set of line segments (Berg et al. 2000, §2): We merge the two lists of tuples $\{(\pi_j \cdot \beta, 1) : j \in [N+1]\}$ and $\{(\pi_{j'} \cdot \beta + \varphi, 2) : j' \in [N+1]\}$ in order of non-decreasing first component; since each list is already sorted, this can be achieved in time $\mathcal{O}(N)$. We initialize the two index counters $j_1 = j_2 = 0$ and loop through the entire merged list of tuples once in sorted order. Whenever we encounter an element $(\pi_j, 1)$, we update $j_1 \leftarrow \pi_j$; whenever we encounter an element $(\pi_{j'}, 2)$, we update $j_2 \leftarrow \pi_{j'}$. After each update, we consider the intersection $A_{jj'} = \Pi_{j_1}(\beta) \cap (\Pi_{j_2}(\beta) + \varphi)$ for possible inclusion in the worst-case event A whenever $(j_1, j_2) \neq (0, 0)$. Since the merged list of tuples has length $2N + 2$, the overall algorithm evidently runs in time $\mathcal{O}(N)$. □

**Proof of Proposition 8** The proof immediately follows from Lemmas 17, 18 and 19. □

### 1.C.3 Bounding P′ and D′ with Non-Uniform Partitions

We next extend the non-uniform upper and lower bounding problems $\mathrm{P}(\boldsymbol{\pi}, \beta)$ and $\mathrm{D}(\boldsymbol{\pi}, \beta)$ of Section 1.4.1 to the data dependent case. We obtain the following upper bound on problem P′,

$$\begin{array}{ll}
\underset{p}{\text{minimize}} & \beta \cdot \sum_{k \in [K]} \omega_k(\beta) \cdot \Big[ \sum_{j \in [N]} c_j(\boldsymbol{\pi}, \beta) \cdot p_k(j) \Big] \\
\text{subject to} & p_k : [N] \mapsto \mathbb{R}_+, \ \sum_{j \in [N]} p_k(j) = 1, \ k \in [K] \\
& \sum_{j \in [N]} p_k(j) \cdot \dfrac{|A \cap \Pi_j(\beta)|}{|\Pi_j(\beta)|} \leq e^{\varepsilon} \cdot \sum_{j \in [N]} p_m(j) \cdot \dfrac{|(A - \varphi) \cap \Pi_j(\beta)|}{|\Pi_j(\beta)|} + \delta \\
& \qquad \forall k, m \in [K], \ \forall (\varphi, A) \in \mathcal{E}'_{km}(L, \beta),
\end{array} \qquad (\mathrm{P}'(\boldsymbol{\pi}, \beta))$$

as well as the following lower bound on problem D′,

$$\begin{aligned}
&\underset{p}{\text{minimize}} && \beta \cdot \sum_{k\in[K]} \omega_k(\beta) \cdot \Big[ \sum_{j\in\mathfrak{N}} \underline{c}_j(\boldsymbol{\pi},\beta) \cdot p_k(j) \Big] \\
&\text{subject to} && p_k : \mathfrak{N} \mapsto \mathbb{R}_+,\ \sum_{j\in\mathfrak{N}} p_k(j) = 1,\ k \in [K] \\
& && \sum_{j\in\mathfrak{N}} p_k(j) \cdot \frac{|A \cap \Pi_j(\beta)|}{|\Pi_j(\beta)|} \leq e^{\varepsilon} \cdot \sum_{j\in\mathfrak{N}} p_m(j) \cdot \frac{|(A-\varphi)\cap \Pi_j(\beta)|}{|\Pi_j(\beta)|} + \delta \\
& && \qquad \forall k, m \in [K],\ \forall(\varphi, A) \in \mathcal{E}'_{km}(L,\beta).
\end{aligned} \tag{D$'(\boldsymbol{\pi},\beta)$}$$

Algorithm 3 extends the ideas of Algorithm 2 to the data dependent setting; as before, extending the domain of the decisions $p$ from $[N]$ to $\mathfrak{N}$ allows us to employ the same algorithm to solve the lower bounding problem D′($\boldsymbol{\pi}, \beta$) as well. Similar arguments as in the previous section show that Algorithm 3 terminates in time $\mathcal{O}(K^2 \cdot N)$. We explain in the GitHub supplement of this work how the bounding problems P′($\boldsymbol{\pi}, \beta$) and D′($\boldsymbol{\pi}, \beta$), as well as Algorithm 3, can be generalized to account for non-uniform partitions of the set of possible query outputs $\Phi$ as well; this reduces computation times when the granularity parameter $\beta$ is small.

**Algorithm 3:** *Identification of a constraint in* $\mathrm{P}'(\boldsymbol{\pi}, \beta)$ *with maximum privacy shortfall*

```
input  : $\boldsymbol{\pi}$, $\beta$, $p$, $\Delta f$
output: constraint $(\varphi^\star, A^\star)$ with maximum privacy shortfall $V(\varphi^\star, A^\star)$
Initialize $V^\star = 0$;
for $k, m \in [K]$ do
    for $\varphi \in \{(m-k-1)\cdot\beta,\ (m-k)\cdot\beta,\ (m-k+1)\cdot\beta\} \cap [-\Delta f, \Delta f]$ do
        Initialize $A = \emptyset$ and $V = 0$;
        for $j = 1, \ldots, N$ do
            Let $A_j = \Pi_j(\beta) \setminus [-L\cdot\beta + \varphi, (L+1)\cdot\beta + \varphi)$ and update
                $$A = A \cup A_j, \quad V = V + |A_j| \cdot \frac{p_k(j)}{|\Pi_j(\beta)|}.$$
            for $j' = 1, \ldots, N$ do
                if $p_k(j)/|\Pi_j(\beta)| > e^\varepsilon \cdot p_m(j')/|\Pi_{j'}(\beta)|$ then
                    Let $A_{jj'} = \Pi_j(\beta) \cap (\Pi_{j'}(\beta) + \varphi)$ and update
                        $$A = A \cup A_{jj'}, \quad V = V + |A_{jj'}| \cdot \left[\frac{p_k(j)}{|\Pi_j(\beta)|} - e^\varepsilon \cdot \frac{p_m(j')}{|\Pi_{j'}(\beta)|}\right].$$
                end
            end
        end
        if $V > V^\star$ then
            Update $\varphi^\star = \varphi$, $A^\star = A$ and $V^\star = V$.
        end
    end
end
return $(\varphi^\star, A^\star)$ and $V^\star(\varphi, A) = V^\star - \delta$.
```

## 1.D Additional Numerical Experiments

Table 1 summarized the suboptimality of the best performing benchmark mechanisms by summing the upper and lower bound gaps. Tables 4 and 5 report these gaps separately.

| | | $\delta$ | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | **0.005** | **0.010** | **0.020** | **0.050** | **0.100** | **0.200** | **0.250** | **0.300** | **0.500** | **0.750** |
| $\varepsilon$ | **0.005** | 0.17% | 0.12% | 0.03% | 0.26% | 0.26% | 0.65% | 0.64% | 0.63% | 0.49% | 8.37% |
| | **0.010** | 0.41% | 0.33% | 0.58% | 0.54% | 0.54% | 0.68% | 0.55% | 0.55% | 0.55% | 5.26% |
| | **0.020** | 0.28% | 0.41% | 0.04% | 0.58% | 0.58% | 0.73% | 0.61% | 1.24% | 0.61% | 8.53% |
| | **0.050** | 0.50% | 0.27% | 0.38% | 0.45% | 0.73% | 0.89% | 0.80% | 1.48% | 0.87% | 8.73% |
| | **0.100** | 0.38% | 0.45% | 0.58% | 0.54% | 0.90% | 1.14% | 1.15% | 1.88% | 1.25% | 9.12% |
| | **0.200** | 0.78% | 0.87% | 1.02% | 0.92% | 0.96% | 1.66% | 1.92% | 2.61% | 1.80% | 10.24% |
| | **0.500** | 1.81% | 1.91% | 2.09% | 2.58% | 2.71% | 3.35% | 3.92% | 4.08% | 4.33% | 12.91% |
| | **1.000** | 4.70% | 4.72% | 4.66% | 4.84% | 5.10% | 5.82% | 6.03% | 6.35% | 8.09% | 16.38% |
| | **2.000** | 7.99% | 8.02% | 8.09% | 8.39% | 8.70% | 9.60% | 10.18% | 10.86% | 13.05% | 19.74% |
| | **5.000** | 12.25% | 12.28% | 12.33% | 12.50% | 12.78% | 13.37% | 13.67% | 13.97% | 14.86% | 16.74% |

Table 4: *Upper bound suboptimality of the best performing benchmark mechanisms on synthetic data independent instances with $\Delta f = 1$, $\ell_1$-loss and various combinations of $\varepsilon$ and $\delta$.*

|  |  | δ |  |  |  |  |  |  |  |  |  |
|---|---|---|---|---|---|---|---|---|---|---|---|
|  |  | **0.005** | **0.010** | **0.020** | **0.050** | **0.100** | **0.200** | **0.250** | **0.300** | **0.500** | **0.750** |
| ε | **0.005** | 1.71% | 1.09% | 1.18% | 5.84% | 18.27% | 2.90% | 48.94% | 22.55% | 49.47% | 24.97% |
|  | **0.010** | 2.48% | 1.09% | 7.42% | 1.77% | 16.55% | 2.40% | 48.63% | 22.48% | 49.36% | 28.09% |
|  | **0.020** | 2.56% | 2.42% | 0.52% | 14.50% | 13.61% | 66.71% | 47.78% | 21.50% | 49.23% | 24.84% |
|  | **0.050** | 1.35% | 1.84% | 2.18% | 17.92% | 5.50% | 65.14% | 45.35% | 20.47% | 48.72% | 24.69% |
|  | **0.100** | 4.40% | 3.73% | 8.10% | 18.83% | 32.42% | 62.71% | 41.72% | 18.97% | 47.92% | 24.38% |
|  | **0.200** | 9.65% | 8.73% | 7.88% | 16.37% | 18.86% | 58.51% | 35.79% | 16.74% | 46.53% | 23.39% |
|  | **0.500** | 21.29% | 19.50% | 23.36% | 13.78% | 35.99% | 39.20% | 25.60% | 14.41% | 41.52% | 20.88% |
|  | **1.000** | 35.41% | 35.01% | 35.57% | 35.56% | 23.52% | 26.71% | 20.59% | 15.08% | 33.71% | 16.98% |
|  | **2.000** | 25.97% | 26.16% | 25.36% | 23.24% | 23.76% | 19.10% | 16.94% | 14.81% | 21.29% | 10.77% |
|  | **5.000** | 7.07% | 7.03% | 6.96% | 6.74% | 6.37% | 5.64% | 5.27% | 4.91% | 3.79% | 2.33% |

Table 5: *Lower bound suboptimality of the best performing benchmark mechanisms on synthetic data independent instances with $\Delta f = 1$, $\ell_1$-loss and various combinations of $\varepsilon$ and $\delta$.*

Table 6 reports results similar to those of Table 1 for the $\ell_2$-loss. Moreover, our GitHub repository includes results analogous to those of Tables 4 and 5. We note that if we replace $\max\{O, 1\}$ with $O$ in the denominator of the optimality gap formula, then the gaps in Table 6 increase to more than 850% for $(\varepsilon, \delta) = (5, 0.75)$.

|  |  | $\delta$ |  |  |  |  |  |  |  |  |  |
|---|---|---|---|---|---|---|---|---|---|---|---|
|  |  | **0.005** | **0.010** | **0.020** | **0.050** | **0.100** | **0.200** | **0.250** | **0.300** | **0.500** | **0.750** |
| $\varepsilon$ | **0.005** | 2.76% | 1.83% | 1.71% | 9.12% | 26.58% | 83.29% | 66.77% | 34.72% | 36.59% | 18.06% |
|  | **0.010** | 4.29% | 2.33% | 12.41% | 3.82% | 24.47% | 83.05% | 66.55% | 34.34% | 39.28% | 15.73% |
|  | **0.020** | 4.04% | 4.39% | 0.32% | 23.37% | 22.76% | 83.06% | 66.06% | 33.60% | 39.25% | 19.53% |
|  | **0.050** | 1.66% | 2.03% | 2.86% | 27.96% | 10.67% | 82.06% | 62.81% | 31.56% | 39.05% | 18.03% |
|  | **0.100** | 4.79% | 3.81% | 11.48% | 29.04% | 47.70% | 78.46% | 60.47% | 27.93% | 38.74% | 18.13% |
|  | **0.200** | 11.32% | 9.61% | 8.40% | 24.36% | 25.68% | 73.75% | 55.80% | 24.00% | 38.23% | 19.51% |
|  | **0.500** | 21.38% | 17.35% | 25.81% | 41.55% | 46.75% | 58.92% | 34.29% | 15.86% | 34.77% | 18.67% |
|  | **1.000** | 37.40% | 38.99% | 43.91% | 55.20% | 27.47% | 31.26% | 20.88% | 12.62% | 27.77% | 17.07% |
|  | **2.000** | 28.11% | 28.60% | 26.62% | 23.70% | 24.67% | 16.97% | 15.78% | 12.87% | 20.87% | 15.76% |
|  | **5.000** | 7.85% | 7.83% | 7.78% | 7.90% | 7.74% | 7.47% | 7.35% | 7.23% | 7.19% | 7.14% |

Table 6: *Suboptimality of the best performing benchmark mechanisms on synthetic data independent instances with $\Delta f = 1$, $\ell_2$-loss and various combinations of $\varepsilon$ and $\delta$.*

# 2 Mixtures of Gaussians in Approximate Differential Privacy

## Abstract

We design a class of additive noise mechanisms that satisfy $(\varepsilon, \delta)$-differential privacy (DP) for scalar, real-valued query functions with known sensitivities. These mechanisms, which we call *mixture mechanisms*, are constructed by mixing multiple Gaussian distributions that share the same variance but differ in their means and mixture weights. The resulting distributions can be interpreted as convex combinations of a zero-mean Gaussian (as used in the analytic Gaussian mechanism (Balle and Wang 2018)) and additional Gaussians whose means depend on the sensitivity of the query function. We derive tight conditions on the variances required for $(\varepsilon, \delta)$-DP and provide efficient algorithms to compute them. Compared to the analytic Gaussian mechanism, our mechanisms yield substantially lower expected noise amplitudes ($l_1$-loss) and variances ($l_2$-loss for zero-mean distributions). In low-privacy regimes, our mechanisms approach optimality, mitigating nearly all of the optimality gaps of the analytic Gaussian mechanism. Finally, we apply our mechanisms to a linear classification task trained via the DP counterpart of proximal coordinate descent. We find that the theoretical improvements in expected losses carry over to superior generalization performance across standard real-world datasets.

## 2.1 Introduction

Differential privacy (DP) (Dwork et al. 2006b, Dwork and Roth 2014) is a formal framework that enables the release of statistics of datasets while providing a quantifiable privacy guarantee to the individuals represented. By definition, DP guarantees hold under any privacy attack, including reconstruction (Dinur and Nissim 2003, Dwork et al. 2017) and de-identification attacks (Sweeney 1997, Heffetz and Ligett 2014). As a result, differentially private (or privacy-preserving) variants of machine learning (ML) algorithms offer protection against a broad range of threats, including membership inference attacks tailored to specific ML tasks and platforms offering ML as a service (see (Shokri et al. 2017, Hu et al. 2022) for a review). We refer the reader to the surveys (Ji et al. 2014, Gong et al. 2020) for DP in ML and (Demelius et al. 2025) for DP in deep learning.

While the definition of DP has inspired a broad range of related notions of computational privacy (Desfontaines and Pejó 2020), we focus on the most widely adopted formulation, known as $(\varepsilon, \delta)$-DP, or approximate differential privacy (Dwork et al. 2006a). Approximate DP is a property of a randomized algorithm $\mathcal{A}$, which maps databases $D \in \mathcal{D}$ to random outputs $\omega \in \Omega$.

Specifically, for any $\varepsilon, \delta \geq 0$, the algorithm $\mathcal{A}$ satisfies $(\varepsilon, \delta)$-DP if

$$\mathbb{P}[\mathcal{A}(D) \in A] \leq e^{\varepsilon} \cdot \mathbb{P}[\mathcal{A}(D') \in A] + \delta \quad \forall (D, D') \in \mathcal{N},\ \forall A \in \mathcal{F},$$

where $(\Omega, \mathcal{F}, \mathbb{P})$ is a probability space and $\mathcal{N} \subset \mathcal{D} \times \mathcal{D}$ denotes the set of neighboring databases that differ in a single individual. Due to the strong privacy guarantees promised to individuals (Dwork 2011), various Bayesian (Vadhan 2017, §1.6) and information-theoretic (Cuff and Yu 2016) interpretations, and a wide range of properties such as composition (Dwork and Roth 2014, Thm 3.16) and post-processing (Dwork and Roth 2014, Prop 2.1), DP found manifold applications in statistics and machine learning (Chaudhuri and Monteleoni 2008, Friedman and Schuster 2010, Chaudhuri et al. 2011, Abadi et al. 2016, Cai and Kou 2019), large language models (Harder et al. 2020) and optimization (Mangasarian 2011, Hsu et al. 2014, Han et al. 2016, Hsu et al. 2016). It also emerged as the *de facto* standard in industry (see (Desfontaines 2021b) for an up-to-date list).

In this work, we study *additive noise mechanisms* that perturb a scalar, real-valued query function $q : \mathcal{D} \mapsto \mathbb{R}$ by returning $\mathcal{A}(D) = q(D) + \tilde{X}$ for a carefully designed random variable $\tilde{X}$, to share an approximate answer to the query $q(D)$ while preserving privacy via an additive noise $\tilde{X}$. Since $\tilde{X}$ is not a function of $D$, such noise mechanisms are data independent (see (Nissim et al. 2007, McSherry and Talwar 2007) for examples of data dependent mechanisms). Typically, the noise $\tilde{X}$ must be a function of the privacy parameters $\varepsilon$ and $\delta$ as well as the sensitivity $\Delta$ of the query $q$ where $\Delta = \sup\{|q(D) - q(D')| : (D, D') \in \mathcal{N}\}$. Arguably, the most common mechanism in this context is the Gaussian mechanism (Dwork and Roth 2014, App A), where $\tilde{X}$ follows a zero-mean Gaussian distribution with standard deviation $\sigma = \sqrt{2 \log(1.25/\delta)(\Delta/\varepsilon)^2}$.

Differential privacy suffers from a *privacy-accuracy tradeoff* (Alvim et al. 2011), where stronger privacy guarantees tend to deteriorate the utility gained from data (*i.e.*, $\mathcal{A}(D)$ may significantly deviate from $q(D)$). This can already be observed in the definition of the Gaussian mechanism, where the standard deviation of the noise increases as $\varepsilon$ and $\delta$ decrease. To this end, an active area of research aims to find optimal distributions that minimize a pre-specified loss, such as the standard deviation, amplitude $\mathbb{E}[|\tilde{X}|]$, and power $\mathbb{E}[\tilde{X}^2]$. The optimal noise distributions are known for the case when $\varepsilon = 0$ (Geng et al. 2019), or $\delta = 0$ (Soria-Comas and Domingo-Ferrer 2013, Geng and Viswanath 2014). For privacy regimes with $\varepsilon, \delta > 0$, the design of optimal mechanisms is less understood, and the existing additive noise mechanisms can be suboptimal. Indeed, (Balle and Wang 2018) show that the Gaussian mechanism may be far from optimal and derive the optimal Gaussian distribution, called the *analytic Gaussian mechanism*, by numerically tuning the standard deviation $\sigma$ for a given sensitivity $\Delta$ and privacy regime $(\varepsilon, \delta)$ with a bisection search method. However, Gaussian distributions are themselves suboptimal as demonstrated by (Geng et al. 2020): a truncated Laplace mechanism, sampling noise from a Laplace distribution whose tails are truncated, can significantly improve the noise amplitude and noise power attained by the analytic Gaussian mechanism in all privacy regimes. Yet, the (analytic) Gaussian mechanism may still be preferred in practice for several reasons: unlike truncated distributions, the Gaussian distribution has an unbounded support and thus does not suffer from distinguishing events (events with $\mathbb{P}[\mathcal{A}(D) \in A] = 0$; (Dwork and Rothblum 2016, Remark 1.3)). Moreover, despite having unbounded supports, the Gaussian distributions enjoy from the empirical rule where *almost all* of the noise sampled is within $[-3\sigma, 3\sigma]$ (Desfontaines 2021a), and they are closed under summations which enables several powerful Gaussian privacy accounting techniques (how $\varepsilon$ and $\delta$ accumulate over time (Bun and Steinke 2016, Dong et al. 2022)) for application areas where one iteratively releases multiple $(\varepsilon, \delta)$-DP statistics.

Recently, (Selvi et al. 2025) proposed a numerical optimization framework to design optimal noise distributions that satisfy $(\varepsilon, \delta)$-DP for $\varepsilon, \delta > 0$. Their method begins by partitioning a bounded interval, centered at zero and of sufficient radius, and assigning a uniform mass to each partition. They then optimize the weights assigned to these intervals. The authors show that, as the partition size decreases, the resulting distributions converge to a lower bound on the optimal value attainable by any $(\varepsilon, \delta)$-DP mechanism. In practice, they solve a sequence of optimization problems via a cutting-plane method, iterating until a mechanism is found to be within an optimality tolerance. They offer three key insights: *(i)* the aforementioned privacy mechanisms (Gaussian, analytic Gaussian, truncated Laplace) can be significantly suboptimal, with suboptimalities reaching up to 700% in low-privacy regimes; *(ii)* optimal distributions are not monotonic, and enforcing monotonicity can significantly increase expected losses; *(iii)* optimal distributions appear to be multimodal with the density peaking at every $\Delta$-length interval and the ratio of density between peaks approximately equal to $e^{\varepsilon}$. Such a numerical optimization framework is not free of drawbacks since the optimal distributions lack closed-form representations, they require iterative numerical optimization with no guarantees on the number of iterations, they are significantly slower to tune than alternative mechanisms, and they have bounded supports. Yet, we build on the insights resulting from such optimized distributions and develop new privacy mechanisms based on mixtures of Gaussians, where each Gaussian is shifted by $\Delta$ from the others and scaled down by a factor of $e^{-\varepsilon}$. Figure 9 shows an example of such distributions.

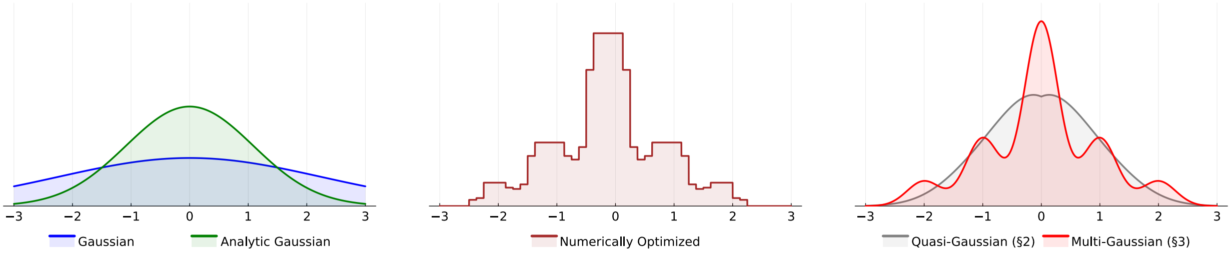


Figure 9: *Different noise distributions that guarantee* $(1, 0.1)$*-DP.*

This section unfolds as follows. Section 2.2 introduces a mechanism that samples additive noise from a mixture of a zero-mean Gaussian and a quasi-Gaussian distribution. It characterizes the parameters required for this distribution to satisfy $(\varepsilon, \delta)$-DP and develops an efficient algorithm to compute them. Section 2.3 presents similar results for a different mechanism, where the additive noise is drawn from a mixture of multiple Gaussian distributions with identical variances but varying means and mixture weights. Section 2.4 presents a broad set of numerical experiments comparing the expected losses of our mechanisms to those of the analytic Gaussian mechanism. We observe that mixtures of Gaussians can substantially improve upon the analytic Gaussian mechanism across both low- and high-privacy regimes. Moreover, the proposed mixture mechanisms are near-optimal in low-privacy regimes, closing nearly all of the optimality gap of the analytic Gaussian mechanism. Finally, through a machine learning experiment (DP proximal coordinate descent to train a regularized classifier), we show that the reductions in expected losses also lead to improved model performance. We conclude the work in Section 2.6. All proofs are relegated to appendices, and all source codes are made available.

**Notation.** We denote by $\log(\cdot)$ the logarithm in base $e = \exp(1)$. The cumulative distribution function of a standard Gaussian distribution is denoted by $\Phi$, and $\Phi'$ is its derivative (probability density function). For a univariate function $f$, we use $\mathrm{d}f(x)/\mathrm{d}x$ to denote its derivative, while for a multivariate function $\partial f(x, y)/\partial x$ denotes the partial derivative. We use the notation $\tilde{X} \sim f(\cdot)$ to denote a random variable whose distribution is specified by the probability density function $f$. Bold lowercase symbols are vectors. The nonnegative part of a real number $x$ is $(x)^+ = \max\{0, x\}$. For a vector $\boldsymbol{x}$, the $l_p$ norm is denoted by $\|\boldsymbol{x}\|_p$, $p \geq 1$. We use $\mathcal{O}(\cdot)$ for the big O notation.

## 2.2 Quasi-Gaussian mixtures

We define a distribution whose density function is the mixture of a zero-mean Gaussian density and a quasi-Gaussian density. We will use this as the additive noise distribution in preserving $(\varepsilon, \delta)$-DP.

**Definition 2** (quasi-Gaussian mixture)**.** *For given privacy parameter $\varepsilon > 0$, query sensitivity $\Delta > 0$, and scale parameter $\sigma > 0$, the* quasi-Gaussian mixture distribution *is defined with the probability density function*

$$f_{\mathrm{q}}(x;\sigma) = \frac{e^{\varepsilon}}{c} \exp\left(-\frac{x^2}{2\sigma^2}\right) + \frac{1}{c} \exp\left(-\frac{(|x|-\Delta)^2}{2\sigma^2}\right), \tag{25}$$

*where $c = \sqrt{2\pi}\sigma(e^{\varepsilon} + 2\Phi(\Delta/\sigma))$ is the normalization constant.*

Appendices 2.A.1-2.A.3 show that Definition 2 specifies a well-defined distribution and plot it, derive its cumulative distribution function, and show how to sample from it. In Appendix 2.A.4, we derive the expectation of $|\tilde{X}|$ (noise amplitude) and $\tilde{X}^2$ (noise power) in closed form where $\tilde{X} \sim f_{\mathrm{q}}(\cdot\,;\sigma)$. Next, we show that this distribution can be used to obtain a feasible additive noise mechanism. That is, for a query function $q : \mathcal{D} \mapsto \mathbb{R}$ with sensitivity $\Delta$, and for privacy parameters $\varepsilon, \delta$, there exists $\sigma > 0$ so that the randomized algorithm $\mathcal{A}(D) := q(D) + \tilde{X},\ D \in \mathcal{D}$, satisfies $(\varepsilon, \delta)$-DP if $\tilde{X} \sim f_{\mathrm{q}}(\cdot\,;\sigma)$.

**Theorem 3** (quasi-Gaussian mechanism)**.** *For privacy parameters $\varepsilon > 0$ and $\delta \in (0,1)$, and query sensitivity $\Delta > 0$, an additive noise mechanism whose noise follows a quasi-Gaussian mixture distribution with density $f_{\mathrm{q}}(x;\sigma)$ satisfies $(\varepsilon, \delta)$-DP if $\sigma \geq \max\{\sigma_1, \sigma_2\}$ where:*

$$\sigma_1 = \min\{\sigma : h_1(\sigma) + h_2(\sigma) \geq 0\} \tag{26a}$$

*for*

$$h_1(\sigma) = e^{2\varepsilon}\Phi\left(-\frac{\varepsilon\sigma}{\Delta} - \frac{\Delta}{\sigma}\right) - \Phi\left(-\frac{\varepsilon\sigma}{\Delta} + \frac{\Delta}{\sigma}\right) \text{ and } h_2(\sigma) = \left(e^{\varepsilon} + 2\Phi\left(\frac{\Delta}{\sigma}\right)\right)\delta \tag{26b}$$

*as well as*

$$\sigma_2 = \min\left\{\sigma : \frac{f_{\mathrm{q,max}}(\sigma)}{f_{\mathrm{q,min}}(\sigma)} \leq e^{\varepsilon}\right\} \tag{27a}$$

*for*

$$f_{\mathrm{q,max}}(\sigma) = \max_{x\in[0,\Delta]} f_{\mathrm{q}}(x;\sigma) \textit{ and } f_{\mathrm{q,min}}(\sigma) = \min_{x\in[0,\Delta]} f_{\mathrm{q}}(x;\sigma). \tag{27b}$$

In Theorem 3, the definition of $\sigma_2$ is independent of $\delta$, hence, for small values of $\delta$, we have $\sigma_1 > \sigma_2$, while for large values of $\delta$ we have $\sigma_1 < \sigma_2$. Since the variance of the quasi-Gaussian mixture distribution scales with $\sigma$, we will be interested in the smallest $\sigma$ satisfying the above result, that is, $\sigma = \max\{\sigma_1, \sigma_2\}$. To this end, the rest of this section is devoted to developing efficient algorithms to find the exact values of $\sigma_1$ and $\sigma_2$. From here on, we fix the privacy parameters $\varepsilon > 0$ and $\delta \in (0,1)$ as well as the query sensitivity $\Delta > 0$ and do not repeat their definitions.

**Lemma 20.** *The function $\sigma \mapsto h_1(\sigma) + h_2(\sigma)$ is monotonically increasing on the region $\sigma \in (0, \sqrt{2(\varepsilon - \log\delta)}\Delta/\varepsilon)$, and nonnegative for all $\sigma \geq \sqrt{2(\varepsilon - \log\delta)}\Delta/\varepsilon$. If additionally $e^{\varepsilon} + 2 \geq \delta^{-1}$, then this function is nonnegative everywhere.*

Lemma 20 shows that $\sigma_1$ as defined in (26a) satisfies $\sigma_1 \in (0, \sqrt{2(\varepsilon - \log\delta)}\Delta/\varepsilon)$ and the constraint in its definition satisfies monotonicity on this region, thus allowing us to search for $\sigma_1$ via bisection search. We next provide a similar result for $\sigma_2$. As a step towards showing this, we will first study the max and min problems of (27b) in the following abridged result (more details are in the appendices).

**Lemma 21** (abridged)**.** *For any $\sigma > 0$, the following statements hold:*

*(i)* *$f_{\mathrm{q}}(\cdot;\sigma)$ is unimodal on the interval*

$$\mathcal{R}_{\max}(\sigma) = \left(0, \frac{\Delta - \sqrt{(\Delta^2 - 4\sigma^2)^+}}{2}\right)$$

*with a unique maximum. The maximizer also solves $\max_{x\in[0,\Delta]} f_{\mathrm{q}}(x;\sigma)$.*

*(ii)* *$f_{\mathrm{q}}(\cdot;\sigma)$ is unimodal on the interval*

$$\mathcal{R}_{\min}(\sigma) = \left(\frac{\Delta}{2}, \frac{\Delta + \sqrt{(\Delta^2 - 4\sigma^2)^+}}{2}\right)$$

*with a unique minimum. If $\mathcal{R}_{\min}(\sigma) = \emptyset$, then $x = \Delta$ solves $\min_{x\in[0,\Delta]} f_{\mathrm{q}}(x;\sigma)$. Otherwise, either $x = \Delta$, or the minimizer of the unimodal region solves $\min_{x\in[0,\Delta]} f_{\mathrm{q}}(x;\sigma)$.*

Lemma 21 allows us to find the values $f_{\mathrm{q,max}}(\sigma)$ and $f_{\mathrm{q,max}}(\sigma)$ values as defined in (27b) for a given $\sigma > 0$ within unimodal regions, allowing the use of a golden-section search method. This lemma also leads us to the following result that will be helpful in computing $\sigma_2$.

**Lemma 22.** *The function $\sigma \mapsto \max_{x\in[0,\Delta]} f_{\mathrm{q}}(x;\sigma)/\min_{x\in[0,\Delta]} f_{\mathrm{q}}(x;\sigma)$ is monotonically nonincreasing in $\sigma$, and is upper bounded by $e^{\varepsilon}$ at the point $\sigma = \sqrt{\Delta^2/(2\varepsilon)}$.*

Lemma 22 allows us to find $\sigma_2$ as defined in (27a) via a bisection search in the region $(0, \sqrt{\Delta^2/(2\varepsilon)})$.

We now present Algorithm 4 that computes a value for $\sigma > 0$ to ensure that the quasi-Gaussian mixture distribution satisfies $(\varepsilon, \delta)$-DP. Note that this does not guarantee that we find the minimum possible such $\sigma$, since the condition for $\sigma \geq \max\{\sigma_1, \sigma_2\}$ satisfying $(\varepsilon, \delta)$-DP (*cf.* Theorem 3) is itself a sufficient condition. However, we find the *exact* values of $\sigma_1$ and $\sigma_2$ in Theorem 3 so that we return the tightest possible $\sigma \geq \max\{\sigma_1, \sigma_2\}$ within the scope of our analysis. In the presentation of Algorithm 4 we use $\texttt{BISECTION}(l, r, \psi(\sigma) = 0)$ to denote the bisection search algorithm to find the root of a monotonic function $\psi$ within the region $(l, r)$. Moreover, Algorithm 5, which is called within Algorithm 4, uses $\texttt{GOLDEN}(l, r, f(x))$ to denote the golden-section search algorithm to find the maximum value of a unimodal function $f$ within the range $(l, r)$. In Appendix 2.A.9, we share plots that visualize various steps of the algorithms presented in this section, for further intuition.

**Theorem 4.** *Algorithm 4 returns the minimum $\sigma > 0$ satisfying the $(\varepsilon, \delta)$-DP condition in Theorem 3 in time $\mathcal{O}(\log \Delta + \frac{\Delta}{\varepsilon} \log \Delta + \log(-\frac{\log \delta}{\varepsilon}))$.*

## 2.3 Multi-Gaussian mixtures

In this section, we work with a distribution that is the mixture of an odd ($2K+1$, $K \in \mathbb{N}$) number of Gaussian distributions with identical variances but varying means and mixture weights. We will use this as the distribution of the additive noise $\tilde{X}$, which preserves $(\varepsilon, \delta)$-DP.

**Definition 3** (multi-Gaussian mixture). *For given privacy parameter $\varepsilon > 0$, query sensitivity $\Delta > 0$, scale parameter $\sigma > 0$, and modality parameter $K \in \mathbb{N}$, the* multi-Gaussian mixture distribution *is defined with the probability density function*

$$f_{\mathrm{m}}(x; \sigma, K) = \frac{1}{c_K} \sum_{k=-K}^{K} e^{-|k|\varepsilon} \phi(x; k\Delta, \sigma), \tag{28}$$

**Algorithm 4:** Tuning $\sigma > 0$ for the quasi-Gaussian mixture.

```
Input: privacy parameters $\varepsilon, \delta > 0$, sensitivity $\Delta > 0$
Output: smallest $\sigma > 0$ (with respect to Theorem 3) of quasi-Gaussian mixture
        distribution (25) that satisfies $(\varepsilon, \delta)$-DP as an additive noise mechanism
if $e^\varepsilon + 2 < \delta^{-1}$ then
    Set $\sigma_1 = \texttt{BISECTION}(l_1, r_1, \psi_1(\sigma_1) = 0)$ where $l_1 = 0,\ r_1 = \sqrt{2(\varepsilon - \log\delta)}\Delta/\varepsilon$ are the
     left and right limits of the bisection search region, and
        $$\psi_1(\sigma) = e^{2\varepsilon}\Phi\left(-\frac{\varepsilon\sigma}{\Delta} - \frac{\Delta}{\sigma}\right) - \Phi\left(-\frac{\varepsilon\sigma}{\Delta} + \frac{\Delta}{\sigma}\right) + \left(e^\varepsilon + 2\Phi\left(\frac{\Delta}{\sigma}\right)\right)\delta$$
     is the monotonically increasing function whose root is sought.
else
    Set $\sigma_1 = 0$.
Set $\sigma_2 = \texttt{BISECTION}(l_2, r_2, \psi_2(\sigma_2) = 0)$ where $l_2 = 0,\ r_2 = \sqrt{\Delta^2/(2\varepsilon)}$ are the left and
 right limits of the bisection search region, and
    $$\psi_2(\sigma) = \max_{x \in [0,\Delta]} f_{\mathrm{q}}(x;\sigma) \Big/ \min_{x \in [0,\Delta]} f_{\mathrm{q}}(x;\sigma) - e^\varepsilon$$
 is the monotonically decreasing function whose root is sought. The computation of
 max- and min-terms are from Algorithm 5.
return $\sigma = \max\{\sigma_1, \sigma_2\}$
```

*where $\phi(x;\mu,\sigma) = \exp(-\frac{(x-\mu)^2}{2\sigma^2})$ is the unnormalized probability density function of a Gaussian with mean $\mu$ and variance $\sigma^2$, and $c_K = \sqrt{2\pi}\sigma\sum_{k=-K}^{K} e^{-|k|\varepsilon}$ is the normalization constant.*

Appendix 2.B.1 reviews that Definition 3 specifies a well-defined distribution and plots it, derives its cumulative distribution function, and discusses how to sample from it. Appendix 2.B.2 derives the expectation of $|\tilde{X}|$ (noise amplitude) and $\tilde{X}^2$ (noise power) in closed form where $\tilde{X} \sim f_{\mathrm{m}}(\cdot;\sigma,K)$. We remark that for $K = 0$, Definition 3 coincides with the Gaussian distribution, and we know that there are feasible $\sigma > 0$ values providing $(\varepsilon,\delta)$-DP guarantees. We generalize this for any $K \in \mathbb{N}$.

**Theorem 5** (multi-Gaussian mechanism)**.** *For privacy parameters $\varepsilon > 0$ and $\delta \in (0,1)$, query sensitivity $\Delta > 0$, and the modality parameter $K \in \mathbb{N}$, consider an additive noise mechanism whose noise follows a multi-Gaussian mixture distribution with density $f_{\mathrm{m}}(x;\sigma,K)$. For a given $\eta \in (0,1)$ and any $\beta \leq \sqrt{2\pi}\eta\sigma\delta$, a sufficient condition for this mechanism to satisfy $(\varepsilon,\delta)$-DP*

**Algorithm 5:** Computing max and min problems (27b) for a given $\sigma > 0$.

```
**Input:** privacy parameters $\varepsilon, \delta > 0$, sensitivity $\Delta > 0$, and $\sigma > 0$ of $f_{\mathrm{q}}(\cdot;\sigma)$
**Output:** $f_{\mathrm{q,max}}(\sigma) = \max_{x\in[0,\Delta]} f_{\mathrm{q}}(x;\sigma)$ and $f_{\mathrm{q,min}}(\sigma) = \min_{x\in[0,\Delta]} f_{\mathrm{q}}(x;\sigma)$
**Initialize:**
    $x_1 = \frac{\Delta - \sqrt{(\Delta^2 - 4\sigma^2)^+}}{2}, \; x_2 = \frac{\Delta + \sqrt{(\Delta^2 - 4\sigma^2)^+}}{2}, \; t = -e^{\varepsilon} + \frac{\Delta - x_2}{x_2} \exp\left(\frac{2x_2\Delta - \Delta^2}{2\sigma^2}\right).$
if $\Delta^2 \leq 4\sigma^2$ then
    Set $x_{\min} = \Delta$ and $x_{\max} = \texttt{GOLDEN}(0, \Delta/2, f_{\mathrm{q}}(x;\sigma))$
else
    if $\Delta^2 > 4\sigma^2$ and $t \leq 0$ then
        Set $x_{\min} = \Delta$ and $x_{\max} = \texttt{GOLDEN}(0, x_1, f_{\mathrm{q}}(x;\sigma))$
    else
        Set $x_{\min} = \Delta$ and $x_{\max} = \texttt{GOLDEN}(0, x_1, f_{\mathrm{q}}(x;\sigma))$
        Set $x_{\mathrm{tmp}} = \texttt{GOLDEN}(\Delta/2, x_2, -f_{\mathrm{q}}(x;\sigma))$
        if $f_{\mathrm{q}}(x_{\mathrm{tmp}};\sigma) < f_{\mathrm{q}}(x_{\min};\sigma)$ then
            Update $x_{\min} = x_{\mathrm{tmp}}$
return $f_{\mathrm{q,max}}(\sigma) = f_{\mathrm{q}}(x_{\max};\sigma)$ and $f_{\mathrm{q,min}}(\sigma) = f_{\mathrm{q}}(x_{\min};\sigma)$
```

*is given by:*

$$\int_{-\infty}^{\infty} \min\{e^{\varepsilon} f_{\mathrm{m}}(x;\sigma,K) - f_{\mathrm{m}}(x+\varphi;\sigma,K), 0\}\mathrm{d}x + (1-\eta)\delta \geq 0 \;\; \forall \varphi \in \{0, \beta, 2\beta, \ldots, \Delta\}. \tag{29}$$

Theorem 5 provides a sufficient condition for $(\varepsilon,\delta)$-DP by ensuring a stronger notion that replaces $\delta$ with $(1-\eta)\delta$, and in return avoids satisfying the definition over an uncountable set of neighbors by checking it only on a discretized grid for $\varphi$. As $\eta$ approaches zero, this condition converges to the standard definition of $(\varepsilon,\delta)$-DP. The next result shows that condition (29) is monotonic in $\sigma > 0$, and we can therefore search for minimum $\sigma > 0$ satisfying this condition via bisection search efficiently.

**Lemma 23.** *For a given $\eta \in (0,1)$, if $\sigma > 0$ satisfies (29), then any $\sigma' \geq \sigma$ must also satisfy (29). Furthermore, by construction, $\sigma = \sigma_{\mathrm{g}}$ satisfies (29) where $\sigma_{\mathrm{g}}$ denotes the standard deviation required by the (standard or analytic) Gaussian mechanism preserving $(\varepsilon,(1-\eta)\delta)$-DP.*

The proof of Lemma 23 establishes a more general result that if a multi-Gaussian mixture mechanism with parameter $\sigma > 0$ satisfies $(\varepsilon,\delta)$-DP, so does one with $\sigma' \geq \sigma$. This proves an analogous result to that of the Gaussian mechanisms: larger values of $\sigma$ correspond to stronger privacy guarantees.

```
Algorithm 6: Tuning $\sigma > 0$ for the multi-Gaussian mixture.
  Input: privacy parameters $\varepsilon, \delta > 0$, sensitivity $\Delta > 0$, modality parameter $K \in \mathbb{N}$ and
         discretization parameter $\eta \in (0, 1)$
  Output: smallest $\sigma > 0$ (with respect to Theorem 5) of multi-Gaussian mixture
          distribution (28) that satisfies $(\varepsilon, \delta)$-DP as an additive noise mechanism
  Set $l = 0$ and $r =$ ANALYTIC_GAUSSIAN$(\varepsilon, (1 - \eta)\delta, \Delta)$
  Set $\sigma =$ BISECTION$(l, r, \psi(\sigma; \eta) = 0)$ where $\psi(\sigma; \eta)$ is the monotonically increasing
    privacy shortfall function whose root is sought. The computation of $\psi(\sigma; \eta)$ is from
    Algorithm 7
  return $\sigma$
```

We now present Algorithm 6 that computes a value for $\sigma > 0$ to ensure that the multi-Gaussian mixture distribution with a given modality parameter $K \in \mathbb{N}$ and discretization parameter $\eta \in (0, 1)$ satisfies $(\varepsilon, \delta)$-DP. Note that this does not guarantee that we find the minimum possible such $\sigma$, since the condition (29) for $\sigma$ satisfying $(\varepsilon, \delta)$-DP (*cf.* Theorem 5) is itself a sufficient condition. However, upon fixing the values of $(K, \eta)$, we find the *smallest* value of $\sigma$ satisfying (29), hence we find the optimal $\sigma$ within the scope of our analysis. In the presentation of Algorithm 6 we use `ANALYTIC_GAUSSIAN`$(\varepsilon, \delta, \Delta)$ to denote the computation of the standard deviation $\sigma_{\text{g}}$ required by the analytic Gaussian mechanism and `BISECTION`$(l, r, \psi(\sigma; \eta) = 0)$ to denote the bisection search algorithm to find the root of a monotonic function $\psi(\cdot; \eta)$ within the region $(l, r)$. Note that Algorithm 7, which is called within Algorithm 6, numerically computes the integral on the left-hand side of (29) multiple times. In our numerical experiments, we use the `QuadGK` package of the Julia programming language and propose some algorithmic improvements for numerical efficiency.

**Theorem 6.** *For any given $K \in \mathbb{N}$ and $\eta \in (0, 1)$, Algorithm 6 returns the minimum $\sigma > 0$ satisfying the $(\varepsilon, \delta)$-DP condition in Theorem 5 in time $\mathcal{O}(\frac{\Delta^2 K^2}{\eta\delta}(\log\frac{\Delta}{\varepsilon} + \log\log\frac{1}{\delta}))$.*

## 2.4 Numerical experiments

In this section, we compare the performance of our mechanisms, namely, the quasi-Gaussian mixture (*cf.* §2.2) and the multi-Gaussian mixture (*cf.* §2.3) mechanisms, with the performance of the optimal Gaussian distribution (the analytic Gaussian). For each pair of $(\varepsilon, \delta)$, we tune the parameters of the quasi-Gaussian and multi-Gaussian mixture distributions via Algorithms 4 and 6 to satisfy $(\varepsilon, \delta)$-DP. We then evaluate and compare the resulting distributions in terms of their expected noise amplitudes, $\mathbb{E}[|\tilde{X}|]$ ($l_1$-loss; results in this section), and noise powers,

**Algorithm 7:** Computing the worst-case left-hand side of (29) for a given $\sigma > 0$.

```
Input: privacy parameters $\varepsilon, \delta > 0$, sensitivity $\Delta > 0$, modality parameter $K \in \mathbb{N}$,
       discretization parameter $\eta \in (0, 1)$ and scale parameter $\sigma > 0$
Output: privacy shortfall $\psi(\sigma; \eta)$ to be called from Algorithm 6
Initialize: $\mathcal{S} = \emptyset$ and $\beta = \Delta / \lceil \Delta / (\sqrt{2\pi}\eta\sigma\delta) \rceil$
for $\varphi \in \{0, \beta, 2\beta, \ldots, \Delta\}$ do
    Numerically compute the one-dimensional integral
        $$s = \int_{-\infty}^{\infty} \min\{e^{\varepsilon} f_{\mathrm{m}}(x; \sigma, K) - f_{\mathrm{m}}(x + \varphi; \sigma, K), 0\} \mathrm{d}x,$$
    and append $\mathcal{S} = \mathcal{S} \cup \{s\}$
Set $s_{\min} = \min\{\mathcal{S}\}$
return $\psi(\sigma; \eta) = s_{\min} + (1 - \eta)\delta$
```

$\mathbb{E}[\tilde{X}^2]$ ($l_2$-loss; results in the appendices), where $\tilde{X}$ is a random variable drawn from each of these privacy-preserving additive noise distributions. All experiments fix the query sensitivity to $\Delta = 1$ without loss of generality since $\Delta$ linearly scales the losses and its impact cancels out when we report the percentage improvements.

Our Julia (MIT license) implementation is provided in the supplementary materials, where we rely on the `Roots` package for the bisection search method in Algorithms 4 and 6, the `Optim` package for the golden-section search method in Algorithm 5, and `QuadGK` for the one-dimensional integral evaluation in Algorithm 7. All experiments were run on Intel Xeon 2.66GHz cluster nodes with 16GB of memory in single-core and single-thread mode with a wall-clock time limit of 1-hour for our algorithms and 24-hour for the numerical-optimization benchmark (Selvi et al. 2025) for obtaining lower bounds.

### 2.4.1 Comparison of the quasi-Gaussian mechanism with the analytic Gaussian mechanism

Recall that the analytic Gaussian mechanism with density $\propto \exp(-x^2/(2\sigma^2))$ is the optimal Gaussian mechanism in the sense that it returns the smallest variance for a zero-mean Gaussian distribution that satisfies $(\varepsilon, \delta)$-DP. The quasi-Gaussian mechanism (25), on the other hand, takes the convex combination of such a Gaussian density with another quasi-Gaussian density $\propto \exp(-(|x| - \Delta)^2/(2\sigma^2))$ using mixture weights (proportional to $e^{\varepsilon}$). Table 7 demonstrates that such an approach can offer significant improvements in the expected $l_1$-loss. The entries of the table represent $100\% \cdot (q - a)/a$ where $q$ and $a$ are the expected losses attained by our quasi-

| $\delta \downarrow \mid \varepsilon \rightarrow$ | 0.1 | 0.25 | 0.5 | 0.75 | 1 | 2 | 3 | 4 | 5 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| 0.0001 | **-0.71%** | **-1.53%** | **-2.51%** | **-3.07%** | **-3.24%** | **-0.97%** | **3.70%** | **8.78%** | **13.69%** | **59.40%** |
| 0.0005 | **-0.80%** | **-1.73%** | **-2.82%** | **-3.41%** | **-3.56%** | **-0.70%** | **5.11%** | **11.56%** | **18.10%** | **56.14%** |
| 0.001 | **-0.84%** | **-1.83%** | **-2.98%** | **-3.59%** | **-3.73%** | **-0.50%** | **6.01%** | **13.35%** | **21.08%** | **54.48%** |
| 0.005 | **-0.92%** | **-2.08%** | **-3.41%** | **-4.09%** | **-4.16%** | **0.41%** | **9.54%** | **20.79%** | **37.45%** | **49.75%** |
| 0.01 | **-0.88%** | **-2.17%** | **-3.62%** | **-4.32%** | **-4.34%** | **1.20%** | **12.41%** | **27.97%** | **56.58%** | **47.20%** |
| 0.02 | **-0.72%** | **-2.20%** | **-3.81%** | **-4.55%** | **-4.48%** | **2.56%** | **17.43%** | **53.82%** | **53.28%** | **44.20%** |
| 0.05 | **0.01%** | **-1.92%** | **-3.91%** | **-4.70%** | **-4.39%** | **6.73%** | **45.29%** | **47.67%** | **47.63%** | **39.27%** |
| 0.1 | **1.46%** | **-1.09%** | **-3.54%** | **-4.29%** | **-3.43%** | **21.24%** | **37.64%** | **41.25%** | **41.83%** | **34.41%** |
| 0.15 | **3.00%** | **-0.03%** | **-2.77%** | **-3.18%** | **-1.25%** | **23.61%** | **31.70%** | **36.35%** | **37.46%** | **30.86%** |
| 0.25 | **-2.94%** | **3.21%** | **1.90%** | **6.11%** | **3.63%** | **10.04%** | **21.68%** | **28.22%** | **30.29%** | **25.22%** |

Table 7: *Improvement (% of $l_1$-loss) of the quasi-Gaussian mixture mechanism (green) over the analytic Gaussian mechanism (red).*

Gaussian mechanism and the analytic Gaussian mechanism, respectively. We observe that in 54 of the cases (out of 100) the quasi-Gaussian mechanism improves over the analytic Gaussian mechanism. Across all instances, the mean improvement is 12.42% (sd 19.52) and the median improvement is 1.68%. The mean and median would change to 4.44% (sd 8.30) and 1.47%, respectively, should we revise our improvement metric to $100\% \cdot (q-a)/\max\{1,a\}$ to eliminate the effect of smaller expected losses.

### 2.4.2 Comparison of the multi-Gaussian mechanism with the analytic Gaussian mechanism

Recall that the multi-Gaussian mechanism constructs a mixture of Gaussians such that the resulting distribution exhibits a peak at every interval of length $\Delta$. When $K = 0$, the distribution is a zero-mean Gaussian distribution, and for $K > 0$ we obtain the mixture of $2K+1$ Gaussians. Table 8 shows that mixing multiple Gaussians can significantly outperform the analytic Gaussian mechanism. The entries of the table represent $100\% \cdot (m-a)/a$ where $m$ and $a$ are the expected losses attained by our multi-Gaussian mechanism (best of $K \in \{1, \ldots, 10\}$; fixes $\eta = 0.01$ in Algorithm 6) and the analytic Gaussian mechanism, respectively. We observe that in 90 of the cases (out of 100) the multi-Gaussian mechanism improves over the analytic Gaussian mechanism. Across all instances, the mean improvement is 47.88% (sd 36.81) and the median improvement is 51.48%. The mean and median would change to 28.36% (sd 21.77) and 27.63%, respectively, should we revise our improvement metric to $100\% \cdot (m-a)/\max\{1,a\}$ to eliminate the effect of smaller expected losses.

Appendix 2.C.1 shows that the multi-Gaussian mechanism is near-optimal in low-privacy

| $\delta \downarrow \mid \varepsilon \rightarrow$ | 0.1 | 0.25 | 0.5 | 0.75 | 1 | 2 | 3 | 4 | 5 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| 0.0001 | **0.08%** | **-0.34%** | **6.20%** | **23.34%** | **62.01%** | **72.86%** | **81.56%** | **88.07%** | **91.67%** | **95.12%** |
| 0.0005 | **-0.16%** | **-0.16%** | **11.01%** | **53.45%** | **55.73%** | **69.36%** | **79.42%** | **91.87%** | **94.78%** | **94.73%** |
| 0.001 | **-0.22%** | **-0.16%** | **15.89%** | **49.97%** | **52.99%** | **67.43%** | **78.26%** | **91.45%** | **94.53%** | **94.53%** |
| 0.005 | **-0.29%** | **0.94%** | **34.86%** | **39.39%** | **42.72%** | **61.43%** | **74.74%** | **90.20%** | **93.78%** | **93.96%** |
| 0.01 | **-0.36%** | **6.48%** | **27.68%** | **33.74%** | **38.03%** | **57.84%** | **79.12%** | **89.54%** | **93.44%** | **93.65%** |
| 0.02 | **-0.40%** | **15.32%** | **18.67%** | **27.76%** | **29.75%** | **53.27%** | **77.18%** | **88.74%** | **92.95%** | **93.29%** |
| 0.05 | **8.05%** | **9.69%** | **9.57%** | **17.49%** | **18.73%** | **44.93%** | **73.74%** | **87.24%** | **92.09%** | **92.70%** |
| 0.1 | **8.05%** | **4.75%** | **-0.13%** | **13.71%** | **13.13%** | **35.70%** | **70.07%** | **85.67%** | **91.33%** | **92.11%** |
| 0.15 | **2.33%** | **3.17%** | **3.47%** | **-0.30%** | **0.85%** | **35.72%** | **67.22%** | **84.48%** | **90.68%** | **91.69%** |
| 0.25 | **1.30%** | **2.26%** | **2.06%** | **10.91%** | **9.64%** | **24.29%** | **62.46%** | **82.50%** | **89.71%** | **91.01%** |

Table 8: *Improvement (% of $l_1$-loss) of the multi-Gaussian mixture mechanism (green) over the analytic Gaussian mechanism (red).*

regimes, while Appendix 2.C.2 shows that in high-privacy regimes (*e.g.*, $(10^{-4}, 10^{-4})$-DP), the quasi-Gaussian mechanism may improve over both the analytic Gaussian and the multi-Gaussian mechanisms.

## 2.5 The price of privacy in machine learning

Thus far, we have compared privacy mechanisms based on their expected losses. We now evaluate the performance of these mechanisms in a machine learning task implemented under $(\varepsilon, \delta)$-DP. Specifically, we compare the in- and out-of-sample performances of privacy-preserving linear classifiers. Each classifier is trained under the same $(\varepsilon, \delta)$-DP guarantee and differs only in the mechanism used to sample the additive noise: `A-G` (analytic Gaussian), `Q-G` (quasi-Gaussian mixture), or `M-G` (multi-Gaussian mixture). This experiment allows us to examine whether privacy mechanisms with smaller noise levels reduce the price of privacy in a machine learning setting, following a similar experiment conducted for the numerical optimization-based algorithm (Selvi et al. 2025, §5.3).

More specifically, we work with a dataset $\{(\boldsymbol{x}^i, y^i)\}_{i=1}^N$ where each data-point $i$ has feature vector $\boldsymbol{x}^i \in \mathbb{R}^d$ and binary label $y^i \in \{-1, 1\}$. We train an $l_1$-regularized logistic classifier $\boldsymbol{h} \in \mathbb{R}^d$ which minimizes the empirical logistic loss $\frac{1}{N}\sum_{i=1}^N \log(1+\exp(-y^i \cdot \boldsymbol{h}^\top \boldsymbol{x}^i)) + \lambda \cdot \|\boldsymbol{h}\|_1$, where $\lambda > 0$ is the regularization parameter. We then predict the label of a new instance with features $\boldsymbol{\chi}$ to be the class that maximizes $[1+\exp(-y \cdot \boldsymbol{h}^\top \boldsymbol{\xi})]^{-1}$. We train the logistic classifier via the proximal coordinate descent method (Friedman et al. 2010, Parikh et al. 2014, Richtárik and Takáč 2014) in $T = 100$ iterations with $P = \lceil d/4 \rceil$ proximal updates per iteration, using a regularization parameter $\lambda = 10^{-8}$. The DP counterpart of this algorithm is due to the work of (Mangold et al. 2022) who inject noise into each proximal update step (which is a scalar update and thus our mechanisms are applicable), and the post processing property (Dwork and Roth 2014, §2.1) of differential privacy guarantees that the resulting classifier is also differentially private. The proximal operator has sensitivity $\Delta = 2$ assuming the feature vectors are pre-processed so that $\|\boldsymbol{x}^i\|_\infty \leq 1,\ i \in [N]$. We impose $(\varepsilon = 1, \delta = 10^{-3})$-DP to each proximal update. Note that there is an extensive literature on how such DP iterations compose cumulatively (Dwork and Roth 2014, Steinke 2022, Altschuler and Talwar 2022) in order to quantify the approximate DP guarantee of the final classifier. We omit the compositions, however, since each iteration uses identical privacy parameters, and we instead compare the performances of the classifiers, *ceteris paribus*.

Table 9 reports the mean in- and out-of-sample errors (standard deviations are available in the appendices) across 500 random training-set/test-set (80/20%) splits of various datasets. We use 13 of the most popular UCI classification datasets (Dua and Graff 2017) (varying licenses); additionally, since the proximal coordinate descent method is commonly used for datasets where $d \gg N$, we also include the colon-cancer dataset of LIBSVM (Chang and Lin 2011) (BSD-3-Clause license). All datasets are included in our implementation supplements. Note that we *(i)* convert labels with more than two classes into a binary label via binning (if the label is ordinal) or distinguishing the majority class from all others (if the label is nominal); *(ii)* apply dummy encoding for the nominal features. For the privacy parameters $(\varepsilon, \delta) = (1, 10^{-3})$ we chose, our earlier experiments suggest that the multi-Gaussian mixture mechanism with $K = 9$ significantly outperforms all the other mechanisms (and is near-optimal) while the quasi-Gaussian mechanism is close to (and slightly worse than) the analytic-Gaussian mechanism. We observe a similar pattern in the in- and out-of-sample errors in Table 9. In 11 out of 14 cases, the classifiers trained via our multi-Gaussian mixture mechanism achieve significantly better errors, whereas the second-best approach alternates between the quasi-Gaussian mixture and analytic Gaussian mechanisms. In 3 out of 14 cases, the smallest errors are achieved by our quasi-Gaussian mixture mechanism despite being noisier than the multi-Gaussian mixture mechanism. Those cases, however, are degenerate: either the non-private model `PCD` is worse than its DP counterparts, or the in-sample errors are better than the out-of-sample errors.

## 2.6 Conclusions

We presented new additive noise mechanisms for $(\varepsilon, \delta)$-differential privacy (approximate DP) whose noise distributions are mixtures of Gaussians and quasi-Gaussians. We subsequently developed efficient algorithms to tune the underlying parameters of these distributions to minimize

| Dataset description | | | In-sample errors | | | | Out-of-sample errors | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Dataset | $N$ | $d$ | A-G | Q-G | M-G | PCD | A-G | Q-G | M-G | PCD |
| post-operative | 86 | 14 | *39.75%* | 39.82% | **35.75%** | 30.02% | *37.44%* | 37.85% | **31.99%** | 21.58% |
| adult | 45,222 | 57 | 22.53% | **22.50%** | *22.53%* | 22.52% | 22.18% | **22.16%** | *22.18%* | 22.18% |
| breast-cancer | 683 | 26 | *4.24%* | 4.25% | **4.17%** | 4.20% | 5.67% | *5.66%* | **5.56%** | 5.59% |
| contraceptive | 1,473 | 18 | 38.03% | *38.02%* | **37.95%** | 37.92% | 41.47% | *41.39%* | **41.29%** | 41.31% |
| dermatology | 366 | 98 | 15.48% | *15.43%* | **13.96%** | 13.27% | 17.60% | *17.53%* | **16.03%** | 15.66% |
| cylinder-bands | 539 | 63 | *28.80%* | 28.87% | **28.36%** | 28.14% | *31.96%* | 31.96% | **31.63%** | 31.18% |
| annealing | 898 | 42 | 16.45% | *16.34%* | **16.19%** | 16.24% | 18.32% | *18.18%* | **18.07%** | 18.13% |
| spect | 160 | 23 | *21.85%* | 21.96% | **18.68%** | 16.55% | *34.43%* | 34.59% | **31.97%** | 30.63% |
| bank | 45,211 | 44 | *12.34%* | **12.32%** | 12.35% | 12.31% | *11.73%* | **11.71%** | 11.74% | 11.70% |
| abalone | 4,177 | 10 | *27.07%* | 27.09% | **27.06%** | 27.04% | 28.55% | *28.54%* | **28.51%** | 28.53% |
| spambase | 4,601 | 58 | 39.07% | **39.04%** | *39.06%* | 39.06% | 39.67% | **39.63%** | *39.65%* | 39.64% |
| ecoli | 336 | 8 | *7.65%* | 7.68% | **7.07%** | 6.75% | *5.43%* | 5.48% | **4.77%** | 4.41% |
| absent | 740 | 70 | *32.85%* | 33.02% | **32.60%** | 32.52% | *35.54%* | 35.68% | **35.33%** | 35.14% |
| colon-cancer | 62 | 2,000 | *12.82%* | 13.68% | **6.09%** | 0.00% | *30.67%* | 30.92% | **29.63%** | 29.00% |

Table 9: *Mean in- and out-of-sample errors of privacy-preserving classifiers under different mechanisms. Bold: best DP mechanism; Italics: second-best DP mechanism in the row.*

their losses subject to pre-specified $(\varepsilon, \delta)$-DP guarantees. Our numerical experiments demonstrate that, compared to the minimum-variance Gaussian distribution that satisfies $(\varepsilon, \delta)$-DP (*i.e.*, the analytic Gaussian mechanism), the mixtures of Gaussians achieve significantly lower levels of expected losses in both high- and low-privacy settings and are near-optimal in low-privacy regimes. We also showed that adopting our mechanisms in privacy-preserving ML may yield better generalizations, with an experiment where we train a logistic classifier via the DP proximal coordinate descent method.

There are several avenues for future research. Firstly, we did not optimize the mixture weights in our distributions, and tuning them could offer further improvements in our mechanisms. Similarly, all the distributions that comprise a mixture distribution share the same variance in our work (*i.e.*, identical $\sigma$). One can allow for varying the scale parameters $\sigma$ that appear in these mixtures. Our work considered scalar queries similar to the recent numerical-optimization based approach Selvi et al. (2025), and extensions to multi-dimensional queries would increase the applicability of our mechanisms.

## 2.A Proofs for Section 2.2

### 2.A.1 The quasi-Gaussian mixture distribution is well-defined

To show that the density function (25) of the quasi-Gaussian mixture distribution is well-defined, fix arbitrary $\sigma > 0$, and note that $f_{\mathrm{q}}(x;\sigma) \geq 0$ for all $x \in \mathbb{R}$ since all terms in its definition are nonnegative. Moreover, the density function integrates to 1 since

$$\begin{aligned}
\int_{-\infty}^{\infty} f_{\mathrm{q}}(x;\sigma)\mathrm{d}x &= \frac{1}{c}\int_{-\infty}^{\infty} e^{\varepsilon}\exp\left(-\frac{x^2}{2\sigma^2}\right) + \exp\left(-\frac{(|x|-\Delta)^2}{2\sigma^2}\right)\mathrm{d}x \\
&= \frac{e^{\varepsilon}}{c}\int_{-\infty}^{\infty} \exp\left(-\frac{x^2}{2\sigma^2}\right)\mathrm{d}x + \frac{2}{c}\int_{0}^{\infty} \exp\left(-\frac{1}{2}\left(\frac{x-\Delta}{\sigma}\right)^2\right)\mathrm{d}x \\
&= \frac{e^{\varepsilon}}{c}\sqrt{2\pi}\sigma + \frac{2}{c}\int_{0}^{\infty} \exp\left(-\frac{1}{2}\left(\frac{x-\Delta}{\sigma}\right)^2\right)\mathrm{d}x \\
&= \frac{e^{\varepsilon}}{c}\sqrt{2\pi}\sigma + \frac{2}{c}\sigma\int_{-\Delta/\sigma}^{\infty} \exp\left(-\frac{1}{2}y^2\right)\mathrm{d}y \\
&= \frac{e^{\varepsilon}}{c}\sqrt{2\pi}\sigma + \frac{2}{c}\sqrt{2\pi}\sigma\Phi\left(\frac{\Delta}{\sigma}\right) \\
&= \frac{1}{c}\sqrt{2\pi}\sigma\left(e^{\varepsilon} + 2\Phi\left(\frac{\Delta}{\sigma}\right)\right) = 1.
\end{aligned}$$

The first equality follows from Definition 2, the second equality applies algebraic manipulations, the third equality recognized the first integrand as a Gaussian density function, the fourth equality applies variable change $y = (x-\Delta)/\sigma$, the fifth equality recognizes the integrand as a Gaussian density function, and the final equations rearrange terms and substitute the definition of $c$.

In Figure 10, we visualize the probability density function (25) of the quasi-Gaussian mixture distribution for $\varepsilon = 1.0$ for smaller and larger values of $\sigma$.

### 2.A.2 The cumulative distribution function of the quasi-Gaussian mixture distribution

For fixed $\varepsilon, \Delta, \sigma > 0$, the cumulative distribution function $F_{\mathrm{q}}(\bar{x};\sigma)$ of the quasi-Gaussian mixture distribution is given by:

$$F_{\mathrm{q}}(\bar{x};\sigma) = \int_{-\infty}^{\bar{x}} f_{\mathrm{q}}(x;\sigma)\mathrm{d}x$$

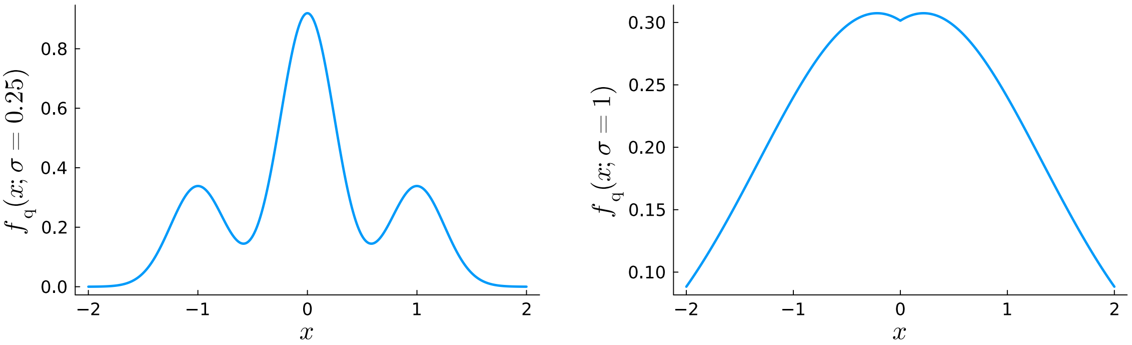


Figure 10: *The probability density function* (25) *of the quasi-Gaussian mixture distribution for* $\varepsilon = 1$ *where* $\sigma$ *is set to* 0.25 *(left) and* 1 *(right).*

$$= \underbrace{\frac{1}{c}\int_{-\infty}^{\bar{x}} e^{\varepsilon} \exp\left(-\frac{x^2}{2\sigma^2}\right) \mathrm{d}x}_{(i)} + \underbrace{\frac{1}{c}\int_{-\infty}^{\bar{x}} \exp\left(-\frac{(|x|-\Delta)^2}{2\sigma^2}\right) \mathrm{d}x}_{(ii)}$$

Here, term $(i)$ can be written in closed form:

$$\frac{1}{c}\int_{-\infty}^{\bar{x}} e^{\varepsilon} \exp\left(-\frac{x^2}{2\sigma^2}\right) \mathrm{d}x = \frac{e^{\varepsilon}}{c}\sqrt{2\pi}\sigma\Phi\left(\frac{\bar{x}}{\sigma}\right) = \frac{e^{\varepsilon}\Phi\left(\frac{\bar{x}}{\sigma}\right)}{e^{\varepsilon}+2\Phi\left(\frac{\Delta}{\sigma}\right)}.$$

Term $(ii)$ can be studied in two cases. For the first case, suppose $\bar{x} < 0$. We have

$$\begin{aligned}\frac{1}{c}\int_{-\infty}^{\bar{x}} \exp\left(-\frac{(|x|-\Delta)^2}{2\sigma^2}\right) \mathrm{d}x &= \frac{1}{c}\int_{-\infty}^{\bar{x}} \exp\left(-\frac{(x+\Delta)^2}{2\sigma^2}\right) \mathrm{d}x \\ &= \frac{\sqrt{2\pi}\sigma}{c}\Phi\left(\frac{\bar{x}+\Delta}{\sigma}\right) = \frac{\Phi\left(\frac{\bar{x}+\Delta}{\sigma}\right)}{e^{\varepsilon}+2\Phi\left(\frac{\Delta}{\sigma}\right)}.\end{aligned}$$

For the case $\bar{x} \geq 0$, we can write term $(ii)$ as

$$\begin{aligned}\frac{1}{c}\int_{-\infty}^{\bar{x}} \exp\left(-\frac{(|x|-\Delta)^2}{2\sigma^2}\right) \mathrm{d}x &= \frac{1}{c}\int_{-\infty}^{0} \exp\left(-\frac{(x+\Delta)^2}{2\sigma^2}\right) \mathrm{d}x + \frac{1}{c}\int_{0}^{\bar{x}} \exp\left(-\frac{(x-\Delta)^2}{2\sigma^2}\right) \mathrm{d}x \\ &= \frac{\sqrt{2\pi}\sigma}{c}\Phi\left(\frac{\Delta}{\sigma}\right) + \frac{1}{c}\int_{0}^{\bar{x}} \exp\left(-\frac{(x-\Delta)^2}{2\sigma^2}\right) \mathrm{d}x\end{aligned}$$

$$
\begin{aligned}
&= \frac{\sqrt{2\pi}\sigma}{c}\Phi\left(\frac{\Delta}{\sigma}\right) + \frac{\sqrt{2\pi}\sigma}{c}\left(\Phi\left(\frac{\bar{x}-\Delta}{\sigma}\right) - \Phi\left(-\frac{\Delta}{\sigma}\right)\right) \\
&= \frac{\Phi\left(\frac{\Delta}{\sigma}\right) + \Phi\left(\frac{\bar{x}-\Delta}{\sigma}\right) - \Phi\left(-\frac{\Delta}{\sigma}\right)}{e^{\varepsilon} + 2\Phi\left(\frac{\Delta}{\sigma}\right)}.
\end{aligned}
$$

We thus have

$$
F_{\mathrm{q}}(\bar{x};\sigma) = \begin{cases} \dfrac{e^{\varepsilon}\Phi\left(\frac{\bar{x}}{\sigma}\right) + \Phi\left(\frac{\bar{x}+\Delta}{\sigma}\right)}{e^{\varepsilon} + 2\Phi\left(\frac{\Delta}{\sigma}\right)} & \text{if } \bar{x} < 0 \\ \dfrac{e^{\varepsilon}\Phi\left(\frac{\bar{x}}{\sigma}\right) + \Phi\left(\frac{\bar{x}-\Delta}{\sigma}\right) + \Phi\left(\frac{\Delta}{\sigma}\right) - \Phi\left(-\frac{\Delta}{\sigma}\right)}{e^{\varepsilon} + 2\Phi\left(\frac{\Delta}{\sigma}\right)} & \text{if } \bar{x} \geq 0. \end{cases}
$$

### 2.A.3 Sampling from the quasi-Gaussian mixture distribution

Since the quasi-Gaussian is a mixture distribution, we decide which of the two distributions to sample noise from via $\tilde{\theta} \sim \text{Bernoulli}(e^{\varepsilon}/(e^{\varepsilon} + 2\Phi(\Delta/\sigma)))$. If $\tilde{\theta} = 1$, then we can simply sample noise from the zero-mean Gaussian distribution with standard deviation $\sigma$. If $\tilde{\theta} = 0$, then we sample noise from the distribution with density $\propto \exp(-(|x| - \Delta)^2)/(2\sigma^2))$. This could be obtained by sampling from $\propto \exp(-(z - \Delta)^2)/(2\sigma^2)) \cdot \mathbb{I}[z \geq 0]$, where $\mathbb{I}$ denotes the indicator function, and then flipping the sign of $z$ with 1/2 probability. The density $\propto \exp(-(z - \Delta)^2)/(2\sigma^2)) \cdot \mathbb{I}[z \geq 0]$ coincides with the truncated Gaussian distribution with mean $\Delta$ and standard deviation $\sigma$, whose inverse cumulative density function is given by $\Delta + \sigma\Phi^{-1}(\Phi(-\Delta/\sigma) + p \cdot \Phi(\mu/\sigma)),\ p \sim \text{Uniform}(0, 1)$. In summary, we can sample noise from the quasi-Gaussian mechanism via Algorithm 8, where Uniform$(0, 1)$ denotes the uniform distribution supported on $(0, 1)$, Bernoulli$(1/2)$ denotes the Bernoulli distribution with success probability 1/2, and $\mathcal{N}(0, \sigma)$ denotes the zero-mean Gaussian with standard deviation $\sigma$.

**Algorithm 8:** Sampling noise from the quasi-Gaussian distribution.

```
Input: privacy parameter $\varepsilon > 0$, sensitivity $\Delta > 0$, variance parameter $\sigma > 0$
Output: a sample from the quasi-Gaussian distribution with density (25)
Sample $\theta \sim \text{Bernoulli}(e^\varepsilon/(e^\varepsilon + 2\Phi(\Delta/\sigma))$
if $\theta = 1$ then
    Sample $x \sim \mathcal{N}(0, \sigma)$
else
    Sample $p \sim \text{Uniform}(0, 1)$
    Set $x = \Delta + \sigma\Phi^{-1}(\Phi(-\Delta/\sigma) + p \cdot \Phi(\Delta/\sigma))$
    Sample $s \sim \text{Bernoulli}(1/2)$
    if $s = 1$ then
        Update $x = -x$
return $x$
```

### 2.A.4 Noise amplitude and power of the quasi-Gaussian mixture distributions

For the *noise amplitude* of the random variable $\tilde{X}$ with density $f_{\text{q}}(\cdot;\sigma)$, we have:

$$\mathbb{E}_{\tilde{X}}[|x|] = \frac{e^\varepsilon}{c} \underbrace{\int_{-\infty}^{\infty} |x| \exp\left(-\frac{x^2}{2\sigma^2}\right) \mathrm{d}x}_{(i)} + \frac{1}{c} \underbrace{\int_{-\infty}^{\infty} |x| \exp\left(-\frac{(|x|-\Delta)^2}{2\sigma^2}\right) \mathrm{d}x}_{(ii)}.$$

Term $(i)$ satisfies:

$$\int_{-\infty}^{\infty} |x| \exp\left(-\frac{x^2}{2\sigma^2}\right) \mathrm{d}x = 2\int_{0}^{\infty} x \exp\left(-\frac{x^2}{2\sigma^2}\right) \mathrm{d}x = 2\sigma^2.$$

Term $(ii)$ satisfies:

$$\begin{aligned}
&\int_{-\infty}^{\infty} |x| \exp\left(-\frac{(|x|-\Delta)^2}{2\sigma^2}\right) \mathrm{d}x \\
=&2\int_{0}^{\infty} x \exp\left(-\frac{(x-\Delta)^2}{2\sigma^2}\right) \mathrm{d}x \\
=&2\int_{-\Delta/\sigma}^{\infty} (\sigma u + \Delta) \exp\left(-\frac{u^2}{2}\right) \sigma \mathrm{d}u \\
=&2\sigma^2 \int_{-\Delta/\sigma}^{\infty} u \exp\left(-\frac{u^2}{2}\right) \mathrm{d}u + 2\sigma\Delta \int_{-\Delta/\sigma}^{\infty} \exp\left(-\frac{u^2}{2}\right) \mathrm{d}u \\
=&2\sigma^2 \exp\left(-\frac{\Delta^2}{2\sigma^2}\right) + 2\sigma\Delta\sqrt{2\pi}\Phi\left(\frac{\Delta}{\sigma}\right)
\end{aligned}$$

where the first equality exploits the symmetry of the absolute value, the second equality applies variable change $u = (x - \Delta)/\sigma$, the third equality exploits the linearity of integrals and the final equality recognizes $\exp(-u^2/2)$ as the (unnormalized) standard Gaussian density. We conclude

$$\begin{aligned}\mathbb{E}_{\tilde{X}}[|x|] &= \frac{e^{\varepsilon}}{c}2\sigma^2 + \frac{1}{c}\left(2\sigma^2 \exp\left(-\frac{\Delta^2}{2\sigma^2}\right) + 2\sigma\Delta\sqrt{2\pi}\Phi\left(\frac{\Delta}{\sigma}\right)\right) \\ &= \frac{2\sigma^2\left(e^{\varepsilon} + \exp\left(-\frac{\Delta^2}{2\sigma^2}\right)\right) + 2\sigma\Delta\sqrt{2\pi}\Phi\left(\frac{\Delta}{\sigma}\right)}{\sqrt{2\pi}\sigma\left(e^{\varepsilon} + 2\Phi\left(\frac{\Delta}{\sigma}\right)\right)} \\ &= \frac{\sqrt{2/\pi}\sigma\left(e^{\varepsilon} + \exp\left(-\frac{\Delta^2}{2\sigma^2}\right)\right) + 2\Delta\Phi\left(\frac{\Delta}{\sigma}\right)}{e^{\varepsilon} + 2\Phi\left(\frac{\Delta}{\sigma}\right)},\end{aligned}$$

where the first equality is derived above, the second equality rearranges terms, and the final equality cancels the common terms from the numerator and the denominator.

For the *noise power* of the random variable $\tilde{X}$ with density $f_{\mathrm{q}}(\cdot;\sigma)$, we have:

$$\mathbb{E}_{\tilde{X}}[x^2] = \frac{e^{\varepsilon}}{c}\underbrace{\int_{-\infty}^{\infty} x^2 \exp\left(-\frac{x^2}{2\sigma^2}\right)\mathrm{d}x}_{(i)} + \frac{1}{c}\underbrace{\int_{-\infty}^{\infty} x^2 \exp\left(-\frac{(|x| - \Delta)^2}{2\sigma^2}\right)\mathrm{d}x}_{(ii)}.$$

Term $(i)$ satisfies

$$\int_{-\infty}^{\infty} x^2 \exp\left(-\frac{x^2}{2\sigma^2}\right)\mathrm{d}x = \sigma^3\sqrt{2\pi}$$

since we recognize $\exp(-x^2/(2\sigma^2))$ as the (unnormalized) Gaussian density with variance $\sigma^2$.

Term $(ii)$ satisfies

$$\begin{aligned}&\int_{-\infty}^{\infty} x^2 \exp\left(-\frac{(|x| - \Delta)^2}{2\sigma^2}\right)\mathrm{d}x \\ =&2\int_{0}^{\infty} x^2 \exp\left(-\frac{(x - \Delta)^2}{2\sigma^2}\right)\mathrm{d}x \\ =&2\sigma\int_{-\Delta/\sigma}^{\infty} (\sigma u + \Delta)^2 \exp\left(-\frac{u^2}{2}\right)\mathrm{d}u\end{aligned}$$

$$
\begin{aligned}
=&2\sigma^3 \int_{-\Delta/\sigma}^{\infty} u^2 \exp\left(-\frac{u^2}{2}\right) \mathrm{d}u + 2\sigma\Delta^2 \underbrace{\int_{-\Delta/\sigma}^{\infty} \exp\left(-\frac{u^2}{2}\right) \mathrm{d}u}_{=\sqrt{2\pi}\Phi(\Delta/\sigma)} \\
&\qquad + 4\sigma^2\Delta \underbrace{\int_{-\Delta/\sigma}^{\infty} u \exp\left(-\frac{u^2}{2}\right) \mathrm{d}u}_{=\exp(-(\Delta^2/(2\sigma^2)))} \\
=&2\sigma^3 \int_{-\Delta/\sigma}^{\infty} u^2 \exp\left(-\frac{u^2}{2}\right) \mathrm{d}u + 2\sigma\Delta^2\sqrt{2\pi}\Phi\left(\frac{\Delta}{\sigma}\right) + 4\sigma^2\Delta\exp\left(-\frac{\Delta^2}{2\sigma^2}\right) \\
=&2\sigma^3 \left[-u\exp\left(-\frac{u^2}{2}\right)\right]_{-\Delta/\sigma}^{\infty} + 2\sigma^3 \underbrace{\int_{-\Delta/\sigma}^{\infty} \exp\left(-\frac{u^2}{2}\right) \mathrm{d}u}_{=\sqrt{2\pi}\Phi(\Delta/\sigma)} \\
&\qquad + 2\sigma\Delta^2\sqrt{2\pi}\Phi\left(\frac{\Delta}{\sigma}\right) + 4\sigma^2\Delta\exp\left(-\frac{\Delta^2}{2\sigma^2}\right) \\
=&-2\sigma^2\Delta\exp\left(-\frac{\Delta^2}{2\sigma^2}\right) + 2\sigma^3\sqrt{2\pi}\Phi\left(\frac{\Delta}{\sigma}\right) + 2\sigma\Delta^2\sqrt{2\pi}\Phi\left(\frac{\Delta}{\sigma}\right) + 4\sigma^2\Delta\exp\left(-\frac{\Delta^2}{2\sigma^2}\right) \\
=&2\sigma^2\Delta\exp\left(-\frac{\Delta^2}{2\sigma^2}\right) + 2\sigma\sqrt{2\pi}\Phi\left(\frac{\Delta}{\sigma}\right)\left(\sigma^2+\Delta^2\right) \\
=&2\sigma\sqrt{2\pi}\left(\Phi\left(\frac{\Delta}{\sigma}\right)\left(\sigma^2+\Delta^2\right) + \frac{\sigma\Delta}{\sqrt{2\pi}}\exp\left(-\frac{\Delta^2}{2\sigma^2}\right)\right).
\end{aligned}
$$

where the first equality exploits the symmetry of the square and absolute value functions, the second equality uses the variable change $u = (x - \Delta)/\sigma$, the third equality breaks down the square term into three components, exploits the linearity of integral, and recognizes $\exp(-u^2/(2\sigma^2))$ as the (unnormalized) Gaussian density with variance $\sigma^2$, the fourth equality substitutes integrals with their closed-form expressions except for the first integral, hence the fifth equality uses integration by parts, the sixth equality writes the expression without an integral, and the final equalities rearrange terms. We conclude:

$$
\begin{aligned}
\mathbb{E}_{\tilde{X}}[x^2] &= \frac{e^\varepsilon}{c}\sigma^3\sqrt{2\pi} + \frac{1}{c}\left(2\sigma\sqrt{2\pi}\left(\Phi\left(\frac{\Delta}{\sigma}\right)\left(\sigma^2+\Delta^2\right) + \frac{\sigma\Delta}{\sqrt{2\pi}}\exp\left(-\frac{\Delta^2}{2\sigma^2}\right)\right)\right) \\
&= \frac{e^\varepsilon\sigma^3\sqrt{2\pi} + 2\sigma\sqrt{2\pi}\left(\Phi\left(\frac{\Delta}{\sigma}\right)\left(\sigma^2+\Delta^2\right) + \frac{\sigma\Delta}{\sqrt{2\pi}}\exp\left(-\frac{\Delta^2}{2\sigma^2}\right)\right)}{\sqrt{2\pi}\sigma\left(e^\varepsilon + 2\Phi\left(\frac{\Delta}{\sigma}\right)\right)}
\end{aligned}
$$

$$= \frac{e^\varepsilon \sigma^2 + 2\left(\Phi\left(\frac{\Delta}{\sigma}\right)(\sigma^2+\Delta^2) + \frac{\sigma\Delta}{\sqrt{2\pi}}\exp\left(-\frac{\Delta^2}{2\sigma^2}\right)\right)}{e^\varepsilon + 2\Phi\left(\frac{\Delta}{\sigma}\right)}.$$

### 2.A.5 Proof of Theorem 3

We first introduce an intermediary result that will be used in the proof of Theorem 3.

**Lemma 24.** *An additive noise with density $f(x)$ satisfies $(\varepsilon, \delta)$-DP if and only if it satisfies*

$$\int_{x\in A} (e^\varepsilon f(x+\varphi) - f(x))\mathrm{d}x \geq -\delta \qquad \forall\varphi \in [0,\Delta],\ \forall A \in \mathcal{F},$$

*or, alternatively, if and only if it satisfies*

$$\int_{x\in A} (e^\varepsilon f(x) - f(x+\varphi))\mathrm{d}x \geq -\delta \qquad \forall\varphi \in [0,\Delta],\ \forall A \in \mathcal{F}.$$

*Proof.* An additive noise with density $f(x)$ achieves $(\varepsilon, \delta)$-DP if and only if it satisfies

$$\int_{x\in A} (e^\varepsilon f(x+\varphi) - f(x))\mathrm{d}x \geq -\delta \qquad \forall\varphi \in [-\Delta,\Delta],\ \forall A \in \mathcal{F}$$

as per the definition. We can replace $\varphi \in [-\Delta, \Delta]$ with $\varphi \in [0, \Delta]$ without loss of generality since:

$$\begin{aligned}
&\int_{x\in A} (e^\varepsilon f(x+\varphi) - f(x))\mathrm{d}x \geq -\delta && \forall\varphi \in [0,\Delta],\ \forall A \in \mathcal{F} \\
\iff &\int_{x\in A} (e^\varepsilon f(-x+\varphi) - f(-x))\mathrm{d}x \geq -\delta && \forall\varphi \in [0,\Delta],\ \forall A \in \mathcal{F} \\
\iff &\int_{x\in A} (e^\varepsilon f(x-\varphi) - f(x))\mathrm{d}x \geq -\delta && \forall\varphi \in [0,\Delta],\ \forall A \in \mathcal{F} \\
\iff &\int_{x\in A} (e^\varepsilon f(x+\varphi) - f(x))\mathrm{d}x \geq -\delta && \forall\varphi \in [-\Delta,0],\ \forall A \in \mathcal{F},
\end{aligned}$$

where the first equivalence follows from substituting $x = -x$ as well as from the fact that $\{A \mid A \in \mathcal{F}\} = \{-A \mid A \in \mathcal{F}\}$, the second equivalence uses the symmetry of $f$, and the final equivalence substitutes $\varphi = -\varphi$. We thus obtained the first representation of the DP constraints presented in the statement of this Lemma. Applying analogous steps to the DP definition

$$\int_{x\in A} (e^\varepsilon f(x) - f(x+\varphi))\mathrm{d}x \geq -\delta \qquad \forall\varphi \in [0,\Delta],\ \forall A \in \mathcal{F}$$

concludes the proof. □

**Proof of Theorem 3.** From Lemma 24 it suffices to show that

$$\int_{x\in A}(e^{\varepsilon}f_{\mathrm{q}}(x+\varphi;\sigma)-f_{\mathrm{q}}(x;\sigma))\mathrm{d}x \geq -\delta \qquad \forall\varphi\in[0,\Delta],\ \forall A\in\mathcal{F}$$

holds. To this end, in the rest of this proof, we will show that for an arbitrary $\varphi \in [0,\Delta]$ and $A \in \mathcal{F}$, the desired inequality holds. We first break down the left-hand side of the inequality into three cases by exploiting the linearity of integrals:

$$\begin{aligned}
&\int_{x\in A}(e^{\varepsilon}f_{\mathrm{q}}(x+\varphi;\sigma)-f_{\mathrm{q}}(x;\sigma))\mathrm{d}x\\
=&\int_{x\in A: x<-\Delta-\varphi}(e^{\varepsilon}f_{\mathrm{q}}(x+\varphi;\sigma)-f_{\mathrm{q}}(x;\sigma))\mathrm{d}x + &&\text{(30a)}\\
&\int_{x\in A\cap[-\Delta-\varphi,\Delta-\varphi]}(e^{\varepsilon}f_{\mathrm{q}}(x+\varphi;\sigma)-f_{\mathrm{q}}(x;\sigma))\mathrm{d}x + &&\text{(30b)}\\
&\int_{x\in A: x>\Delta-\varphi}(e^{\varepsilon}f_{\mathrm{q}}(x+\varphi;\sigma)-f_{\mathrm{q}}(x;\sigma))\mathrm{d}x. &&\text{(30c)}
\end{aligned}$$

To complete the proof, we show that the sum of (30a), (30b) and (30c) is greater than or equal to $-\delta$. The first terms are studied next:

(30a): For any $x < 0$, we have

$$\frac{\partial}{\partial x}f_{\mathrm{q}}(x;\sigma) = \frac{e^{\varepsilon}}{c}\left(\frac{-2x}{2\sigma^2}\right)\exp\left(-\frac{x^2}{2\sigma^2}\right)+\frac{1}{c}\left(-\frac{2(x+\Delta)}{2\sigma^2}\right)\exp\left(-\frac{(x+\Delta)^2}{2\sigma^2}\right)$$

where all terms are nonnegative as we have $x < -\Delta-\varphi \leq 0$. This concludes that $f_{\mathrm{q}}(x;\sigma)$ is increasing in the region $(-\infty,-\Delta)$ and we thus have:

$$e^{\varepsilon}f_{\mathrm{q}}(x+\varphi;\sigma)-f_{\mathrm{q}}(x;\sigma) \geq f_{\mathrm{q}}(x+\varphi;\sigma)-f_{\mathrm{q}}(x;\sigma) \geq 0.$$

(30b): We have

$$e^{\varepsilon}f_{\mathrm{q}}(x+\varphi;\sigma)-f_{\mathrm{q}}(x;\sigma) \geq e^{\varepsilon}f_{\mathrm{q,min}}(\sigma)-f_{\mathrm{q}}(x;\sigma) \geq f_{\mathrm{q,max}}(\sigma)-f_{\mathrm{q}}(x;\sigma) \geq 0,$$

where the first inequality holds since $x+\varphi \in [-\Delta,\Delta]$ and $f_{\mathrm{q}}(\cdot\,;\sigma)$ is symmetric, the second inequality holds since $\sigma \geq \sigma_2$ as specified in the statement of this theorem, and the final inequality is due to $x \leq \Delta-\varphi \leq \Delta$.

This concludes that (30a) + (30b) $\geq 0$, and it will thus be sufficient to show (30c) $\geq -\delta$.

(30c): We have

$$\begin{aligned}
&\int_{x\in A:x>\Delta-\varphi}(e^{\varepsilon}f_{\mathrm{q}}(x+\varphi;\sigma)-f_{\mathrm{q}}(x;\sigma))\mathrm{d}x=(30c)\\
\geq&\int_{x\in A:x>\Delta-\varphi}(e^{\varepsilon}f_{\mathrm{q}}(x+\Delta;\sigma)-f_{\mathrm{q}}(x;\sigma))\mathrm{d}x\\
\geq&\int_{x\in A:x>\Delta-\varphi}(e^{\varepsilon}f_{\mathrm{q}}(x+\Delta;\sigma)-f_{\mathrm{q}}(x;\sigma))^{-}\mathrm{d}x\\
\geq&\int_{x\geq 0}(e^{\varepsilon}f_{\mathrm{q}}(x+\Delta;\sigma)-f_{\mathrm{q}}(x;\sigma))^{-}\mathrm{d}x \qquad (31)
\end{aligned}$$

where $z^{-}$ denotes the negative part of $z$, (*i.e.*, $z^{-}=\min\{0,z\}$). Here, the first inequality follows since, analogously to the case of (30a), $\frac{\mathrm{d}}{\mathrm{d}x}f_{\mathrm{q}}(x;\sigma)$ is decreasing in region $x\in(\Delta,\infty)$ and as a result $f_{\mathrm{q}}(x+\varphi)\geq f_{\mathrm{q}}(x+\Delta)$ holds. The second and third inequalities follow as any $z\in\mathbb{R}$ satisfies $z\geq z^{-}$ and $0\geq z^{-}$, respectively. The domain $\{x:x\in\mathbb{R}\}$ of integration in (31) can be replaced with $\{x:x>\varepsilon\sigma^{2}/\Delta\}$ without loss of generality since the term whose negative part is taken satisfies:

$$\begin{aligned}
&e^{\varepsilon}f_{\mathrm{q}}(x+\Delta)-f_{\mathrm{q}}(x;\sigma)\\
=&\frac{e^{\varepsilon}}{c}\left[\exp\left(\varepsilon-\frac{(x+\Delta)^{2}}{2\sigma^{2}}\right)+\exp\left(-\frac{x^{2}}{2\sigma^{2}}\right)-\exp\left(-\frac{x^{2}}{2\sigma^{2}}\right)-\exp\left(-\varepsilon-\frac{(x-\Delta)^{2}}{2\sigma^{2}}\right)\right]\\
=&\frac{e^{\varepsilon}}{c}\left[\exp\left(\varepsilon-\frac{(x+\Delta)^{2}}{2\sigma^{2}}\right)-\exp\left(-\varepsilon-\frac{(x-\Delta)^{2}}{2\sigma^{2}}\right)\right],
\end{aligned}$$

which is nonnegative if and only if

$$\varepsilon-\frac{(x+\Delta)^{2}}{2\sigma^{2}}\geq-\varepsilon-\frac{(x-\Delta)^{2}}{2\sigma^{2}}\iff 2\varepsilon\geq\frac{(x+\Delta)^{2}-(x-\Delta)^{2}}{2\sigma^{2}}\iff x\leq\frac{\varepsilon\sigma^{2}}{\Delta}.$$

We can thus conclude:

$$\begin{aligned}
(30c)\geq(31)\geq&\int_{x>\varepsilon\sigma^{2}/\Delta}(e^{\varepsilon}f_{\mathrm{q}}(x+\Delta;\sigma)-f_{\mathrm{q}}(x;\sigma))\mathrm{d}x\\
=&\frac{1}{c}\int_{\varepsilon\sigma^{2}/\Delta}^{\infty}e^{2\varepsilon}\exp\left(-\frac{(x+\Delta)^{2}}{2\sigma^{2}}\right)-\exp\left(-\frac{(x-\Delta)^{2}}{2\sigma^{2}}\right)\mathrm{d}x\\
=&\frac{e^{2\varepsilon}\sqrt{2\pi}\sigma}{c}\Phi\left(-\frac{\varepsilon\sigma}{\Delta}-\frac{\Delta}{\sigma}\right)-\frac{1}{c}\int_{\varepsilon\sigma^{2}/\Delta}^{\infty}\exp\left(-\frac{(x-\Delta)^{2}}{2\sigma^{2}}\right)\mathrm{d}x
\end{aligned}$$

$$
\begin{aligned}
&= \frac{e^{2\varepsilon}\sqrt{2\pi}\sigma}{c}\Phi\left(-\frac{\varepsilon\sigma}{\Delta}-\frac{\Delta}{\sigma}\right)-\frac{\sqrt{2\pi}\sigma}{c}\Phi\left(-\frac{\varepsilon\sigma}{\Delta}+\frac{\Delta}{\sigma}\right)\\
&= \frac{1}{e^{\varepsilon}+2\Phi\left(\frac{\Delta}{\sigma}\right)}\left(e^{2\varepsilon}\Phi\left(-\frac{\varepsilon\sigma}{\Delta}-\frac{\Delta}{\sigma}\right)-\Phi\left(-\frac{\varepsilon\sigma}{\Delta}+\frac{\Delta}{\sigma}\right)\right)\\
&= \frac{h_1(\sigma)}{h_2(\sigma)/\delta}\geq -\delta.
\end{aligned}
$$

Here, the first two inequalities were proved earlier, the four equalities that follow derive the integral in closed form by recognizing the integrands as Gaussian density functions, the final equality substitutes $h_1(\sigma)$ and $h_2(\sigma)$ as defined in (26b), and the final inequality follows from $\sigma \geq \sigma_1$ where $\sigma_1$ is defined in (26a). Note that while the final inequality is straightforward for $\sigma = \sigma_1$ by the definition of $\sigma_1$, to have the same conclusion for any $\sigma \geq \sigma_1$, we rely on Lemma 20.

We conclude the proof since the left-hand side of the DP constraints, that needs to be greater than or equal to $-\delta$, can be written as (30a) + (30b) + (30c), where (30a), (30b) $\geq 0$ and (30c) $\geq -\delta$. □

### 2.A.6 Proof of Lemma 20

To prove that the desired monotonicity, we investigate the derivative of the function of interest. To this end, note that

$$
\begin{aligned}
&\frac{\mathrm{d}}{\mathrm{d}\sigma}h_1(\sigma)\\
&= e^{2\varepsilon}\left(-\frac{\varepsilon}{\Delta}+\frac{\Delta}{\sigma^2}\right)\Phi'\left(-\frac{\varepsilon\sigma}{\Delta}-\frac{\Delta}{\sigma}\right)-\left(-\frac{\varepsilon}{\Delta}-\frac{\Delta}{\sigma^2}\right)\Phi'\left(-\frac{\varepsilon\sigma}{\Delta}+\frac{\Delta}{\sigma}\right)\\
&= \frac{e^{2\varepsilon}}{\sqrt{2\pi}}\left(-\frac{\varepsilon}{\Delta}+\frac{\Delta}{\sigma^2}\right)\exp\left(-\frac{1}{2}\left(\frac{\varepsilon\sigma}{\Delta}+\frac{\Delta}{\sigma}\right)^2\right)+\frac{1}{\sqrt{2\pi}}\left(\frac{\varepsilon}{\Delta}+\frac{\Delta}{\sigma^2}\right)\exp\left(-\frac{1}{2}\left(-\frac{\varepsilon\sigma}{\Delta}+\frac{\Delta}{\sigma}\right)^2\right)\\
&= \frac{2\Delta}{\sqrt{2\pi}\sigma^2}\exp\left(-\frac{1}{2}\left(-\frac{\varepsilon\sigma}{\Delta}+\frac{\Delta}{\sigma}\right)^2\right),
\end{aligned}
$$

where the first equality follows from the chain rule, the second equality uses the definition of the Gaussian probability density function, and the final equality is by using the algebraic property

$(a+b)^2 = (a-b)^2 + 4ab$. Moreover, we have

$$\frac{\mathrm{d}}{\mathrm{d}\sigma} h_2(\sigma) = -2\delta \frac{\Delta}{\sigma^2} \Phi'\left(\frac{\Delta}{\sigma}\right) = -\frac{2\Delta}{\sqrt{2\pi\sigma^2}} \delta \exp\left(-\frac{1}{2}\left(\frac{\Delta}{\sigma}\right)^2\right)$$

which concludes that

$$\begin{aligned}
&\frac{\mathrm{d}}{\mathrm{d}\sigma}\left[h_1(\sigma) + h_2(\sigma)\right] \\
=&\frac{2\Delta}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2}\left(-\frac{\varepsilon\sigma}{\Delta} + \frac{\Delta}{\sigma}\right)^2\right) - \frac{2\Delta}{\sqrt{2\pi\sigma^2}} \delta \exp\left(-\frac{1}{2}\left(\frac{\Delta}{\sigma}\right)^2\right) \\
=&\frac{2\Delta}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2}\left(\frac{\Delta}{\sigma}\right)^2\right)\left(\exp\left(\varepsilon - \frac{1}{2}\left(\frac{\varepsilon\sigma}{\Delta}\right)^2\right) - \delta\right),
\end{aligned}$$

which is positive whenever $\varepsilon - \varepsilon^2\sigma^2/(2\Delta^2) > \log\delta$ which is equivalent to $\sigma < \sqrt{2(\varepsilon - \log\delta)}\Delta/\varepsilon$.

To conclude the function is nonnegative for $\sigma \geq \sqrt{2(\varepsilon - \log\delta)}\Delta/\varepsilon$, note that

$$\begin{aligned}
h_1(\sigma) + h_2(\sigma) \geq& h_1(\sigma) + 2\delta \\
\geq& -\Phi\left(-\frac{\varepsilon\sigma}{\Delta} + \frac{\Delta}{\sigma}\right) + 2\delta, \qquad (32)
\end{aligned}$$

where the first inequality holds since $e^\varepsilon \geq 1$ and $\Phi(\Delta/\sigma) \geq \Phi(0) = 0.5$ and the second inequality follows from removing the positive part from the definition of $h_1(\sigma)$. The expression (32) is trivially nonnegative for $\delta \geq 0.5$. To conclude the proof, we next show that (32) is nonnegative for $\delta < 0.5$. To this end, suppose $\delta < 0.5$. Note that $\varepsilon\sigma/\Delta - \Delta/\sigma$ is monotonically increasing with respect to $\sigma$, so the fact that $\sigma \geq \sqrt{2(\varepsilon - \log\delta)}\Delta/\varepsilon$ implies

$$\frac{\varepsilon\sigma}{\Delta} - \frac{\Delta}{\sigma} \geq \sqrt{2(\varepsilon - \log\delta)} - \frac{\varepsilon}{\sqrt{2(\varepsilon - \log\delta)}} = \frac{\varepsilon - 2\log\delta}{\sqrt{2(\varepsilon - \log\delta)}} > 0. \qquad (33)$$

By using the substitution $t = (\varepsilon - 2\log\delta)/\sqrt{2(\varepsilon - \log\delta)}$, we can further bound (32) via

$$\begin{aligned}
\Phi\left(-\frac{\varepsilon\sigma}{\Delta} + \frac{\Delta}{\sigma}\right) \leq \Phi(-t) \leq& \frac{1}{\sqrt{2\pi}t} \exp\left(-\frac{\varepsilon^2 - 4(\varepsilon - \log\delta)\log\delta}{4(\varepsilon - \log\delta)}\right) \qquad (34) \\
=& \frac{\delta}{\sqrt{2\pi}t} \exp\left(-\frac{\varepsilon^2}{4(\varepsilon - \log\delta)}\right), \qquad (35)
\end{aligned}$$

where the first inequality is due to (33) and the second inequality is due to Mill's inequality, asserting that for $t > 0$ we have $\Phi(-t) \leq (1/\sqrt{2\pi}t)\exp\left(-t^2/2\right)$. This allows us to conclude the proof as:

$$\begin{aligned}
(32) &\geq 2\delta - \frac{\delta}{\sqrt{2\pi}t}\exp\left(-\frac{\varepsilon^2}{4(\varepsilon - \log\delta)}\right) \\
&\geq 2\delta - \frac{1}{\sqrt{\pi\log 2}}\exp\left(-\frac{\varepsilon^2}{4(\varepsilon - \log\delta)}\right)\delta \\
&\geq \delta > 0.
\end{aligned}$$

Here the first inequality follows from (34), the second inequality follows from

$$t = \frac{\varepsilon - 2\log\delta}{\sqrt{2(\varepsilon - \log\delta)}} \geq \frac{\varepsilon - \log\delta}{\sqrt{2(\varepsilon - \log\delta)}} = \sqrt{\frac{\varepsilon - \log\delta}{2}} \geq \sqrt{\frac{\log 2}{2}},$$

and the third inequality is due to the fact that $1/\sqrt{\pi\log 2} \leq 1$.

Finally, to conclude that $e^\varepsilon + 2 \geq \delta^{-1}$ implies the function is nonnegative everywhere, we investigate the limiting case where

$$\lim_{\sigma\to 0} h_1(\sigma) = -1$$

and

$$\lim_{\sigma\to 0} h_2(\sigma) = (e^\varepsilon + 2)\delta,$$

hold, hence the function $h_1(\sigma) + h_2(\sigma) \geq (e^\varepsilon + 2)\delta - 1 \geq 0$ is nonnegative everywhere under this condition.

### 2.A.7 Proof of Lemma 21

We first state the unabridged lemma here.

**Lemma 21.** *For any $\sigma > 0$ let*

$$x_1 := \frac{\Delta - \sqrt{\Delta^2 - 4\sigma^2}}{2},\ x_2 := \frac{\Delta + \sqrt{\Delta^2 - 4\sigma^2}}{2},\ g(x) := -e^\varepsilon + \frac{\Delta - x}{x}\exp\left(\frac{2x\Delta - \Delta^2}{2\sigma^2}\right).$$

*The following statements hold for $\min_{x\in[0,\Delta]} f_{\mathrm{q}}(x;\sigma)$ and $\max_{x\in[0,\Delta]} f_{\mathrm{q}}(x;\sigma)$:*

*(i) If $\Delta^2 \leq 4\sigma^2$, then $x = \Delta$ solves $\min_{x\in[0,\Delta]} f_{\mathrm{q}}(x;\sigma)$. Moreover, $f_{\mathrm{q}}(x;\sigma)$ is unimodal on*

*the interval* $(0, \Delta/2)$ *with a unique maximum. The maximizer of this unimodal region also solves* $\max_{x\in[0,\Delta]} f_{\mathrm{q}}(x;\sigma)$*.*

*(ii) If* $\Delta^2 > 4\sigma^2$ *and* $g(x_2) \le 0$*, then* $x = \Delta$ *solves* $\min_{x\in[0,\Delta]} f_{\mathrm{q}}(x;\sigma)$*. Moreover,* $f_{\mathrm{q}}(x;\sigma)$ *is unimodal on the interval* $(0, x_1)$ *with a unique maximum. The maximizer of this unimodal region also solves* $\max_{x\in[0,\Delta]} f_{\mathrm{q}}(x;\sigma)$*.*

*(iii) If* $\Delta^2 > 4\sigma^2$ *and* $g(x_2) > 0$*, then* $f_{\mathrm{q}}(x;\sigma)$ *is unimodal on the interval* $(\Delta/2, x_2)$ *with a unique minimum. Either* $x = \Delta$ *or the minimizer of this unimodal region solves* $\min_{x\in[0,\Delta]} f_{\mathrm{q}}(x;\sigma)$*. Moreover,* $f_{\mathrm{q}}(x;\sigma)$ *is unimodal on the interval* $(0, x_1)$ *with a unique maximum. The maximizer of this unimodal region also solves* $\max_{x\in[0,\Delta]} f_{\mathrm{q}}(x;\sigma)$*.*

Since this lemma explores the minimizer and maximizer of $f_{\mathrm{q}}(x;\sigma)$ over a nonnegative interval $x \in [0,\Delta]$, we can rewrite it without an absolute value as

$$f_{\mathrm{q}}(x;\sigma) = \frac{e^{\varepsilon}}{c}\exp\left(-\frac{x^2}{2\sigma^2}\right) + \frac{1}{c}\exp\left(-\frac{(x-\Delta)^2}{2\sigma^2}\right), \tag{36}$$

and its derivative as

$$\begin{aligned}\frac{\partial}{\partial x} f_{\mathrm{q}}(x;\sigma) &= \frac{e^{\varepsilon}}{c}\left(-\frac{x}{\sigma^2}\right)\exp\left(-\frac{x^2}{2\sigma^2}\right) + \frac{1}{c}\left(-\frac{x-\Delta}{\sigma^2}\right)\exp\left(-\frac{(x-\Delta)^2}{2\sigma^2}\right)\\ &= \frac{1}{c\sigma^2}\exp\left(-\frac{x^2}{2\sigma^2}\right)\left[-e^{\varepsilon}x + (\Delta - x)\exp\left(\frac{2x\Delta-\Delta^2}{2\sigma^2}\right)\right],\end{aligned} \tag{37}$$

where the first equality follows from the chain rule and the second equality rearranges terms. We now state and prove an intermediary lemma that will be used in the proof of Lemma 21.

**Lemma 25.** *For any* $x \in [0, \Delta/2)$*, we have* $f_{\mathrm{q}}(x;\sigma) > f_{\mathrm{q}}(\Delta - x;\sigma)$*.*

*Proof.* We can use (36) and show

$$f_{\mathrm{q}}(x;\sigma) - f_{\mathrm{q}}(\Delta - x;\sigma) = \frac{e^{\varepsilon}-1}{c}\left(\exp\left(-\frac{x^2}{2\sigma^2}\right) - \exp\left(-\frac{(x-\Delta)^2}{2\sigma^2}\right)\right).$$

For $x \in [0, \Delta/2)$, we have $(x-\Delta)^2 > x^2$, which implies $\exp(-\frac{x^2}{2\sigma^2}) > \exp(-\frac{(x-\Delta)^2}{2\sigma^2})$ and as a result concludes that the expression above is positive. □

We can now prove the unabridged version of Lemma 21.

**Proof of Lemma 21.** Since we fix $\sigma > 0$, we will use the shorthand $f_{\mathrm{q}}(x)$ and $f'_{\mathrm{q}}(x)$ for $f_{\mathrm{q}}(x;\sigma)$ and $\partial f_{\mathrm{q}}(x;\sigma)/\partial x$, respectively. We will study the sign of the derivative (37) over $[0,\Delta]$

in order to conclude the proof. To this end, we first plug in $x \in \{0, \Delta/2, \Delta\}$ in (37) and note:

$$f_{\mathrm{q}}'(0) > 0,\ f_{\mathrm{q}}'(\Delta/2) < 0,\ f_{\mathrm{q}}'(\Delta) < 0. \tag{38}$$

From the continuity of $f_{\mathrm{q}}'$, there exists some small enough constant $\xi > 0$ such that $f_{\mathrm{q}}'(x) > 0$ for all $x \in [0, \xi]$. For $x \in [\xi, \Delta]$, we can thus multiply the expression (37) by $x/x$ and obtain:

$$f_{\mathrm{q}}'(x) = \underbrace{\frac{x}{c\sigma^2} \exp\left(-\frac{x^2}{2\sigma^2}\right)}_{=:g^+(x)} \underbrace{\left[-e^{\varepsilon} + \frac{\Delta - x}{x} \exp\left(\frac{2x\Delta - \Delta^2}{2\sigma^2}\right)\right]}_{=:g(x)}.$$

Here, we have $g^+(x) > 0$ for all $x \in [0, \xi]$. Thus, the sign of $f_{\mathrm{q}}'(x)$ coincides the sign of $g(x)$ in the region $x \in [\xi, \Delta]$. Then, (38) implies that:

$$g(\xi) > 0,\ g(\Delta/2) < 0,\ g(\Delta) < 0. \tag{39}$$

In order to explore the sign of $g(x)$, we further compute its derivative

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}x} g(x) =: g'(x) &= \exp\left(\frac{2x\Delta - \Delta^2}{2\sigma^2}\right)\left(\frac{\Delta - x}{x} \cdot \frac{\Delta}{\sigma^2} - \frac{\Delta}{x^2}\right) \\ &= \underbrace{\frac{\Delta}{\sigma^2 x^2} \exp\left(\frac{2x\Delta - \Delta^2}{2\sigma^2}\right)}_{=:g'^1(x)} \underbrace{\left(x(\Delta - x) - \sigma^2\right)}_{=:g'^2(x)}, \end{aligned}$$

where the first equality follows from the chain rule and the second equality rearranges terms. Since $g'^1(x) > 0$ for all $x \in [\xi, \Delta]$, we investigate the sign of $g'^2(x)$ by investigating three cases.

**Case 1 ($\Delta^2 \leq 4\sigma^2$):** Since $x(\Delta - x) \leq \Delta^2/4$ for all $x \in \mathbb{R}$, this case implies that $g'^2(x) \leq 0$ for all $x \in [\xi, \Delta]$, and allows us to conclude that $g$ is monotonically decreasing for $x \in [\xi, \Delta]$. Such monotonicity implies in (39) that there exists some $x_0 \in (\xi, \Delta/2)$ satisfying $g(x) > 0$ for all $x \in (\xi, x_0)$ and $g(x) < 0$ for all $x \in (x_0, \Delta]$. Thus, $f_{\mathrm{q}}'$ is positive on $[0, x_0)$ and negative on $(x_0, \Delta]$, hence $x_0$ is the maximizer of $f_{\mathrm{q}}$ and the minimizer is at the extreme point. From Lemma 25 we can conclude that $\Delta$ is the minimizer. This discussion also shows that $f_{\mathrm{q}}(x)$ is unimodal on $(0, \Delta/2)$. We thus proved item *(i)* of this lemma.

**Case 2 ($\Delta^2 > 4\sigma^2$):** Since $g'(x) = 0$ if and only if $g'^2(x) = 0$, and since $g'^2(x)$ is a quadratic function, we can solve $g'(x) = 0$ and obtain two roots

$$x_1 = \frac{\Delta - \sqrt{\Delta^2 - 4\sigma^2}}{2}, \quad x_2 = \frac{\Delta + \sqrt{\Delta^2 - 4\sigma^2}}{2}. \tag{40}$$

These roots, along with the fact that $g'(\xi) < 0$ and $g'(\Delta/2) > 0$ imply that $g(x)$ is decreasing in $x \in [\xi, x_1]$ and increasing in $x \in [x_1, x_2]$. This monotonicity implies

$$g(x_1) < g(\Delta/2) < 0, \tag{41}$$

since $\Delta/2 \in (x_1, x_2)$ and $g(\Delta/2) < 0$.

Thus far we discussed that $g(\xi) > 0$ (*cf.* equation 39), $g(x_1) < 0$ (*cf.* equation 41) and $g'(x) < 0$ in $(\xi, x_1)$ (*cf.* equation 40). This concludes that there exists some $y_{\max} \in (\xi, x_1)$ such that $g(x) > 0$ for $x \in (\xi, y_{\max})$ and $g(x) < 0$ for $x \in (y_{\max}, x_1)$, and $f'_{\mathrm{q}}(x)$ exhibits a similar pattern as its sign coincides with the sign of $g(x)$. This shows that $f_{\mathrm{q}}(x)$ is unimodal in the region $(0, x_1)$ and has a local maximizer $y_{\max}$. To investigate the behavior of $f_{\mathrm{q}}(x)$ in the region $[x_1, \Delta]$, we further focus on two cases:

- Suppose that we have $g(x_2) \leq 0$. Given that $g(x)$ is increasing in $[x_1, x_2]$, this implies that $g(x) \leq 0$ for all $x \in [x_1, x_2]$. Furthermore, since we have no roots in $(x_2, \Delta]$, the fact that $g(\Delta) < 0$ implies that $g(x) \leq 0$ for all $x \in (x_2, \Delta]$. Hence, $f'_{\mathrm{q}}(x) \leq 0$ holds for all $x \in [x_1, \Delta]$, which implies that the function $f_{\mathrm{q}}(x)$ is decreasing in $[x_1, \Delta]$. We can conclude that $f_{\mathrm{q}}(x)$ achieves minimum at the extreme point $\Delta$ since $f_{\mathrm{q}}(\Delta) < f_{\mathrm{q}}(0)$ (*cf.* Lemma 25), and a maximum at $y_{\max} \in (0, x_1)$. We thus proved item *(ii)* of this lemma.

- Suppose now that $g(x_2) > 0$. We investigate the behavior of the function in four regions: $\mathcal{R}_1 = [0, x_1]$, $\mathcal{R}_2 = [x_1, \Delta/2]$, $\mathcal{R}_3 = [\Delta/2, x_2]$, $\mathcal{R}_4 = [x_2, \Delta]$. Note that Lemma 25 implies that the maximizer must lie in $[0, \Delta/2) = \mathcal{R}_1 \cup \mathcal{R}_2$ and the minimizer in $[\Delta/2, \Delta] = \mathcal{R}_3 \cup \mathcal{R}_4$. We already discussed that $f_q(x)$ is unimodal on $\mathcal{R}_1$ with a maximizer $y_{\max} \in (\xi, x_1)$ for an arbitrarily small $\xi > 0$. We can discard $\mathcal{R}_2$ since we also discussed that $g(x) < 0$ for $x \in \mathcal{R}_2$ (*i.e.*, the function $f_q(x)$ is decreasing in $\mathcal{R}_2$). Since $g(\Delta/2) < 0$, $g(x_2) > 0$, and $g'(x) > 0,\ x \in \mathcal{R}_3$ simultaneously hold, there exists a local minimum $y_{\min} \in (\Delta/2, x_2)$ in $\mathcal{R}_3$. Analogously, for $\mathcal{R}_4$, we have $g(x_2) > 0$ and $g(\Delta) < 0$, and since there is no root of $g'(x) = 0$ in $(x_2, \Delta)$, the function $g(x)$ is decreasing in $\mathcal{R}_4$. This allows us to conclude that there exists a local maximizer of $f_{\mathrm{q}}(x)$ in $\mathcal{R}_4$, but we can ignore this as we only seek minimizers in this region. In this case, the minimizer of $\mathcal{R}_4$ is one of the extreme points ($x_2$

or $\Delta$), but since we already found a local minimum in $\mathcal{R}_3$, the only candidate minimizer for $f_{\mathrm{q}}(x)$ is $\Delta$. We thus proved item *(iii)* of this lemma.

Since we proved each of the three items in the statement of this lemma, we conclude the proof. □

### 2.A.8 Proof of Lemma 22

Let $f_{\mathrm{q,max}}(\sigma) = \max_{x \in [0,\Delta]} f_{\mathrm{q}}(x;\sigma)$ and $f_{\mathrm{q,min}}(\sigma) = \min_{x \in [0,\Delta]} f_{\mathrm{q}}(x;\sigma)$. We consider the three cases studied in Lemma 21 and show that $f_{\mathrm{q,max}}(\sigma)/f_{\mathrm{q,min}}(\sigma)$ is decreasing in each of these cases.

For the *first case*, suppose $\Delta^2 \le 4\sigma^2$. In this case, Lemma 21 implies:

$$f_{\mathrm{q,min}}(\sigma) = f_{\mathrm{q}}(\Delta;\sigma) = \frac{1}{c} e^{\varepsilon} \exp\left(-\frac{\Delta^2}{2\sigma^2}\right). \tag{42}$$

Its derivative satisfies:

$$\frac{\mathrm{d}}{\mathrm{d}\sigma} f_{\mathrm{q,min}}(\sigma) = \frac{1}{c} e^{\varepsilon} \frac{\Delta^2}{\sigma^3} \exp\left(-\frac{\Delta^2}{2\sigma^2}\right). \tag{43}$$

Moreover, for $x_{\mathrm{max}} \in (0, \Delta/2)$ being the solution that satisfies $f_{\mathrm{q,max}}(\sigma) = f_{\mathrm{q}}(x_{\mathrm{max}};\sigma)$, we have

$$\begin{aligned}
&\frac{\partial}{\partial x}\Big|_{x=x_{\mathrm{max}}} f_{\mathrm{q}}(x;\sigma) = \frac{e^{\varepsilon}}{c}\left(-\frac{x_{\mathrm{max}}}{\sigma^2}\right)\exp\left(-\frac{x_{\mathrm{max}}^2}{2\sigma^2}\right) + \frac{1}{c}\left(-\frac{x_{\mathrm{max}}-\Delta}{\sigma^2}\right)\exp\left(-\frac{(x_{\mathrm{max}}-\Delta)^2}{2\sigma^2}\right) = 0 \\
&\implies -e^{\varepsilon} x_{\mathrm{max}} \exp\left(-\frac{x_{\mathrm{max}}^2}{2\sigma^2}\right) + (\Delta - x_{\mathrm{max}})\exp\left(-\frac{(x_{\mathrm{max}}-\Delta)^2}{2\sigma^2}\right) = 0 \\
&\implies e^{\varepsilon} \frac{x_{\mathrm{max}}}{\Delta - x_{\mathrm{max}}} \exp\left(-\frac{x_{\mathrm{max}}^2}{2\sigma^2}\right) = \exp\left(-\frac{(x_{\mathrm{max}}-\Delta)^2}{2\sigma^2}\right),
\end{aligned}$$

where the first equality follows from the definition of the partial derivative of $f_{\mathrm{q}}(x;\sigma)$ in $x$ as derived in Section 2.A.7 (which is equal to 0; *cf.* Lemma 21), the implication that follows multiplies the expression by the positive quantity $c\sigma^2$, and the final implication divides the expression by the positive quantity $\Delta - x_{\mathrm{max}}$. The final equality can be substituted in the definition of $f_{\mathrm{q}}(x_{\mathrm{max}};\sigma)$ to conclude:

$$\begin{aligned}
f_{\mathrm{q}}(x_{\mathrm{max}};\sigma) &= \frac{e^{\varepsilon}}{c}\exp\left(-\frac{x_{\mathrm{max}}^2}{2\sigma^2}\right) + \frac{1}{c}\exp\left(-\frac{(x_{\mathrm{max}}-\Delta)^2}{2\sigma^2}\right), \\
&= \frac{1}{c} e^{\varepsilon} \frac{\Delta}{\Delta - x_{\mathrm{max}}} \exp\left(-\frac{x_{\mathrm{max}}^2}{2\sigma^2}\right).
\end{aligned} \tag{44}$$

Note that $x_{\max}$ is a function of $\sigma$ since it is the solution that maximizes $f_{\mathrm{q}}(x;\sigma)$ on $(0,\Delta/2)$. However, it can be treated as constant when computing the derivative of $f_{\mathrm{q}}(x_{\max};\sigma)$ in $\sigma$, since $x_{\max}$ is a local minimum. To better reflect this, use a generic $x(\sigma)$ instead of $x_{\max}$, we can use the chain rule and show:

$$\frac{\mathrm{d}}{\mathrm{d}\sigma} f_{\mathrm{q}}(x(\sigma);\sigma) = \underbrace{\frac{\partial}{\partial x} f_{\mathrm{q}}(x;\sigma)}_{=0 \text{ when } x(\sigma)=x_{\max}} \cdot \frac{\mathrm{d}}{\mathrm{d}\sigma} x(\sigma) + \frac{\partial}{\partial \sigma'} f_{\mathrm{q}}(x(\sigma);\sigma'),$$

and evaluating this derivative at $x(\sigma) = x_{\max}$ gives us

$$\left.\frac{\mathrm{d}}{\mathrm{d}\sigma}\right|_{x(\sigma)=x_{\max}} f_{\mathrm{q}}(x(\sigma);\sigma) = \frac{\partial}{\partial \sigma} f_{\mathrm{q}}(x_{\max};\sigma),$$

that is, $x_{\max}$ can be treated as a constant. We thus have:

$$\frac{\mathrm{d}}{\mathrm{d}\sigma} f_{\mathrm{q,max}}(\sigma) = \frac{\partial}{\partial \sigma} f_{\mathrm{q}}(x_{\max};\sigma) = \frac{1}{c} e^{\varepsilon} \frac{\Delta}{\Delta - x_{\max}} \frac{x_{\max}^2}{\sigma^3} \exp\left(-\frac{x_{\max}^2}{2\sigma^2}\right). \tag{45}$$

We will conclude the desired result for this case if we can show

$$\left(\frac{f_{\mathrm{q,max}}(\sigma)}{f_{\mathrm{q,min}}(\sigma)}\right)' \leq 0. \tag{46}$$

We have

$$\begin{aligned}
\left(\frac{f_{\mathrm{q,max}}(\sigma)}{f_{\mathrm{q,min}}(\sigma)}\right)' &\propto \quad f'_{\mathrm{q,max}}(\sigma) f_{\mathrm{q,min}}(\sigma) - f'_{\mathrm{q,min}}(\sigma) f_{\mathrm{q,max}}(\sigma) \\
&= \left[\frac{1}{c} e^{\varepsilon} \frac{\Delta}{\Delta - x_{\max}} \frac{x_{\max}^2}{\sigma^3} \exp\left(-\frac{x_{\max}^2}{2\sigma^2}\right) \cdot \frac{1}{c} e^{\varepsilon} \exp\left(-\frac{\Delta^2}{2\sigma^2}\right)\right] - \\
&\qquad \left[\frac{1}{c} e^{\varepsilon} \frac{\Delta^2}{\sigma^3} \exp\left(-\frac{\Delta^2}{2\sigma^2}\right) \cdot \frac{1}{c} e^{\varepsilon} \frac{\Delta}{\Delta - x_{\max}} \exp\left(-\frac{x_{\max}^2}{2\sigma^2}\right)\right] \\
&\propto \frac{\Delta}{\Delta - x_{\max}} \exp\left(-\frac{\Delta^2}{2\sigma^2}\right) \exp\left(-\frac{x_{\max}^2}{2\sigma^2}\right) \left[\frac{x_{\max}^2}{\sigma^3} - \frac{\Delta^2}{\sigma^3}\right] \\
&\propto \frac{x_{\max}^2}{\sigma^3} - \frac{\Delta^2}{\sigma^3} \leq 0.
\end{aligned}$$

Here, the first step uses the quotient rule and ignores the positive quantity $f_{\mathrm{q,min}}(\sigma)^{-2}$, the second step uses equations (42), (43), (44) and (45), the third step ignores the positive term $(e^{\varepsilon}/c)^2$, the fourth step ignores the exponential terms (as they are positive) as well as $\Delta/(\Delta - x_{\max})$ (since $x_{\max} \in (0,\Delta/2)$), and the final inequality follows from $x_{\max} \in (0,\Delta/2)$.

For the *second case*, suppose $\Delta^2 > 4\sigma^2$ and $g(x_2) \leq 0$ where $x_2$ and $g$ are defined as in Lemma 21. Similarly to the first case the minimizer of $f_{\mathrm{q}(x;\sigma)}$ over $[0, \Delta]$ (for fixed $\sigma$) is $x = \Delta$. The maximizer is the local maximizer in the unimodal region $(0, x_1)$, which is a smaller region than the one in the first case ($x_1 \leq \Delta/2$), hence the above analysis applies analogously here.

For the *third case*, suppose $\Delta^2 > 4\sigma^2$ and $g(x_2) > 0$ where $x_2$ and $g$ are defined as in Lemma 21. Let $x_{\max}$ be the solution that satisfies $f_{\mathrm{q},\max}(\sigma) = f_{\mathrm{q}}(x_{\max};\sigma)$, and $x_{\min}$ be the solution that satisfies $f_{\mathrm{q},\min}(\sigma) = f_{\mathrm{q}}(x_{\min};\sigma)$. From Lemma 21 it follows that $x_{\max}$ is the local maximum of $f_{\mathrm{q}}(x;\sigma)$ over the unimodal region $x \in (0, x_1)$ for $x_1$ as defined in Lemma 21. On the other hand, $x_{\min}$ is either equal to $\Delta$, or it is the local minimum of $f_{\mathrm{q}}(x;\sigma)$ over the unimodal region $x \in (\Delta/2, x_2)$ for $x_2$ as defined in Lemma 21. If $x_{\min} = \Delta$, the proof follows analogously as in the first two cases. Hence, for the rest of the proof, we assume without loss of generality that:

*(i)* $x_{\min}$ is the local minimum of $f_{\mathrm{q}}(x;\sigma)$ over the unimodal region $x \in (\Delta/2, x_2)$ for $x_2 = \Delta/2 + \sqrt{\Delta^2 - 4\sigma^2}/2$;

*(ii)* $x_{\max}$ is the local maximum of $f_{\mathrm{q}}(x;\sigma)$ over the unimodal region $x \in (0, x_1)$ for $x_1 = \Delta/2 - \sqrt{\Delta^2 - 4\sigma^2}/2$.

Since $f_{\mathrm{q},\max}(\sigma)$ and $f_{\mathrm{q},\min}(\sigma)$ are positive functions, a sufficient condition for $f_{\mathrm{q},\max}(\sigma)/f_{\mathrm{q},\min}(\sigma)$ being decreasing in $\sigma$ is $\log(f_{\mathrm{q},\max}(\sigma)/f_{\mathrm{q},\min}(\sigma))$ being decreasing in $\sigma$. We have:

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}\sigma}\log\left(\frac{f_{\mathrm{q},\max}(\sigma)}{f_{\mathrm{q},\min}(\sigma)}\right) &= \frac{\mathrm{d}}{\mathrm{d}\sigma}\left[\log(f_{\mathrm{q},\max}(\sigma)) - \log(f_{\mathrm{q},\min}(\sigma))\right] \\ &= \underbrace{\frac{f'_{\mathrm{q},\max}(\sigma)}{f_{\mathrm{q},\max}(\sigma)}}_{(i)} - \underbrace{\frac{f'_{\mathrm{q},\min}(\sigma)}{f_{\mathrm{q},\min}(\sigma)}}_{(ii)}.\end{aligned}$$

We will thus show that *(i)* $\leq$ *(ii)* and conclude the desired result. From (45) and (44), we have:

$$\frac{f'_{\mathrm{q},\max}(\sigma)}{f_{\mathrm{q},\max}(\sigma)} = \frac{\frac{1}{c}e^{\varepsilon}\frac{\Delta}{\Delta - x_{\max}}\frac{x_{\max}^2}{\sigma^3}\exp\left(-\frac{x_{\max}^2}{2\sigma^2}\right)}{\frac{1}{c}e^{\varepsilon}\frac{\Delta}{\Delta - x_{\max}}\exp\left(-\frac{x_{\max}^2}{2\sigma^2}\right)} = \frac{x_{\max}^2}{\sigma^3}.$$

With analogous calculations, we have:

$$\frac{f'_{\mathrm{q},\min}(\sigma)}{f_{\mathrm{q},\min}(\sigma)} = \frac{x_{\min}^2}{\sigma^3}.$$

Since $x_{\min} \in (\Delta/2, \Delta/2 + \sqrt{\Delta^2 - 4\sigma^2}/2)$ and $x_{\max} \in (0, \Delta/2 - \sqrt{\Delta^2 - 4\sigma^2}/2)$, we always have $x_{\max} < x_{\min}$, hence $(i) \leq (ii)$.

We conclude the proof of monotonicity since we showed that, regardless of the case, the derivative of $f_{\mathrm{q},\max}(\sigma)/f_{\mathrm{q},\min}(\sigma)$ is negative. We can ignore the end-points of the cases since all the functions and their derivatives investigated in this proof are continuous in.

Finally, to see that the function of interest is upper bounded by $e^{\varepsilon}$ at $\sigma = \sqrt{\Delta^2/(2\varepsilon)}$, note that

$$\begin{aligned}
\max_{x\in[0,\Delta]} f_{\mathrm{q}}(x;\sigma) &\leq \frac{1}{\sqrt{2\pi}\sigma} \max\left\{ \max_{x\in[0,\Delta]} \exp\left(-\frac{x^2}{2\sigma^2}\right), \max_{x\in[0,\Delta]} \exp\left(-\frac{(|x|-\Delta)^2}{2\sigma^2}\right)\right\} \\
&= \frac{1}{\sqrt{2\pi}\sigma} \max\left\{ \exp\left(-\frac{0}{2\sigma^2}\right), \exp\left(-\frac{0}{2\sigma^2}\right)\right\} \\
&= \frac{1}{\sqrt{2\pi}\sigma}
\end{aligned}$$

where the inequality holds since $f_{\mathrm{q}}(\cdot;\sigma)$ is the convex combination of two distributions, hence the maximum of $f_{\mathrm{q}}(\cdot;\sigma)$ is upper bounded by one of the maxima of each of these two distributions, the first equality holds by solving the inner max-problems, and the second equality holds by noting that both inner max problems have the same optimal value. Analogously, we can show

$$\begin{aligned}
\min_{x\in[0,\Delta]} f_{\mathrm{q}}(x;\sigma) &\geq \frac{1}{\sqrt{2\pi}\sigma} \min\left\{ \min_{x\in[0,\Delta]} \exp\left(-\frac{x^2}{2\sigma^2}\right), \min_{x\in[0,\Delta]} \exp\left(-\frac{(|x|-\Delta)^2}{2\sigma^2}\right)\right\} \\
&= \frac{1}{\sqrt{2\pi}\sigma} \min\left\{ \exp\left(-\frac{\Delta^2}{2\sigma^2}\right), \exp\left(-\frac{\Delta^2}{2\sigma^2}\right)\right\} \\
&= \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{\Delta^2}{2\sigma^2}\right).
\end{aligned}$$

We thus have

$$\begin{aligned}
\frac{\max_{x\in[0,\Delta]} f_{\mathrm{q}}(x;\sigma)}{\min_{x\in[0,\Delta]} f_{\mathrm{q}}(x;\sigma)} &\leq \frac{\frac{1}{\sqrt{2\pi}\sigma}}{\frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{\Delta^2}{2\sigma^2}\right)} \\
&= \exp\left(\frac{\Delta^2}{2\sigma^2}\right)
\end{aligned}$$

and plugging in $\sigma = \sqrt{\Delta^2}/(2\varepsilon)$ sets the above quantity to $e^{\varepsilon}$.

### 2.A.9 Plots for Section 2.2

We visualize the monotonicity of the constraint of $\sigma_1$ in (26a), as discussed in Lemma 20. We visualize a case where $\Delta = 1$, and plot two cases: the left plot uses $(\varepsilon, \delta) = (1, 0.1)$ and the right plot uses $(\varepsilon, \delta) = (1, 0.25)$. In the first plot, we observe that the function of interest is monotonically increasing on $(0\sqrt{2(\varepsilon - \log\delta)}\Delta/\varepsilon)$ and is nonnegative beyond this region. The point where the function hits 0 is also visualized as $\sigma_1$. In the second case, we have $e^{\varepsilon} + 2 \geq \delta^{-1}$, and thus the function is nonnegative (hence $\sigma_1 = 0$).

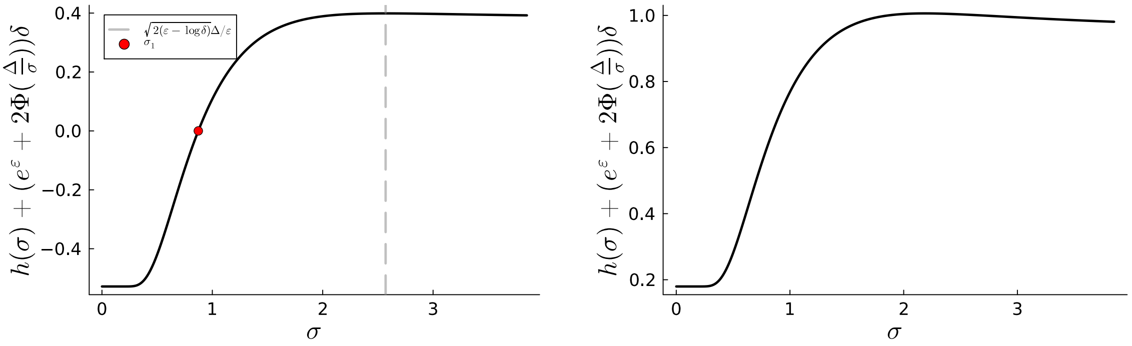


Figure 11: *The constraint of* (26a) *as a function of* $\sigma$ *under* $\Delta = 1$, $(\varepsilon, \delta) = (1, 0.1)$ *(left) and* $(\varepsilon, \delta) = (1, 0.25)$ *(right).*

Finally, in Figure 12, we visualize Lemmas 21 and (22) that are used in finding $\sigma_2$ in (27a), for an instance with $\Delta = 1$ and $\varepsilon = 1$. The left figure plots $f_{\mathrm{q}}(x; \sigma)$ on $x \in [0, \Delta]$ for fixed $\sigma = 0.2$ and marks the maximizer and minimizer of this function over this range. We can observe what Lemma 21 proves: the maximizer $x_{\max}$ is the local maximum of the function on the unimodal region $(0, x_1)$, and the minimizer $x_{\min}$ is the local minimum of the function on the unimodal region $(\Delta/2, x_2)$. The term $f_{\mathrm{q}}(x_{\max}; \sigma)/f_{\mathrm{q}}(x_{\min}; \sigma) - e^{\varepsilon}$, as a function of $\sigma$, is monotonically decreasing as proved in Lemma (22), which is displayed in the right figure.

### 2.A.10 Proof of Theorem 4

For the correctness of the algorithm, note that the $\sigma_1$ computed coincides with the $\sigma_1$ in Theorem 3, which is found via a bisection search method due to the monotonicity presented in Lemma 20. This lemma also gives an upper bound for $\sigma$ as beyond this upper bound the function is nonnegative, and Algorithm 4 accordingly restricts the search to a bounded region. Note that the if condition in this algorithm also ensures that we apply the bisection search method

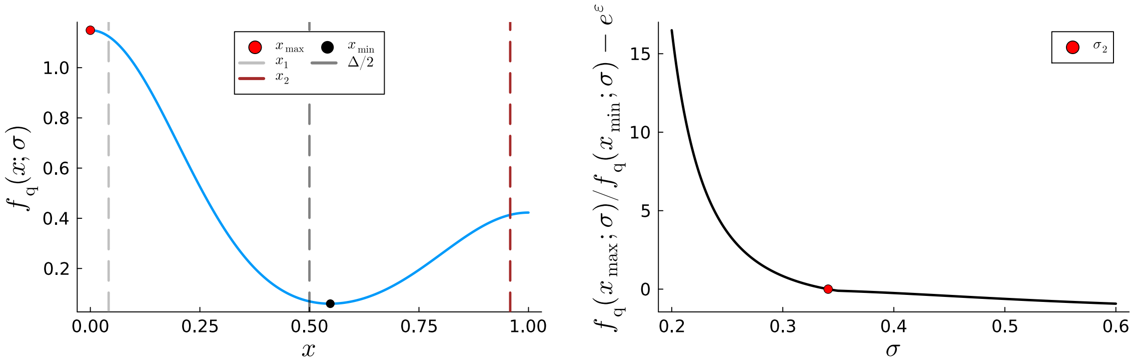


Figure 12: *(Left) The density* $f_{\mathrm{q}}(x;\sigma)$ *on* $x \in [0,\Delta]$ *for* $\Delta = 1$*,* $\varepsilon = 1$*, and* $\sigma = 0.2$*. The notation* $x_1, x_2, x_{\max}$ *and* $x_{\min}$ *comes from Algorithm 5. (Right) The term* $f_{\mathrm{q}}(x_{\max};\sigma)/f_{\mathrm{q}}(x_{\min};\sigma) - e^{\varepsilon}$ *displayed as a function of* $\sigma$ *where* $\Delta = 1$ *and* $\varepsilon = 1$*. The point that sets this expression to* 0 *is the* $\sigma_2$ *of* (27a)*.*

only in the case if the function is not nonnegative everywhere, as presented by Lemma 20. The $\sigma_2$ computed in this algorithm, similarly, coincides with $\sigma_2$ of Theorem 3. This value can be found via a bisection search restricted to $(0, \sqrt{\Delta^2/(2\varepsilon)})$ due to Lemma 22. Each evaluation of the bisection search relies on Lemma 21 where the if-else if-else condition shows the cases presented in this lemma (the cases are elaborated in the unabridged version of the lemma). Finally, Algorithm 4 returns $\sigma$ as the maximum of $\sigma_1$ and $\sigma_2$, and therefore from Theorem 3 we conclude the feasibility of the returned value.

For the runtime complexity, first note that we treat the tolerances in the bisection and golden-section search methods as constants. Note that one needs to be careful in implementation since numerically the tolerances should be reflected in the presentation of $\delta$ to ensure feasibility.

Now, note that Algorithm 4 has two bisection search methods implemented. The first one has a search radius of $\sqrt{2(\varepsilon - \log\delta)}\Delta/\varepsilon$ and the evaluation of $\psi_1$ is constant assuming computation of $\Phi(\cdot)$ is constant. The complexity of the first bisection search is thus:

$$\begin{aligned}\mathcal{O}\left(\log(\sqrt{2(\varepsilon - \log\delta)}\Delta/\varepsilon)\right) &= \mathcal{O}\left(\log(\varepsilon - \log\delta)\right) + \mathcal{O}(\log\Delta) - \mathcal{O}(\log\varepsilon)\\ &= \mathcal{O}\left(\log\left(\frac{\varepsilon - \log\delta}{\varepsilon}\right)\right) + \mathcal{O}(\log\Delta)\\ &= \mathcal{O}\left(\log\left(-\frac{\log\delta}{\varepsilon}\right)\right) + \mathcal{O}(\log\Delta).\end{aligned}$$

On the other hand, the second bisection search has a radius of $\sqrt{\Delta^2/(2\varepsilon)}$ where each iteration relies on Algorithm 5 which, in the worst case, calls a golden-section search method whose search region is a subset of $(0, \Delta)$. The computation of $f_{\mathrm{q}}(x; \sigma)$ during the golden-section search evaluations is constant, hence the golden section has complexity $\mathcal{O}(\log \Delta)$. The complexity of the bisection search is overall:

$$\mathcal{O}\left(\log \Delta \cdot \sqrt{\frac{\Delta^2}{2\varepsilon}}\right) = \mathcal{O}\left(\frac{\Delta \log \Delta}{\varepsilon}\right).$$

The sum of the complexities of both bisection search methods concludes the complexity presented in the statement of this theorem.

## 2.B Proofs for Section 2.3

### 2.B.1 The multi-Gaussian mixture distribution is well-defined

To show that the density function (28) of the multi-Gaussian mixture distribution is well-defined, fix arbitrary $\sigma > 0$ and $k \in \mathbb{N}$, and note that $f_{\mathrm{m}}(x; \sigma, K) \geq 0$ for all $x \in \mathbb{R}$ since all terms in its definition are nonnegative. Moreover, observe that the density function integrates to 1 since each $\phi$-term is the unnormalized Gaussian density function where only the mean varies, that is,

$$\int_{-\infty}^{\infty} \phi(x; k\Delta, \sigma) \mathrm{d}x = \sqrt{2\pi}\sigma,$$

hence by the linearity of integrals, the density (28) integrates to 1. We can also simply represent the cumulative distribution function $F_{\mathrm{m}}(\bar{x}; \sigma)$ via Gaussian cumulative distribution functions:

$$\begin{aligned}
F_{\mathrm{m}}(\bar{x}; \sigma) &= \frac{1}{c_K} \sum_{k=-K}^{K} e^{-|k|\varepsilon} \int_{-\infty}^{\bar{x}} \phi(x; k\Delta, \sigma) \mathrm{d}x \\
&= \frac{1}{c_K} \sum_{k=-K}^{K} e^{-|k|\varepsilon} \sqrt{2\pi}\sigma \Phi\left(\frac{\bar{x} - k\Delta}{\sigma}\right) \\
&= \frac{1}{\sqrt{2\pi}\sigma \sum_{k=-K}^{K} e^{-|k|\varepsilon}} \sum_{k=-K}^{K} e^{-|k|\varepsilon} \sqrt{2\pi}\sigma \Phi\left(\frac{\bar{x} - k\Delta}{\sigma}\right) \\
&= \frac{1}{\sum_{k=-K}^{K} e^{-|k|\varepsilon}} \sum_{k=-K}^{K} e^{-|k|\varepsilon} \Phi\left(\frac{\bar{x} - k\Delta}{\sigma}\right),
\end{aligned}$$

where the first equality exploits the linearity of integrals, the second equality replaces the integral via a closed-form term using the Gaussian cumulative distribution, the third equality

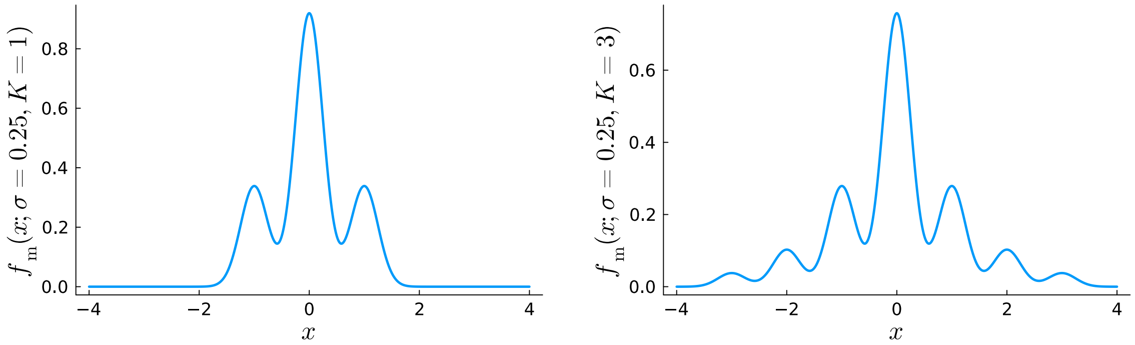


Figure 13: *The probability density function* (28) *of the multi-Gaussian mixture distribution for* $\varepsilon = 1$ *and* $\sigma = 0.25$ *where* $K$ *is set to* 1 *(left) and* 3 *(right).*

substitutes the definition of $c_K$, and the final equality cancels the $\sqrt{2\pi}\sigma$ terms.

Figure 13 visualizes the probability density function (28) of the multi-Gaussian mixture distribution for $\varepsilon = 1.0$ and $\sigma = 0.25$ for $K = 1$ and $K. = 3$. For a better intuition, one can rewrite (28) as:

$$f_{\mathrm{m}}(x; \sigma, K) = \frac{1}{c_K}\left[\phi(x; 0, \sigma) + \sum_{k=1}^{K} e^{-k\varepsilon}\left(\phi(x; k\Delta, \sigma) + \phi(x; -k\Delta, \sigma)\right)\right].$$

Note that sampling from this distribution is straightforward since for $k \in \{-K, \dots, K\}$, we sample from the $k\Delta$-mean Gaussian distribution with standard deviation $\sigma$ with probability $e^{-|k|\varepsilon} / \sum_{k=-K}^{K} e^{-|k|\varepsilon}$.

### 2.B.2 Noise amplitude and power of the multi-Gaussian mixture distributions

For the *noise amplitude* of the random variable $\tilde{X}$ with density $f_{\mathrm{m}}(\cdot; \sigma, K)$, we have:

$$\mathbb{E}_{\tilde{X}}[|x|] = \frac{1}{c_K} \sum_{k=-K}^{K} e^{-|k|\varepsilon} \int_{-\infty}^{\infty} |x| \phi(x; k\Delta, \sigma) \mathrm{d}x.$$

The integral can be written as

$$\begin{aligned} &\int_{-\infty}^{\infty} |x| \exp\left(-\frac{(x - k\Delta)^2}{2\sigma^2}\right) \mathrm{d}x \\ =& \int_{-\infty}^{\infty} |y + k\Delta| \exp\left(-\frac{y^2}{2\sigma^2}\right) \mathrm{d}y \end{aligned}$$

$$
\begin{aligned}
&= - \int_{-\infty}^{-k\Delta} (y + k\Delta) \exp\left(-\frac{y^2}{2\sigma^2}\right) \mathrm{d}y + \int_{-k\Delta}^{\infty} (y + k\Delta) \exp\left(-\frac{y^2}{2\sigma^2}\right) \mathrm{d}y \\
&= - \int_{-\infty}^{-k\Delta} y \exp\left(-\frac{y^2}{2\sigma^2}\right) \mathrm{d}y - k\Delta \int_{-\infty}^{-k\Delta} \exp\left(-\frac{y^2}{2\sigma^2}\right) \mathrm{d}y \\
&\qquad + \int_{-k\Delta}^{\infty} y \exp\left(-\frac{y^2}{2\sigma^2}\right) \mathrm{d}y + k\Delta \int_{-k\Delta}^{\infty} \exp\left(-\frac{y^2}{2\sigma^2}\right) \mathrm{d}y \\
&= - \int_{-\infty}^{-k\Delta} y \exp\left(-\frac{y^2}{2\sigma^2}\right) \mathrm{d}y + \int_{-k\Delta}^{\infty} y \exp\left(-\frac{y^2}{2\sigma^2}\right) \mathrm{d}y + k\Delta\sqrt{2\pi}\sigma \left(\Phi\left(\frac{k\Delta}{\sigma}\right) - \Phi\left(-\frac{k\Delta}{\sigma}\right)\right) \\
&= \int_{-\infty}^{\infty} y \exp\left(-\frac{y^2}{2\sigma^2}\right) \mathrm{d}y - 2\int_{-\infty}^{-k\Delta} y \exp\left(-\frac{y^2}{2\sigma^2}\right) \mathrm{d}y + k\Delta\sqrt{2\pi}\sigma \left(\Phi\left(\frac{k\Delta}{\sigma}\right) - \Phi\left(-\frac{k\Delta}{\sigma}\right)\right) \\
&= -2\int_{-\infty}^{-k\Delta} y \exp\left(-\frac{y^2}{2\sigma^2}\right) \mathrm{d}y + k\Delta\sqrt{2\pi}\sigma \left(\Phi\left(\frac{k\Delta}{\sigma}\right) - \Phi\left(-\frac{k\Delta}{\sigma}\right)\right) \\
&= -2\left[-\sigma^2 \exp\left(-\frac{y^2}{2\sigma^2}\right)\right]_{y=-\infty}^{-k\Delta} + k\Delta\sqrt{2\pi}\sigma \left(\Phi\left(\frac{k\Delta}{\sigma}\right) - \Phi\left(-\frac{k\Delta}{\sigma}\right)\right) \\
&= 2\sigma^2 \exp\left(-\frac{k^2\Delta^2}{2\sigma^2}\right) + k\Delta\sqrt{2\pi}\sigma \left(1 - 2\Phi\left(-\frac{k\Delta}{\sigma}\right)\right).
\end{aligned}
$$

Here, the first equality is due to variable change, the second equality partitions the integral region into two regions, the third equality exploits the linearity of integrals, the fourth equality replaces the Gaussian density integrals with their closed form representation, the fifth equality breaks down its second integral (from $-k\Delta$ to $\infty$) as the subtraction of two integrals (from $-\infty$ to $\infty$ and from $-\infty$ to $-k\Delta$), the sixth equality notes that the first integral is the expected value of a 0-mean Gaussian, and the following two equalities compute the definite integral. Hence, we have

$$
\mathbb{E}_{\tilde{X}}[|x|] = \frac{1}{c_K} \sum_{k=-K}^{K} e^{-|k|\varepsilon} \left[2\sigma^2 \exp\left(-\frac{k^2\Delta^2}{2\sigma^2}\right) + k\Delta\sqrt{2\pi}\sigma \left(1 - 2\Phi\left(-\frac{k\Delta}{\sigma}\right)\right)\right].
$$

For the *noise power* of the random variable $\tilde{X}$ with density $f_{\mathrm{m}}(\cdot;\sigma,K)$, we similarly have:

$$
\mathbb{E}_{\tilde{X}}[x^2] = \frac{1}{c_K} \sum_{k=-K}^{K} e^{-|k|\varepsilon} \int_{-\infty}^{\infty} x^2 \phi(x; k\Delta, \sigma) \mathrm{d}x.
$$

The integral can be written as

$$
\int_{-\infty}^{\infty} x^2 \exp\left(-\frac{(x-k\Delta)^2}{2\sigma^2}\right) \mathrm{d}x
$$

$$
\begin{aligned}
&= \int_{-\infty}^{\infty} (y + k\Delta)^2 \exp\left(-\frac{y^2}{2\sigma^2}\right) \mathrm{d}y \\
&= \underbrace{\int_{-\infty}^{\infty} y^2 \exp\left(-\frac{y^2}{2\sigma^2}\right) \mathrm{d}y}_{=\sqrt{2\pi}\sigma^3} + k^2\Delta^2 \underbrace{\int_{-\infty}^{\infty} \exp\left(-\frac{y^2}{2\sigma^2}\right) \mathrm{d}y}_{=\sqrt{2\pi}\sigma} + 2k\Delta \underbrace{\int_{-\infty}^{\infty} y \exp\left(-\frac{y^2}{2\sigma^2}\right) \mathrm{d}y}_{=0} \\
&= \sqrt{2\pi}\sigma^3 + k^2\Delta^2\sqrt{2\pi}\sigma
\end{aligned}
$$

where the first equality is due to variable change, the second equality expands the squared term, and the final equality uses the fact that $\exp(-y^2/2\sigma^2)$ is the (unnormalized) density function of a Gaussian distribution with 0 mean and $\sigma^2$ variance. This concludes that:

$$
\begin{aligned}
\mathbb{E}_{\tilde{X}}[x^2] =& \frac{1}{c_K} \sum_{k=-K}^{K} e^{-|k|\varepsilon} (\sqrt{2\pi}\sigma^3 + k^2\Delta^2\sqrt{2\pi}\sigma) \\
=& \frac{1}{\sum_{k=-K}^{K} e^{-|k|\varepsilon}} \sum_{k=-K}^{K} e^{-|k|\varepsilon} (\sigma^2 + k^2\Delta^2).
\end{aligned}
$$

### 2.B.3 Proof of Theorem 5

We first state and prove an intermediary result.

**Lemma 26.** *For a multi-Gaussian mixture distribution with parameters $\varphi, \Delta, \sigma > 0$ and $K \in \mathbb{N}$, and for any $0 \leq \varphi_1 \leq \varphi_2 \leq \Delta$, we have:*

$$
\int_{-\infty}^{\infty} |f_{\mathrm{m}}(x + \varphi_1; \sigma, K) - f_{\mathrm{m}}(x + \varphi_2; \sigma, K)| \mathrm{d}x \leq \frac{\sqrt{2/\pi}}{\sigma} (\varphi_2 - \varphi_1).
$$

*Proof.* Since $\sigma > 0$ and $K \in \mathbb{N}$ are fixed, we denote by $f'_{\mathrm{m}}(x; \sigma, K)$ the derivative of $f_{\mathrm{m}}$ in $x$. By using this notation, the left-hand side of the inequality in the statement can be upper bounded by

$$
\begin{aligned}
\int_{-\infty}^{\infty} |f_{\mathrm{m}}(x + \varphi_1; \sigma, K) - f_{\mathrm{m}}(x + \varphi_2; \sigma, K)| \mathrm{d}x &= \int_{-\infty}^{\infty} \left| \int_{\varphi_1}^{\varphi_2} f'_{\mathrm{m}}(x + \varphi; \sigma, K) \mathrm{d}\varphi \right| \mathrm{d}x \\
&\leq \int_{-\infty}^{\infty} \int_{\varphi_1}^{\varphi_2} \left| f'_{\mathrm{m}}(x + \varphi; \sigma, K) \right| \mathrm{d}\varphi \, \mathrm{d}x \\
&= \int_{\varphi_1}^{\varphi_2} \underbrace{\int_{-\infty}^{\infty} \left| f'_{\mathrm{m}}(x + \varphi; \sigma, K) \right| \mathrm{d}x}_{(i)} \, \mathrm{d}\varphi,
\end{aligned}
$$

where the first equality follows from the fundamental theorem of calculus, the inequality follows from the triangle inequality of the absolute value, and the final equality changes the order of integrals via Fubini's theorem. Finally, term $(i)$ satisfies

$$\begin{aligned}
(i) &= \int_{-\infty}^{\infty} \left| f_{\mathrm{m}}'(x;\sigma,K) \right| \mathrm{d}x \\
&\leq \frac{1}{c_K} \sum_{k=-K}^{K} e^{-|k|\varepsilon} \int_{-\infty}^{\infty} \left| \frac{\partial}{\partial x} \exp\left( -\frac{(x-k\Delta)^2}{2\sigma^2} \right) \right| \mathrm{d}x \\
&= \frac{1}{c_K} \sum_{k=-K}^{K} e^{-|k|\varepsilon} \int_{-\infty}^{\infty} \frac{|x-k\Delta|}{\sigma^2} \exp\left( -\frac{(x-k\Delta)^2}{2\sigma^2} \right) \mathrm{d}x \\
&= \frac{1}{c_K} \sum_{k=-K}^{K} e^{-|k|\varepsilon} \int_{-\infty}^{\infty} \frac{|y|}{\sigma^2} \exp\left( -\frac{y^2}{2\sigma^2} \right) \mathrm{d}y \\
&= \frac{2}{\sqrt{2\pi}\sigma \sum_{k=-K}^{K} e^{-|k|\varepsilon}} \sum_{k=-K}^{K} e^{-|k|\varepsilon} = \frac{\sqrt{2/\pi}}{\sigma},
\end{aligned}$$

where the first equality follows from variable change $x = x + \varphi$, the inequality uses the definition of $f_{\mathrm{m}}(x;\sigma,K)$ and exploits the triangle inequality of the absolute value, the second equality explicitly writes the derivative term, the third equality follows from variable change $y = x - k\Delta$, the fourth equality explicitly writes the expression for $c_K$ as well as notes that the integral computes the expected absolute value of a random variable that follows a Gaussian distribution with 0 mean and $\sigma^2$ variance (*cf.* term $(i)$ of Appendix 2.A.4), and the final equality cancels common terms. Using this upper bound for $(i)$ back allows us to conclude

$$\int_{\varphi_1}^{\varphi_2} \int_{-\infty}^{\infty} \left| f_{\mathrm{m}}'(x+\varphi;\sigma,K) \right| \mathrm{d}x \, \mathrm{d}\varphi \ \leq \ \int_{\varphi_1}^{\varphi_2} \frac{\sqrt{2/\pi}}{\sigma} \mathrm{d}\varphi \ = \ \frac{\sqrt{2/\pi}}{\sigma}(\varphi_2 - \varphi_1)$$

which coincides with the right-hand side presented in the lemma. □

We can now prove Theorem 5.

**Proof of Theorem 5.** Given $\eta \in (0,1)$ fix any $\beta \leq \sqrt{2\pi}\eta\sigma\delta$, and assume the condition in the statement of this theorem holds:

$$\int_{-\infty}^{\infty} \min\{e^{\varepsilon} f_{\mathrm{m}}(x;\sigma,K) - f_{\mathrm{m}}(x+\varphi;\sigma,K), 0\} \mathrm{d}x + (1-\eta)\delta \geq 0 \quad \forall \varphi \in \{0, \beta, 2\beta, \ldots, \Delta\}. \tag{47}$$

To show the desired result, we should show that this condition guarantees the following definition of $(\varepsilon, \delta)$-DP (*cf.* Lemma 24):

$$\begin{aligned}
&\int_{x \in A} (e^{\varepsilon} f_{\mathrm{m}}(x; \sigma, K) - f_{\mathrm{m}}(x + \varphi; \sigma, K)) \mathrm{d}x \geq -\delta && \forall \varphi \in [0, \Delta],\ \forall A \in \mathcal{F} \\
\iff\ & \inf_{A \in \mathcal{F}} \left\{ \int_{x \in A} (e^{\varepsilon} f_{\mathrm{m}}(x; \sigma, K) - f_{\mathrm{m}}(x + \varphi; \sigma, K)) \mathrm{d}x \right\} \geq -\delta && \forall \varphi \in [0, \Delta] \\
\iff\ & \int_{-\infty}^{\infty} \min\{e^{\varepsilon} f_{\mathrm{m}}(x; \sigma, K) - f_{\mathrm{m}}(x + \varphi; \sigma, K), 0\} \mathrm{d}x \geq -\delta && \forall \varphi \in [0, \Delta]. \qquad (48)
\end{aligned}$$

Here, the first equivalence follows from taking the infimum of the left hand side over $A \in \mathcal{F}$, and the second equivalence follows since the worst-case event $A$ does not include points $x$ with $e^{\varepsilon} f_{\mathrm{m}}(x; \sigma, K) - f_{\mathrm{m}}(x + \varphi; \sigma, K)) > 0$. We will show (47) $\implies$ (48).

Fix an arbitrary $\varphi \in [0, \Delta]$, let $k' \in \mathbb{N}$ be the index that satisfies $|\varphi - k'\beta| \leq \beta/2$, and note that the left-hand side of (48) satisfies:

$$\begin{aligned}
&\int_{-\infty}^{\infty} \min\{(e^{\varepsilon} f_{\mathrm{m}}(x; \sigma, K) - f_{\mathrm{m}}(x + \varphi; \sigma, K)), 0\} \mathrm{d}x \\
\geq &\int_{-\infty}^{\infty} \min\{f_{\mathrm{m}}(x; \sigma, K) - f_{\mathrm{m}}(x + \varphi; \sigma, K), 0\} \mathrm{d}x \\
= &\int_{-\infty}^{\infty} \min\{(f_{\mathrm{m}}(x; \sigma, K) - f_{\mathrm{m}}(x + k'\beta; \sigma, K)) - (f_{\mathrm{m}}(x + \varphi; \sigma, K) - f_{\mathrm{m}}(x + k'\beta; \sigma, K)), 0\} \mathrm{d}x \\
\geq &\int_{-\infty}^{\infty} \min\{f_{\mathrm{m}}(x; \sigma, K) - f_{\mathrm{m}}(x + k'\beta; \sigma, K), 0\} - |f_{\mathrm{m}}(x + \varphi; \sigma, K) - f_{\mathrm{m}}(x + k'\beta; \sigma, K)| \mathrm{d}x \\
= &\underbrace{\int_{-\infty}^{\infty} \min\{f_{\mathrm{m}}(x; \sigma, K) - f_{\mathrm{m}}(x + k'\beta; \sigma, K), 0\} \mathrm{d}x}_{(i)} \\
&\qquad\qquad - \underbrace{\int_{-\infty}^{\infty} |f_{\mathrm{m}}(x + \varphi; \sigma, K) - f_{\mathrm{m}}(x + k'\beta; \sigma, K)| \mathrm{d}x}_{(ii)}.
\end{aligned}$$

Here, the first inequality follows from $e^{\varepsilon} \geq 1$, the equality that follows adds and subtracts the common term $f_{\mathrm{m}}(x + k'\beta; \sigma, K)$ in the min-term, the second inequality is due to the fact that for any $a, b \in \mathbb{R}$ we have $\min\{a - b, 0\} \geq \min\{a, 0\} - |b|$, and the final equality exploits the linearity of integrals. To conclude this proof, we will now show that $(i) - (ii) \geq -\delta$.

The fact that $\varphi \in [0, \Delta]$ implies that $k'\beta \in \{0, \beta, 2\beta, \ldots, \Delta\}$, hence from (47) we have:

$$(i) \geq -(1 - \eta)\delta.$$

Moreover, term $(ii)$ can be upper bounded by

$$(ii) \leq \frac{\sqrt{2/\pi}}{\sigma}|\varphi - k'\beta| \leq \frac{1}{\sqrt{2\pi}\sigma}\beta \leq \eta\delta,$$

where the first inequality follows from Lemma 26, the second inequality is due to the definition of $k' \in \mathbb{N}$, and the final inequality is due to the definition of $\beta$.

Putting these terms together finally concludes the proof since $(i) - (ii) \geq -(1-\eta)\delta - \eta\delta = -\delta$. □

### 2.B.4 Proof of Lemma 23

Throughout this proof, we will use the representation

$$f_{\mathrm{m}}(x; \sigma, K) = \frac{1}{\sum_{k=-K}^{K} e^{-|k|\varepsilon}} \sum_{k=-K}^{K} \mathcal{N}(x; k\Delta, \sigma),$$

where

$$\mathcal{N}(x; \mu, \sigma) := \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

denotes the normalized Gaussian density function. Using such notation, we first provide an intermediary result.

**Lemma 27.** *For fixed $\varepsilon > 0$, $\delta \in (0,1)$, $\Delta > 0$, $K \in \mathbb{N}$, and $\varphi \in [0, \Delta]$, define*

$$\ell(x; \sigma) := e^{\varepsilon} f_{\mathrm{m}}(x; \sigma, K) - f_{\mathrm{m}}(x + \varphi; \sigma, K)$$

*as the integrand of the left-hand side of the constraint (29) for any $\sigma > 0$. For all $\sigma' \geq \sigma > 0$, we have*

$$\ell(x; \sigma') := (\ell(\cdot; \sigma) \otimes \mathcal{N}(\cdot; 0, \sqrt{\sigma'^2 - \sigma^2}))(x),$$

*where $\otimes$ denotes the convolution operator defined as $(f \otimes g)(x) := \int_{-\infty}^{\infty} f(y) g(x-y) \mathrm{d}y$.*

*Proof.* For $\sigma' \geq \sigma$, if we denote by $\tau = \sqrt{\sigma'^2 - \sigma^2}$, then from (Bromiley 2003) the following identity holds:

$$\mathcal{N}(x; k\Delta, \sigma') = \mathcal{N}(x; k\Delta, \sigma) \otimes \mathcal{N}(x; 0, \tau).$$

Note that we have a slight breach of notation since with $\mathcal{N}(x;k\Delta,\sigma)\otimes\mathcal{N}(x;0,\tau)$ we mean $(\mathcal{N}(\cdot;k\Delta,\sigma)\otimes\mathcal{N}(\cdot;0,\tau))(x)$. We can now show

$$\begin{aligned}
f_{\mathrm{m}}(x;\sigma',K) &= \frac{1}{\sum_{k=-K}^{K} e^{-|k|\varepsilon}} \sum_{k=-K}^{K} \mathcal{N}(x;k\Delta,\sigma') \\
&= \frac{1}{\sum_{k=-K}^{K} e^{-|k|\varepsilon}} \sum_{k=-K}^{K} \mathcal{N}(x;k\Delta,\sigma)\otimes\mathcal{N}(x;0,\tau) \\
&= \frac{1}{\sum_{k=-K}^{K} e^{-|k|\varepsilon}} \left[\left(\sum_{k=-K}^{K} \mathcal{N}(x;k\Delta,\sigma)\right)\otimes\mathcal{N}(x;0,\tau)\right] \\
&= \left(\frac{1}{\sum_{k=-K}^{K} e^{-|k|\varepsilon}} \sum_{k=-K}^{K} \mathcal{N}(x;k\Delta,\sigma)\right)\otimes\mathcal{N}(x;0,\tau) \\
&= f_{\mathrm{m}}(x;\sigma,K)\otimes\mathcal{N}(x;0,\tau),
\end{aligned}$$

where the first equality is due to the definition of $f_{\mathrm{m}}$, the second equality is due to the convolution property, the third equality is due to the distributivity property of the convolution operator, the fourth equality is due to the associativity property of the convolution operator with scalar multiplication, and the final equality substitutes the definition of $f_{\mathrm{m}}(x;\sigma,K)$. Thanks to this identity, we can derive

$$\begin{aligned}
\ell(x;\sigma') &= e^{\varepsilon} f_{\mathrm{m}}(x;\sigma',K) - f_{\mathrm{m}}(x+\varphi;\sigma',K) \\
&= e^{\varepsilon} f_{\mathrm{m}}(x;\sigma,K)\otimes\mathcal{N}(x;0,\tau) - f_{\mathrm{m}}(x+\varphi;\sigma,K)\otimes\mathcal{N}(x;0,\tau) \\
&= (e^{\varepsilon} f_{\mathrm{m}}(x;\sigma,K) - f_{\mathrm{m}}(x+\varphi;\sigma,K))\otimes\mathcal{N}(x;0,\tau) \\
&= \ell(x;\sigma)\otimes\mathcal{N}(x;0,\tau),
\end{aligned}$$

which coincides with the statement of this intermediary lemma. □

We now prove Lemma 23.

**Proof of Lemma 23.** For the given $\eta\in(0,1)$, select any $\beta\leq\sqrt{2\pi}\eta\sigma\delta$ and assume a given $\sigma>0$ satisfies (29). Such selection of $\beta$ is also valid for $\sigma'\geq\sigma$. Hence, for any fixed $\varphi\in\{0,\beta,2\beta,\dots,\Delta\}$, if we denote by $\ell(x;\sigma) := e^{\varepsilon} f_{\mathrm{m}}(x;\sigma,K) - f_{\mathrm{m}}(x+\varphi;\sigma,K)$, then we should show that

$$\int_{-\infty}^{\infty} \min\{\ell(x;\sigma'),0\}\mathrm{d}x \geq \int_{-\infty}^{\infty} \min\{\ell(x;\sigma),0\}\mathrm{d}x \tag{49}$$

holds to conclude the proof. The min-term on the left-hand side of inequality (49) satisfies

$$\begin{aligned}\min\{\ell(x;\sigma'),0\} &= \min\left\{\int_{-\infty}^{\infty}\ell(x-t;\sigma,K)\mathcal{N}(t;0,\tau)\mathrm{d}t, 0\right\}\\ &\geq \int_{-\infty}^{\infty}\min\{\ell(x-t;\sigma,K),0\}\mathcal{N}(t;0,\tau)\mathrm{d}t, \qquad (50)\end{aligned}$$

where the equality follows from Lemma 27 as well as the definition of the convolution operator, and the inequality follows from Jensen's inequality applied to the concave function $t \mapsto \min\{t,0\}$ which is applicable since $\mathcal{N}(t;0,\tau)$ is a density function. Hence, the desired inequality is shown as:

$$\begin{aligned}\int_{-\infty}^{\infty}\min\{\ell(x;\sigma'),0\}\mathrm{d}x &\geq \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\min\{\ell(x-t;\sigma,K),0\}\mathcal{N}(t;0,\tau)\mathrm{d}t\mathrm{d}x\\ &= \int_{-\infty}^{\infty}\left[\int_{-\infty}^{\infty}\min\{\ell(x-t;\sigma,K),0\}\mathcal{N}(t;0,\tau)\mathrm{d}x\right]\mathrm{d}t\\ &= \int_{-\infty}^{\infty}\left[\int_{-\infty}^{\infty}\min\{\ell(x;\sigma),0\}\mathcal{N}(t;0,\tau)\mathrm{d}x\right]\mathrm{d}t\\ &= \int_{-\infty}^{\infty}\mathcal{N}(t;0,\tau)\left[\int_{-\infty}^{\infty}\min\{\ell(x;\sigma),0\}\mathrm{d}x\right]\mathrm{d}t\\ &= \underbrace{\left[\int_{-\infty}^{\infty}\mathcal{N}(t;0,\tau)\mathrm{d}t\right]}_{=1}\left[\int_{-\infty}^{\infty}\min\{\ell(x;\sigma),0\}\mathrm{d}x\right]\\ &= \int_{-\infty}^{\infty}\min\{\ell(x;\sigma),0\}\mathrm{d}x.\end{aligned}$$

Here, the inequality follows from (50), the equalities that follow change the order of the integral, apply variable change $x = x - t$, rearrange constants within integrals, note that $\mathcal{N}(t;0,\tau)$ is a density function integrating to 1. The monotonicity proof is complete since we proved the desired inequality (49).

Now, to see that $\sigma = \sigma_{\mathrm{g}}$ satisfies (29), consider the stronger condition

$$\int_{-\infty}^{\infty}\min\{e^{\varepsilon}f_{\mathrm{m}}(x;\sigma,K) - f_{\mathrm{m}}(x+\varphi;\sigma,K),0\}\mathrm{d}x + (1-\eta)\delta \geq 0 \quad \forall\varphi\in[0,\Delta],$$

which coincides with the definition of $(\varepsilon,(1-\eta)\delta)$-DP (*cf.* proof of Theorem 5). A sufficient condition for this is when each mixture density $\frac{1}{\sqrt{2\pi}\sigma}\exp(-(x-\mu)^2/(2\sigma^2))$ of (28) satisfies $(\varepsilon,(1-\eta)\delta)$-DP since approximate DP is closed under mixtures (Selvi et al. 2025, §2.1). Moreover, since approximate DP is also closed under shifting distributions (*cf.* proof of Lemma 24), it is sufficient to show the zero-mean Gaussian density $\frac{1}{\sqrt{2\pi}\sigma}\exp(-x^2/(2\sigma^2))$ satisfies $(\varepsilon, (1-\eta)\delta)$-DP. We can therefore borrow the standard deviation $\sigma_\mathrm{g}$ needed for the Gaussian mechanism (Dwork and Roth 2014, Thm 3.22) or the analytic Gaussian mechanism (Balle and Wang 2018), and conclude that a sufficient condition for (29) is $\sigma = \sigma_\mathrm{g}$. □

### 2.B.5 Proof of Theorem 6

The correctness of Algorithm 6 follows directly from Lemma 23. The first step of this algorithm is to compute the analytic Gaussian mechanism, which sets $\sigma_\mathrm{g}$ in Lemma 23. In light of the monotonicity presented in this lemma, the bisection method finds the smallest $\sigma$ value so that any $\sigma' \geq \sigma$ satisfies condition (29). The computation for checking whether a given $\sigma$ satisfies this condition is given in Algorithm 7, which simply evaluates the left-hand side of the constraint (29) for every $\varphi$ in the grid. Note that this grid is of $\varphi$ is a function of $\beta$. To make sure that *(i)* $\beta \leq \sqrt{2\pi}\eta\sigma\delta$ and also *(ii)* $\beta$ divides $\Delta$, Algorithm 6 sets $\beta$ as $\Delta/\lceil \Delta/(\sqrt{2\pi}\eta\sigma\delta)\rceil$.

For the runtime, similarly to the proof of Theorem 3, we assume that the optimality tolerances are fixed, independent of the $\delta$ value. Note that Algorithm 6 first finds the standard deviation $r = \sigma_\mathrm{g}$ needed by the analytic Gaussian mechanism. The implementation of the analytic Gaussian (Balle and Wang 2018) is via a bisection search method where the root of a sought in a region with an upper bound for $\sigma_\mathrm{g}$ as $\mathcal{O}((\Delta/\varepsilon)\sqrt{\log(\delta^{-1})})$. The function, however, can be evaluated in constant time. Given that Algorithm 6 applies a bisection search within $(0, r)$, the runtime of the rest of this algorithm dominates the computation of the analytic Gaussian. We borrow the upper bound $r = \mathcal{O}((\Delta/\varepsilon)\sqrt{\log(\delta^{-1})})$ for the analysis of our bisection search method next.

Our bisection method is restricted to an interval of radius $\mathcal{O}((\Delta/\varepsilon)\sqrt{\log(\delta^{-1})})$, and the number of iterations is therefore:

$$\mathcal{O}\left(\log\left(\frac{\Delta}{\varepsilon}\sqrt{\log\delta^{-1}}\right)\right). \tag{51}$$

The iterations of this bisection search, as shown in Algorithm 7, computes an integral, in total

$$\mathcal{O}\left(\frac{\Delta}{\eta\delta}\right) \tag{52}$$

times. The computation of this integral has complexity

$$\mathcal{O}\left((K\Delta)\cdot K\right) \tag{53}$$

| $\delta \downarrow \mid \varepsilon \rightarrow$ | 0.1 | 0.25 | 0.5 | 0.75 | 1 | 2 | 3 | 4 | 5 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| 0.0001 | **0.16%** | **-0.59%** | **10.31%** | **37.51%** | **96.51%** | **98.09%** | **95.34%** | **95.89%** | **97.22%** | **95.42%** |
| 0.0005 | **-0.40%** | **-0.33%** | **20.86%** | **95.53%** | **95.13%** | **97.01%** | **94.64%** | **88.29%** | **92.43%** | **95.05%** |
| 0.001 | **-0.62%** | **-0.37%** | **32.53%** | **95.21%** | **95.64%** | **96.81%** | **94.27%** | **87.56%** | **92.02%** | **94.87%** |
| 0.005 | **-1.41%** | **3.16%** | **93.47%** | **93.14%** | **92.16%** | **95.71%** | **92.32%** | **85.40%** | **89.76%** | **94.33%** |
| 0.01 | **-2.30%** | **27.67%** | **88.0%** | **91.71%** | **92.49%** | **94.90%** | **82.22%** | **82.67%** | **88.96%** | **94.04%** |
| 0.02 | **-3.14%** | **85.54%** | **74.29%** | **90.40%** | **84.15%** | **95.75%** | **80.50%** | **81.07%** | **88.18%** | **93.69%** |
| 0.05 | **62.64%** | **73.57%** | **53.51%** | **76.82%** | **68.49%** | **99.27%** | **76.32%** | **79.06%** | **87.23%** | **93.12%** |
| 0.1 | **54.21%** | **38.45%** | **-0.94%** | **77.78%** | **59.67%** | **90.37%** | **76.08%** | **78.29%** | **86.86%** | **92.54%** |
| 0.15 | **18.53%** | **24.45%** | **26.63%** | **-1.87%** | **4.45%** | **96.96%** | **77.97%** | **78.85%** | **87.25%** | **92.11%** |
| 0.25 | **7.69%** | **22.86%** | **15.72%** | **76.32%** | **57.09%** | **70.02%** | **85.53%** | **81.96%** | **88.92%** | **91.41%** |

Table 10: *Optimality gap (% of $l_1$-loss) of the analytic-Gaussian mixture mechanism closed by the multi-Gaussian mixture mechanism.*

because the integral can be approximated on a fixed support of radius $\mathcal{O}(K\Delta)$ since the Gaussian tails vanish infinitely fast, and the evaluation of $f_{\mathrm{m}}(x;\sigma,K)$ takes time $\mathcal{O}(K)$. The multiplication (51) × (52) × (53) concludes the runtime represented in the statement of this theorem.

## 2.C Omitted details in numerical experiments

### 2.C.1 Optimality gaps closed

In the numerical experiments, we compared our mechanisms against the analytic Gaussian mechanism in terms of expected losses of the noise. Here, we focus on the suboptimalities of mechanisms, that is, we investigate optimality gaps compared to the best possible loss any additive noise mechanism can attain under $(\varepsilon,\delta)$-DP. To this end, we use the work of (Selvi et al. 2025), who propose a sequential numerical optimization approach to construct a distribution, where in each iteration the upper bounds (the losses attained by the distribution) and the lower bounds (an associated lower bound that ) provably converge. Since in practice this convergence can take a significant amount of time, we run the algorithm until the upper and lower bounds are almost equal with a tolerance of $10^{-3}$, and then use the lower bound as a tight estimation of the ideal loss one can attain under $(\varepsilon,\delta)$-DP. If we denote by $o$ this estimation, than the analytic Gaussian mechanism has an optimality gap of $100\% \cdot (a-o)/o$. Table 10 reports how much of this optimality gap is closed with the multi-Gaussian mixture mechanism, that is, each entry of this table presents $100\% \cdot (a-m)/(a-o)$. While the mean and median gaps closed are 67.67% (sd 34.78) and 85.47%, respectively, we observe that the multi-Gaussian mixture distribution is closer to the lower bounds in low-privacy regimes.

| $\delta \downarrow \mid \varepsilon \rightarrow$ | 0.1 | 0.25 | 0.5 | 0.75 | 1 | 2 | 3 | 4 | 5 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| 0.0001 | **4** | **10** | **10** | **10** | **9** | **8** | **8**$\star$ | **4**$\star$ | **4**$\star$ | **5** |
| 0.0005 | **3** | **1** | **10** | **9** | **7** | **8** | **8** | **9** | **9**$\star$ | **5**$\star$ |
| 0.001 | **5** | **2** | **10** | **8** | **9** | **8** | **8** | **9** | **9** | **5** |
| 0.005 | **3** | **10** | **8** | **6** | **5** | **8** | **8** | **9** | **9** | **5** |
| 0.01 | **2** | **10** | **7** | **5** | **4** | **8** | **9** | **9** | **9** | **5** |
| 0.02 | **6** | **8** | **6** | **4** | **4** | **8** | **9** | **9**$\star$ | **9** | **5** |
| 0.05 | **7** | **5** | **4** | **3** | **3** | **8**$\star$ | **9** | **9** | **9** | **5** |
| 0.1 | **4** | **3** | **2** | **2**$\star$ | **2**$\star$ | **8** | **9** | **9** | **9** | **5** |
| 0.15 | **2**$\star$ | **2**$\star$ | **2**$\star$ | **2**$\star$ | **2** | **9** | **9** | **9** | **9** | **5** |
| 0.25 | **1** | **1** | **1** | **1** | **1** | **9** | **9** | **9** | **9** | **5** |

Table 11: $K \in [10]$ *values that attain the best* $l_1$*-loss for the multi-Gaussian mixture. The star sign indicates instances where quasi-Gaussian outperforms multi-Gaussian with* $K = 1$.

### 2.C.2 Comparing the quasi-Gaussian mechanism with the multi-Gaussian mechanism

Table 11 reports the values of $K \in \{1, \dots, 10\}$ that yield the smallest $l_1$-loss for the multi-Gaussian mechanism across the 100 combinations of $(\varepsilon, \delta)$ considered. The results show that all values within the grid are being chosen, demonstrating the importance of tuning $K$. Although both the multi-Gaussian mechanism with $K = 1$ and the quasi-Gaussian mechanism produce distributions with three modes, there is no clear hierarchy between them. In fact, entries marked with a red star indicate cases where the quasi-Gaussian mechanism outperforms the multi-Gaussian mechanism with $K = 1$.

Moreover, Sections 2.4.1, 2.4.2, and 2.C.1 may suggest that if we tune $K \in [10]$ (not $K = 1$ fixed), then the multi-Gaussian mixture mechanism consistently outperforms the quasi-Gaussian mechanism. However, this is inaccurate. Table 12 compares the quasi-Gaussian, multi-Gaussian, and the analytic Gaussian mechanisms, simultaneously, on a revised grid $\varepsilon \in \{0.0001, 0.0005, 0.001, 0.005, 0.01, 0.05\}$. We can observe that, for smaller $\varepsilon$ and $\delta$, the quasi-Gaussian mechanism outperforms the multi-Gaussian mechanisms (for all $K \in [10]$).

### 2.C.3 Additional tables for the numerical experiments

Tables 13-17 are the $l_2$-counterparts of Tables 7-12, respectively, and one can draw similar conclusions. Table 18 presents the standard deviations corresponding to the errors displayed in Table 9.

| $\delta \downarrow \mid \varepsilon \rightarrow$ | 0.0001 | 0.0005 | 0.001 | 0.005 | 0.01 | 0.05 |
|---|---|---|---|---|---|---|
| 0.0001 | **0.09%** | **-0.01%** | **-0.01%** | **-0.05%** | **-0.09%** | **-0.38%** |
| 0.0005 | **0.02%** | **0.01%** | **0.01%** | **-0.04%** | **-0.09%** | **-0.43%** |
| 0.001 | **0.04%** | **0.03%** | **0.02%** | **-0.03%** | **-0.09%** | **-0.26%** |
| 0.005 | **0.21%** | **0.20%** | **0.19%** | **0.10%** | **0.02%** | **-0.19%** |
| 0.01 | **0.42%** | **0.41%** | **0.40%** | **0.30%** | **0.21%** | **-0.12%** |
| 0.02 | **-0.09%** | **0.84%** | **0.82%** | **0.71%** | **0.60%** | **0.36%** |
| 0.05 | **17.33%** | **17.07%** | **16.75%** | **14.28%** | **12.79%** | **11.66%** |
| 0.1 | **13.12%** | **13.00%** | **12.84%** | **11.63%** | **10.14%** | **10.59%** |
| 0.15 | **16.65%** | **16.58%** | **16.49%** | **15.75%** | **14.83%** | **8.17%** |
| 0.25 | **1.14%** | **1.24%** | **1.26%** | **1.00%** | **1.27%** | **1.70%** |

Table 12: *Improvement (% of $l_1$-loss) of the quasi-Gaussian (blue) and multi-Gaussian (green) mixture mechanisms over the analytic Gaussian mechanism (red).*

| $\delta \downarrow \mid \varepsilon \rightarrow$ | 0.1 | 0.25 | 0.5 | 0.75 | 1 | 2 | 3 | 4 | 5 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| 0.0001 | **-1.20%** | **-2.62%** | **-4.33%** | **-5.29%** | **-5.58%** | **-1.39%** | **7.23%** | **16.50%** | **25.21%** | **83.49%** |
| 0.0005 | **-1.33%** | **-2.91%** | **-4.81%** | **-5.85%** | **-6.11%** | **-0.86%** | **9.75%** | **21.29%** | **32.47%** | **80.73%** |
| 0.001 | **-1.38%** | **-3.06%** | **-5.05%** | **-6.13%** | **-6.36%** | **-0.49%** | **11.31%** | **24.28%** | **37.14%** | **79.24%** |
| 0.005 | **-1.38%** | **-3.35%** | **-5.65%** | **-6.85%** | **-6.99%** | **1.14%** | **17.26%** | **35.96%** | **59.67%** | **74.71%** |
| 0.01 | **-1.19%** | **-3.37%** | **-5.88%** | **-7.14%** | **-7.22%** | **2.50%** | **21.84%** | **46.18%** | **79.05%** | **72.08%** |
| 0.02 | **-0.67%** | **-3.17%** | **-5.99%** | **-7.35%** | **-7.32%** | **4.77%** | **29.44%** | **74.39%** | **75.75%** | **68.82%** |
| 0.05 | **1.30%** | **-2.03%** | **-5.58%** | **-7.14%** | **-6.82%** | **11.36%** | **62.07%** | **67.11%** | **69.54%** | **63.06%** |
| 0.1 | **4.98%** | **0.52%** | **-4.01%** | **-5.69%** | **-4.67%** | **31.45%** | **50.72%** | **58.54%** | **62.41%** | **56.91%** |
| 0.15 | **8.86%** | **3.54%** | **-1.62%** | **-2.95%** | **-0.43%** | **30.00%** | **40.89%** | **51.34%** | **56.55%** | **52.12%** |
| 0.25 | **-0.53%** | **12.26%** | **9.46%** | **14.04%** | **5.36%** | **2.92%** | **22.28%** | **38.11%** | **46.01%** | **43.98%** |

Table 13: *Improvement (% of $l_2$-loss) of the quasi-Gaussian mixture mechanism (green) over the analytic Gaussian mechanism (red).*

### 2.C.4 Implementation supplements

In addition to our submission, we are submitting implementation supplements that document our Julia codes, provide further mathematical and experimental details behind the proximal coordinate descent experiments, compare our mechanisms with the truncated Laplace mechanism, and provide detailed reports on the runtimes of our algorithms.

| $\delta \downarrow \mid \varepsilon \rightarrow$ | 0.1 | 0.25 | 0.5 | 0.75 | 1 | 2 | 3 | 4 | 5 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| 0.0001 | **0.16%** | **-0.67%** | **11.37%** | **39.88%** | **80.96%** | **87.07%** | **91.81%** | **95.32%** | **97.45%** | **99.72%** |
| 0.0005 | **-0.33%** | **-0.31%** | **19.44%** | **72.53%** | **74.79%** | **83.51%** | **89.79%** | **94.84%** | **97.30%** | **99.67%** |
| 0.001 | **-0.43%** | **-0.31%** | **27.07%** | **68.58%** | **70.94%** | **81.38%** | **88.61%** | **94.29%** | **97.04%** | **99.65%** |
| 0.005 | **-0.58%** | **1.65%** | **50.19%** | **55.34%** | **58.92%** | **73.88%** | **84.62%** | **92.50%** | **96.18%** | **99.57%** |
| 0.01 | **-0.72%** | **11.42%** | **40.01%** | **48.12%** | **53.14%** | **68.78%** | **83.43%** | **91.37%** | **95.65%** | **99.53%** |
| 0.02 | **1.53%** | **26.77%** | **26.20%** | **40.59%** | **39.50%** | **61.66%** | **80.21%** | **89.88%** | **94.97%** | **99.47%** |
| 0.05 | **23.31%** | **21.65%** | **14.70%** | **25.82%** | **22.16%** | **46.75%** | **73.80%** | **87.01%** | **93.68%** | **99.37%** |
| 0.1 | **25.14%** | **15.19%** | **2.24%** | **22.87%** | **15.91%** | **30.38%** | **65.96%** | **83.62%** | **92.21%** | **99.27%** |
| 0.15 | **14.18%** | **11.55%** | **9.57%** | **-0.31%** | **1.57%** | **35.73%** | **59.17%** | **80.78%** | **91.00%** | **99.19%** |
| 0.25 | **4.49%** | **7.72%** | **8.68%** | **20.98%** | **13.65%** | **15.03%** | **46.33%** | **75.55%** | **88.83%** | **99.05%** |

Table 14: *Improvement (% of $l_2$-loss) of the multi-Gaussian mixture mechanism (green) over the analytic Gaussian mechanism (red).*

| $\delta \downarrow \mid \varepsilon \rightarrow$ | 0.1 | 0.25 | 0.5 | 0.75 | 1 | 2 | 3 | 4 | 5 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| 0.0001 | **0.24%** | **-0.90%** | **14.46%** | **49.51%** | **98.57%** | **99.72%** | **98.48%** | **98.13%** | **98.24%** | **99.90%** |
| 0.0005 | **-0.57%** | **-0.48%** | **27.40%** | **97.89%** | **97.77%** | **99.53%** | **98.10%** | **96.8%** | **97.44%** | **99.89%** |
| 0.001 | **-0.86%** | **-0.52%** | **40.72%** | **97.53%** | **97.04%** | **99.61%** | **97.90%** | **96.44%** | **97.18%** | **99.88%** |
| 0.005 | **-1.73%** | **3.77%** | **95.30%** | **94.96%** | **94.22%** | **99.70%** | **97.10%** | **95.35%** | **96.07%** | **99.86%** |
| 0.01 | **-2.50%** | **31.38%** | **87.92%** | **93.30%** | **94.41%** | **98.47%** | **95.23%** | **94.45%** | **95.56%** | **99.84%** |
| 0.02 | **5.62%** | **87.90%** | **68.94%** | **91.95%** | **80.18%** | **95.02%** | **95.17%** | **93.74%** | **94.92%** | **99.83%** |
| 0.05 | **75.69%** | **80.03%** | **48.95%** | **74.65%** | **56.30%** | **81.80%** | **95.78%** | **92.63%** | **93.94%** | **99.83%** |
| 0.1 | **70.63%** | **53.03%** | **8.41%** | **78.71%** | **49.02%** | **62.79%** | **98.85%** | **92.33%** | **93.34%** | **99.85%** |
| 0.15 | **36.97%** | **37.54%** | **36.22%** | **-1.14%** | **5.38%** | **79.27%** | **96.05%** | **93.11%** | **93.33%** | **99.89%** |
| 0.25 | **11.18%** | **37.16%** | **33.56%** | **79.26%** | **51.84%** | **38.72%** | **80.72%** | **96.64%** | **94.30%** | **100.00%** |

Table 15: *Optimality gap (% of $l_2$-loss) of the analytic-Gaussian mixture mechanism closed by the multi-Gaussian mixture mechanism.*

| $\delta \downarrow \mid \varepsilon \rightarrow$ | 0.1 | 0.25 | 0.5 | 0.75 | 1 | 2 | 3 | 4 | 5 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| 0.0001 | 4 | 10 | 10 | 10 | 9 | 8⋆ | 8⋆ | 4⋆ | 4⋆ | 5 |
| 0.0005 | 3 | 1 | 10 | 9 | 7 | 8 | 8 | 9⋆ | 9⋆ | 5⋆ |
| 0.001 | 5 | 2 | 10 | 8 | 7 | 8 | 8 | 9 | 9 | 5⋆ |
| 0.005 | 3 | 10 | 8 | 6 | 5 | 8 | 8 | 9 | 9 | 5 |
| 0.01 | 4 | 10 | 7 | 5 | 4 | 8 | 9 | 9 | 9 | 5 |
| 0.02 | 10 | 8 | 6 | 4 | 4 | 8 | 9 | 9⋆ | 9 | 5 |
| 0.05 | 7 | 5 | 4 | 3 | 3 | 8⋆ | 9 | 9 | 9 | 5 |
| 0.1 | 4⋆ | 3⋆ | 2⋆ | 2⋆ | 2⋆ | 1 | 9 | 9 | 9 | 5 |
| 0.15 | 3 | 2 | 2 | 1 | 1 | 1 | 9 | 9 | 9 | 5 |
| 0.25 | 1⋆ | 1⋆ | 1⋆ | 1⋆ | 1 | 1 | 9 | 9 | 9 | 5 |

Table 16: *$K \in [10]$ values that attain the best $l_2$-loss for the multi-Gaussian mixture. The star sign indicates instances where quasi-Gaussian outperforms multi-Gaussian with $K = 1$.*

| $\delta \downarrow \mid \varepsilon \rightarrow$ | 0.0001 | 0.0005 | 0.001 | 0.005 | 0.01 | 0.05 |
|---|---|---|---|---|---|---|
| 0.0001 | **0.01%** | **-0.01%** | **-0.01%** | **-0.07%** | **-0.14%** | **-0.64%** |
| 0.0005 | **0.05%** | **0.03%** | **0.02%** | **-0.06%** | **-0.14%** | **-0.70%** |
| 0.001 | **0.10%** | **0.08%** | **0.06%** | **-0.03%** | **-0.12%** | **-0.53%** |
| 0.005 | **0.49%** | **0.47%** | **0.45%** | **0.31%** | **0.18%** | **-0.33%** |
| 0.01 | **0.99%** | **0.97%** | **0.95%** | **0.78%** | **0.62%** | **-0.22%** |
| 0.02 | **-0.15%** | **1.95%** | **1.93%** | **1.74%** | **1.54%** | **1.74%** |
| 0.05 | **41.91%** | **41.53%** | **41.06%** | **37.29%** | **32.64%** | **31.23%** |
| 0.1 | **35.74%** | **35.55%** | **35.30%** | **33.38%** | **30.96%** | **27.20%** |
| 0.15 | **40.69%** | **40.58%** | **40.44%** | **39.32%** | **37.91%** | **26.97%** |
| 0.25 | **14.45%** | **14.36%** | **14.26%** | **13.42%** | **12.38%** | **4.91%** |

Table 17: *Improvement (% of $l_2$-loss) of the quasi-Gaussian (blue) and multi-Gaussian (green) mixture mechanisms over the analytic Gaussian mechanism (red).*

| Dataset description | | | In-sample errors | | | | Out-of-sample errors | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Dataset | $N$ | $d$ | A-G | Q-G | M-G | PCD | A-G | Q-G | M-G | PCD |
| post-operative | 86 | 14 | 5.42% | 5.52% | 4.40% | 2.41% | 13.73% | 14.24% | 11.15% | 5.14% |
| adult | 45,222 | 57 | 1.43% | 1.44% | 1.41% | 1.44% | 1.44% | 1.46% | 1.42% | 1.45% |
| breast-cancer | 683 | 26 | 0.80% | 0.82% | 0.79% | 0.75% | 1.59% | 1.58% | 1.49% | 1.44% |
| contraceptive | 1,473 | 18 | 1.62% | 1.55% | 1.62% | 1.62% | 2.60% | 2.64% | 2.65% | 2.62% |
| dermatology | 366 | 98 | 2.83% | 2.79% | 2.51% | 2.52% | 4.48% | 4.77% | 4.25% | 4.36% |
| cylinder-bands | 539 | 63 | 2.37% | 2.29% | 2.21% | 2.20% | 3.62% | 3.74% | 3.66% | 3.46% |
| annealing | 898 | 42 | 2.13% | 2.10% | 2.08% | 2.09% | 2.77% | 2.79% | 2.73% | 2.77% |
| spect | 160 | 23 | 3.77% | 3.90% | 3.38% | 2.79% | 5.04% | 5.21% | 4.07% | 3.23% |
| bank | 45,211 | 44 | 0.57% | 0.48% | 0.66% | 0.50% | 0.55% | 0.48% | 0.66% | 0.49% |
| abalone | 4,177 | 10 | 0.57% | 0.57% | 0.57% | 0.55% | 0.42% | 0.40% | 0.43% | 0.41% |
| spambase | 4,601 | 58 | 1.86% | 1.87% | 1.86% | 1.88% | 1.82% | 1.82% | 1.86% | 1.83% |
| ecoli | 336 | 8 | 2.11% | 2.18% | 1.72% | 1.51% | 2.61% | 2.48% | 1.98% | 1.70% |
| absent | 740 | 70 | 2.57% | 2.59% | 2.49% | 2.54% | 4.14% | 4.15% | 3.97% | 3.95% |
| colon-cancer | 62 | 2,000 | 4.50% | 4.58% | 4.69% | 0.09% | 12.93% | 13.39% | 11.04% | 11.88% |

Table 18: *Standard deviations of the in- and out-of-sample errors of privacy-preserving classifiers under different mechanisms.*

# Chapter II

# Robust Machine Learning

This chapter presents my work on the intersection of robust optimization and machine learning (ML), which develops ML models that are robust against overfitting and adversarial attacks.

In the following, Section 3 is based on the following work (Belbasi et al. 2025):

**Reza Belbasi, Aras Selvi, Wolfram Wiesemann** (2025). It's all in the mix: Wasserstein classification and regression with mixed features. *Under Review.*

- A preliminary version of this work (Selvi et al. 2022a), focusing solely on logistic regression and additionally co-authored by Martin Haugh, appeared in **NeurIPS** (2022).

Section 4 is based on the following work (Selvi et al. 2022a) that I completed during a research internship at JP Morgan AI Research:

**Aras Selvi, Eleonora Kreacic, Mohsen Ghassemi, Vamsi Potluru, Tucker Balch, Manuela Veloso** (2025). Distributionally and adversarially robust logistic regression via intersecting Wasserstein balls. **UAI (oral paper)**.

- 2024 INFORMS Data Mining Society Best Student Paper Competition (Finalist)
- 2024 INFORMS Workshop on Data Science (Accepted)

# 3 It's All in the Mix: Wasserstein Classification and Regression with Mixed Features

## Abstract

*Problem definition:* A key challenge in supervised learning is data scarcity, which can cause prediction models to overfit to the training data and perform poorly out of sample. A contemporary approach to combat overfitting is offered by distributionally robust problem formulations that consider all data-generating distributions close to the empirical distribution derived from historical samples, where 'closeness' is determined by the Wasserstein distance. While such formulations show significant promise in prediction tasks where all input features are continuous, they scale exponentially when discrete features are present.

*Methodology/results:* We demonstrate that distributionally robust mixed-feature classification and regression problems can indeed be solved in polynomial time. Our proof relies on classical ellipsoid method-based solution schemes that do not scale well in practice. To overcome this limitation, we develop a practically efficient (yet, in the worst case, exponential time) cutting plane-based algorithm that admits a polynomial time separation oracle, despite the presence of exponentially many constraints. We compare our method against alternative techniques both theoretically and empirically on standard benchmark instances.

*Managerial implications:* Data-driven operations management problems often involve prediction models with discrete features. We develop and analyze distributionally robust prediction models that faithfully account for the presence of discrete features, and we demonstrate that our models can significantly outperform existing methods that are agnostic to the presence of discrete features, both theoretically and on standard benchmark instances.

## 3.1 Introduction

The recent application of machine learning tools across all areas of operations management has led to a plethora of data-driven and end-to-end approaches that blend predictive models from machine learning with optimization frameworks from operations research and operations management. Notable examples include inventory management (Ban and Rudin 2019, Bertsimas and Kallus 2020), logistics (Bertsimas et al. 2019a, Behrendt et al. 2023) and supply chain management (Glaeser et al. 2019), assortment optimization (Kallus and Udell 2020, Feldman et al. 2022) and revenue management (Ferreira et al. 2016, Alley et al. 2023) as well as healthcare operations (Bertsimas et al. 2016b, Bastani and Bayati 2020, Bertsimas and Pauphilet 2023).

The machine learning algorithms used for prediction are prone to overfitting the available data. Overfitted models perform well on the training data used to calibrate the model, but their performance deteriorates when exposed to new, unseen data. This undesirable effect is amplified if the output of a machine learning model is used as input to a downstream optimization model; this phenomenon is known by different communities as the *Optimizer's Curse* (Smith and Winkler 2006) or the *Error-Maximization Effect of Optimization* (Michaud 1989). Traditionally, overfitting is addressed with regularization techniques that penalize complex models characterized by large and/or dense model parameters (Hastie et al. 2009, Murphy 2022). A contemporary alternative from the robust optimization community frames machine learning problems as Stackelberg leader-follower games where the learner selects a model that performs best against a worst-case data-generating distribution selected by a conceptual adversary ('nature') from a predefined ambiguity set (Ben-Tal et al. 2009, Rahimian and Mehrotra 2022, Bertsimas and den Hertog 2022). We talk about Wasserstein machine learning problems when the ambiguity set constitutes a Wasserstein ball centered around the empirical distribution of the available historical observations (Mohajerin Esfahani and Kuhn 2018, Blanchet and Murthy 2019, Gao and Kleywegt 2016). Over the last few years, Wasserstein machine learning problems have attracted enormous attention in the machine learning and optimization communities; we refer to Kuhn et al. (2019, 2025) for recent reviews of the literature. Interestingly, Wasserstein learning problems admit dual characterizations as regularized learning problems (Shafieezadeh-Abadeh et al. 2015, Shafieezadeh-Abadeh et al. 2019, Gao et al. 2022), and they thus contribute to a deeper understanding of the impact of regularization in machine learning. We note that other classes of ambiguity sets have been explored as well, such as moment ambiguity sets and those based on $\phi$-divergences (such as the Kullback-Leibler divergence). We will not delve into the comparative advantages of different ambiguity sets, and we instead refer the interested reader to the existing literature (see, *e.g.*, Van Parys et al. 2021, Kuhn et al. 2019, 2025 and Lam 2019).

Although Wasserstein formulations of many classical machine learning tasks admit formulations as convex optimization problems, these formulations scale exponentially in the discrete input features. This limitation has, thus far, confined the use of Wasserstein machine learning models primarily to datasets with exclusively continuous features. This constitutes a major restriction in operations management, where estimation problems frequently include discrete features. Recent examples include Qi et al. (2022), who apply Wasserstein-based quantile regression to a bike sharing inventory management problem characterized by numeric and discrete features (*e.g.*, the weather conditions, the hour of the day as well as the day of the week); Samorani et al. (2022), who study appointment scheduling problems where most features are discrete (*e.g.*, the day of the week, the patient's marital status and her insurance type); Li et al. (2023), who detect human trafficking from user review websites (here, the discrete features describe the presence or absence of indicative words and phrases); Chan et al. (2023), who predict the macronutrient content of human milk donations using discrete features such as the infant status (term vs preterm); and Duchi et al. (2023), who enforce fairness in offender recidivism prediction through the use of discrete features such as the offender's race, gender and the existence of prior misdemeanour charges. More broadly, at the time of writing, 240 of the 496 classification and 64 of the 159 regression problems in the popular UCI machine learning repository contain discrete input features (Dua and Graff 2017).

We will demonstrate that naïvely replacing discrete features with unbounded continuous features leads to pathological ambiguity sets in the Wasserstein learning problem whose worst-case distributions lack theoretical appeal and whose resulting prediction models underperform in practice. We also show that replacing discrete features with continuous features that are supported on suitably chosen convex hulls of the discrete supports leads to ambiguity sets that are *equivalent* to those resulting from faithfully accounting for discrete features. Unfortunately, however, tractable reformulations of the Wasserstein learning problem over such bounded continuous features are only available for piece-wise affine convex loss functions, where we will demonstrate empirically that their solution times can be prohibitively long.

This work studies linear Wasserstein classification and regression problems with mixed (continuous and discrete) features from a theoretical, computational and numerical perspective. Our work makes several contributions to the state-of-the-art:

*(i)* From a *theoretical perspective*, we demonstrate that while linear Wasserstein classification and regression with mixed features are inherently NP-hard, a wide range of problems can be solved in polynomial time. Also, contrary to problems with exclusively continuous features, we establish that mixed-feature linear Wasserstein classification and regression do not reduce to regularized problems. On the other hand, we demonstrate that mixed-feature linear Wasserstein classification and regression problems are equivalent to bounded continuous-feature problems with suitably chosen supports.

*(ii)* From a *computational perspective*, we propose a cutting plane scheme that solves progressively refined relaxations of the linear Wasserstein classification and regression problems as convex optimization problems. While our overall scheme is not guaranteed to terminate

in polynomial time, we show that the key step of our algorithm—the identification of the most violated constraint—can be implemented efficiently for broad classes of problems, despite the presence of exponentially many constraints.

*(iii)* From a *numerical perspective*, we show that our cutting plane scheme is substantially faster than an equivalent polynomial-size problem reformulation via bounded continuous features. We also show that our model can perform favorably against classical, regularized and alternative robust problem formulations on standard benchmark instances.

Our work is most closely related to the recent work of Shafieezadeh-Abadeh et al. (2015) and Shafieezadeh-Abadeh et al. (2019). Shafieezadeh-Abadeh et al. (2015) formulate the Wasserstein logistic regression problem as a convex optimization problem, they discuss the out-of-sample guarantees of their model, and they report numerical results on simulated and benchmark instances. Shafieezadeh-Abadeh et al. (2019) extend their previous work to a wider class of Wasserstein classification and regression problems. Both works focus on problems with exclusively continuous features, and their proposed formulations would scale exponentially in any discrete features. In contrast, our work studies linear Wasserstein classification and regression problems with mixed features: we examine the theoretical properties of such problems, we develop a practically efficient solution scheme, and we report numerical results. Our work also relates closely to a recent stream of literature that characterizes Wasserstein learning problems as regularized learning problems (Shafieezadeh-Abadeh et al. 2015, Shafieezadeh-Abadeh et al. 2019, Blanchet et al. 2019, Gao et al. 2022). In particular, we demonstrate that our mixed-feature Wasserstein learning problems do not admit an equivalent representation as regularized learning problems, which forms a notable contrast to the existing findings from the literature.

The present work constitutes a completely revised and substantially expanded version of a conference paper (Selvi et al. 2022a). While that work focuses on logistic regression, the present work studies broad classes of linear Wasserstein classification and regression problems. This expansion necessitates significant adaptations of the proof for the computational complexity of the Wasserstein learning problem (*cf.* Theorem 9 in Section 3.4), an entirely new proof for the absence of regularized problem formulations (*cf.* Theorem 10 in Section 3.4) that applies to any loss function (as opposed to only the log-loss function in Selvi et al. 2022a) and any non-trivial Wasserstein learning instance, as well as a substantially generalized cutting plane scheme (*cf.* Algorithms 9 and 10 as well as Theorem 8 in Section 3.3). We also show that the mixed-feature Wasserstein learning problem is equivalent to a bounded continuous-feature formulation, and we present a considerably augmented set of numerical results that encompass both classification and regression problems (*cf.* Section 3.6).

The remainder of this work is organized as follows. Section 3.2 introduces the mixed-feature linear Wasserstein classification and regression problems of interest, and it develops a unified representation of both problems as convex programs of exponential size. Section 3.3 presents and analyzes our cutting plane solution approach for this exponential-size convex program. Section 3.4 shows that while mixed-feature linear Wasserstein classification and regression problems are generically NP-hard, important special cases can be solved in polynomial time. We also show that mixed-feature linear Wasserstein classification and regression problems do not reduce to regularized problems. Section 3.5 contrasts our mixed-feature model against formulations that replace discrete features with (bounded or unbounded) continuous ones. We report on numerical experiments in Section 3.6, and we offer concluding remarks in Section 3.7. Exponential-size convex reformulations of the mixed-feature linear Wasserstein classification and regression problems, which are unified by our representation in Section 3.2, as well as all proofs are relegated to the e-companion. All datasets and source codes accompanying this work are available open source.[5]

**Notation.** We denote by $\mathbb{R}$ ($\mathbb{R}_+$, $\mathbb{R}_-$) the set of (non-negative, non-positive) real numbers, by $\mathbb{N}$ the set of positive integers, and we define $\mathbb{B} = \{0, 1\}$ as well as $[N] = \{1, \ldots, N\}$ for $N \in \mathbb{N}$. For a proper cone $\mathcal{C} \subseteq \mathbb{R}^n$, we write $\boldsymbol{x} \preccurlyeq_{\mathcal{C}} \boldsymbol{x}'$ and $\boldsymbol{x} \prec_{\mathcal{C}} \boldsymbol{x}'$ to abbreviate $\boldsymbol{x}' - \boldsymbol{x} \in \mathcal{C}$ and $\boldsymbol{x}' - \boldsymbol{x} \in \operatorname{int} \mathcal{C}$, respectively. The dual norm of $\|\cdot\|$ is $\|\boldsymbol{x}\|_* = \sup_{\boldsymbol{x}' \in \mathbb{R}^n} \{\boldsymbol{x}^\top \boldsymbol{x}' \, : \, \|\boldsymbol{x}'\| \leq 1\}$, and the cone dual to a cone $\mathcal{C}$ is $\mathcal{C}^* = \{\boldsymbol{x}' \, : \, \boldsymbol{x}^\top \boldsymbol{x}' \geq 0 \;\; \forall \boldsymbol{x} \in \mathcal{C}\}$. The support function of a set $\mathbb{X} \subseteq \mathbb{R}^n$ is $\mathcal{S}_{\mathbb{X}}(\boldsymbol{x}) = \sup\{\boldsymbol{x}^\top \boldsymbol{x}' \, : \, \boldsymbol{x}' \in \mathbb{X}\}$. For a function $L : \mathbb{X} \to \mathbb{R}$, we define the Lipschitz modulus as $\operatorname{lip}(L) = \sup\{|L(\boldsymbol{x}) - L(\boldsymbol{x}')| \,/\, \|\boldsymbol{x} - \boldsymbol{x}'\| \, : \, \boldsymbol{x}, \boldsymbol{x}' \in \mathbb{X}, \; \boldsymbol{x} \neq \boldsymbol{x}'\}$. The set $\mathcal{P}_0(\Xi)$ contains all probability distributions supported on $\Xi$, and the Dirac distribution $\delta_{\boldsymbol{x}} \in \mathcal{P}_0(\mathbb{R}^n)$ places unit probability mass on $\boldsymbol{x} \in \mathbb{R}^n$. The indicator function $\mathbb{1}[\mathcal{E}]$ attains the value 1 (0) whenever the logical expression $\mathcal{E}$ is (not) satisfied.

## 3.2 Mixed-Feature Wasserstein Classification and Regression

We study learning problems over $N$ data points $\boldsymbol{\xi}^n = (\boldsymbol{x}^n, \boldsymbol{z}^n, y^n) \in \Xi = \mathbb{X} \times \mathbb{Z} \times \mathbb{Y}$, $n \in [N]$, where $\boldsymbol{x}^n$, $\boldsymbol{z}^n$ and $y^n$ represent the numerical features, the discrete features and the output variable, respectively. We assume that the support $\mathbb{X}$ of the numerical features is a closed and convex subset of $\mathbb{R}^{M_\mathrm{x}}$. The support $\mathbb{Z}$ of the $K$ discrete features satisfies $\mathbb{Z} = \mathbb{Z}(k_1) \times \ldots \times \mathbb{Z}(k_K)$, where $k_m \in \mathbb{N} \setminus \{1\}$ denotes the number of values that the $m$-th discrete feature can attain,

[5] Website: https://anonymous.4open.science/r/Wasserstein-Mixed-Features-088D/.

$m \in [K]$, and $\mathbb{Z}(s) = \{\boldsymbol{z} \in \mathbb{B}^{s-1} : \sum_{i \in [s-1]} z_i \leq 1\}$ is the one-hot feature encoding. We let $M_{\mathrm{z}} = \sum_{m \in [K]}(k_m - 1)$ denote the number of coefficients associated with the discrete features. The support $\mathbb{Y}$ of the output variable is $\{-1, +1\}$ for classification and a closed and convex subset of $\mathbb{R}$ for regression problems, respectively. We wish to solve the Wasserstein learning problem

$$\begin{array}{ll} \underset{\boldsymbol{\beta}}{\text{minimize}} & \sup\limits_{\mathbb{Q} \in \mathfrak{B}_\epsilon(\widehat{\mathbb{P}}_N)} \mathbb{E}_{\mathbb{Q}}\left[l_{\boldsymbol{\beta}}(\boldsymbol{x}, \boldsymbol{z}, y)\right] \\ \text{subject to} & \boldsymbol{\beta} = (\beta_0, \boldsymbol{\beta}_{\mathrm{x}}, \boldsymbol{\beta}_{\mathrm{z}}) \in \mathbb{R}^{1+M_{\mathrm{x}}+M_{\mathrm{z}}}, \end{array} \tag{54}$$

where the ambiguity set $\mathfrak{B}_\epsilon(\widehat{\mathbb{P}}_N) = \{\mathbb{Q} \in \mathcal{P}_0(\Xi) : \mathrm{W}(\mathbb{Q}, \widehat{\mathbb{P}}_N) \leq \epsilon\}$ represents the Wasserstein ball of radius $\epsilon > 0$ that is centered at the empirical distribution $\widehat{\mathbb{P}}_N = \frac{1}{N}\sum_{n \in [N]} \delta_{\boldsymbol{\xi}^n}$ placing equal probability mass on the $N$ data points $\boldsymbol{\xi}^n$, $n \in [N]$, as per the following definition.

**Definition 4** (Wasserstein Distance)**.** *The type-1* Wasserstein *(*Kantorovich-Rubinstein*, or* earth mover's*)* distance *between two distributions* $\mathbb{P} \in \mathcal{P}_0(\Xi)$ *and* $\mathbb{Q} \in \mathcal{P}_0(\Xi)$ *is defined as*

$$\mathrm{W}(\mathbb{P}, \mathbb{Q}) := \inf_{\Pi \in \mathcal{P}_0(\Xi^2)} \left\{ \int_{\Xi^2} d(\boldsymbol{\xi}, \boldsymbol{\xi}')\, \Pi(\mathrm{d}\boldsymbol{\xi}, \mathrm{d}\boldsymbol{\xi}') \ : \ \Pi(\mathrm{d}\boldsymbol{\xi}, \Xi) = \mathbb{P}(\mathrm{d}\boldsymbol{\xi}),\, \Pi(\Xi, \mathrm{d}\boldsymbol{\xi}') = \mathbb{Q}(\mathrm{d}\boldsymbol{\xi}') \right\},$$

*where the ground metric $d$ on $\Xi$ satisfies*

$$d(\boldsymbol{\xi}, \boldsymbol{\xi}') = \|\boldsymbol{x} - \boldsymbol{x}'\| + \kappa_{\mathrm{z}} d_{\mathrm{z}}(\boldsymbol{z}, \boldsymbol{z}') + \kappa_{\mathrm{y}} d_{\mathrm{y}}(y, y') \quad \forall \boldsymbol{\xi} = (\boldsymbol{x}, \boldsymbol{z}, y) \in \Xi,\ \boldsymbol{\xi}' = (\boldsymbol{x}', \boldsymbol{z}', y') \in \Xi \tag{55a}$$

*with $\kappa_{\mathrm{z}}, \kappa_{\mathrm{y}} > 0$ as well as, for some $p \geq 1$,*

$$d_{\mathrm{z}}(\boldsymbol{z}, \boldsymbol{z}') = \left( \sum_{m \in [K]} \mathbb{1}[\boldsymbol{z}_m \neq \boldsymbol{z}'_m] \right)^{1/p} \quad \textit{and} \quad d_{\mathrm{y}}(y, y') = \begin{cases} \mathbb{1}[y \neq y'] & \textit{if } \mathbb{Y} = \{-1, +1\}, \\ |y - y'| & \textit{otherwise.} \end{cases} \tag{55b}$$

The loss function $l_{\boldsymbol{\beta}}(\boldsymbol{x}, \boldsymbol{z}, y) : \mathbb{X} \times \mathbb{Z} \times \mathbb{Y} \to \mathbb{R}_+$ in problem (54) satisfies

$$l_{\boldsymbol{\beta}}(\boldsymbol{x}, \boldsymbol{z}, y) = \begin{cases} L(y \cdot [\beta_0 + {\boldsymbol{\beta}_{\mathrm{x}}}^\top \boldsymbol{x} + {\boldsymbol{\beta}_{\mathrm{z}}}^\top \boldsymbol{z}]) & \text{if } \mathbb{Y} = \{-1, +1\}, \\ L(\beta_0 + {\boldsymbol{\beta}_{\mathrm{x}}}^\top \boldsymbol{x} + {\boldsymbol{\beta}_{\mathrm{z}}}^\top \boldsymbol{z} - y) & \text{otherwise,} \end{cases}$$

where $L : \mathbb{R} \to \mathbb{R}_+$ measures the similarity between the prediction $\beta_0 + {\boldsymbol{\beta}_{\mathrm{x}}}^\top \boldsymbol{x} + {\boldsymbol{\beta}_{\mathrm{z}}}^\top \boldsymbol{z}$ and the output $y$. With a slight abuse of terminology, we also refer to $L$ as the loss function whenever the context is clear. For both classification and regression problems, we consider two settings:

(i) $L$ is convex and Lipschitz continuous with Lipschitz modulus $\mathrm{lip}(L)$, $\mathbb{X} = \mathbb{R}^{M_\mathrm{x}}$ and $\mathbb{Y} = \{-1, +1\}$ (for classification problems) or $\mathbb{Y} = \mathbb{R}$ (for regression problems);

(ii) $L$ satisfies $L(e) = \max_{j \in [J]}\{a_j e + b_j\}$, and $\mathbb{X} \subseteq \mathbb{R}^{M_\mathrm{x}}$ and $\mathbb{Y} \subseteq \mathbb{R}$ are closed and convex.

In either case, we assume that $L$ is not constant.

The Wasserstein learning problem (54) offers attractive generalization guarantees. While the classical choice of Wasserstein radii suffers from the curse of dimensionality (Mohajerin Esfahani and Kuhn 2018), recent work has developed asymptotic (Blanchet et al. 2019, Blanchet and Kang 2021) as well as finite sample guarantees (Shafieezadeh-Abadeh et al. 2019, Gao 2023) that apply to Wasserstein radii of the order $\mathcal{O}(1/\sqrt{N})$.

We next review a result that expresses the Wasserstein learning problem (54) as a convex optimization problem.

**Observation 6.** *The Wasserstein learning problem* (54) *admits the equivalent formulation*

$$\begin{array}{lll} \underset{\boldsymbol{\beta}, \lambda, \boldsymbol{s}}{\text{minimize}} & \lambda \epsilon + \dfrac{1}{N} \displaystyle\sum_{n \in [N]} s_n & \\ \text{subject to} & \displaystyle\sup_{(\boldsymbol{x}, y) \in \mathbb{X} \times \mathbb{Y}} \{ l_{\boldsymbol{\beta}}(\boldsymbol{x}, \boldsymbol{z}, y) - \lambda \| \boldsymbol{x} - \boldsymbol{x}^n \| - \lambda \kappa_\mathrm{y} d_\mathrm{y}(y, y^n) \} - \lambda \kappa_\mathrm{z} d_\mathrm{z}(\boldsymbol{z}, \boldsymbol{z}^n) \leq s_n & \\ & \qquad\qquad\qquad\qquad\qquad\qquad \forall n \in [N],\ \forall \boldsymbol{z} \in \mathbb{Z} & \\ & \boldsymbol{\beta} = (\beta_0, \boldsymbol{\beta}_\mathrm{x}, \boldsymbol{\beta}_\mathrm{z}) \in \mathbb{R}^{1 + M_\mathrm{x} + M_\mathrm{z}}, \quad \lambda \in \mathbb{R}_+, \quad \boldsymbol{s} \in \mathbb{R}^N_+. & \end{array} \tag{56}$$

Problem (56) contains embedded maximization problems, and it comprises exponentially many constraints. We next prove the existence of equivalent reformulations of (56) for classification and regression problems, respectively, that do not contain embedded optimization problems. Since the resulting reformulations exhibit an exponential number of constraints, the next section will develop a cutting plane approach to introduce these constraints iteratively.

Our equivalent reformulations of (56) leverage the unified problem representation

$$\begin{array}{lll} \underset{\boldsymbol{\theta}, \boldsymbol{\sigma}}{\text{minimize}} & f_0(\boldsymbol{\theta}, \boldsymbol{\sigma}) & \\ \text{subject to} & f_{ni}(g_{ni}(\boldsymbol{\theta}, \boldsymbol{\xi}^n_{-\mathbf{z}}; \boldsymbol{z})) - h_{ni}(\boldsymbol{\theta}, \boldsymbol{\xi}^n_{-\mathbf{z}}; d_\mathrm{z}(\boldsymbol{z}, \boldsymbol{z}^n)) \leq \sigma_n & \forall n \in [N],\ \forall i \in \mathcal{I},\ \forall \boldsymbol{z} \in \mathbb{Z} \\ & \boldsymbol{\theta} \in \Theta, \quad \boldsymbol{\sigma} \in \mathbb{R}^N_+, & \end{array} \tag{57}$$

where $\boldsymbol{\xi}^n_{-\mathbf{z}} = (\boldsymbol{x}^n, y^n)$ and $\mathcal{I}$ is a finite index set. To ensure that (57) is convex, we stipulate that $f_0 : \mathbb{R}^{M_{\boldsymbol{\theta}}} \times \mathbb{R}^N_+ \to \mathbb{R}$ and $f_{ni} : \mathbb{R} \to \mathbb{R}$ are convex, $g_{ni} : \mathbb{R}^{M_{\boldsymbol{\theta}}} \times (\mathbb{X} \times \mathbb{Y}) \times \mathbb{R}^{M_\mathrm{z}} \to \mathbb{R}$ is bi-affine in $\boldsymbol{\theta}$ and $\boldsymbol{z}$ for every fixed $\boldsymbol{\xi}^n_{-\mathbf{z}}$, $h_{ni} : \mathbb{R}^{M_{\boldsymbol{\theta}}} \times (\mathbb{X} \times \mathbb{Y}) \times \mathbb{R} \to \mathbb{R}$ is concave in $\boldsymbol{\theta}$ for every fixed $\boldsymbol{\xi}^n_{-\mathbf{z}}$

and every fixed value of its last component, $n \in [N]$ and $i \in \mathcal{I}$, and $\Theta \subseteq \mathbb{R}^{M_{\boldsymbol{\theta}}}$ is a convex set. We further assume that $f_0$ is non-decreasing in $\boldsymbol{\sigma}$ so that (57) is not unbounded.

**Theorem 7.** *The following problems admit equivalent reformulations in the form of* (57)*:*

(i) *the mixed-feature Wasserstein classification and regression problems with convex and Lipschitz continuous loss functions $L$ as well as the continuous feature support $\mathbb{X} = \mathbb{R}^{M_\mathrm{x}}$.*

(ii) *the mixed-feature Wasserstein classification and regression problems with piece-wise affine convex loss functions $L(e) = \max_{j \in [J]}\{a_j e + b_j\}$ as well as the continuous feature support*

$$\mathbb{X} = \{\boldsymbol{x} \in \mathbb{R}^{M_\mathrm{x}} \, : \, \boldsymbol{C}\boldsymbol{x} \preccurlyeq_{\mathcal{C}} \boldsymbol{d}\} \text{ for some } \boldsymbol{C} \in \mathbb{R}^{r \times M_\mathrm{x}}, \boldsymbol{d} \in \mathbb{R}^r \\ \text{and proper convex cone } \mathcal{C} \subseteq \mathbb{R}^r$$

*for classification problems as well as the support*

$$\mathbb{X} \times \mathbb{Y} = \left\{(\boldsymbol{x}, y) \in \mathbb{R}^{M_\mathrm{x}+1} \, : \, \boldsymbol{C}_\mathrm{x}\boldsymbol{x} + \boldsymbol{c}_\mathrm{y} \cdot y \preccurlyeq_{\mathcal{C}} \boldsymbol{d}\right\} \text{ for some } \boldsymbol{C}_\mathrm{x} \in \mathbb{R}^{r \times M_\mathrm{x}}, \boldsymbol{c}_\mathrm{y} \in \mathbb{R}^r, \boldsymbol{d} \in \mathbb{R}^r \\ \text{and proper convex cone } \mathcal{C} \subseteq \mathbb{R}^r$$

*for regression problems, assuming that the supports admit Slater points.*

The Appendices A and B in the e-companion prove Theorem 7 for classification and regression problems, respectively. The proofs follow similar arguments as those of Shafieezadeh-Abadeh et al. (2019), adapted to the presence of discrete features as well as our ground metric.

### 3.3 Cutting Plane Solution Scheme

The unified problem representation (57) of the mixed-feature Wasserstein regression and classification problems (*cf.* Theorem 7) is convex, but it comprises constraints whose number scales exponentially in $K$, the number of discrete features. We therefore develop a cutting plane approach that iteratively introduces only those constraints that are most violated by a sequence of incumbent solutions.

Algorithm 9 outlines our cutting plane scheme, which follows the standard methodology of this class of techniques. The algorithm iteratively solves a sequence of relaxations of problem (57), where each relaxation only includes a subset $\mathcal{W}$ of the constraints indexed by $[N] \times \mathcal{I} \times \mathbb{Z}$. Consequently, any optimal solution $(\boldsymbol{\theta}^\star, \boldsymbol{\sigma}^\star)$ obtained from the relaxation gives rise to a valid lower bound for the optimal value of problem (57). At each iteration, the algorithm identifies and incorporates the most violated constraint from problem (57) with respect to the current

```
Algorithm 9: Cutting Plane Scheme for Problem (57)
  Input: (Possibly empty) initial constraint set $\mathcal{W} \subseteq [N] \times \mathcal{I} \times \mathbb{Z}$
  Output: Optimal solution $(\boldsymbol{\theta}^\star, \boldsymbol{\sigma}^\star)$ to problem (57)
  Initialize $(\mathrm{LB}, \mathrm{UB}) = (-\infty, +\infty)$ as lower and upper bounds for problem (57)
  while $\mathrm{LB} < \mathrm{UB}$ do
      Let $(\boldsymbol{\theta}^\star, \boldsymbol{\sigma}^\star)$ be optimal in the relaxation of (57) involving the constraints
        $(n, i, \boldsymbol{z}) \in \mathcal{W}$
      for $n \in [N]$ do
          Identify, for each $i \in \mathcal{I}$, a most violated constraint

              $\boldsymbol{z}(n,i) \in \arg\max_{\boldsymbol{z} \in \mathbb{Z}} \left\{ f_{ni}(g_{ni}(\boldsymbol{\theta}^\star, \boldsymbol{\xi}^n_{-\mathbf{z}}; \boldsymbol{z})) - h_{ni}(\boldsymbol{\theta}^\star, \boldsymbol{\xi}^n_{-\mathbf{z}}; d_{\mathrm{z}}(\boldsymbol{z}, \boldsymbol{z}^n)) - \sigma^\star_n \right\}$

            associated with $(\boldsymbol{\theta}^\star, \boldsymbol{\sigma}^\star)$ and $(n, i)$, and denote the constraint violation by
            $\vartheta(n, i)$
          Let $i(n) \in \arg\max\{\vartheta(n, i) : i \in \mathcal{I}\}$ and add $(n, i(n), \boldsymbol{z}(n, i(n)))$ to $\mathcal{W}$ if
            $\vartheta(n, i(n)) > 0$
      Define $\boldsymbol{\vartheta}^\star \in \mathbb{R}^N$ via $\vartheta^\star_n = \max\{\vartheta(n, i(n)), 0\}$, $n \in [N]$
      Update $\mathrm{LB} = f_0(\boldsymbol{\theta}^\star, \boldsymbol{\sigma}^\star)$ and $\mathrm{UB} = \min\{\mathrm{UB},\ f_0(\boldsymbol{\theta}^\star, \boldsymbol{\sigma}^\star + \boldsymbol{\vartheta}^\star)\}$
```

solution $(\boldsymbol{\theta}^\star, \boldsymbol{\sigma}^\star)$. This procedure ensures the generation of a monotonically non-decreasing sequence of lower bounds, which is guaranteed to converge to the optimal value of problem (57) in a finite number of iterations. A common challenge in cutting plane techniques is the derivation of upper bounds, which allow us to quantify the optimality gap, particularly when the algorithm is terminated prematurely, such as when reaching a predefined time limit. In our case, the special structure of problem (57) enables us to obtain such upper bounds efficiently: We adjust the optimal solution $(\boldsymbol{\theta}^\star, \boldsymbol{\sigma}^\star)$ of the relaxation by increasing the slack variables $\boldsymbol{\sigma}^\star$ so as to account for the maximum constraint violations $\boldsymbol{\vartheta}^\star$.

The next result formalizes our intuition.

**Proposition 9.** *Algorithm 9 terminates in finite time with an optimal solution* $(\boldsymbol{\theta}^\star, \boldsymbol{\sigma}^\star)$ *to problem* (57). *Moreover,* LB *and* UB *constitute monotonic sequences of lower and upper bounds on the optimal value of* (57) *throughout the execution of the algorithm.*

A key step in Algorithm 9 concerns the identification of a constraint $(n, i, \boldsymbol{z}) \in [N] \times \mathcal{I} \times \mathbb{Z}$ that is most violated by the incumbent solution $(\boldsymbol{\theta}^\star, \boldsymbol{\sigma}^\star)$. We next argue that the underlying combinatorial problem can be solved efficiently by Algorithm 10.

Algorithm 10 determines a most violated constraint index, denoted as $\boldsymbol{z}(n, i)$, for a given incumbent solution $(\boldsymbol{\theta}^\star, \boldsymbol{\sigma}^\star)$ and a given constraint group $(n, i) \in [N] \times \mathcal{I}$ in the optimization

**Algorithm 10:** Identification of a Most Violated Constraint in Problem (57)

```
**Input:** Incumbent solution $(\boldsymbol{\theta}^\star, \boldsymbol{\sigma}^\star)$ and constraint group index $(n, i) \in [N] \times \mathcal{I}$.
**Output:** A most violated constraint index $\boldsymbol{z}(n, i)$ in constraint group $(n, i)$.
Initialize the candidate constraint index set $\mathcal{Z} = \emptyset$;
Let $(\boldsymbol{w}, w_0) \in \mathbb{R}^{M_{\mathrm{z}}} \times \mathbb{R}$ be such that $g_{ni}(\boldsymbol{\theta}^\star, \boldsymbol{\xi}^n_{-\mathbf{z}}; \boldsymbol{z}) = \boldsymbol{w}^\top \boldsymbol{z} + w_0$;
**for** $\mu \in \{\pm 1\}$ **do**
    Compute $\boldsymbol{z}_m^\star \in \arg\max\{\mu \cdot {\boldsymbol{w}_m}^\top \boldsymbol{z}_m \, : \, \boldsymbol{z}_m \in \mathbb{Z}(k_m) \setminus \{\boldsymbol{z}_m^n\}\}$ for all $m \in [K]$;
    Compute a permutation $\pi : [K] \rightarrow [K]$ such that
        $\mu \cdot {\boldsymbol{w}_{\pi(m)}}^\top (\boldsymbol{z}^\star_{\pi(m)} - \boldsymbol{z}^n_{\pi(m)}) \;\geq\; \mu \cdot {\boldsymbol{w}_{\pi(m')}}^\top (\boldsymbol{z}^\star_{\pi(m')} - \boldsymbol{z}^n_{\pi(m')}) \qquad \forall 1 \leq m \leq m' \leq K.$
    **for** $\delta \in [K] \cup \{0\}$ **do**
        Add $\boldsymbol{z}(\delta)$ to $\mathcal{Z}$, where $\boldsymbol{z}_m(\delta) = \boldsymbol{z}_m^\star$ if $\pi(m) \leq \delta$; $= \boldsymbol{z}_m^n$ otherwise, $m \in [K]$;
Select $\boldsymbol{z}(n, i) \in \arg\max\{f_{ni}(g_{ni}(\boldsymbol{\theta}^\star, \boldsymbol{\xi}^n_{-\mathbf{z}}; \boldsymbol{z})) - h_{ni}(\boldsymbol{\theta}^\star, \boldsymbol{\xi}^n_{-\mathbf{z}}; d_{\mathrm{z}}(\boldsymbol{z}, \boldsymbol{z}^n)) - \sigma_n^\star \, : \, \boldsymbol{z} \in \mathcal{Z}\}$;
```

problem (57). We obtain such a constraint by maximizing the constraint left-hand side in (57),

$$f_{ni}(g_{ni}(\boldsymbol{\theta}^\star, \boldsymbol{\xi}^n_{-\mathbf{z}}; \boldsymbol{z})) - h_{ni}(\boldsymbol{\theta}^\star, \boldsymbol{\xi}^n_{-\mathbf{z}}; d_{\mathrm{z}}(\boldsymbol{z}, \boldsymbol{z}^n)),$$

over $\boldsymbol{z} \in \mathbb{Z}$. Note that the constraint function comprises two distinct terms. The second term, $h_{ni}(\boldsymbol{\theta}^\star, \boldsymbol{\xi}^n_{-\mathbf{z}}; d_{\mathrm{z}}(\boldsymbol{z}, \boldsymbol{z}^n))$, depends on $\boldsymbol{z}$ only through the distance metric $d_{\mathrm{z}}(\boldsymbol{z}, \boldsymbol{z}^n)$ within the ground metric $d$ (*cf.* Definition 4). Consequently, we can decompose the maximization of the difference of both terms into two steps. In a first step, we maximize the first term $f_{ni}(g_{ni}(\boldsymbol{\theta}^\star, \boldsymbol{\xi}^n_{-\mathbf{z}}; \boldsymbol{z}))$ over all $\boldsymbol{z} \in \mathbb{Z}$ that satisfy $d_{\mathrm{z}}(\boldsymbol{z}, \boldsymbol{z}^n) = \delta$ for any fixed $\delta \in [K] \cup \{0\}$. This is what the inner for-loop in Algorithm 10 accomplishes, and it results in $K+1$ candidate solutions $\boldsymbol{z}(\delta)$. In a second step, we can then select an optimal solution $\boldsymbol{z}(\delta^\star)$ that maximizes

$$f_{ni}(g_{ni}(\boldsymbol{\theta}^\star, \boldsymbol{\xi}^n_{-\mathbf{z}}; \boldsymbol{z}(\delta))) - h_{ni}(\boldsymbol{\theta}^\star, \boldsymbol{\xi}^n_{-\mathbf{z}}; \delta)$$

across all $\delta \in [K] \cup \{0\}$ by evaluating the constraint left-hand side for the $K+1$ candidate solutions $\boldsymbol{z}(\delta)$. This is what the final step in Algorithm 10 implements.

To maximize $f_{ni}(g_{ni}(\boldsymbol{\theta}^\star, \boldsymbol{\xi}^n_{-\mathbf{z}}; \boldsymbol{z}))$ over all $\boldsymbol{z} \in \mathbb{Z}$ satisfying $d_{\mathrm{z}}(\boldsymbol{z}, \boldsymbol{z}^n) = \delta$, note first that $f_{ni}$ is assumed to be univariate and convex. Thus, its maximum over any compact set $\mathcal{X} \subseteq \mathbb{R}$ is attained either at $\min\{x : x \in \mathcal{X}\}$ or $\max\{x : x \in \mathcal{X}\}$. Hence, to maximize $f_{ni}(g_{ni}(\boldsymbol{\theta}^\star, \boldsymbol{\xi}^n_{-\mathbf{z}}; \boldsymbol{z}))$ over all $\boldsymbol{z} \in \mathbb{Z}$ with $d_{\mathrm{z}}(\boldsymbol{z}, \boldsymbol{z}^n) = \delta$, we only need to maximize and minimize the inner function $g_{ni}(\boldsymbol{\theta}^\star, \boldsymbol{\xi}^n_{-\mathbf{z}}; \boldsymbol{z})$ over all $\boldsymbol{z} \in \mathbb{Z}$ with $d_{\mathrm{z}}(\boldsymbol{z}, \boldsymbol{z}^n) = \delta$. The maximization and minimization of $g_{ni}$ is unified by the outer for-loop in Algorithm 10, which maximizes $\mu \cdot g_{ni}(\boldsymbol{\theta}^\star, \boldsymbol{\xi}^n_{-\mathbf{z}}; \boldsymbol{z})$ for $\mu \in \{\pm 1\}$.

We claim that the optimization of $g_{ni}(\boldsymbol{\theta}^\star, \boldsymbol{\xi}^n_{-\mathbf{z}}; \boldsymbol{z})$ over $\boldsymbol{z} \in \mathbb{Z}$ with $d_{\text{z}}(\boldsymbol{z}, \boldsymbol{z}^n) = \delta$ can be decomposed into separate optimizations over $\boldsymbol{z}_m \in \mathbb{Z}(k_m)$, $m \in [K]$. Indeed, note that $g_{ni}$ is affine in $\boldsymbol{z}$ for fixed $\boldsymbol{\theta}^\star$, which allows us to express the mapping $\boldsymbol{z} \mapsto g_{ni}(\boldsymbol{\theta}^\star, \boldsymbol{\xi}^n_{-\mathbf{z}}; \boldsymbol{z})$ as $\boldsymbol{w}^\top \boldsymbol{z} + w_0$ for some $\boldsymbol{w} \in \mathbb{R}^{M_{\text{z}}}$ and $w_0 \in \mathbb{R}$. Moreover, the distance metric $d_{\text{z}}(\boldsymbol{z}, \boldsymbol{z}^n)$ only counts the number of subvectors $m \in [K]$ where $\boldsymbol{z}_m$ differs from $\boldsymbol{z}^n_m$, without considering the magnitude of the difference (as both vectors are one-hot encoded). Finally, we recall that $\mathbb{Z}$ is rectangular, that is, $\mathbb{Z} = \mathbb{Z}(k_1) \times \ldots \times \mathbb{Z}(k_K)$. Taken together, the aforementioned three observations imply that the optimization of $g_{ni}(\boldsymbol{\theta}^\star, \boldsymbol{\xi}^n_{-\mathbf{z}}; \boldsymbol{z})$ over $\boldsymbol{z} \in \mathbb{Z}$ with $d_{\text{z}}(\boldsymbol{z}, \boldsymbol{z}^n) = \delta$ can be broken into two interconnected steps: We need to select the $\delta$ subvectors $m_1 < \ldots < m_\delta$ where $\boldsymbol{z}$ differs from $\boldsymbol{z}^n$, and for each differing subvector $m$, we need to select a solution $\boldsymbol{z}^\star_m$ that optimizes ${\boldsymbol{w}_m}^\top (\boldsymbol{z}_m - \boldsymbol{z}^n_m)$ across all $\boldsymbol{z}_m \neq \boldsymbol{z}^n_m$. To this end, Algorithm 10 first computes for every $m \in [K]$ a solution $\boldsymbol{z}^\star_m$ that optimizes ${\boldsymbol{w}_m}^\top (\boldsymbol{z}_m - \boldsymbol{z}^n_m)$ over $\boldsymbol{z}_m \neq \boldsymbol{z}^n_m$ (first step inside the outer for-loop), it subsequently sorts the contributions ${\boldsymbol{w}_m}^\top (\boldsymbol{z}^\star_m - \boldsymbol{z}^n_m)$, $m \in [K]$ (second step inside the outer for-loop), and it finally selects the $\delta$ largest (smallest) contributions to maximize (minimize) $g_{ni}$ (inner for-loop).

We formalize the above intuition in the next result.

**Theorem 8.** *For a given incumbent solution $(\boldsymbol{\theta}^\star, \boldsymbol{\sigma}^\star)$ and a constraint group index $(n, i) \in [N] \times \mathcal{I}$ in problem (57), Algorithm 10 identifies a most violated constraint index $\mathbf{z}(n, i)$ in time $\mathcal{O}(M_{\mathbf{z}} + KT + K \log K)$, where $T$ is the time required to compute $f_{ni}$, $g_{ni}$ and $h_{ni}$.*

### 3.4 Complexity Analysis

Despite its exponential size, our unified problem representation (57) admits a polynomial time solution for the classes of loss functions that we consider in this work. In fact, it follows from Theorem 8 that Algorithm 10 provides a polynomial time separation oracle for problem (57).

**Theorem 9** (Complexity of the Unified Problem Representation (57))**.** *Assume that the subgradients of the functions $f_{ni}$ and $-h_{ni}$ in problem (57) can be computed in polynomial time and that $\Theta$ is compact, full-dimensional and admits a polynomial time weak separation oracle. Then problem (57) can be solved to $\delta$-accuracy in polynomial time.*

Recall that an optimization problem is solved to $\delta$-accuracy if a $\delta$-suboptimal solution is identified that satisfies all constraints modulo a violation of at most $\delta$. The consideration of $\delta$-accurate solutions is standard in the numerical solution of nonlinear programs where an optimal solution may be irrational. We show in the Appendices A and B that the subgradients of $f_{ni}$

and $-h_{ni}$ can be computed in polynomial time and that $\Theta$ can be assumed to be compact in the Wasserstein classification and regression problems that we consider. The regularity assumptions imposed on problem (57) are crucial in this regard. In fact, it follows from Theorem 2 of Selvi et al. (2022a) that if the functions $f_{ni}$ in problem (57) were allowed to be non-convex, then the problem would be strongly NP-hard even if $M_{\boldsymbol{\theta}} = 0$, $N = |\mathcal{I}| = 1$ and $h_{11}$ vanished.

While Theorem 9 principally allows us to solve problem (57) in polynomial time, the available solution algorithms (see, *e.g.*, Lee et al. 2015) rely on variants of the ellipsoid method and are not competitive on practical problem instances. For this reason, we rely on our cutting plane algorithm throughout our numerical experiments.

A by now classical result shows that when $K = 0$ (absence of discrete features) and $\mathbb{X} = \mathbb{R}^{M_\mathrm{x}}$, the Wasserstein learning problem (54) reduces to a classical learning problem with an additional regularization term in the objective function whenever the output weight $\kappa_\mathrm{y}$ in Definition 4 approaches $\infty$ (Shafieezadeh-Abadeh et al. 2015, Shafieezadeh-Abadeh et al. 2019, Blanchet et al. 2019, Gao et al. 2022). It turns out that this reduction no longer holds when discrete features are present.

**Theorem 10** (Absence of Regularizers)**.** *Fix any convex Lipschitz continuous or piece-wise affine convex loss function $L$ that is not constant, any Wasserstein classification or regression instance (*cf.* Theorem 7 as well as Appendices A and B) with any number $N \geq 1$ training samples, non-zero numbers $M_\mathrm{x}$ of continuous and $K$ of discrete features, $\mathbb{X} = \mathbb{R}^{M_\mathrm{x}}$, as well as any $\kappa_\mathrm{z} > 0$ and any $p \geq 1$ in the definition of the Wasserstein ground metric. For sufficiently large radii $\epsilon$ of the ambiguity set, the objective function*

$$\sup_{\mathbb{Q} \in \mathfrak{B}_\epsilon(\widehat{\mathbb{P}}_N)} \quad \mathbb{E}_{\mathbb{Q}} \left[ l_{\boldsymbol{\beta}}(\boldsymbol{x}, \boldsymbol{z}, y) \right],$$

*of the Wasserstein learning problem does not admit an equivalent reformulation as a classical regularized learning problem,*

$$\mathbb{E}_{\widehat{\mathbb{P}}_N} \left[ l_{\boldsymbol{\beta}}(\boldsymbol{x}, \boldsymbol{z}, y) \right] + \mathfrak{R}(\boldsymbol{\beta}) \qquad \textit{for any } \mathfrak{R} : \mathbb{R}^{1+M_\mathrm{x}+M_\mathrm{z}} \to \mathbb{R},$$

*even when the weight $\kappa_\mathrm{y}$ of the output distance $d_\mathrm{y}$ approaches $\infty$.*

We emphasize that Theorem 10 applies to *any* loss function $L$ and *any* regualrizer $\mathfrak{R}$. We are not aware of any prior results of this form in the literature.

## 3.5 Comparison with Continuous-Feature Formulations

We next contrast the mixed-feature Wasserstein learning problem (56) with an *unbounded* continuous-feature formulation that replaces the support $\mathbb{Z}$ of the discrete features $\boldsymbol{z}$ with $\mathbb{R}^{M_{\mathrm{z}}}$ (Section 3.5.1), as well as with a *bounded* continuous-feature formulation that replaces $\mathbb{Z}$ with its convex hull $\mathrm{conv}(\mathbb{Z})$ (Section 3.5.2). This section focuses on a qualitative comparison of the three formulations; a quantitative comparison in terms of runtimes and generalization errors on benchmark instances is relegated to our numerical experiments (Section 3.6).

### 3.5.1 Comparison with Unbounded Continuous-Feature Formulation

Our reformulation (56) of the mixed-feature Wasserstein learning problem scales exponentially in the discrete features. It may thus be tempting to replace the support $\mathbb{Z}$ with $\mathbb{R}^{M_{\mathrm{z}}}$, which would allow us to solve problem (56) in polynomial time using the reformulations proposed by Shafieezadeh-Abadeh et al. (2015) and Shafieezadeh-Abadeh et al. (2019). We next present a stylized example to illustrate that disregarding the discrete nature of the features $\boldsymbol{z}$ can result in pathological worst-case distributions that lack relevance to the problem.

Consider a classification problem where the single feature is binary and follows the Bernoulli distribution $z \sim \frac{1}{2}\delta_{-1} + \frac{1}{2}\delta_1$, and where the output variable $y \in \{-1, +1\}$ is related to the feature $z$ via $\mathrm{Prob}(y = z \mid z) = 0.8$. Note that in slight deviation from our earlier convention, the binary feature is supported on $\{-1, +1\}$ instead of $\{0, 1\}$; this allows us to present the example without the use of an intercept $\beta_0$, which in turn will simplify our exposition. We set the label mismatch cost to $\kappa_{\mathrm{y}} = 1$, and we use a non-smooth Hinge loss function. We generate random datasets comprising $N = 10$ samples, and we compare the following three formulations:

*(i)* *Empirical risk model.* This model replaces the worst-case expectation in the Wasserstein learning problem (54) with an expectation over the empirical distribution $\widehat{\mathbb{P}}_N$.

*(ii)* *Mixed-feature Wasserstein model.* We solve the Wasserstein learning problem (56) with the discrete support $\mathbb{Z} = \{-1, +1\}$.

*(iii)* *Unbounded continuous-feature Wasserstein model.* We solve the Wasserstein learning problem (56) with the unbounded continuous support $\mathbb{Z} = \mathbb{R}$.

For the fixed hypothesis $\beta_{\mathrm{z}} = 1$ and across 10,000 randomly generated datasets, the mean expected Hinge loss of the empirical risk model evaluates to 0.41. The worst-case expected Hinge

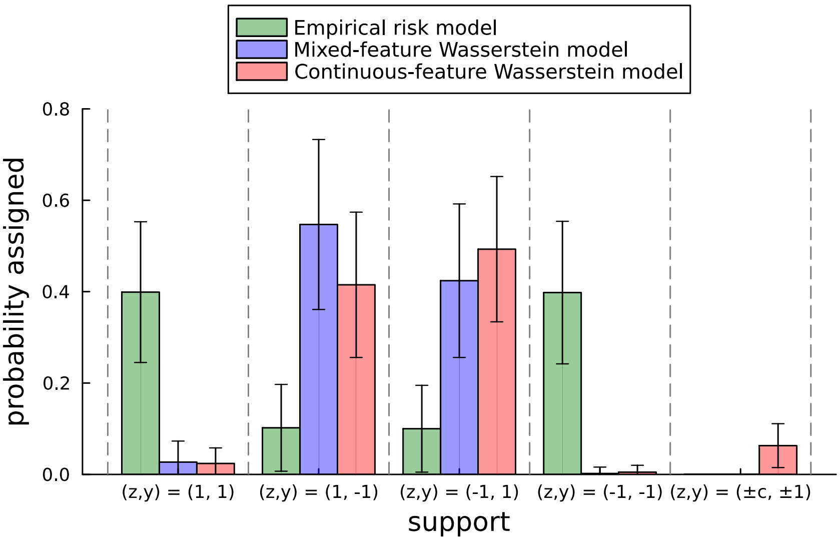


Figure 14: *(Worst-case) distributions for our three formulations. The bars represent the mean probabilities, and the whiskers indicate the corresponding standard deviations, computed over 10,000 statistically independent simulations.*

loss of the mixed-feature Wasserstein learning problem varies from 0.41 to 2; the largest worst-case expected loss is attained for large Wasserstein radii $\epsilon$, where the worst-case distributions approach any mixture of $\delta_{(+1,-1)}$ and $\delta_{(-1,+1)}$. The worst-case expected Hinge loss of the unbounded continuous-feature Wasserstein learning problem, on the other hand, diverges as the Wasserstein radii $\epsilon$ increase. In fact, for sufficiently large $\epsilon$ the worst-case distributions place a probability mass of 0.1 on a non-sensical atom $(\pm c, \pm 1)$, where $c$ increases with $\epsilon$.

Figure 14 illustrates the (worst-case) distributions of the three formulations for an intermediate Wasserstein radius of $\epsilon = 0.85$. As can be seen from the figure, the bounded continuous-feature Wasserstein learning problem places a mean probability mass of 0.063 (std. dev. 0.048) on an atom $(\pm c, \pm 1)$, where $c$ is instance-specific and where $|c|$ lies in $[1.5, 7.5]$ (mean 2.29).

One might argue that an illogical worst-case distribution is less concerning in practical applications, where the primary role of Wasserstein learning may be viewed as providing a regularizing effect through the worst-case formulation. However, our numerical results will demonstrate that the unbounded continuous-feature formulation is consistently outperformed by our mixed-feature formulation in terms of generalization error, even when the Wasserstein radii $\epsilon$ are chosen via cross-validation.

### 3.5.2 Equivalence to Bounded Continuous-Feature Formulation

It is natural to ask whether the mixed-feature Wasserstein learning problem (54) is equivalent to its *bounded* continuous-feature counterpart that replaces the support $\mathbb{Z}$ of the discrete features with its convex hull, $\text{conv}(\mathbb{Z})$. We show in the following that this is indeed the case, but we argue that this equivalence bears little computational consequence.

To construct the bounded continuous-feature model, fix any $s \geq 2$ and define $\mathbb{U}(s)$ as the convex hull of the one-hot discrete feature encoding $\mathbb{Z}(s)$. One readily observes that

$$\mathbb{U}(s) \quad = \quad \left\{ \boldsymbol{u} \in [0,1]^{s-1} : \sum_{i \in [s-1]} u_i \leq 1 \right\}.$$

Indeed, we have $\mathbb{Z}(s) \subseteq \mathbb{U}(s)$ by construction. To see that $\mathbb{U}(s) \subseteq \text{conv}(\mathbb{Z}(s))$, on the other hand, we note that any $\boldsymbol{u} \in \mathbb{U}(s)$ can be represented as $\boldsymbol{u} = \lambda_0 \cdot \mathbf{0} + \sum_{i \in [s-1]} \lambda_i \cdot \mathbf{e}_i$, where $\mathbf{0} \in \mathbb{Z}(s)$ is the vector of all zeros and where $\mathbf{e}_i \in \mathbb{Z}(s)$ is the $i$-th canonic basis vector in $\mathbb{R}^{s-1}$, with $\lambda_i = u_i$ and $\lambda_0 = 1 - \sum_{i \in [s-1]} \lambda_i$. Defining $\mathbb{U}$ as the convex hull of $\mathbb{Z}$, we observe that

$$\mathbb{U} \quad = \quad \text{conv}(\mathbb{Z}) \quad = \quad \bigtimes_{m \in [K]} \text{conv}(\mathbb{Z}(k_m)) \quad = \quad \bigtimes_{m \in [K]} \text{conv}(\mathbb{U}(k_m)),$$

where the second identity follows from Exercise 1.19 of Bertsekas (2009). Since the ground metric $d_{\text{z}}$ from Definition 4 is only suitable for pairs of discrete feature vectors, we replace it with the ground metric

$$d_{\text{u}}(\boldsymbol{u}, \boldsymbol{u}') = \left( \sum_{m \in [K]} \frac{1}{2} \|\boldsymbol{u}_m - \boldsymbol{u}'_m\|_1 + \frac{1}{2} |\mathbf{1}^\top (\boldsymbol{u}_m - \boldsymbol{u}'_m)| \right)^{1/p}$$

that extends to pairs of continuous feature vectors. One readily verifies that $d_{\text{u}}$ is a metric, that is, it satisfies the identity of indiscernibles, positivity, symmetry and the triangle inequality. The proof of Proposition 10 below will also show that $d_{\text{z}}$ and $d_{\text{u}}$ coincide over all pairs of discrete features. We finally define the bounded continuous-feature model as

$$\begin{array}{lll} \underset{\boldsymbol{\beta},\lambda,\boldsymbol{s}}{\text{minimize}} & \lambda\epsilon + \dfrac{1}{N}\displaystyle\sum_{n\in[N]} s_n & \\ \text{subject to} & \displaystyle\sup_{(\boldsymbol{x},y)\in\mathbb{X}\times\mathbb{Y}} \{l_{\boldsymbol{\beta}}(\boldsymbol{x},\boldsymbol{u},y) - \lambda\|\boldsymbol{x}-\boldsymbol{x}^n\| - \lambda\kappa_{\text{y}} d_{\text{y}}(y,y^n)\} - \lambda\kappa_{\text{z}} d_{\text{u}}(\boldsymbol{u},\boldsymbol{z}^n) \leq s_n & \\ & \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \forall n\in[N],\ \forall \boldsymbol{u}\in\mathbb{U} & \\ & \boldsymbol{\beta} = (\beta_0,\boldsymbol{\beta}_{\text{x}},\boldsymbol{\beta}_{\text{z}}) \in \mathbb{R}^{1+M_{\text{x}}+M_{\text{z}}},\ \ \lambda\in\mathbb{R}_+,\ \ \boldsymbol{s}\in\mathbb{R}^N_+, & \end{array} \tag{56'}$$

where $\boldsymbol{u}$ and $\mathbb{U}$ serve as the continuous representations of the discrete features $\boldsymbol{z}$ and their support $\mathbb{Z}$, respectively. Note that in this formulation, the features $\boldsymbol{u}\in\mathbb{U}$ can be readily absorbed in the continuous feature set $\boldsymbol{x}\in\mathbb{X}$; we keep the distinction for improved clarity.

We next state the main result of this section.

**Proposition 10.** *The mixed-feature Wasserstein learning problem* (56) *and its corresponding bounded continuous-feature counterpart* (56') *share the same optimal value and the same set of optimal solutions.*

The proof of Proposition 10 crucially relies on the concavity of the family of mappings $\boldsymbol{u}\mapsto d_{\text{u}}(\boldsymbol{u},\boldsymbol{u}')$, $\boldsymbol{u}'\in\mathbb{Z}$. Indeed, the equivalence stated in the proposition would *not* hold if we were to replace $d_{\text{u}}$ with $d'_{\text{u}}(\boldsymbol{u},\boldsymbol{u}') = \left(\sum_{m\in[K]}\|\boldsymbol{u}_m - \boldsymbol{u}'_m\|_\infty\right)^{1/p}$, despite $d'_{\text{u}}$ being a metric on $\mathbb{U}$.

Unfortunately, the equivalence of Proposition 10 does not lead to more efficient solution schemes for the mixed-feature Wasserstein learning problem (56). Indeed, we are not aware of any finite-dimensional convex reformulations of the bounded continuous-feature counterpart (56') when the loss function $L$ is convex and Lipschitz continuous. When $L$ is piece-wise affine and convex, on the other hand, problem (56') indeed admits a finite-dimensional convex reformulation of polynomial size. We will see in our numerical experiments, however, that our cutting plane scheme from Section 3.3 is much faster than the monolithic solution of this reformulation. This is due to the fact that the reformulation comprises a large number of decision variables as well as a constraint set that lacks desirable sparsity patterns.

### 3.6 Numerical Results

We compare empirically the performance of our mixed-feature Wasserstein learning problem (54) with classical and regularized learning methods as well as continuous-feature formulations of problem (54). To this end, Section 3.6.1 compares the out-of-sample losses of our mixed-feature

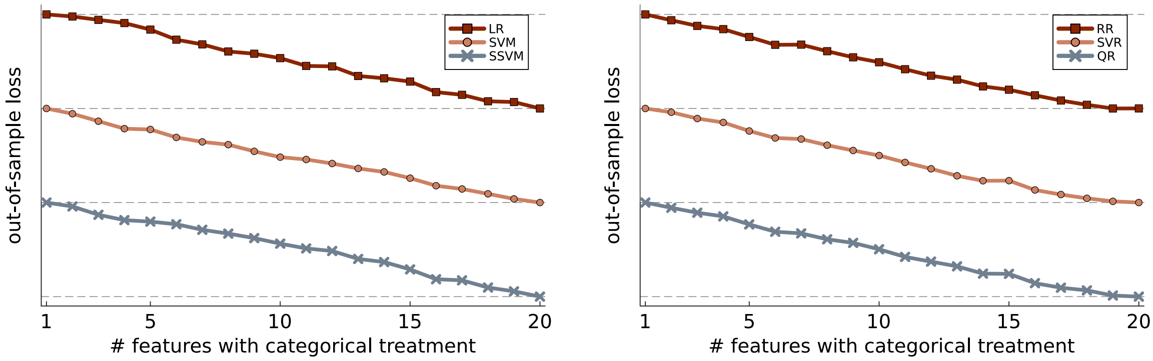


Figure 15: *Mean out-of-sample losses for various classification (left) and regression (right) tasks when the number of discrete features that are treated as such varies. All losses are scaled to* $[0, 1]$ *and shifted so that the curves do not overlap.*

Wasserstein learning problem (54) with those of the unbounded continuous-feature approximation when the Wasserstein radius is selected via cross-validation. Section 3.6.2 compares the runtime of our cutting plane solution scheme for the mixed-feature Wasserstein learning problem with that of solving the equivalent bounded continuous-feature formulation monolithically. Finally, Section 3.6.3 compares the out-of-sample performance of our formulation (54) with that of alternative methods on standard benchmark instances.

All algorithms were implemented in Julia v1.9.2 using the JuMP package (Lubin et al. 2023) and MOSEK v10.0, and all experiments were run on Intel Xeon 2.66GHz cluster nodes with 8GB memory in single-core and single-thread mode (unless otherwise specified). All implementations, datasets and experimental results are available on the GitHub repository accompanying this work.

### 3.6.1 Comparison with Unbounded Continuous-Feature Formulation

Section 3.5.1 demonstrated that modeling discrete features in the Wasserstein learning problem (54) as continuous and subsequently solving an unbounded continuous-feature formulation may inadvertently hedge against pathological worst-case distributions.

We now explore whether the findings from Section 3.5.1 have implications on the out-of-sample performance of the mixed-feature and unbounded continuous-feature formulations when the Wasserstein radius $\epsilon$ is selected via cross-validation. We generate synthetic problem instances to be in full control of the problem dimensions. In particular, we generate 100 synthetic instances for each of the 6 loss functions considered in Appendices A and B: log-loss (LR), non-smooth Hinge Loss (SVM), smooth Hinge loss (SSVM), Huber loss (RR), $\tau$-insensitive loss

(SVR) and pinball loss (QR), assuming that the loss functions explain a large part (but not all) of the variability in the data. All instances comprise $N = 20$ data points, no numerical features, and $K = 20$ binary features. The small number of data points ensures that distributional robustness is required to obtain small out-of-sample losses. For each instance, we solve a mixed-feature Wasserstein learning problem that treats some of the features as binary, whereas the other features are treated as continuous and unbounded. In particular, the extreme cases of modeling all and none of the features as binary correspond to our mixed-feature Wasserstein learning problem (54) and its unbounded continuous-feature approximation, respectively. We cross-validate the Wasserstein radius from the set $\epsilon \in \{0.005, 0.01, 0.015, \ldots, 0.1\}$, and we choose $(\kappa_{\mathrm{z}}, \kappa_{\mathrm{y}}, p) = (1, 1, 1)$ as well as $\|\cdot\| = \|\cdot\|_1$ in our ground metric (55). We refer to the GitHub repository for further details of the instance generation procedure. Figure 15 reports the average out-of-sample losses across 100 randomly generated problem instances for the different loss functions. The figure reveals an overall trend of improved results when the number of binary features that are treated as such increases. Qualitatively, we observe that this conclusion is robust to different choices of the $\epsilon$-grid used for cross-validation; we refer to the GitHub repository for further details.

### 3.6.2 Comparison with Bounded Continuous-Feature Formulation

In Section 3.5.2, we showed that our mixed-feature Wasserstein learning problem (54) is equivalent to a bounded continuous-feature model that replaces the support $\mathbb{Z}$ of the discrete features with its convex hull $\mathrm{conv}(\mathbb{Z})$. In this section, we compare the runtimes of solving the mixed-feature formulation via our cutting plane method with those of solving the bounded continuous-feature formulation via an off-the-shelf solver.

When the loss function $L$ in the mixed-feature Wasserstein learning problem (54) is piecewise affine and convex, the equivalent bounded continuous-feature model admits a reformulation as a polynomial-size convex optimization problem. However, solving this reformulation is computationally expensive since it involves $\mathcal{O}(N \cdot J \cdot K)$ decision variables, where $N$, $J$ and $K$ denote the numbers of data points, affine pieces in the loss function and discrete features, respectively. Moreover, the reformulation lacks desirable sparsity features. The consequences are illustrated in Table 19, which compares the monolithic solution of the bounded continuous-feature reformulation with the solution of the mixed-feature model using our cutting plane approach. The table reports median runtimes over 100 datasets with $N \in \{500, 1{,}000, 2{,}500\}$ data points. All instances contain 30 features, out of which 25%, 50%, 75% are discrete with 4 possible values, whereas the rest are continuous. We set the Wasserstein radius $\epsilon = 10^{-2}$ and use $(\kappa_{\text{z}}, \kappa_{\text{y}}, p) = (1, K, 1)$ for our ground metric. Further details of the implementation are available on the GitHub repository. We observe that our cutting plane scheme is significantly faster than the solution of the equivalent bounded continuous-feature model.

| | **N = 500** | | | **N = 1,000** | | | **N = 2,500** | | |
|---|---|---|---|---|---|---|---|---|---|
| **Method** | %25 | %50 | %75 | %25 | %50 | %75 | %25 | %50 | %75 |
| SVM-cut | 0.05 | 0.20 | 0.25 | 0.13 | 0.21 | 0.45 | 0.29 | 0.61 | 1.23 |
| SVM-piece | 6.35 | 11.80 | 15.25 | 12.48 | 22.99 | 40.97 | 30.4 | 109.78 | 146.62 |
| QR-cut | 0.15 | 0.24 | 0.36 | 0.28 | 0.57 | 1.18 | 1.08 | 1.64 | 3.08 |
| QR-piece | 3.35 | 5.59 | 9.24 | 8.71 | 19.95 | 26.1 | 22.83 | 48.34 | 114.19 |
| SVR-cut | 0.12 | 0.22 | 0.24 | 0.26 | 0.44 | 0.55 | 0.76 | 0.92 | 2.73 |
| SVR-piece | 4.35 | 9.20 | 17.29 | 10.2 | 16.78 | 35.46 | 27.13 | 52.02 | 95.08 |

Table 19: *Median runtimes in secs to solve the mixed-feature model (-cut; using our cutting plane method) and the bounded continuous-feature model (-piece; solving a polynomial-size convex reformulation) for the hinge (SVM), pinball (QR; with* $\tau = 0.5$*) and* $\tau$*-insensitive (SVR; with* $\tau = 10^{-2}$*) loss functions.*

Although no polynomial-size convex reformulations are known for the equivalent bounded continuous-feature model when the loss function $L$ in the mixed-feature Wasserstein learning problem (54) is Lipschitz continuous, we can approximate such loss functions to any desired accuracy via piece-wise affine convex functions. To this end, we take the pointwise maximum of tangent lines at equally spaced breakpoints and subsequently solve the bounded continuous-feature formulation for piece-wise affine convex loss functions. This provides a lower bound on the original problem, whose value approaches the optimal value of the original problem as the number of affine pieces in the approximated loss function increases. Table 20 reports the median runtimes for 100 datasets; we use the same problem parameters as in our previous experiment, and we report results for loss function approximations with $J \in \{5, 10, 25\}$ affine pieces. Similar to our previous experiments, the table shows that our cutting plane scheme is significantly faster than the solution of the equivalent bounded continuous-feature model. We note that the approximated loss functions would require more than 25 pieces to guarantee a suboptimality of 0.1% or less; details are relegated to the GitHub repository.

|  | **N = 500** |  |  | **N = 1,000** |  |  | **N = 2,500** |  |  |
|---|---|---|---|---|---|---|---|---|---|
| **Method** | %25 | %50 | %75 | %25 | %50 | %75 | %25 | %50 | %75 |
| LR-cut | 0.16 | 0.61 | 1.32 | 0.22 | 0.60 | 1.48 | 0.56 | 0.67 | 1.87 |
| LR-piece [5] | 42.08 | 108.0 | 157.06 | 127.97 | 382.6 | 608.76 | 743.01 | 2,371.29* | **NaN** |
| LR-piece [10] | 50.32 | 133.0 | 186.56 | 482.51 | 236.94* | 339.82* | 3,107.19* | **NaN** | 2,595.02* |
| LR-piece [25] | 526.72 | 280.11 | 499.72 | 2,178.79* | 1,634.96* | 1,137.18* | **NaN** | **NaN** | **NaN** |
| SSVM-cut | 0.14 | 0.31 | 0.54 | 0.22 | 0.46 | 0.63 | 0.6 | 0.61 | 1.57 |
| SSVM-piece [5] | 20.54 | 59.6 | 84.11 | 44.43 | 125.33 | 201.43 | 460.29 | 349.74 | 612.18* |
| SSVM-piece [10] | 34.85 | 80.19 | 114.63 | 82.52 | 166.36 | 277.17* | 1,924.21* | 439.99* | 730.46* |
| SSVM-piece [25] | 90.25 | 198.32* | 332.86* | 221.74* | 406.68 | 763.53* | **NaN** | 1,161.48* | 1,996.73* |
| RR-cut | 0.16 | 0.19 | 0.3 | 0.27 | 0.3 | 0.52 | 0.7 | 0.78 | 1.31 |
| RR-piece [5] | 9.87 | 23.23 | 38.57 | 23.13 | 55.95 | 87.75 | 73.02 | 171.16 | 269.98 |
| RR-piece [10] | 16.52 | 38.92 | 60.38 | 37.69 | 87.57 | 143.39 | 105.4 | 281.6 | 464.49 |
| RR-piece [25] | 43.16 | 95.53 | 155.26 | 93.21 | 222.79 | 373.44* | 296.18* | 824.93* | 1,584.85* |

Table 20: *Median runtimes in secs to solve the mixed-feature model (-cut; using our cutting plane method) and the bounded continuous-feature model (-piece [J]; solving a polynomial-size convex reformulation) for the logistic (LR), smooth hinge (SSVM) and Huber (RR; with $\delta = 0.05$) loss functions. A* ***NaN*** *indicates that more than half of the instances could not be solved within an hour of runtime. An asterisk (*) indicates that at least one instance could not be solved within an hour of runtime.*

### 3.6.3 Performance on Benchmark Instances

In Section 3.6.1, we observed that the mixed-feature Wasserstein learning problem (54) can outperform its unbounded continuous-feature approximation in terms of out-of-sample losses. In practice, however, loss functions merely serve as surrogates for the misclassification rate (in classification problems) or mean squared error (in regression problems). Moreover, Section 3.6.1 considered synthetically generated problem instances, which may lack some of the intricate structure of real-life datasets. This section therefore compares the out-of-sample misclassification rates and mean squared errors of problem (54) with those of alternative methods on standard benchmark instances. In particular, we selected 8 of the most popular classification and 7 of the most popular regression datasets from the UCI machine learning repository (Dua and Graff 2017). We processed the datasets to *(i)* handle missing values and inconsistencies in the data, *(ii)* employ a one-hot encoding for discrete variables, and *(iii)* convert the output variable (to a binary value for classification tasks and a $[-1, 1]$-interval for regression tasks). All datasets, processing scripts as well as further results on other UCI datasets can be found in the GitHub repository.

Tables 21–26 report averaged results over 100 random splits into 80% training set and 20% test set for both the original datasets as well as parsimonious variants where only half of the data points are available. In the tables, the column groups report (from left to right) the problem instances' names and dimensions; the results for the unregularized implementations of the nominal problem as well as the mixed-feature Wasserstein learning problem (54) using $(\kappa_{\mathrm{z}}, p) = (1, 1)$, $\|\cdot\| = \|\cdot\|_1$ and $\kappa_{\mathrm{y}} \in \{1, K\}$ in our ground metric (55); the results for the corresponding $l_2$-regularized versions; and the results for two unbounded continuous-feature approximations of problem (54). For the unregularized mixed-feature and continuous-feature Wasserstein learning problems, we cross-validate the hyperparameter $\epsilon$ from the set $\{0, 10^{-5}, 10^{-3}, 10^{-1}\}$. For the regularized nominal model, we cross-validate the regularization penalty $\alpha$ from the set $\{0\} \cup \{c \cdot 10^{-p} : c \in \{1, 5\},\ p = 1, \ldots, 6\}$. For the regularized mixed-feature Wasserstein learning problem, finally, we cross-validate the hyperparameters $(\epsilon, \alpha)$ from the Cartesian product of the previous two sets. The method with the smallest error in each column group is highlighted in bold, and the method with the smallest error across all column groups has a grey background. For instances where a version of the mixed-feature Wasserstein learning problem attains the smallest error across all column groups, the symbols † and ‡ indicate a statistically significant improvement of that model over the nominal and regularized nominal learning problem as well as over the better of the two continuous-feature approximations, respectively, at a $p$-value of 0.05. Most nominal models were solved within seconds, most continuous-feature models were solved within minutes, and most of the mixed-feature models were solved within 10 minutes. We relegate the details of the statistical significance tests and runtimes to the GitHub repository.

Overall, we observe from the tables that the mixed-feature Wasserstein learning problems outperform both the (regularized) nominal problems and the unbounded continuous-feature approximations in the classification and regression tasks. The outperformance of the mixed-feature Wasserstein learning problems over the continuous-feature approximations tends to be more substantial for problem instances with many discrete features, which further confirms our findings from Sections 3.5.1 and 3.6.1. Also, the mixed-feature Wasserstein learning problems tend to perform better on the parsimonious versions of the problem instances. This is intuitive as we expect robust optimization to be particularly effective in data sparse environments. Regularizing the mixed-feature Wasserstein learning problem does not yield significant advantages in the classification tasks, but it does help in the regression tasks. Finally, while we do not observe any significant differences among the classification loss functions, the Huber loss function performs best on the regression problem instances.

## 3.7 Conclusions

Wasserstein learning offers a principled approach to mitigating overfitting in supervised learning. While Wasserstein learning has been studied extensively for datasets with exclusively continuous features, the existing solution techniques cannot handle discrete features faithfully without sacrificing performance in real-world applications.

In this work, we propose a cutting plane algorithm for Wasserstein classification and regression problems that involve both continuous and discrete features. While the worst-case complexity of our method is exponential, it performs well on both synthetic and benchmark datasets. The core of our algorithm is an efficient identification of the most violated constraints with respect to the current solution. Despite the discrete nature of the problem, we show that it can be solved in polynomial time. We also study the complexity of mixed-feature Wasserstein learning more broadly, and we investigate its connections to regularized as well as different continuous-feature formulations.

Our work opens several promising directions for future research. Most notably, it would be instructive to explore whether similar algorithmic techniques can be developed for other supervised learning models, including decision trees, ensemble methods, and neural networks. Furthermore, it would be valuable to investigate the development of polynomial-time solution schemes that avoid reliance on the ellipsoid method and that overcome the scalability limitations of monolithic reformulations. For example, the fact that our algorithm for the mixed-feature problem relies on a sorting argument to identify the most violated constraint could potentially pave the way to developing robust counterparts using standard convex reformulations of sorting-type constraints.

| Dataset | $N$ | $M_x$ | $M_z$ | $K$ | Reduced? | Nom | MixF ($\kappa_y = 1$) | MixF ($\kappa_y = K$) | r-Nom | r-MixF ($\kappa_y = 1$) | r-MixF ($\kappa_y = K$) | ConF ($\kappa_y = 1$) | ConF ($\kappa_y = K$) |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| balance-scale | 625 | 0 | 16 | 4 | ✖ | 0.56% | 0.44% | **0.40% †‡** | 0.52% | 0.52% | **0.48%** | 0.48% | 0.44% |
| | | | | | ✓ | **1.65%** | **1.65% ‡** | **1.65% ‡** | **1.92%** | **1.92%** | **1.92%** | 1.73% | 1.91% |
| breast-cancer | 277 | 0 | 42 | 9 | ✖ | 30.10% | **28.91%** | 29.64% | 29.46% | **28.64% †‡** | 29.36% | 29.18% | 29.55% |
| | | | | | ✓ | 31.87% | 30.93% | **30.30%** | 29.58% | 28.68% | **28.46% †‡** | 31.69% | 36.45% |
| credit-approval | 690 | 6 | 36 | 9 | ✖ | 13.59% | 13.48% | **13.01% †‡** | 13.70% | 13.59% | **13.12%** | 13.95% | 13.23% |
| | | | | | ✓ | 16.30% | **15.00%** | 15.10% | 14.58% | 14.94% | **14.43% ‡** | 14.92% | 15.64% |
| cylinder-bands | 539 | 19 | 43 | 14 | ✖ | 22.85% | 22.52% | **22.38%** | 22.90% | **22.24% †** | 23.18% | 22.90% | 22.52% |
| | | | | | ✓ | 27.37% | 28.55% | **26.39% †** | 27.18% | 27.25% | **27.12%** | 26.81% | 26.47% |
| lymphography | 148 | 0 | 42 | 18 | ✖ | 17.24% | **16.90%** | 17.24% | 15.86% | 16.90% | **14.83% †‡** | 18.97% | 17.10% |
| | | | | | ✓ | 23.98% | 22.22% | **19.10%** | 17.96% | **17.67% †‡** | 18.30% | 25.51% | 19.43% |
| primacy | 339 | 0 | 25 | 17 | ✖ | **13.51%** | **13.51%** | 14.40% | 14.25% | **13.73%** | 14.33% | 13.81% | 14.55% |
| | | | | | ✓ | 15.89% | **13.99% †** | 15.30% | 14.70% | **14.34%** | 14.73% | 14.09% | 15.10% |
| spect | 267 | 0 | 22 | 22 | ✖ | 20.28% | 19.43% | **19.25%** | 20.09% | 19.25% | **17.35% †‡** | 20.28% | 20.00% |
| | | | | | ✓ | 23.22% | 21.13% | **19.97% †‡** | 23.34% | 22.44% | **20.34%** | 22.50% | 20.88% |
| tic-tac-toe | 958 | 0 | 18 | 9 | ✖ | 1.94% | **1.70%** | **1.70%** | **1.70%** | **1.70%** | **1.70%** | 1.70% | 1.70% |
| | | | | | ✓ | 2.75% | **1.70% †** | **1.70% †** | 1.81% | 1.67% | **1.66%** | 1.70% | 1.70% |

Table 21: *Mean classification error for the* ***logistic regression loss function****. $N$, $M_x$, $M_z$ and $K$ refer to the problem parameters from Section 3.2. The column 'Reduced?' indicates whether the full or sparse version of the dataset is being considered. Nom, MixF and ConF refer to the nominal problem formulation, the mixed-feature Wasserstein learning problem (54) and its unbounded continuous-feature approximation, respectively. The prefix 'r-' indicates the use of an $l_2$-regularization.*

| Dataset | $N$ | $M_x$ | $M_z$ | $K$ | Reduced? | Nom | MixF ($\kappa_y = 1$) | MixF ($\kappa_y = K$) | r-Nom | r-MixF ($\kappa_y = 1$) | r-MixF ($\kappa_y = K$) | ConF ($\kappa_y = 1$) | ConF ($\kappa_y = K$) |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| balance-scale | 625 | 0 | 16 | 4 | ✖ | 4.20% | **4.08%** | 4.20% | 3.36% | **3.08% †‡** | 3.28% | 3.88% | 4.20% |
| | | | | | ✓ | 5.91% | **5.56% †** | 5.88% | 5.69% | 6.23% | **5.64%** | 5.64% | 5.68% |
| breast-cancer | 277 | 0 | 42 | 9 | ✖ | 30.27% | **30.09% ‡** | 30.82% | 30.46% | 30.82% | **30.18%** | 35.18% | 36.00% |
| | | | | | ✓ | 31.66% | 31.15% | **30.69%** | 31.05% | 30.93% | **29.58% †‡** | 31.96% | 36.78% |
| credit-approval | 690 | 6 | 36 | 9 | ✖ | 13.84% | 13.88% | **13.51%** | 14.06% | 14.02% | **13.59%** | 13.62% | 14.06% |
| | | | | | ✓ | 15.86 % | 16.21% | **15.81%** | 14.84% | 15.98% | **14.83% ‡** | 15.22% | 18.62% |
| cylinder-bands | 539 | 19 | 43 | 14 | ✖ | 23.22% | 23.08% | **22.52% †** | 23.08% | 23.04% | **22.85%** | 22.85% | 24.11% |
| | | | | | ✓ | 28.47% | 27.86% | **27.29% †‡** | 28.55% | 28.99% | **28.05%** | 27.97% | 27.88% |
| lymphography | 148 | 0 | 42 | 18 | ✖ | 19.14% | **18.97%** | 20.52% | 18.79% | 17.41% | **17.24% †** | 18.28% | 19.83% |
| | | | | | ✓ | **20.28%** | 20.91% | 21.25% | 19.49% | 19.77% | **18.64% †‡** | 23.92% | 21.31% |
| primacy | 339 | 0 | 25 | 17 | ✖ | 14.03% | 13.73% | **13.28%** | 13.36% | 13.73% | **13.13% ‡** | 13.81% | 13.66% |
| | | | | | ✓ | 15.96% | **15.05%** | 15.47% | 14.53% | 14.68% | **14.51% ‡** | 14.78% | 15.00% |
| spect | 267 | 0 | 22 | 22 | ✖ | 20.19% | 20.76% | **18.96% †‡** | 21.04% | 20.85% | **19.34%** | 21.23% | 20.28% |
| | | | | | ✓ | 24.16% | **21.50% †** | 21.97% | 22.031% | 22.06% | **21.91%** | 23.28% | 21.53% |
| tic-tac-toe | 958 | 0 | 18 | 9 | ✖ | **1.70%** | **1.70%** | **1.70%** | **1.70%** | **1.70%** | **1.70%** | 1.70% | 1.70% |
| | | | | | ✓ | 2.58% | **1.65%** | **1.65%** | **1.65%** | **1.65%** | **1.65%** | 1.65% | 1.65% |

Table 22: *Mean classification error for the* ***hinge loss function****. We use the same abbreviations as in Table 21.*

| Dataset | $N$ | $M_{\mathrm{x}}$ | $M_{\mathrm{z}}$ | $K$ | Reduced? | Nom | MixF ($\kappa_{\mathrm{y}} = 1$) | MixF ($\kappa_{\mathrm{y}} = K$) | r-Nom | r-MixF ($\kappa_{\mathrm{y}} = 1$) | r-MixF ($\kappa_{\mathrm{y}} = K$) | ConF ($\kappa_{\mathrm{y}} = 1$) | ConF ($\kappa_{\mathrm{y}} = K$) |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| balance-scale | 625 | 0 | 16 | 4 | ✖ | 1.36% | **0.40%** | **0.40% †‡** | 1.52% | **0.48%** | **0.48%** | 1.32% | 1.36% |
| | | | | | ✓ | 3.15% | 2.15% | **2.01%** | 3.25% | 2.08% | **1.81% †‡** | 3.29% | 3.31% |
| breast-cancer | 277 | 0 | 42 | 9 | ✖ | 29.55% | **28.82% †** | 29.18% | 30.18% | 32.27% | **29.64%** | 30.36% | 29.18% |
| | | | | | ✓ | 31.30% | 31.36% | **30.36%** | 28.55% | 28.52% | **27.41% †‡** | 29.88% | 29.94% |
| credit-approval | 690 | 6 | 36 | 9 | ✖ | 13.44% | **13.37%** | 13.41% | 13.44% | **13.37% ‡** | 13.41% | 13.66% | 14.46% |
| | | | | | ✓ | 15.59% | **15.16%** | 15.52% | 14.65% | 14.53% | **14.52% †‡** | 15.30% | 15.75% |
| cylinder-bands | 539 | 19 | 43 | 14 | ✖ | 23.27% | **21.78% †‡** | 22.43% | 23.04% | **22.52%** | 22.85% | 22.57% | 22.71% |
| | | | | | ✓ | 27.43% | **26.50% †‡** | 27.57% | **27.09%** | 27.40% | 27.37 % | 27.11% | 27.40% |
| lymphography | 148 | 0 | 42 | 18 | ✖ | 19.31% | **17.76%** | 18.97% | 16.38% | 16.55% | **15.52% †** | 20.52% | 15.86% |
| | | | | | ✓ | 21.31% | 21.93% | **20.91%** | 18.47% | **18.35% ‡** | 18.47% | 20.11% | 20.34% |
| primacy | 339 | 0 | 25 | 17 | ✖ | 13.51% | **12.99% ‡** | 13.51% | **13.13%** | 13.43% | 13.21% | 13.66% | 14.18% |
| | | | | | ✓ | 15.10% | **14.80%** | 15.03% | **14.38%** | 14.58% | 14.41% | **14.06%** | 14.66% |
| spect | 267 | 0 | 22 | 22 | ✖ | 20.28% | 18.77% | **18.30%** | 19.25% | 19.25% | **17.08% †‡** | 21.04% | 21.13% |
| | | | | | ✓ | 23.16% | 21.97% | **20.53%** | 23.16% | 22.34% | **20.34% †‡** | 21.09% | 21.50% |
| tic-tac-toe | 958 | 0 | 18 | 9 | ✖ | 1.73% | **1.70%** | **1.70%** | **1.70%** | **1.70%** | **1.70%** | 1.70% | 1.70% |
| | | | | | ✓ | 2.77% | **1.65%** | **1.65%** | **1.65%** | **1.65%** | **1.65%** | 1.65% | 1.65% |

Table 23: *Mean classification error for the* ***smooth hinge loss function****. We use the same abbreviations as in Table 21.*

| Dataset | $N$ | $M_{\mathrm{x}}$ | $M_{\mathrm{z}}$ | $K$ | Reduced? | Nom | MixF ($\kappa_{\mathrm{y}} = 1$) | MixF ($\kappa_{\mathrm{y}} = K$) | r-Nom | r-MixF ($\kappa_{\mathrm{y}} = 1$) | r-MixF ($\kappa_{\mathrm{y}} = K$) | ConF ($\kappa_{\mathrm{y}} = 1$) | ConF ($\kappa_{\mathrm{y}} = K$) |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| bike | 17,379 | 4 | 31 | 5 | ✖ | **210.23** | **210.23** | **210.23** | **210.23** | **210.23** | **210.23** | 210.23 | 210.23 |
| | | | | | ✓ | 210.82 | **210.81 †‡** | 210.81 | 210.84 | **210.84** | 210.84 | 213.47 | 213.60 |
| fire | 517 | 10 | 31 | 4 | ✖ | **123.33** | **123.33 †‡** | **123.33 †‡** | 123.48 | **123.33 †‡** | 124.74 | 149.03 | 149.04 |
| | | | | | ✓ | 132.71 | 117.42 | **111.50 †‡** | 114.39 | 113.56 | **112.98** | 120.41 | 118.30 |
| flare | 1,066 | 0 | 21 | 9 | ✖ | 198.10 | **198.10** | 198.10 | 198.87 | **198.86** | **198.86** | **198.10** | 198.10 |
| | | | | | ✓ | 199.85 | **199.84** | **199.84** | 200.09 | **200.09** | **200.09** | 199.61 | 199.60 |
| garments | 1,197 | 7 | 22 | 4 | ✖ | **363.11** | 363.40 | 366.48 | **362.55** | 362.57 | 362.60 | 363.36 | 364.38 |
| | | | | | ✓ | 890.81 | **367.76** | 368.95 | 368.15 | **367.29 †** | 367.84 | 367.64 | 368.97 |
| imports | 193 | 14 | 45 | 10 | ✖ | 442.71 | **289.01** | 302.02 | 311.08 | **269.97 †‡** | 270.94 | 289.19 | 300.18 |
| | | | | | ✓ | 660.04 | 377.26 | **368.77** | 352.79 | **334.54 †‡** | 336.18 | 384.48 | 370.93 |
| student | 395 | 13 | 26 | 17 | ✖ | **379.13** | 379.20 | 383.07 | 378.66 | **378.12 †** | 378.66 | 379.15 | 383.80 |
| | | | | | ✓ | **411.98** | 412.32 | 413.51 | 384.20 | **384.07 †‡** | 384.77 | 412.11 | 396.66 |
| vegas | 504 | 5 | 106 | 14 | ✖ | 232,330.88 | 533.63 | **514.45** | **479.91** | 479.92 | **479.91 ‡** | 520.72 | 482.06 |
| | | | | | ✓ | 287,615.23 | 605.27 | **554.18** | **493.60** | **493.60 ‡** | 493.79 | 581.52 | 494.27 |

Table 24: *Mean squared errors for the* ***Huber loss function*** *(with* $\delta = 0.5$*). We use the same abbreviations as in Table 21.*

| Dataset | $N$ | $M_x$ | $M_z$ | $K$ | Reduced? | Nom | MixF ($\kappa_y = 1$) | MixF ($\kappa_y = K$) | r-Nom | r-MixF ($\kappa_y = 1$) | r-MixF ($\kappa_y = K$) | ConF ($\kappa_y = 1$) | ConF ($\kappa_y = K$) |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| bike | 17,379 | 4 | 31 | 5 | ✖ | 217.87 | 217.87 | **217.87** | 217.85 | **217.82** | 217.87 | 217.86 | 217.87 |
| | | | | | ✓ | 218.17 | 218.17 | **218.16** | 218.17 | 218.18 | **218.17** | 218.16 | 218.21 |
| fire | 517 | 10 | 31 | 4 | ✖ | 124.36 | 124.36 | **124.29 †‡** | 130.45 | **124.36** | 124.45 | 142.59 | 124.33 |
| | | | | | ✓ | 120.54 | 110.76 | **109.71** | 116.67 | **109.60 †‡** | **109.60** | 109.74 | 109.81 |
| flare | 1,066 | 0 | 21 | 9 | ✖ | **218.56** | 218.62 | 218.87 | 218.27 | 219.26 | **204.93 †‡** | 218.55 | 218.54 |
| | | | | | ✓ | 2,230.29 | **750.89** | 752.05 | 215.22 | 215.22 | **202.78 †‡** | 215.98 | 215.80 |
| garments | 1,197 | 7 | 22 | 4 | ✖ | 374.08 | **373.77** | 377.32 | **373.32** | 373.76 | 374.70 | 373.79 | 377.65 |
| | | | | | ✓ | 894.07 | 376.12 | **375.77 †** | 377.78 | 377.06 | **376.58** | 376.05 | 377.07 |
| imports | 193 | 14 | 45 | 10 | ✖ | 671.67 | **299.63** | 305.48 | 351.66 | 292.86 | **289.26 †‡** | 311.48 | 319.25 |
| | | | | | ✓ | 726.46 | **379.16** | 408.11 | 371.56 | **348.24 †‡** | 349.712 | 406.18 | 389.03 |
| student | 395 | 13 | 26 | 17 | ✖ | 399.65 | 399.65 | **397.63** | **383.57** | **383.57 ‡** | 386.14 | 399.66 | 405.33 |
| | | | | | ✓ | 451.85 | 451.89 | **397.59 †‡** | 403.51 | 398.50 | **393.65** | 451.95 | 408.27 |
| vegas | 504 | 5 | 106 | 14 | ✖ | 232,332.44 | 529.55 | **500.85** | 503.37 | 503.37 | **490.76 †‡** | 538.10 | 503.46 |
| | | | | | ✓ | 287,597.53 | 580.67 | **505.11** | 507.74 | 507.74 | **500.22 †‡** | 592.46 | 511.80 |

Table 25: *Mean squared errors for the* ***pinball loss function*** *(with* $\tau = 0.5$*). We use the same abbreviations as in Table 21.*

| Dataset | $N$ | $M_x$ | $M_z$ | $K$ | Reduced? | Nom | MixF ($\kappa_y = 1$) | MixF ($\kappa_y = K$) | r-Nom | r-MixF ($\kappa_y = 1$) | r-MixF ($\kappa_y = K$) | ConF ($\kappa_y = 1$) | ConF ($\kappa_y = K$) |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| bike | 17,379 | 4 | 31 | 5 | ✖ | **217.87** | 217.87 | **217.87** | 217.88 | **217.87** | 217.88 | 217.86 | 217.87 |
| | | | | | ✓ | 218.18 | 218.17 | **218.16** | **218.15** | **218.15** | 218.15 | 218.14 | 218.20 |
| fire | 517 | 10 | 31 | 4 | ✖ | 133.10 | 133.02 | **132.64** | 136.26 | 132.90 | **132.67 †‡** | 133.05 | 132.82 |
| | | | | | ✓ | 185.04 | 128.05 | **124.38 ‡** | 125.23 | 125.30 | **125.02** | 126.67 | 124.93 |
| flare | 1,066 | 0 | 21 | 9 | ✖ | **212.93** | 212.95 | 213.20 | 215.49 | 215.40 | **211.44 †‡** | 212.92 | 212.94 |
| | | | | | ✓ | 1,562.88 | 532.37 | **359.82** | 232.67 | 210.53 | **207.05 †‡** | 209.49 | 209.28 |
| garments | 1,197 | 7 | 22 | 4 | ✖ | 369.10 | **368.65 †** | 371.72 | 370.08 | **369.07** | 370.17 | 368.65 | 371.67 |
| | | | | | ✓ | 871.46 | **370.90** | 372.19 | 370.72 | 370.77 | **370.68** | 371.17 | 372.03 |
| imports | 193 | 14 | 45 | 10 | ✖ | 621.38 | **288.81 †‡** | 300.38 | 331.36 | 300.22 | **299.10** | 306.61 | 322.70 |
| | | | | | ✓ | 761.05 | **370.96** | 394.00 | 370.97 | 338.80 | **338.07 †‡** | 392.69 | 380.75 |
| student | 395 | 13 | 26 | 17 | ✖ | 399.65 | 399.66 | **397.60** | 387.07 | 386.27 | **385.75 ‡** | 399.68 | 405.34 |
| | | | | | ✓ | 451.85 | 451.89 | **397.59** | 403.51 | 398.50 | **393.65 †‡** | 451.95 | 408.27 |
| vegas | 504 | 5 | 106 | 14 | ✖ | 232,323.47 | 530.00 | **500.36** | 493.28 | 493.29 | **489.17 †‡** | 538.51 | 503.48 |
| | | | | | ✓ | 287,607.59 | 580.52 | **504.85** | 517.85 | 517.86 | **497.64 †‡** | 592.31 | 511.76 |

Table 26: *Mean squared errors for the* $\tau$***-insensitive loss function*** *(with* $\tau = 0.1$*). We use the same abbreviations as in Table 21.*

## 3.A Wasserstein Classification with Mixed Features

This section derives exponential-size convex reformulations of the mixed-feature Wasserstein learning problem (56) for classification problems with convex and Lipschitz continuous as well as piece-wise affine and convex loss functions $L$. One readily confirms that the resulting reformulations are special cases of our unified representation (57), which implies that our results in this appendix prove Theorem 7 for the case of classification problems.

**Proposition 11.** *Consider a classification problem with a convex and Lipschitz continuous loss function as well as $\mathbb{X} = \mathbb{R}^{M_\mathrm{x}}$. In this case, the Wasserstein learning problem (56) is equivalent to*

$$\begin{array}{lll} \underset{\boldsymbol{\beta},\lambda,\boldsymbol{s}}{\text{minimize}} & \lambda\epsilon + \dfrac{1}{N}\displaystyle\sum_{n\in[N]} s_n & \\ \text{subject to} & \left.\begin{array}{l} l_{\boldsymbol{\beta}}(\boldsymbol{x}^n, \boldsymbol{z}, y^n) - \lambda\kappa_\mathrm{z} d_\mathrm{z}(\boldsymbol{z}, \boldsymbol{z}^n) \leq s_n \\ l_{\boldsymbol{\beta}}(\boldsymbol{x}^n, \boldsymbol{z}, -y^n) - \lambda\kappa_\mathrm{z} d_\mathrm{z}(\boldsymbol{z}, \boldsymbol{z}^n) - \lambda\kappa_\mathrm{y} \leq s_n \end{array}\right\} \forall n \in [N],\ \forall \boldsymbol{z} \in \mathbb{Z} & \\ & \mathrm{lip}(L) \cdot \|\boldsymbol{\beta}_\mathrm{x}\|_* \leq \lambda & \\ & \boldsymbol{\beta} = (\beta_0, \boldsymbol{\beta}_\mathrm{x}, \boldsymbol{\beta}_\mathrm{z}) \in \mathbb{R}^{1+M_\mathrm{x}+M_\mathrm{z}},\ \ \lambda \in \mathbb{R}_+,\ \ \boldsymbol{s} \in \mathbb{R}^N_+. & \end{array} \tag{58}$$

Formulation (58) emerges as a special case of our unified representation (57) if we set $\mathcal{I} = \{\pm 1\}$, $\boldsymbol{\theta} = (\boldsymbol{\beta}, \lambda)$, $\boldsymbol{\sigma} = \boldsymbol{s}$, $f_0(\boldsymbol{\theta}, \boldsymbol{\sigma}) = \lambda\epsilon + \dfrac{1}{N}\sum_{n\in[N]} s_n$, $f_{ni}(e) = L(e)$, $g_{ni}(\boldsymbol{\theta}, \boldsymbol{\xi}^n_{-\mathbf{z}}; \boldsymbol{z}) = iy^n \cdot [\beta_0 + {\boldsymbol{\beta}_\mathrm{x}}^\top \boldsymbol{x}^n + {\boldsymbol{\beta}_\mathrm{z}}^\top \boldsymbol{z}]$, $h_{ni}(\boldsymbol{\theta}, \boldsymbol{\xi}^n_{-\mathbf{z}}; d_\mathrm{z}(\boldsymbol{z}, \boldsymbol{z}^n)) = \lambda\kappa_\mathrm{z} d_\mathrm{z}(\boldsymbol{z}, \boldsymbol{z}^n) + \lambda\kappa_\mathrm{y} \cdot \mathbb{1}[i = -1]$ and $\Theta = \{\boldsymbol{\theta} : \mathrm{lip}(L) \cdot \|\boldsymbol{\beta}_\mathrm{x}\|_* \leq \lambda,\ \lambda \in \mathbb{R}_+\}$. One verifies that this choice of $\mathcal{I}$, $\boldsymbol{\theta}$, $\boldsymbol{\sigma}$, $f_0$, $f_{ni}$, $g_{ni}$, $h_{ni}$ and $\Theta$ satisfies the regularity conditions imposed on problem (57) in the main text.

**Proposition 12.** *If $\boldsymbol{\beta}$ in problem (58) is restricted to a bounded hypothesis set $\mathcal{H}$, then $\lambda$ can be restricted to a bounded set as well.*

Proposition 11 covers Wasserstein support vector machines with a smooth Hinge loss,

$$L(e) = \begin{cases} 1/2 - e & \text{if } e \leq 0, \\ (1/2) \cdot (1-e)^2 & \text{if } e \in (0,1), \\ 0 & \text{otherwise,} \end{cases}$$

where the Lipschitz modulus is $\mathrm{lip}(L) = 1$, and logistic regression with a log-loss,

$$L(e) = \log(1 + \exp[-e]),$$

where again $\mathrm{lip}(L) = 1$. We provide the corresponding reformulations next.

**Corollary 2.** *The first set of inequality constraints in* (58) *can be reformulated as follows.*

*(i) For the smooth Hinge loss function:*

$$\left.\begin{aligned}&\frac{1}{2}\left(w^{+}_{\boldsymbol{z},n} - y^n\cdot(\beta_0 + {\boldsymbol{\beta}_{\mathrm{x}}}^\top \boldsymbol{x}^n + {\boldsymbol{\beta}_{\mathrm{z}}}^\top \boldsymbol{z})\right)^2 + 1 - w^{+}_{\boldsymbol{z},n} - \lambda\kappa_{\mathrm{z}} d_{\mathrm{z}}(\boldsymbol{z},\boldsymbol{z}^n) \leq s_n \\ &\frac{1}{2}\left(w^{+}_{\boldsymbol{z},n} - y^n\cdot(\beta_0 + {\boldsymbol{\beta}_{\mathrm{x}}}^\top \boldsymbol{x}^n + {\boldsymbol{\beta}_{\mathrm{z}}}^\top \boldsymbol{z})\right)^2 - \lambda\kappa_{\mathrm{z}} d_{\mathrm{z}}(\boldsymbol{z},\boldsymbol{z}^n) \leq s_n\end{aligned}\right\}\ \forall n\in[N],\ \forall \boldsymbol{z}\in\mathbb{Z}$$

*(ii) For the log-loss function:*

$$\log\left(1+\exp\left[-y^n\cdot(\beta_0 + {\boldsymbol{\beta}_{\mathrm{x}}}^\top \boldsymbol{x}^n + {\boldsymbol{\beta}_{\mathrm{z}}}^\top \boldsymbol{z})\right]\right) - \lambda\kappa_{\mathrm{z}} d_{\mathrm{z}}(\boldsymbol{z},\boldsymbol{z}^n) \leq s_n \quad \forall n\in[N],\ \forall \boldsymbol{z}\in\mathbb{Z}$$

*Here, $w^{+}_{\boldsymbol{z},n}\in\mathbb{R}$ are auxiliary decision variables. The second set of inequality constraints in* (58) *follows similarly if we replace $w^{+}_{\boldsymbol{z},n}\in\mathbb{R}$ with additional auxiliary decision variables $w^{-}_{\boldsymbol{z},n}\in\mathbb{R}$, replace $-y^n$ with $+y^n$ and subtract the expression $\lambda\kappa_{\mathrm{y}}$ from the constraint left-hand sides.*

We now provide a reformulation of problem (56) without embedded maximizations when the loss function $L$ is piece-wise affine and convex.

**Proposition 13.** *Consider a classification problem with a piece-wise affine and convex loss function $L(e) = \max_{j\in[J]}\{a_j e + b_j\}$, and assume that $\mathbb{X} = \{\boldsymbol{x}\in\mathbb{R}^{M_{\mathrm{x}}} : \boldsymbol{C}\boldsymbol{x} \preccurlyeq_{\mathcal{C}} \mathbf{d}\}$ for some $\boldsymbol{C}\in\mathbb{R}^{r\times M_{\mathrm{x}}}$, $\mathbf{d}\in\mathbb{R}^r$ and proper convex cone $\mathcal{C}\subseteq\mathbb{R}^r$. If $\mathbb{X}$ admits a Slater point $\boldsymbol{x}^{\mathrm{s}}\in\mathbb{R}^{M_{\mathrm{x}}}$ such that $\boldsymbol{C}\boldsymbol{x}^{\mathrm{s}} \prec_{\mathcal{C}} \mathbf{d}$, then the Wasserstein learning problem* (56) *is equivalent to*

$$\begin{array}{ll}\underset{\boldsymbol{\beta},\lambda,\boldsymbol{s},\boldsymbol{q}^{+}_{nj},\boldsymbol{q}^{-}_{nj}}{\text{minimize}} & \lambda\epsilon + \dfrac{1}{N}\displaystyle\sum_{n\in[N]} s_n \\ \text{subject to} & \left.\begin{aligned}&{\boldsymbol{q}^{+}_{nj}}^\top(\mathbf{d}-\boldsymbol{C}\boldsymbol{x}^n) + a_j y^n\cdot(\beta_0 + {\boldsymbol{\beta}_{\mathrm{x}}}^\top \boldsymbol{x}^n + {\boldsymbol{\beta}_{\mathrm{z}}}^\top \boldsymbol{z}) - \lambda\kappa_{\mathrm{z}} d_{\mathrm{z}}(\boldsymbol{z},\boldsymbol{z}^n) + b_j \leq s_n \\ &{\boldsymbol{q}^{-}_{nj}}^\top(\mathbf{d}-\boldsymbol{C}\boldsymbol{x}^n) - a_j y^n\cdot(\beta_0 + {\boldsymbol{\beta}_{\mathrm{x}}}^\top \boldsymbol{x}^n + {\boldsymbol{\beta}_{\mathrm{z}}}^\top \boldsymbol{z}) - \lambda\kappa_{\mathrm{z}} d_{\mathrm{z}}(\boldsymbol{z},\boldsymbol{z}^n) - \lambda\kappa_{\mathrm{y}} + b_j \leq s_n\end{aligned}\right\} \\ & \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \forall n\in[N],\ \forall j\in[J],\ \forall\boldsymbol{z}\in\mathbb{Z} \\ & \|a_j y^n\cdot\boldsymbol{\beta}_{\mathrm{x}} - \boldsymbol{C}^\top\boldsymbol{q}^{+}_{nj}\|_* \leq \lambda,\ \ \|a_j y^n\cdot\boldsymbol{\beta}_{\mathrm{x}} + \boldsymbol{C}^\top\boldsymbol{q}^{-}_{nj}\|_* \leq \lambda \qquad \forall n\in[N],\ \forall j\in[J] \\ & \boldsymbol{\beta} = (\beta_0,\boldsymbol{\beta}_{\mathrm{x}},\boldsymbol{\beta}_{\mathrm{z}})\in\mathbb{R}^{1+M_{\mathrm{x}}+M_{\mathrm{z}}},\ \ \lambda\in\mathbb{R}_+,\ \ \boldsymbol{s}\in\mathbb{R}^N_+ \\ & \boldsymbol{q}^{+}_{nj},\boldsymbol{q}^{-}_{nj}\in\mathcal{C}^*,\ n\in[N] \textit{ and } j\in[J].\end{array} \tag{59}$$

Formulation (59) emerges as a special case of our unified representation (57) if we set $i = (j,t)$, $\mathcal{I} = [J]\times\{\pm 1\}$, $\boldsymbol{\theta} = (\boldsymbol{\beta},\lambda,\boldsymbol{q}_{ni})$, $\boldsymbol{\sigma} = \boldsymbol{s}$, $f_0(\boldsymbol{\theta},\boldsymbol{\sigma}) = \lambda\epsilon + \frac{1}{N}\sum_{n\in[N]} s_n$, $f_{ni}(e) = a_j e + b_j$, $g_{ni}(\boldsymbol{\theta},\boldsymbol{\xi}^n_{-\mathbf{z}};\boldsymbol{z}) = t y^n\cdot[\beta_0 + {\boldsymbol{\beta}_{\mathrm{x}}}^\top\boldsymbol{x}^n + {\boldsymbol{\beta}_{\mathrm{z}}}^\top\boldsymbol{z}]$, $h_{ni}(\boldsymbol{\theta},\boldsymbol{\xi}^n_{-\mathbf{z}}; d_{\mathrm{z}}(\boldsymbol{z},\boldsymbol{z}^n)) = -{\boldsymbol{q}_{ni}}^\top(\mathbf{d}-\boldsymbol{C}\boldsymbol{x}^n) +$

$\lambda\kappa_{\mathrm{z}} d_{\mathrm{z}}(\boldsymbol{z}, \boldsymbol{z}^n) + \lambda\kappa_{\mathrm{y}} \cdot \mathbb{1}[t = -1]$ and $\Theta = \{\boldsymbol{\theta} : \|a_j y^n \cdot \boldsymbol{\beta}_{\mathrm{x}} - t \cdot \boldsymbol{C}^\top \boldsymbol{q}_{ni}\|_* \leq \lambda \, \forall (n,i) \in [N] \times \mathcal{I},\ \boldsymbol{q}_{ni} \in \mathcal{C}^* \, \forall (n,i) \in [N] \times \mathcal{I},\ \lambda \in \mathbb{R}_+\}$. One verifies that this choice of $\mathcal{I}$, $\boldsymbol{\theta}$, $\boldsymbol{\sigma}$, $f_0$, $f_{ni}$, $g_{ni}$, $h_{ni}$ and $\Theta$ satisfies the regularity conditions imposed on problem (57) in the main text.

**Proposition 14.** *If $\boldsymbol{\beta}$ in problem (59) is restricted to a bounded hypothesis set $\mathcal{H}$, then $(\boldsymbol{q}_{ni}^+, \boldsymbol{q}_{ni}^-)_{n,i}$ and $\lambda$ can be restricted to a bounded set as well.*

Proposition 13 covers Wasserstein support vector machines with a (non-smooth) Hinge loss,

$$L(e) = \max\{1 - e,\ 0\},$$

which is a piece-wise affine convex function with $J = 2$, $a_1 = -1$, $b_1 = 1$, $a_2 = 0$ and $b_2 = 0$. We provide the corresponding reformulation next.

**Corollary 3.** *For the (non-smooth) Hinge loss function, the first set of inequality constraints in (59) can be reformulated as*

$$\boldsymbol{q}_n^{+\top}(\boldsymbol{d} - \boldsymbol{C}\boldsymbol{x}^n) - y^n(\beta_0 + {\boldsymbol{\beta}_{\mathrm{x}}}^\top \boldsymbol{x}^n + {\boldsymbol{\beta}_{\mathrm{z}}}^\top \boldsymbol{z}) - \lambda\kappa_{\mathrm{z}} d_{\mathrm{z}}(\boldsymbol{z}, \boldsymbol{z}^n) + 1 \leq s_n \quad \forall n \in [N],\ \forall \boldsymbol{z} \in \mathbb{Z}.$$

*The second set of inequality constraints in (59) follows similarly if we replace $-y^n$ with $+y^n$ as well as $\boldsymbol{q}_n^+$ with $\boldsymbol{q}_n^-$ and subtract the expression $\lambda\kappa_{\mathrm{y}}$ from the constraint left-hand sides.*

## 3.B Wasserstein Regression with Mixed Features

This section derives exponential-size convex reformulations of the mixed-feature Wasserstein learning problem (56) for regression problems with convex and Lipschitz continuous as well as piece-wise affine and convex loss functions $L$. One readily confirms that the resulting reformulations are special cases of our unified representation (57), which implies that our results in this appendix prove Theorem 7 for the case of regression problems.

**Proposition 15.** *Consider a regression problem with a convex and Lipschitz continuous loss* *function $L$ and $(\mathbb{X},\mathbb{Y}) = (\mathbb{R}^{M_\mathrm{x}},\mathbb{R})$. In this case, the Wasserstein learning problem (56) equals*

$$\begin{array}{lll} \underset{\boldsymbol{\beta},\lambda,\boldsymbol{s}}{\text{minimize}} & \lambda\epsilon + \dfrac{1}{N}\displaystyle\sum_{n\in[N]} s_n & \\ \text{subject to} & l_{\boldsymbol{\beta}}(\boldsymbol{x}^n,\boldsymbol{z},y^n) - \lambda\kappa_\mathrm{z} d_\mathrm{z}(\boldsymbol{z},\boldsymbol{z}^n) \le s_n & \forall n\in[N],\ \forall \boldsymbol{z}\in\mathbb{Z} \\ & \mathrm{lip}(L)\cdot\|\boldsymbol{\beta}_\mathrm{x}\|_* \le \lambda & \\ & \mathrm{lip}(L) \le \lambda\kappa_\mathrm{y} & \\ & \boldsymbol{\beta} = (\beta_0,\boldsymbol{\beta}_\mathrm{x},\boldsymbol{\beta}_\mathrm{z}) \in \mathbb{R}^{1+M_\mathrm{x}+M_\mathrm{z}},\ \ \lambda\in\mathbb{R}_+,\ \ \boldsymbol{s}\in\mathbb{R}^N_+. & \end{array} \tag{60}$$

Formulation (60) emerges as a special case of our unified representation (57) if we set $\boldsymbol{\theta} = (\boldsymbol{\beta},\lambda)$, $\boldsymbol{\sigma} = \boldsymbol{s}$, $f_0(\boldsymbol{\theta},\boldsymbol{\sigma}) = \lambda\epsilon + \frac{1}{N}\sum_{n\in[N]} s_n$, $f_{ni}(e) = L(e)$, $g_{ni}(\boldsymbol{\theta},\boldsymbol{\xi}^n_{-\mathbf{z}};\boldsymbol{z}) = \beta_0 + {\boldsymbol{\beta}_\mathrm{x}}^\top\boldsymbol{x}^n + {\boldsymbol{\beta}_\mathrm{z}}^\top\boldsymbol{z} - y^n$, $h_{ni}(\boldsymbol{\theta},\boldsymbol{\xi}^n_{-\mathbf{z}}; d_\mathrm{z}(\boldsymbol{z},\boldsymbol{z}^n)) = \lambda\kappa_\mathrm{z} d_\mathrm{z}(\boldsymbol{z},\boldsymbol{z}^n)$ and $\Theta = \{\boldsymbol{\theta} : \mathrm{lip}(L)\cdot\|\boldsymbol{\beta}_\mathrm{x}\|_* \le \lambda,\ \mathrm{lip}(L)\le\lambda\kappa_\mathrm{y},\ \lambda\in\mathbb{R}_+\}$. One verifies that this choice of $\boldsymbol{\theta}$, $\boldsymbol{\sigma}$, $f_0$, $f_{ni}$, $g_{ni}$, $h_{ni}$ and $\Theta$ satisfies the regularity conditions imposed on problem (57) in the main text.

**Corollary 4.** *If $\boldsymbol{\beta}$ in problem (60) is restricted to a bounded hypothesis set $\mathcal{H}$, then $\lambda$ can be restricted to a bounded set as well.*

Proposition 15 covers Wasserstein regression with a Huber loss,

$$L(e) = \begin{cases} (1/2)\cdot e^2 & \text{if } |e|\le\delta, \\ \delta\cdot(|e| - (1/2)\cdot\delta) & \text{otherwise,} \end{cases}$$

where $\delta\in\mathbb{R}_+$ determines the boundary between the quadratic and the absolute loss, implying that $\mathrm{lip}(L) = \delta$. We provide the corresponding reformulation next.

**Corollary 5.** *For the Huber loss function, the first set of inequality constraints in (60) can be reformulated as*

$$\left.\begin{array}{l} \frac{1}{2}(\beta_0 + {\boldsymbol{\beta}_\mathrm{x}}^\top\boldsymbol{x}^n + {\boldsymbol{\beta}_\mathrm{z}}^\top\boldsymbol{z} - y^n - p_{\boldsymbol{z},n})^2 + \delta p_{\boldsymbol{z},n} - \lambda\kappa_\mathrm{z} d_\mathrm{z}(\boldsymbol{z},\boldsymbol{z}^n) \le s_n \\ \frac{1}{2}(\beta_0 + {\boldsymbol{\beta}_\mathrm{x}}^\top\boldsymbol{x}^n + {\boldsymbol{\beta}_\mathrm{z}}^\top\boldsymbol{z} - y^n - p_{\boldsymbol{z},n})^2 - \delta p_{\boldsymbol{z},n} - \lambda\kappa_\mathrm{z} d_\mathrm{z}(\boldsymbol{z},\boldsymbol{z}^n) \le s_n \end{array}\right\} \forall n\in[N],\ \forall\boldsymbol{z}\in\mathbb{Z},$$

*where $p_{\boldsymbol{z},n}\in\mathbb{R}$ are auxiliary decision variables.*

We now provide a reformulation of problem (56) without embedded maximizations when the loss function $L$ is piece-wise affine and convex.

**Proposition 16.** *Consider a regression problem with a piece-wise affine and convex loss function* $L(e) = \max_{j\in[J]}\{a_j e + b_j\}$*, and assume that* $\mathbb{X}\times\mathbb{Y} = \big\{(\boldsymbol{x}, y) \in \mathbb{R}^{M_\mathrm{x}+1} : \boldsymbol{C}_\mathrm{x}\boldsymbol{x} + \boldsymbol{c}_\mathrm{y}\cdot y \preccurlyeq_{\mathcal{C}} \boldsymbol{d}\big\}$ *for some* $\boldsymbol{C}_\mathrm{x} \in \mathbb{R}^{r\times M_\mathrm{x}}$*,* $\boldsymbol{c}_\mathrm{y} \in \mathbb{R}^r$*,* $\boldsymbol{d} \in \mathbb{R}^r$ *and proper convex cone* $\mathcal{C} \subseteq \mathbb{R}^r$*. If this set admits a Slater point* $(\boldsymbol{x}^\mathrm{s}, y^\mathrm{s}) \in \mathbb{R}^{M_\mathrm{x}+1}$ *such that* $\boldsymbol{C}_\mathrm{x}\boldsymbol{x}^\mathrm{s} + \boldsymbol{c}_\mathrm{y}\cdot y^\mathrm{s} \prec_{\mathcal{C}} \boldsymbol{d}$*, then problem* (56) *is equivalent to*

$$\begin{array}{ll}\underset{\boldsymbol{\beta},\lambda,\boldsymbol{s},\boldsymbol{q}_{nj}}{\text{minimize}} & \lambda\epsilon + \dfrac{1}{N}\displaystyle\sum_{n\in[N]} s_n \\ \text{subject to} & \boldsymbol{q}_{nj}^\top(\boldsymbol{d} - \boldsymbol{C}_\mathrm{x}\boldsymbol{x}^n - \boldsymbol{c}_\mathrm{y}\cdot y^n) + a_j\cdot(\beta_0 + \boldsymbol{\beta}_\mathrm{x}^\top\boldsymbol{x}^n + \boldsymbol{\beta}_\mathrm{z}{}^\top\boldsymbol{z} - y^n) - \lambda\kappa_\mathrm{z}d_\mathrm{z}(\boldsymbol{z}, \boldsymbol{z}^n) + b_j \le s_n \\ & \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \forall n \in [N],\ \forall j \in [J],\ \forall \boldsymbol{z} \in \mathbb{Z} \\ & \|a_j\cdot\boldsymbol{\beta}_\mathrm{x} - \boldsymbol{C}_\mathrm{x}{}^\top\boldsymbol{q}_{nj}\|_* \le \lambda,\ \ |-a_j - \boldsymbol{c}_\mathrm{y}^\top\boldsymbol{q}_{nj}| \le \lambda\kappa_\mathrm{y} \qquad\quad \forall n \in [N],\ \forall j \in [J] \\ & \boldsymbol{\beta} = (\beta_0, \boldsymbol{\beta}_\mathrm{x}, \boldsymbol{\beta}_\mathrm{z}) \in \mathbb{R}^{1+M_\mathrm{x}+M_\mathrm{z}},\ \ \lambda \in \mathbb{R}_+,\ \ \boldsymbol{s} \in \mathbb{R}^N_+ \\ & \boldsymbol{q}_{nj} \in \mathcal{C}^\star,\ n \in [N] \textit{ and } j \in [J].\end{array} \tag{61}$$

Formulation (61) emerges as a special case of our unified representation (57) if we set $\mathcal{I} = [J]$, $\boldsymbol{\theta} = (\boldsymbol{\beta}, \lambda, \boldsymbol{q}_{ni})$, $\boldsymbol{\sigma} = \boldsymbol{s}$, $f_0(\boldsymbol{\theta}, \boldsymbol{\sigma}) = \lambda\epsilon + \frac{1}{N}\sum_{n\in[N]} s_n$, $f_{ni}(e) = a_i e + b_i$, $g_{ni}(\boldsymbol{\theta}, \boldsymbol{\xi}^n_{-\mathbf{z}}; \boldsymbol{z}) = \beta_0 + \boldsymbol{\beta}_\mathrm{x}{}^\top\boldsymbol{x}^n + \boldsymbol{\beta}_\mathrm{z}{}^\top\boldsymbol{z} - y^n$, $h_{ni}(\boldsymbol{\theta}, \boldsymbol{\xi}^n_{-\mathbf{z}}; d_\mathrm{z}(\boldsymbol{z}, \boldsymbol{z}^n)) = -\boldsymbol{q}_{ni}{}^\top(\boldsymbol{d} - \boldsymbol{C}_\mathrm{x}\boldsymbol{x}^n - \boldsymbol{c}_\mathrm{y}\cdot y^n) + \lambda\kappa_\mathrm{z}d_\mathrm{z}(\boldsymbol{z}, \boldsymbol{z}^n)$ and $\Theta = \{\boldsymbol{\theta} : \|a_i\cdot\boldsymbol{\beta}_\mathrm{x} - \boldsymbol{C}^\top\boldsymbol{q}_{ni}\|_* \le \lambda\,\forall(n,i) \in [N]\times\mathcal{I},\ |-a_i - \boldsymbol{c}_\mathrm{y}^\top\boldsymbol{q}_{ni}| \le \lambda\kappa_\mathrm{y}\,\forall(n,i) \in [N]\times\mathcal{I},\ \boldsymbol{q}_{ni} \in \mathcal{C}^*\,\forall(n,i) \in [N]\times\mathcal{I},\ \lambda \in \mathbb{R}_+\}$. One verifies that this choice of $\mathcal{I}$, $\boldsymbol{\theta}$, $\boldsymbol{\sigma}$, $f_0$, $f_{ni}$, $g_{ni}$, $h_{ni}$ and $\Theta$ satisfies the regularity conditions imposed on problem (57) in the main text.

**Corollary 6.** *If* $\boldsymbol{\beta}$ *in problem* (61) *is restricted to a bounded hypothesis set* $\mathcal{H}$*, then* $(\boldsymbol{q}_{ni})_{n,i}$ *and* $\lambda$ *can be restricted to a bounded set as well.*

Proposition 16 covers Wasserstein support vector regression with an $\tau$-insensitive loss function,

$$L(e) = \max\{|e| - \tau,\ 0\}$$

with robustness parameter $\tau \in \mathbb{R}_+$, which is a piece-wise affine convex function with $J = 3$, $a_1 = 1$, $b_1 = -\tau$, $a_2 = -1$, $b_2 = -\tau$, $a_3 = 0$ and $b_3 = 0$. It also covers Wasserstein quantile regression with a pinball loss function,

$$L(e) = \max\{-\tau e,\ (1-\tau)e\}$$

with robustness parameter $0 \le \tau \le 1$, which is a piece-wise affine convex function with $J = 2$, $a_1 = -\tau$, $b_1 = 0$, $a_2 = 1 - \tau$ and $b_2 = 0$. We provide the corresponding reformulations next.

**Corollary 7.** *The first set of inequality constraints in* (61) *can be reformulated as follows.*

*(i) For the $\tau$-insensitive loss function:*

$$\left.\begin{aligned}&\boldsymbol{t}_{n1}{}^\top(\boldsymbol{d}_1-\boldsymbol{C}_1\boldsymbol{x}^n)+\boldsymbol{v}_{n1}{}^\top(\boldsymbol{d}_2-\boldsymbol{C}_2y^n)+(\beta_0+\boldsymbol{\beta}_{\mathrm{x}}{}^\top\boldsymbol{x}^n+\boldsymbol{\beta}_{\mathrm{z}}{}^\top\boldsymbol{z}-y^n)\\&\qquad\qquad-\tau-\lambda\kappa_{\mathrm{z}}d_{\mathrm{z}}(\boldsymbol{z},\boldsymbol{z}^n)\leq s_n\\&\boldsymbol{t}_{n2}{}^\top(\boldsymbol{d}_1-\boldsymbol{C}_1\boldsymbol{x}^n)+\boldsymbol{v}_{n2}{}^\top(\boldsymbol{d}_2-\boldsymbol{C}_2y^n)-(\beta_0+\boldsymbol{\beta}_{\mathrm{x}}{}^\top\boldsymbol{x}^n+\boldsymbol{\beta}_{\mathrm{z}}{}^\top\boldsymbol{z}-y^n)\\&\qquad\qquad-\tau-\lambda\kappa_{\mathrm{z}}d_{\mathrm{z}}(\boldsymbol{z},\boldsymbol{z}^n)\leq s_n\end{aligned}\right\}\ \forall n\in[N],\ \forall\boldsymbol{z}\in\mathbb{Z}$$

*(ii) For the pinball loss function:*

$$\left.\begin{aligned}&\boldsymbol{t}_{n1}{}^\top(\boldsymbol{d}_1-\boldsymbol{C}_1\boldsymbol{x}^n)+\boldsymbol{v}_{n1}{}^\top(\boldsymbol{d}_2-\boldsymbol{C}_2y^n)-\tau(\beta_0+\boldsymbol{\beta}_{\mathrm{x}}{}^\top\boldsymbol{x}^n+\boldsymbol{\beta}_{\mathrm{z}}{}^\top\boldsymbol{z}-y^n)\\&\qquad\qquad-\lambda\kappa_{\mathrm{z}}d_{\mathrm{z}}(\boldsymbol{z},\boldsymbol{z}^n)\leq s_n\\&\boldsymbol{t}_{n2}{}^\top(\boldsymbol{d}_1-\boldsymbol{C}_1\boldsymbol{x}^n)+\boldsymbol{v}_{n2}{}^\top(\boldsymbol{d}_2-\boldsymbol{C}_2y^n)+(1-\tau)(\beta_0+\boldsymbol{\beta}_{\mathrm{x}}{}^\top\boldsymbol{x}^n+\boldsymbol{\beta}_{\mathrm{z}}{}^\top\boldsymbol{z}-y^n)\\&\qquad\qquad-\lambda\kappa_{\mathrm{z}}d_{\mathrm{z}}(\boldsymbol{z},\boldsymbol{z}^n)\leq s_n\end{aligned}\right\}\ \forall n\in[N],\ \forall\boldsymbol{z}\in\mathbb{Z}$$

## 3.C Proofs

**Proof of Observation 6.** The statement follows from similar arguments as in the proof of Theorem 1 by Shafieezadeh-Abadeh et al. (2015). Details are omitted for the sake of brevity. □

**Proof of Proposition 9.** We first show that LB and UB constitute lower and upper bounds on the optimal value of (57) throughout the execution of the algorithm, and then we conclude that Algorithm 9 terminates in finite time with an optimal solution $(\boldsymbol{\theta}^\star,\boldsymbol{\sigma}^\star)$ to problem (57).

Algorithm 9 updates LB to $f_0(\boldsymbol{\theta}^\star,\boldsymbol{\sigma}^\star)$ in each iteration of the while-loop. Since $(\boldsymbol{\theta}^\star,\boldsymbol{\sigma}^\star)$ is an optimal solution to the relaxation of problem (57) that only contains the constraints $\mathcal{W}\subseteq[N]\times\mathcal{I}\times\mathbb{Z}$, LB indeed constitutes a lower bound on the optimal value of (57). Moreover, since no elements are ever removed from $\mathcal{W}$, the sequence of lower bounds LB is monotonic.

To see that UB constitutes an upper bound on the optimal value of (57) throughout the execution of Algorithm 9, we claim that in each iteration, $(\boldsymbol{\theta}^\star,\boldsymbol{\sigma}^\star+\boldsymbol{\vartheta}^\star)$ constitutes a feasible solution to problem (57). Indeed, we have $(\boldsymbol{\theta}^\star,\boldsymbol{\sigma}^\star)\in\Theta\times\mathbb{R}^N_+$ and $\boldsymbol{\vartheta}^\star\in\mathbb{R}^N_+$ by construction, while for all $n\in[N]$, $i\in\mathcal{I}$ and $\boldsymbol{z}\in\mathbb{Z}$, we have that

$$f_{ni}(g_{ni}(\boldsymbol{\theta}^\star,\boldsymbol{\xi}^n_{-\mathbf{z}};\boldsymbol{z}))-h_{ni}(\boldsymbol{\theta}^\star,\boldsymbol{\xi}^n_{-\mathbf{z}};d_{\mathrm{z}}(\boldsymbol{z},\boldsymbol{z}^n))\ \leq\ \sigma^\star_n+\vartheta(n,i)\ \leq\ \sigma^\star_n+\vartheta(n,i(n))\ \leq\ \sigma^\star_n+\vartheta^\star_n,$$

that is, $(\boldsymbol{\theta}^\star,\boldsymbol{\sigma}^\star+\boldsymbol{\vartheta}^\star)$ is indeed feasible in problem (57). Moreover, the sequence of upper bounds UB is monotonic by construction of the updates.

To see that Algorithm 9 terminates in finite time, finally, note that each iteration of the while-loop either adds a new constraint index $(n, i(n), \boldsymbol{z}(n, i(n)))$ to $\mathcal{W}$, or we have $\vartheta(n, i(n)) \leq 0$ for all $n \in [N]$. In the latter case, however, we have $\boldsymbol{\vartheta}^\star = \mathbf{0}$ and thus LB = UB at the end of the iteration. The claim now follows from the fact that the index set $[N] \times \mathcal{I} \times \mathbb{Z}$ is finite. □

**Proof of Theorem 8.** For fixed $(\boldsymbol{\theta}^\star, \boldsymbol{\sigma}^\star)$, finding the most violated constraint index $\boldsymbol{z}(n, i)$ in constraint group $(n, i)$ amounts to solving the combinatorial optimization problem

$$\begin{array}{ll} \underset{\boldsymbol{z}}{\text{maximize}} & f_{ni}(g_{ni}(\boldsymbol{\theta}^\star, \boldsymbol{\xi}^n_{-\mathbf{z}}; \boldsymbol{z})) - h_{ni}(\boldsymbol{\theta}^\star, \boldsymbol{\xi}^n_{-\mathbf{z}}; d_{\mathrm{z}}(\boldsymbol{z}, \boldsymbol{z}^n)) - \sigma^\star_n \\ \text{subject to} & \boldsymbol{z} \in \mathbb{Z}. \end{array}$$

We can solve this problem by solving the $K + 1$ problems

$$\left[ \begin{array}{ll} \underset{\boldsymbol{z}}{\text{maximize}} & f_{ni}(g_{ni}(\boldsymbol{\theta}^\star, \boldsymbol{\xi}^n_{-\mathbf{z}}; \boldsymbol{z})) - h_{ni}(\boldsymbol{\theta}^\star, \boldsymbol{\xi}^n_{-\mathbf{z}}; \delta) - \sigma^\star_n \\ \text{subject to} & d_{\mathrm{z}}(\boldsymbol{z}, \boldsymbol{z}^n) = \delta \\ & \boldsymbol{z} \in \mathbb{Z} \end{array} \right] \quad \forall \delta \in [K] \cup \{0\},$$

where each problem conditions on a fixed number $\delta$ of discrepancies between $\boldsymbol{z}$ and $\boldsymbol{z}^n$, and subsequently choosing any solution $\boldsymbol{z}(\delta)$ that attains the maximum optimal objective value among those $K+1$ problems. Removing constant terms from those $K+1$ problems, we observe that the $\delta$-th problem shares its set of optimal solutions with the problem

$$\begin{array}{ll} \underset{\boldsymbol{z}}{\text{maximize}} & f_{ni}(g_{ni}(\boldsymbol{\theta}^\star, \boldsymbol{\xi}^n_{-\mathbf{z}}; \boldsymbol{z})) \\ \text{subject to} & d_{\mathrm{z}}(\boldsymbol{z}, \boldsymbol{z}^n) = \delta \\ & \boldsymbol{z} \in \mathbb{Z}. \end{array}$$

Since the outer function $f_{ni}$ in the objective function is convex, the objective function is maximized whenever the inner function $g_{ni}$ in the objective function is either maximized or minimized. We thus conclude that the $\delta$-th problem is solved by solving the two problems

$$\left[ \begin{array}{ll} \underset{\boldsymbol{z}}{\text{maximize}} & \mu \cdot g_{ni}(\boldsymbol{\theta}^\star, \boldsymbol{\xi}^n_{-\mathbf{z}}; \boldsymbol{z}) \\ \text{subject to} & d_{\mathrm{z}}(\boldsymbol{z}, \boldsymbol{z}^n) = \delta \\ & \boldsymbol{z} \in \mathbb{Z} \end{array} \right] \quad \forall \mu \in \{\pm 1\}$$

and subsequently choosing the solution $\boldsymbol{z}(\mu, \delta)$ that attains the larger value $f_{ni}(g_{ni}(\boldsymbol{\theta}^\star, \boldsymbol{\xi}^n_{-\mathbf{z}}; \boldsymbol{z}(\mu, \delta))$ among those two solutions (with ties broken arbitrarily). Fixing $\mu$ to either value, adopting the notation for $\boldsymbol{w}$ and $w_0$ of Algorithm 10 and ignoring constant terms, we can write the problem as

$$\begin{array}{ll} \underset{\boldsymbol{z}}{\text{maximize}} & \mu \cdot \Big[ \sum_{m \in [K]} \boldsymbol{w}_m{}^\top \boldsymbol{z}_m \Big] \\ \text{subject to} & d_{\mathrm{z}}(\boldsymbol{z}, \boldsymbol{z}^n) = \delta \\ & \boldsymbol{z} \in \mathbb{Z}. \end{array}$$

The rectangularity of $\mathbb{Z}$ implies that the decisions of this problem admit a decomposition into the selection $\mathcal{M} \subseteq [K]$ of $\delta$ discrete features $m \in [K]$ along which $\boldsymbol{z}_m$ differs from $\boldsymbol{z}_m^n$ and, for those features $m \in [K]$ where $\boldsymbol{z}_m \neq \boldsymbol{z}_m^n$, the choice of $\boldsymbol{z}_m \in \mathbb{Z}(k_m) \setminus \{\boldsymbol{z}_m^n\}$:

$$\begin{array}{ll} \underset{\mathcal{M}}{\text{maximize}} & \max_{\boldsymbol{z}} \left\{ \mu \cdot \Big[ \sum_{m \in [K]} \boldsymbol{w}_m{}^\top \boldsymbol{z}_m \Big] \; : \; \begin{bmatrix} \boldsymbol{z}_m \in \mathbb{Z}(k_m) \setminus \{\boldsymbol{z}_m^n\} & \forall m \in \mathcal{M} \\ \boldsymbol{z}_m = \boldsymbol{z}_m^n & \forall m \in [K] \setminus \mathcal{M} \end{bmatrix} \right\} \\ \text{subject to} & \mathcal{M} \subseteq [K], \;\; |\mathcal{M}| = \delta. \end{array}$$

Noticing that the embedded maximization problem decomposes along the discrete features $m \in [K]$, we can adopt the notation for $\boldsymbol{z}_m^\star$ of Algorithm 10 to obtain the equivalent formulation

$$\begin{array}{ll} \underset{\mathcal{M}}{\text{maximize}} & \Big[ \sum_{m \in \mathcal{M}} \mu \cdot \boldsymbol{w}_m{}^\top \boldsymbol{z}_m^\star \Big] + \Big[ \sum_{m \in [K] \setminus \mathcal{M}} \mu \cdot \boldsymbol{w}_m{}^\top \boldsymbol{z}_m^n \Big] \\ \text{subject to} & \mathcal{M} \subseteq [K], \;\; |\mathcal{M}| = \delta. \end{array}$$

The two summations in the objective function of this problem admit the reformulation

$$\Big[ \sum_{m \in \mathcal{M}} \mu \cdot \boldsymbol{w}_m{}^\top \boldsymbol{z}_m^\star \Big] + \Big[ \sum_{m \in [K] \setminus \mathcal{M}} \mu \cdot \boldsymbol{w}_m{}^\top \boldsymbol{z}_m^n \Big] \; = \; \Big[ \sum_{m \in [K]} \mu \cdot \boldsymbol{w}_m{}^\top \boldsymbol{z}_m^n \Big] + \Big[ \sum_{m \in \mathcal{M}} \mu \cdot \boldsymbol{w}_m{}^\top (\boldsymbol{z}_m^\star - \boldsymbol{z}_m^n) \Big],$$

and ignoring constant terms once more simplifies our optimization problem to

$$\begin{array}{ll} \underset{\mathcal{M}}{\text{maximize}} & \sum_{m \in \mathcal{M}} \mu \cdot \boldsymbol{w}_m{}^\top (\boldsymbol{z}_m^\star - \boldsymbol{z}_m^n) \\ \text{subject to} & \mathcal{M} \subseteq [K], \;\; |\mathcal{M}| = \delta. \end{array}$$

This problem is solved by identifying $\mathcal{M}$ with the indices of the $\delta$ largest elements of the sequence $\mu \cdot \boldsymbol{w}_1{}^\top (\boldsymbol{z}_1^\star - \boldsymbol{z}_1^n), \ldots, \mu \cdot \boldsymbol{w}_K{}^\top (\boldsymbol{z}_K^\star - \boldsymbol{z}_K^n)$, and this problem can be solved by a simple sorting algorithm. One readily verifies that Algorithm 10 adopts the solution approach just described to determine a maximally violated constraint index $\boldsymbol{z}(n, i)$.

The runtime of Algorithm 10, finally, is dominated by determining the $2K$ maximizers $\boldsymbol{z}^\star_m$, $m \in [K]$ and $\mu \in \{\pm 1\}$, which takes time $\mathcal{O}(M_{\mathbf{z}})$ due to the one-hot encoding employed by $\mathbb{Z}$, sorting the $2K$ values $\mu \cdot {\boldsymbol{w}_m}^\top (\boldsymbol{z}^\star_m - \boldsymbol{z}^n_m)$, which takes time $\mathcal{O}(K \log K)$, as well as determining a maximally violated constraint among the $2K+2$ candidates $\boldsymbol{z} \in \mathcal{Z}$, which takes time $\mathcal{O}(KT)$. □

**Proof of Theorem 9.** Recall from Grötschel et al. (1988, Corollary 4.2.7) that problem (57) can be solved to $\delta$-accuracy in polynomial time if *(i)* the problem admits a polynomial time weak separation oracle and if *(ii)* the feasible region of the problem is a circumscribed convex body. We prove both of these properties next.

Fix a convex and compact set $\mathcal{K} \subseteq \mathbb{R}^n$ and a rational number $\delta > 0$. Grötschel et al. (1988, Definition 2.1.13) define a *weak* separation oracle for $\mathcal{K}$ as an algorithm which for any vector $\boldsymbol{q} \in \mathbb{R}^n$ either confirms that $\boldsymbol{q} \in \mathcal{S}(\mathcal{K}, \delta)$, where $\mathcal{S}(\mathcal{K}, \delta) = \{\boldsymbol{r} \in \mathbb{R}^n \, : \, \|\boldsymbol{r} - \boldsymbol{r}'\|_2 \leq \delta \text{ for some } \boldsymbol{r}' \in \mathcal{K}\}$ is the $\delta$-enclosure around $\mathcal{K}$ (*i.e.*, $\boldsymbol{q}$ is *almost* in $\mathcal{K}$), or finds a vector $\boldsymbol{c} \in \mathbb{R}^n$ with $\|\boldsymbol{c}\|_\infty = 1$ such that $\boldsymbol{c}^\top \boldsymbol{p} \leq \boldsymbol{c}^\top \boldsymbol{q} + \delta$ for all $\boldsymbol{p} \in \mathcal{S}(\mathcal{K}, -\delta)$, where $\mathcal{S}(\mathcal{K}, -\delta) = \{\boldsymbol{r} \in \mathcal{K} \, : \, \mathcal{S}(\{\boldsymbol{r}\}, \delta) \subseteq \mathcal{K}\}$ is the $\delta$-interior of $\mathcal{K}$ (*i.e.*, $\boldsymbol{c}$ is an *almost* separating hyperplane). We construct a weak separation oracle for problem (57) as follows. Denote the feasible region of problem (57) by $\mathcal{K} \subseteq \Theta \times \mathbb{R}^N_+$, and fix any $\boldsymbol{q}' = (\boldsymbol{\theta}', \boldsymbol{\sigma}') \in \mathbb{R}^{M_{\boldsymbol{\theta}}} \times \mathbb{R}^N$. If $\sigma'_n < 0$ for any $n \in [N]$, then our algorithm returns the separating hyperplane $\boldsymbol{c}^\top = (\mathbf{0}^\top \;\; -{\mathbf{e}_n}^\top)$. If $\boldsymbol{\theta}' \notin \Theta$, then we can augment the weakly separating hyperplane $\boldsymbol{c} \in \mathbb{R}^{M_{\boldsymbol{\theta}}}$ for the set $\Theta$ to a weakly separating hyperplane $(\boldsymbol{c}^\top \;\; \mathbf{0}^\top)$ for problem (57). Otherwise, we leverage Algorithm 10 to obtain a *strong* separation oracle for $\mathcal{K}$, which Grötschel et al. (1988, Definition 2.1.4) define as an algorithm which for any vector $\boldsymbol{q} \in \mathbb{R}^n$ either confirms that $\boldsymbol{q} \in \mathcal{K}$ or finds a vector $\boldsymbol{c} \in \mathbb{R}^n$ such that $\boldsymbol{c}^\top \boldsymbol{p} < \boldsymbol{c}^\top \boldsymbol{q}$ for all $\boldsymbol{p} \in \mathcal{K}$. To this end, we execute Algorithm 10 for all constraint group indices $(n, i) \in [N] \times \mathcal{I}$ to identify most violated constraint indices $\boldsymbol{z}(n, i)$ along with the constraint violations $\vartheta(n, i)$. If $\vartheta(n, i) \leq 0$ for all $(n, i) \in [N] \times \mathcal{I}$, then $\boldsymbol{q}'$ is feasible in problem (57) and our oracle terminates. If $\vartheta(n^\star, i^\star) > 0$ for some $(n^\star, i^\star) \in [N] \times \mathcal{I}$, on the other hand, then our oracle computes a subgradient $\boldsymbol{c}$ of the function

$$r(\boldsymbol{\theta}, \boldsymbol{\sigma}) \;\;=\;\; f_{n^\star i^\star}(g_{n^\star i^\star}(\boldsymbol{\theta}, \boldsymbol{\xi}^{n^\star}_{-\mathbf{z}}; \boldsymbol{z}(n^\star, i^\star))) - h_{n^\star i^\star}(\boldsymbol{\theta}, \boldsymbol{\xi}^{n^\star}_{-\mathbf{z}}; d_{\mathrm{z}}(\boldsymbol{z}(n^\star, i^\star), \boldsymbol{z}^{n^\star})) - \sigma_{n^\star}$$

at $(\boldsymbol{\theta}, \boldsymbol{\sigma}) = (\boldsymbol{\theta}', \boldsymbol{\sigma}')$. We have

$$\boldsymbol{c}^\top (\boldsymbol{\theta}, \boldsymbol{\sigma}) \;\;\leq\;\; \boldsymbol{c}^\top (\boldsymbol{\theta}', \boldsymbol{\sigma}') + r(\boldsymbol{\theta}, \boldsymbol{\sigma}) - r(\boldsymbol{\theta}', \boldsymbol{\sigma}') \;\;<\;\; \boldsymbol{c}^\top (\boldsymbol{\theta}', \boldsymbol{\sigma}') \qquad \forall (\boldsymbol{\theta}, \boldsymbol{\sigma}) \in \mathcal{K},$$

where the first inequality is due to the definition of subgradients and the fact that $r$ is convex, and the second inequality holds since $r(\boldsymbol{\theta}', \boldsymbol{\sigma}') > 0$ while $r(\boldsymbol{\theta}, \boldsymbol{\sigma}) \leq 0$ for all $(\boldsymbol{\theta}, \boldsymbol{\sigma}) \in \mathcal{K}$. In conclusion, $\boldsymbol{c}$ is a strong separating hyperplane for $(\boldsymbol{\theta}', \boldsymbol{\sigma}')$ as desired. Note that by construction, a strong separation oracle is also a weak separation oracle. Finally, the assumptions in the statement of our theorem, together with Theorem 8, imply that our oracle runs in polynomial time.

We next turn to the claim that the feasible region $\mathcal{K} \subseteq \Theta \times \mathbb{R}_+^N$ of problem (57) is a circumscribed convex body. According to Grötschel et al. (1988, Definition 2.1.16), this is the case whenever $\mathcal{K}$ is a finite-dimensional, full-dimensional, closed and convex subset of a ball whose finite radius we can specify. By construction, $\mathcal{K}$ is finite-dimensional, closed and convex. Moreover, $\mathcal{K}$ is full-dimensional since $\Theta$ is full-dimensional by construction and the constraints have an epigraphical structure. To see that $\mathcal{K}$ can be circumscribed by a ball whose finite radius we can specify, we note that $\Theta$ is bounded by assumption and $\boldsymbol{\sigma}$ is non-negative by construction. Since the objective function $f_0$ in (57) is non-decreasing in $\boldsymbol{\sigma}$, we can without loss of generality include in (57) the additional constraints $\sigma_n \leq \overline{\sigma}_n$ for

$$\overline{\sigma}_n \;=\; \max_{\boldsymbol{\theta} \in \Theta} \max_{i \in \mathcal{I}} \max_{\boldsymbol{z} \in \mathbb{Z}} \left\{ f_{ni}(g_{ni}(\boldsymbol{\theta}, \boldsymbol{\xi}^n_{-\mathbf{z}}; \boldsymbol{z})) - h_{ni}(\boldsymbol{\theta}, \boldsymbol{\xi}^n_{-\mathbf{z}}; d_{\mathrm{z}}(\boldsymbol{z}, \boldsymbol{z}^n)) \right\}, \qquad n \in [N].$$

Note that all $\overline{\sigma}_n$ are finite since $\Theta$ is compact, $\mathcal{I}$ and $\mathbb{Z}$ are finite sets and the objective function is continuous. We thus conclude that $\boldsymbol{\sigma}$ can be bounded as well, and thus $\mathcal{K}$ can indeed be circumscribed by a ball whose finite radius we can specify. □

**Proof of Theorem 10.** Fix any Wasserstein classification or regression instance as described in the statement of the theorem, fix any ambiguity radius $\epsilon > \kappa_{\mathrm{z}} K^{1/p}$, any non-zero numbers $M_{\mathrm{x}}$ of continuous and $K$ of discrete features, and denote the training set as $\{\boldsymbol{\xi}^n\}_{n \in [N]}$ with $\boldsymbol{\xi}^n = (\boldsymbol{x}^n, \boldsymbol{z}^n, y^n)$, $n \in [N]$. Our proof proceeds in three steps. We first derive a closed-form expression for the objective function of the Wasserstein learning problem (54) at a judiciously chosen model $\hat{\boldsymbol{\beta}}$. Our derivation will show that this objective function constitutes the sum of the empirical loss $N^{-1} \cdot \sum_{n \in [N]} l_{\hat{\boldsymbol{\beta}}}(\boldsymbol{\xi}^n)$ and a function $h_{\hat{\boldsymbol{\beta}}}(\{\boldsymbol{\xi}^n\}_{n \in [N]})$. We then construct two sets of data points $\{\hat{\boldsymbol{\xi}}^n\}_{n \in [N]}$ and $\{\check{\boldsymbol{\xi}}^n\}_{n \in [N]}$ at which $h_{\hat{\boldsymbol{\beta}}}(\{\hat{\boldsymbol{\xi}}^n\}_{n \in [N]}) \neq h_{\hat{\boldsymbol{\beta}}}(\{\check{\boldsymbol{\xi}}^n\}_{n \in [N]})$, showing that $h_{\hat{\boldsymbol{\beta}}}(\{\boldsymbol{\xi}^n\}_{n \in [N]})$ exhibits a dependence on the dataset $\{\boldsymbol{\xi}^n\}_{n \in [N]}$ that cannot be replicated by any data-agnostic regularizer $\mathfrak{R}(\hat{\boldsymbol{\beta}})$.

Fix the model $\hat{\boldsymbol{\beta}} = (\hat{\beta}_0, \hat{\boldsymbol{\beta}}_{\mathrm{x}}, \hat{\boldsymbol{\beta}}_{\mathrm{z}})$ with $\hat{\beta}_0 \in \mathbb{R}$ and $\hat{\boldsymbol{\beta}}_{\mathrm{x}} \neq \mathbf{0}$ selected arbitrarily. Fix any discrete feature realization $\boldsymbol{z}^\star \in \mathbb{Z} \setminus \{\boldsymbol{z}^1\}$ as well as the discrete feature coefficients

$$\hat{\beta}_{\mathrm{z},mi} = \begin{cases} -\dfrac{2}{d'} \cdot \kappa_{\mathrm{z}} \mathrm{lip}(L) \cdot \|\hat{\boldsymbol{\beta}}_{\mathrm{x}}\|_* \cdot d_{\mathrm{z}}(\boldsymbol{z}^\star, \boldsymbol{z}^1) & \text{if } z^1_{mi} = 1, \\ \dfrac{2}{d'} \cdot \kappa_{\mathrm{z}} \mathrm{lip}(L) \cdot \|\hat{\boldsymbol{\beta}}_{\mathrm{x}}\|_* \cdot d_{\mathrm{z}}(\boldsymbol{z}^\star, \boldsymbol{z}^1) & \text{otherwise} \end{cases} \qquad \text{for all } m \in [K] \text{ and } i \in [k_m],$$

where $d' = (\mathrm{d}/\mathrm{d}e) L(e) \big|_{e=x'}$ is the derivative of the loss function $L$ at any point $x' \in \mathbb{R}$ where the derivative does not vanish. Such points $x'$ exist due to Rademacher's theorem, which ensures that a Lipschitz continuous function is differentiable almost everywhere, as well as the assumption that the loss function $L$ is non-constant. We make two observations that we will leverage later on in this proof:

*(i)* We have $d' \cdot \hat{\beta}_{\mathrm{z},mi} \cdot (z_{mi} - z^1_{mi}) \geq 0$ for all $\boldsymbol{z} \in \mathbb{Z}$, $m \in [K]$ and $i \in [k_m]$. Indeed, fix any $\boldsymbol{z} \in \mathbb{Z}$, $m \in [K]$ and $i \in [k_m]$. If $z^1_{mi} = 0$, then $d' \cdot \hat{\beta}_{\mathrm{z},mi} = 2 \cdot \kappa_{\mathrm{z}} \mathrm{lip}(L) \cdot \|\hat{\boldsymbol{\beta}}_{\mathrm{x}}\|_* \cdot d_{\mathrm{z}}(\boldsymbol{z}^\star, \boldsymbol{z}^1) \geq 0$ and $z_{mi} - z^1_{mi} \geq 0$. If $z^1_{mi} = 1$, on the other hand, then $d' \cdot \hat{\beta}_{\mathrm{z},mi} = -2 \cdot \kappa_{\mathrm{z}} \mathrm{lip}(L) \cdot \|\hat{\boldsymbol{\beta}}_{\mathrm{x}}\|_* \cdot d_{\mathrm{z}}(\boldsymbol{z}^\star, \boldsymbol{z}^1) \leq 0$ and $z_{mi} - z^1_{mi} \leq 0$. In particular, we have $d' \cdot \hat{\boldsymbol{\beta}}_{\mathrm{z}}{}^\top (\boldsymbol{z} - \boldsymbol{z}^1) \geq 0$ for all $\boldsymbol{z} \in \mathbb{Z}$.

*(ii)* We have $d' \cdot \hat{\boldsymbol{\beta}}_{\mathrm{z}}{}^\top (\boldsymbol{z}^\star - \boldsymbol{z}^1) > \kappa_{\mathrm{z}} \mathrm{lip}(L) \cdot \|\hat{\boldsymbol{\beta}}_{\mathrm{x}}\|_* \cdot d_{\mathrm{z}}(\boldsymbol{z}^\star, \boldsymbol{z}^1)$. Indeed, since $\boldsymbol{z}^\star \neq \boldsymbol{z}^1$, there is at least one $m^\star \in [K]$ and $i^\star \in [k_{m^\star}]$ where $z^\star_{m^\star i^\star} - z^1_{m^\star i^\star} \neq 0$, and thus $d' \cdot \hat{\beta}_{\mathrm{z},m^\star i^\star} \cdot (z^\star_{m^\star i^\star} - z^1_{m^\star i^\star}) = 2 \cdot \kappa_{\mathrm{z}} \mathrm{lip}(L) \cdot \|\hat{\boldsymbol{\beta}}_{\mathrm{x}}\|_* \cdot d_{\mathrm{z}}(\boldsymbol{z}^\star, \boldsymbol{z}^1) > \kappa_{\mathrm{z}} \mathrm{lip}(L) \cdot \|\hat{\boldsymbol{\beta}}_{\mathrm{x}}\|_* \cdot d_{\mathrm{z}}(\boldsymbol{z}^\star, \boldsymbol{z}^1)$. The claim now follows from the fact that $d' \cdot \hat{\beta}_{\mathrm{z},mi} \cdot (z_{mi} - z^1_{mi}) \geq 0$ for all other $m \in [K]$ and $i \in [k_m]$ thanks to our previous observation *(i)*.

Note that piece-wise affine loss functions are Lipschitz continuous. For our selected problem instance, Proposition 11 from Appendix A therefore implies that the objective function of the Wasserstein classification problem becomes

$$\begin{array}{ll} \underset{\lambda, \boldsymbol{s}}{\text{minimize}} & \lambda \epsilon + \dfrac{1}{N} \displaystyle\sum_{n \in [N]} s_n \\ \text{subject to} & \left. \begin{array}{l} l_{\hat{\boldsymbol{\beta}}}(\boldsymbol{x}^n, \boldsymbol{z}, y^n) - \lambda \kappa_{\mathrm{z}} d_{\mathrm{z}}(\boldsymbol{z}, \boldsymbol{z}^n) \leq s_n \\ l_{\hat{\boldsymbol{\beta}}}(\boldsymbol{x}^n, \boldsymbol{z}, -y^n) - \lambda \kappa_{\mathrm{z}} d_{\mathrm{z}}(\boldsymbol{z}, \boldsymbol{z}^n) - \lambda \kappa_{\mathrm{y}} \leq s_n \end{array} \right\} \forall n \in [N],\ \forall \boldsymbol{z} \in \mathbb{Z} \\ & \mathrm{lip}(L) \cdot \|\hat{\boldsymbol{\beta}}_{\mathrm{x}}\|_* \leq \lambda \\ & \lambda \in \mathbb{R}_+,\ \ \boldsymbol{s} \in \mathbb{R}^N_+, \end{array}$$

and Proposition 15 from Appendix B implies that the objective function of the Wasserstein regression problem becomes

$$\begin{array}{lll}
\underset{\lambda,\boldsymbol{s}}{\text{minimize}} & \lambda\epsilon + \dfrac{1}{N}\displaystyle\sum_{n\in[N]} s_n & \\
\text{subject to} & l_{\hat{\boldsymbol{\beta}}}(\boldsymbol{x}^n, \boldsymbol{z}, y^n) - \lambda\kappa_{\mathrm{z}} d_{\mathrm{z}}(\boldsymbol{z}, \boldsymbol{z}^n) \le s_n & \forall n \in [N],\ \forall \boldsymbol{z} \in \mathbb{Z} \\
& \mathrm{lip}(L) \cdot \|\hat{\boldsymbol{\beta}}_{\mathrm{x}}\|_* \le \lambda & \\
& \mathrm{lip}(L) \le \lambda\kappa_{\mathrm{y}} & \\
& \lambda \in \mathbb{R}_+,\ \ \boldsymbol{s} \in \mathbb{R}^N_+. &
\end{array}$$

When $\kappa_{\mathrm{y}} \to \infty$, the second set of constraints in the classification problem and the third constraint in the regression problem become redundant since $\lambda \ge \mathrm{lip}(L) \cdot \|\hat{\boldsymbol{\beta}}_{\mathrm{x}}\|_* > 0$ because $\hat{\boldsymbol{\beta}}_{\mathrm{x}} \ne \mathbf{0}$. For all $n \in [N]$, we can then replace each $s_n$ with a maximum of the left-hand side of the first constraint set over all $\boldsymbol{z} \in \mathbb{Z}$ in either problem to obtain the unified formulation

$$\begin{array}{ll}
\underset{\lambda}{\text{minimize}} & \lambda\epsilon + \dfrac{1}{N}\displaystyle\sum_{n\in[N]} \max_{\boldsymbol{z}\in\mathbb{Z}} \Big\{ l_{\hat{\boldsymbol{\beta}}}(\boldsymbol{x}^n, \boldsymbol{z}, y^n) - \lambda\kappa_{\mathrm{z}} d_{\mathrm{z}}(\boldsymbol{z}, \boldsymbol{z}^n) \Big\} \\
\text{subject to} & \mathrm{lip}(L) \cdot \|\hat{\boldsymbol{\beta}}_{\mathrm{x}}\|_* \le \lambda \\
& \lambda \in \mathbb{R}_+
\end{array}$$

of the objective function of both the classification and the regression problem. Note that the non-negativity of $s_n$ for $n \in [N]$ is preserved in the unified formulation since $d_{\mathrm{z}}(\boldsymbol{z}^n, \boldsymbol{z}^n) = 0$ and $l_{\hat{\boldsymbol{\beta}}}(\boldsymbol{x}^n, \boldsymbol{z}^n, y^n) \ge 0$ for all $n \in [N]$ by definition of the loss function $L$ that underlies $l_{\hat{\boldsymbol{\beta}}}$. We claim that $\lambda^\star = \mathrm{lip}(L) \cdot \|\hat{\boldsymbol{\beta}}_{\mathrm{x}}\|_*$ at optimality. Indeed, any increment $\Delta\lambda > 0$ in $\lambda$ will cause an increase of $\Delta\lambda \cdot \epsilon$ and a maximum decrease of $\Delta\lambda \cdot \kappa_{\mathrm{z}} K^{1/p}$ in the objective function, and we have $\epsilon > \kappa_{\mathrm{z}} K^{1/p}$ by assumption. Hence, the objective function of the Wasserstein classification and regression problem simplifies to

$$\begin{aligned}
f_1(\hat{\boldsymbol{\beta}}, \{\boldsymbol{\xi}^n\}_{n\in[N]}) \;&=\; \lambda^\star\epsilon + \frac{1}{N}\sum_{n\in[N]} \max_{\boldsymbol{z}\in\mathbb{Z}} \Big\{ l_{\hat{\boldsymbol{\beta}}}(\boldsymbol{x}^n, \boldsymbol{z}, y^n) - \lambda^\star\kappa_{\mathrm{z}} d_{\mathrm{z}}(\boldsymbol{z}, \boldsymbol{z}^n) \Big\} \\
&=\; \frac{1}{N}\sum_{n\in[N]} l_{\hat{\boldsymbol{\beta}}}(\boldsymbol{\xi}^n) + h_{\hat{\boldsymbol{\beta}}}(\{\boldsymbol{\xi}^n\}_{n\in[N]}),
\end{aligned}$$

where

$$h_{\hat{\boldsymbol{\beta}}}(\{\boldsymbol{\xi}^n\}_{n\in[N]}) \;=\; \lambda^\star\epsilon + \frac{1}{N}\sum_{n\in[N]} \max_{\boldsymbol{z}\in\mathbb{Z}} \Big\{ l_{\hat{\boldsymbol{\beta}}}(\boldsymbol{x}^n, \boldsymbol{z}, y^n) - l_{\hat{\boldsymbol{\beta}}}(\boldsymbol{x}^n, \boldsymbol{z}^n, y^n) - \lambda^\star\kappa_{\mathrm{z}} d_{\mathrm{z}}(\boldsymbol{z}, \boldsymbol{z}^n) \Big\}.$$

In contrast, the objective function of a generic regularized learning problem has the form

$$f_2(\hat{\boldsymbol{\beta}}, \{\boldsymbol{\xi}^n\}_{n\in[N]}) \;=\; \frac{1}{N}\sum_{n\in[N]} l_{\hat{\boldsymbol{\beta}}}(\boldsymbol{\xi}^n) + \mathfrak{R}(\hat{\boldsymbol{\beta}}).$$

By construction, $\mathfrak{R}(\hat{\boldsymbol{\beta}})$ does not vary with $\{\boldsymbol{\xi}^n\}_{n\in[N]}$. In contrast, we claim that $h_{\hat{\boldsymbol{\beta}}}(\{\boldsymbol{\xi}^n\}_{n\in[N]})$ varies with $\{\boldsymbol{\xi}^n\}_{n\in[N]}$. To this end, we will construct two sets of data points $\{\hat{\boldsymbol{\xi}}^n\}_{n\in[N]}$ and $\{\check{\boldsymbol{\xi}}^n\}_{n\in[N]}$ at which $h_{\hat{\boldsymbol{\beta}}}(\{\hat{\boldsymbol{\xi}}^n\}_{n\in[N]}) \neq h_{\hat{\boldsymbol{\beta}}}(\{\check{\boldsymbol{\xi}}^n\}_{n\in[N]})$. The two datasets $\{\hat{\boldsymbol{\xi}}^n\}_{n\in[N]}$ and $\{\check{\boldsymbol{\xi}}^n\}_{n\in[N]}$ are identical to $\{\boldsymbol{\xi}^n\}_{n\in[N]}$ except for the realization $\boldsymbol{x}^1$ of the continuous features in sample 1. In other words, we have $(\hat{\boldsymbol{x}}^n, \hat{\boldsymbol{z}}^n, \hat{y}^n) = (\check{\boldsymbol{x}}^n, \check{\boldsymbol{z}}^n, \check{y}^n) = (\boldsymbol{x}^n, \boldsymbol{z}^n, y^n)$ for all $n \in [N] \setminus 1$ as well as $(\hat{\boldsymbol{z}}^1, \hat{y}^1) = (\check{\boldsymbol{z}}^1, \check{y}^1) = (\boldsymbol{z}^1, y^1)$.

We select $\hat{\boldsymbol{x}}^1$ such that the maximum in the definition of $h_{\hat{\boldsymbol{\beta}}}$ is strictly positive at sample $n = 1$. To achieve this, we choose $\hat{\boldsymbol{x}}^1$ such that $\hat{y}^1 \cdot [\hat{\beta}_0 + \hat{\boldsymbol{\beta}}_{\rm x}{}^\top \hat{\boldsymbol{x}}^1 + \hat{\boldsymbol{\beta}}_{\rm z}{}^\top \hat{\boldsymbol{z}}^1] = x'$ for classification problems and $\hat{\beta}_0 + \hat{\boldsymbol{\beta}}_{\rm x}{}^\top \hat{\boldsymbol{x}}^1 + \hat{\boldsymbol{\beta}}_{\rm z}{}^\top \hat{\boldsymbol{z}}^1 - \hat{y}^1 = x'$ for regression problems, respectively, where $x'$ was defined earlier as a point where the derivative of the loss function does not vanish. Note that this is always possible since $\hat{\boldsymbol{\beta}}_{\rm x} \neq \mathbf{0}$. We then have

$$\begin{aligned}
&\max_{\boldsymbol{z}\in\mathbb{Z}} l_{\hat{\boldsymbol{\beta}}}(\hat{\boldsymbol{x}}^1, \boldsymbol{z}, \hat{y}^1) - l_{\hat{\boldsymbol{\beta}}}(\hat{\boldsymbol{x}}^1, \hat{\boldsymbol{z}}^1, \hat{y}^1) - \lambda^\star \kappa_{\rm z} d_{\rm z}(\boldsymbol{z}, \hat{\boldsymbol{z}}^1) \\
&\qquad\geq\; l_{\hat{\boldsymbol{\beta}}}(\hat{\boldsymbol{x}}^1, \boldsymbol{z}^\star, \hat{y}^1) - l_{\hat{\boldsymbol{\beta}}}(\hat{\boldsymbol{x}}^1, \hat{\boldsymbol{z}}^1, \hat{y}^1) - \lambda^\star \kappa_{\rm z} d_{\rm z}(\boldsymbol{z}^\star, \hat{\boldsymbol{z}}^1) \\
&\qquad=\; L(x' + \hat{\boldsymbol{\beta}}_{\rm z}{}^\top(\boldsymbol{z}^\star - \hat{\boldsymbol{z}}^1)) - L(x') - \lambda^\star \kappa_{\rm z} d_{\rm z}(\boldsymbol{z}^\star, \hat{\boldsymbol{z}}^1) \\
&\qquad\geq\; \hat{\boldsymbol{\beta}}_{\rm z}{}^\top(\boldsymbol{z}^\star - \hat{\boldsymbol{z}}^1) \cdot \frac{\rm d}{{\rm d}e} L(e) \,\big|_{e=x'} - \lambda^\star \kappa_{\rm z} d_{\rm z}(\boldsymbol{z}^\star, \hat{\boldsymbol{z}}^1) \\
&\qquad=\; d' \cdot \hat{\boldsymbol{\beta}}_{\rm z}{}^\top(\boldsymbol{z}^\star - \hat{\boldsymbol{z}}^1) - \kappa_{\rm z} {\rm lip}(L) \cdot \|\hat{\boldsymbol{\beta}}_{\rm x}\|_* \cdot d_{\rm z}(\boldsymbol{z}^\star, \hat{\boldsymbol{z}}^1) > 0,
\end{aligned}$$

where the first inequality holds since $\boldsymbol{z}^\star \in \mathbb{Z}$, the first identity uses the definitions of $l_{\hat{\boldsymbol{\beta}}}$ and $\hat{\boldsymbol{x}}^1$, the second inequality exploits the convexity of $L$, the second identity uses the definition of $d'$ as well as $\lambda^\star$, and the third inequality follows from our earlier observation *(ii)*. Since the term inside the maximum in the definition of $h_{\hat{\boldsymbol{\beta}}}$ is strictly positive for $\boldsymbol{z} = \boldsymbol{z}^\star$ at sample $n = 1$, the maximum is guaranteed to be positive at sample $n = 1$ as well.

We choose $\check{\boldsymbol{x}}^1$ such that the term inside the maximum in the definition of $h_{\hat{\boldsymbol{\beta}}}$ is strictly negative at sample $n = 1$ for all $\boldsymbol{z} \in \mathbb{Z} \setminus \{\check{\boldsymbol{z}}^1\}$. Consider the case where $d' > 0$; the alternative case where $d' < 0$ follows from analogous arguments. Since $L$ is non-negative and convex, we have that

$$\lim_{x\to-\infty} \frac{\rm d}{{\rm d}e} L(e) \,\big|_{e=x} \leq 0, \tag{62}$$

which in turn implies that for all $\boldsymbol{z} \in \mathbb{Z} \setminus \{\check{\boldsymbol{z}}^1\}$, we have

$$\lim_{x \to -\infty} L(x + \hat{\boldsymbol{\beta}}_{\mathrm{z}}{}^\top(\boldsymbol{z} - \check{\boldsymbol{z}}^1)) - L(x) \;\leq\; \lim_{x \to -\infty} \hat{\boldsymbol{\beta}}_{\mathrm{z}}{}^\top(\boldsymbol{z} - \check{\boldsymbol{z}}^1) \cdot \frac{\mathrm{d}}{\mathrm{d}e} L(e) \,\big|_{e = x + \hat{\boldsymbol{\beta}}_{\mathrm{z}}{}^\top(\boldsymbol{z} - \check{\boldsymbol{z}}^1)} \;\leq\; 0, \tag{63}$$

where the first inequality exploits the fact that the derivative of a convex function is non-decreasing and the second inequality holds because $d' \cdot \hat{\boldsymbol{\beta}}_{\mathrm{z}}{}^\top(\boldsymbol{z} - \check{\boldsymbol{z}}^1) \geq 0$ due to our earlier observation *(i)* and the fact that $\lim_{x \to -\infty} \frac{\mathrm{d}}{\mathrm{d}e} L(e) \,\big|_{e = x + \hat{\boldsymbol{\beta}}_{\mathrm{z}}{}^\top(\boldsymbol{z} - \check{\boldsymbol{z}}^1)} = \lim_{x \to -\infty} \frac{\mathrm{d}}{\mathrm{d}e} L(e) \,\big|_{e = x} \leq 0$ by equation (62). Moreover, for $n = 1$ and for all $\boldsymbol{z} \in \mathbb{Z} \setminus \{\check{\boldsymbol{z}}^1\}$, we have

$$\begin{aligned} &\lim_{\hat{\boldsymbol{\beta}}_{\mathrm{x}}{}^\top \check{\boldsymbol{x}}^1 \to -\infty} l_{\hat{\boldsymbol{\beta}}}(\check{\boldsymbol{x}}^1, \boldsymbol{z}, \check{y}^1) - l_{\hat{\boldsymbol{\beta}}}(\check{\boldsymbol{x}}^1, \check{\boldsymbol{z}}^1, \check{y}^1) - \lambda^\star \kappa_{\mathrm{z}} d_{\mathrm{z}}(\boldsymbol{z}, \check{\boldsymbol{z}}^1) \\ &\qquad = \lim_{x'' \to -\infty} L(x'' + \hat{\boldsymbol{\beta}}_{\mathrm{z}}{}^\top(\boldsymbol{z} - \check{\boldsymbol{z}}^1)) - L(x'') - \lambda^\star \kappa_{\mathrm{z}} d_{\mathrm{z}}(\boldsymbol{z}, \check{\boldsymbol{z}}^1) \\ &\qquad \leq -\lambda^\star \kappa_{\mathrm{z}} d_{\mathrm{z}}(\boldsymbol{z}, \check{\boldsymbol{z}}^1) \;<\; 0, \end{aligned}$$

where the identity applies the change of variables $\check{y}^1 \cdot [\hat{\beta}_0 + \hat{\boldsymbol{\beta}}_{\mathrm{x}}{}^\top \check{\boldsymbol{x}}^1 + \hat{\boldsymbol{\beta}}_{\mathrm{z}}{}^\top \check{\boldsymbol{z}}^1] = x''$ for classification problems and $\hat{\beta}_0 + \hat{\boldsymbol{\beta}}_{\mathrm{x}}{}^\top \check{\boldsymbol{x}}^1 + \hat{\boldsymbol{\beta}}_{\mathrm{z}}{}^\top \check{\boldsymbol{z}}^1 - \check{y}^1 = x''$ for regression problems, respectively, the first inequality uses (63), and the last inequality is due to the fact that $\lambda^\star$, $\kappa_{\mathrm{z}}$ and $d_{\mathrm{z}}(\boldsymbol{z}, \check{\boldsymbol{z}}^1)$ are all strictly positive. Since the term inside the maximum in the definition of $h_{\hat{\boldsymbol{\beta}}}$ is strictly negative at sample $n = 1$ for all $\boldsymbol{z} \in \mathbb{Z} \setminus \{\check{\boldsymbol{z}}^1\}$, the maximum is guaranteed to evaluate to 0 for $n = 1$. Given that the datasets $\{\hat{\boldsymbol{\xi}}^n\}_{n \in [N]}$ and $\{\check{\boldsymbol{\xi}}^n\}_{n \in [N]}$ are identical for all other samples $n \in [N] \setminus 1$, this implies that $h_{\hat{\boldsymbol{\beta}}}(\{\hat{\boldsymbol{\xi}}^n\}_{n \in [N]}) \neq h_{\hat{\boldsymbol{\beta}}}(\{\check{\boldsymbol{\xi}}^n\}_{n \in [N]})$ and confirms that $h_{\hat{\boldsymbol{\beta}}}(\{\boldsymbol{\xi}^n\}_{n \in [N]})$ has a dependency on the dataset $\{\boldsymbol{\xi}^n\}_{n \in [N]}$ that cannot be replicated by any data-agnostic regularizer $\mathfrak{R}(\hat{\boldsymbol{\beta}})$. □

**Proof of Proposition 10.** We prove the statement of the proposition in three steps. We first show that our revised ground metric $d_{\mathrm{u}}$ coincides with the original ground metric $d_{\mathrm{z}}$ over all pairs of discrete feature vectors $\boldsymbol{z}, \boldsymbol{z}' \in \mathbb{Z}$. This implies that the left-hand sides of the constraints indexed by $\boldsymbol{u} \in \mathbb{Z}$ in (56') reduce to their counterparts in (56). In other words, problem (56') is a restriction of (56) since it contains all of the constraints of (56). We then prove that $d_{\mathrm{u}}(\boldsymbol{u}, \boldsymbol{z}')$ is concave in $\boldsymbol{u}$ for every fixed $\boldsymbol{z}' \in \mathbb{Z}$. Our third step leverages this intermediate result to show that every feasible solution to our mixed-feature Wasserstein learning problem (56) is also feasible in the bounded continuous-feature formulation (56').

In view of the first step, fix any $\boldsymbol{z}, \boldsymbol{z}' \in \mathbb{Z}$ and observe that for any $m \in [K]$, we have

$$\frac{1}{2}\|\boldsymbol{z}_m - \boldsymbol{z}'_m\|_1 + \frac{1}{2}|\mathbf{1}^\top(\boldsymbol{z}_m - \boldsymbol{z}'_m)| \;=\; \frac{1}{2}\left\|\left(\begin{smallmatrix}\boldsymbol{z}_m\\ 1-\mathbf{1}^\top\boldsymbol{z}_m\end{smallmatrix}\right) - \left(\begin{smallmatrix}\boldsymbol{z}'_m\\ 1-\mathbf{1}^\top\boldsymbol{z}'_m\end{smallmatrix}\right)\right\|_1 \;=\; \begin{cases} 0 & \text{if } \boldsymbol{z}_m = \boldsymbol{z}'_m, \\ 1 & \text{otherwise.} \end{cases}$$

Here, the first identity applies basic algebraic manipulations, and the second identity holds since $(\boldsymbol{z}_m^\top, 1 - \mathbf{1}^\top\boldsymbol{z}_m)$ and $(\boldsymbol{z}_m'^\top, 1 - \mathbf{1}^\top\boldsymbol{z}'_m)$ are canonic basis vectors in $\mathbb{R}^{k_m}$. We thus conclude that

$$d_{\mathrm{u}}(\boldsymbol{z}, \boldsymbol{z}') = \left(\sum_{m\in[K]} \frac{1}{2}\|\boldsymbol{z}_m - \boldsymbol{z}'_m\|_1 + \frac{1}{2}|\mathbf{1}^\top(\boldsymbol{z}_m - \boldsymbol{z}'_m)|\right)^{1/p} = \sum_{m\in[K]} (\mathbb{1}[\boldsymbol{z}_m \neq \boldsymbol{z}'_m])^{1/p} = d_{\mathrm{z}}(\boldsymbol{z}, \boldsymbol{z}')$$

as claimed.

As for the second statement, fix any $\boldsymbol{z}' \in \mathbb{Z}$. We then find that

$$\begin{aligned} d_{\mathrm{u}}(\boldsymbol{u}, \boldsymbol{z}') \;=\; \Bigg(&\frac{1}{2}\sum_{m\in[K]}\sum_{j\in[k_m-1]} \Big[\mathbb{1}[z'_{mj} = 0]\cdot u_{mj} + \mathbb{1}[z'_{mj} = 1]\cdot(1 - u_{mj})\Big] \;+ \\ &\qquad\qquad \frac{1}{2}\sum_{m\in[K]} \Big[\mathbb{1}[\boldsymbol{z}'_m = \mathbf{0}]\cdot\mathbf{1}^\top\boldsymbol{u}_m + \mathbb{1}[\boldsymbol{z}'_m \neq \mathbf{0}]\cdot(1 - \mathbf{1}^\top\boldsymbol{u}_m)\Big]\Bigg)^{1/p}, \end{aligned}$$

where the first summation evaluates the 1-norm differences between $\boldsymbol{u}_m$ and $\boldsymbol{z}'_m$ and the second summation computes the absolute values in $d_{\mathrm{u}}$, respectively. The above reformulation shows that the mapping $\boldsymbol{u} \mapsto d_{\mathrm{u}}(\boldsymbol{u}, \boldsymbol{z}')$ can be represented as $f(g(\boldsymbol{u}))$, where $f(x) = x^{1/p}$ and $g : \mathbb{U} \rightarrow \mathbb{R}$ is affine. It then follows from §3.2.2 of Boyd and Vandenberghe (2004) that the mapping $\boldsymbol{u} \mapsto d_{\mathrm{u}}(\boldsymbol{u}, \boldsymbol{z}')$ is concave.

In view of the third statement, finally, fix any feasible solution $(\boldsymbol{\beta}, \lambda, \boldsymbol{s})$ to problem (56). Since the objective functions of (56) and (56') coincide, we only need to show that $(\boldsymbol{\beta}, \lambda, \boldsymbol{s})$

satisfies the constraints of problem (56'). To this end, we observe that for all $n \in [N]$, we have

$$\begin{aligned}
& \sup_{(\boldsymbol{x},y)\in\mathbb{X}\times\mathbb{Y}} \{l_{\boldsymbol{\beta}}(\boldsymbol{x},\boldsymbol{u},y) - \lambda\|\boldsymbol{x}-\boldsymbol{x}^n\| - \lambda\kappa_{\mathrm{y}} d_{\mathrm{y}}(y,y^n)\} - \lambda\kappa_{\mathrm{z}} d_{\mathrm{u}}(\boldsymbol{u},\boldsymbol{z}^n) \leq s_n && \forall \boldsymbol{u}\in\mathbb{U} \\
\iff\; & \sup_{\boldsymbol{u}\in\mathbb{U}} \{l_{\boldsymbol{\beta}}(\boldsymbol{x},\boldsymbol{u},y) - \lambda\|\boldsymbol{x}-\boldsymbol{x}^n\| - \lambda\kappa_{\mathrm{y}} d_{\mathrm{y}}(y,y^n)\} - \lambda\kappa_{\mathrm{z}} d_{\mathrm{u}}(\boldsymbol{u},\boldsymbol{z}^n) \leq s_n && \forall (\boldsymbol{x},y)\in\mathbb{X}\times\mathbb{Y} \\
\iff\; & \sup_{\boldsymbol{u}\in\mathbb{U}} \{l_{\boldsymbol{\beta}}(\boldsymbol{x},\boldsymbol{u},y) - \lambda\kappa_{\mathrm{z}} d_{\mathrm{u}}(\boldsymbol{u},\boldsymbol{z}^n)\} - \lambda\|\boldsymbol{x}-\boldsymbol{x}^n\| - \lambda\kappa_{\mathrm{y}} d_{\mathrm{y}}(y,y^n)\} \leq s_n && \forall (\boldsymbol{x},y)\in\mathbb{X}\times\mathbb{Y} \\
\iff\; & \sup_{\boldsymbol{z}\in\mathbb{Z}} \{l_{\boldsymbol{\beta}}(\boldsymbol{x},\boldsymbol{z},y) - \lambda\kappa_{\mathrm{z}} d_{\mathrm{u}}(\boldsymbol{z},\boldsymbol{z}^n)\} - \lambda\|\boldsymbol{x}-\boldsymbol{x}^n\| - \lambda\kappa_{\mathrm{y}} d_{\mathrm{y}}(y,y^n)\} \leq s_n && \forall (\boldsymbol{x},y)\in\mathbb{X}\times\mathbb{Y} \\
\iff\; & \sup_{\boldsymbol{z}\in\mathbb{Z}} \{l_{\boldsymbol{\beta}}(\boldsymbol{x},\boldsymbol{z},y) - \lambda\kappa_{\mathrm{z}} d_{\mathrm{z}}(\boldsymbol{z},\boldsymbol{z}^n)\} - \lambda\|\boldsymbol{x}-\boldsymbol{x}^n\| - \lambda\kappa_{\mathrm{y}} d_{\mathrm{y}}(y,y^n)\} \leq s_n && \forall (\boldsymbol{x},y)\in\mathbb{X}\times\mathbb{Y} \\
\iff\; & \sup_{(\boldsymbol{x},y)\in\mathbb{X}\times\mathbb{Y}} \{l_{\boldsymbol{\beta}}(\boldsymbol{x},\boldsymbol{z},y) - \lambda\kappa_{\mathrm{z}} d_{\mathrm{z}}(\boldsymbol{z},\boldsymbol{z}^n)\} - \lambda\|\boldsymbol{x}-\boldsymbol{x}^n\| - \lambda\kappa_{\mathrm{y}} d_{\mathrm{y}}(y,y^n)\} \leq s_n && \forall \boldsymbol{z}\in\mathbb{Z}.
\end{aligned}$$

Here, the first, the second and the last equivalences reorder terms. As for the third equivalence, recall that $d_{\mathrm{u}}(\boldsymbol{u},\boldsymbol{z}^n)$ is concave in $\boldsymbol{u}$ for every fixed $\boldsymbol{z}^n \in \mathbb{Z}$. Thus, the expression inside the supremum in the third row is convex in $\boldsymbol{u}$, which implies that it attains its maximum at an extreme point of $\mathbb{U}$. The equivalence then follows from the fact that $\mathbb{Z}$ coincides with the extreme points of $\mathbb{U}$. The fourth equivalence, finally, leverages the first step of this proof, which showed that $d_{\mathrm{u}}$ and $d_{\mathrm{z}}$ agree over all pairs of discrete feature vectors $\boldsymbol{z},\boldsymbol{z}' \in \mathbb{Z}$. □

The proof of Proposition 11 utilizes the following lemma, which we state and prove first.

**Lemma 28.** *Assume that the loss function $L$ is convex and Lipschitz continuous. For fixed $\boldsymbol{\alpha},\boldsymbol{x}^n \in \mathbb{R}^{M_{\mathrm{x}}}$, $\alpha_0 \in \mathbb{R}$ and $\lambda \in \mathbb{R}_+$, we have*

$$\sup_{\boldsymbol{x}\in\mathbb{R}^{M_{\mathrm{x}}}} L(\boldsymbol{\alpha}^\top\boldsymbol{x}+\alpha_0) - \lambda\|\boldsymbol{x}-\boldsymbol{x}^n\| = \begin{cases} L(\boldsymbol{\alpha}^\top\boldsymbol{x}^n+\alpha_0) & \text{if } \operatorname{lip}(L)\cdot\|\boldsymbol{\alpha}\|_* \leq \lambda, \\ +\infty & \text{otherwise.} \end{cases} \tag{64}$$

Lemma 28 generalizes Lemma 47 of Shafieezadeh-Abadeh et al. (2019) in that it includes a constant $\alpha_0$ in the argument of the loss function $L$ and that it extends to the case where $\lambda = 0$.

**Proof of Lemma 28.** Consider first the case where $\boldsymbol{\alpha} \neq \boldsymbol{0}$. We conduct the change of variables $\boldsymbol{w} = \boldsymbol{x} + \boldsymbol{d}$, where $\boldsymbol{d} \in \mathbb{R}^{M_{\mathrm{x}}}$ is any vector such that $\boldsymbol{\alpha}^\top\boldsymbol{d} = \alpha_0$. Note that $\boldsymbol{d}$ is guaranteed to exist since $\boldsymbol{\alpha} \neq \boldsymbol{0}$. Setting $\boldsymbol{w}^n = \boldsymbol{x}^n + \boldsymbol{d}$, the left-hand side of (64) can then be written as

$$\begin{aligned}
\sup_{\boldsymbol{x}\in\mathbb{R}^{M_{\mathrm{x}}}} L(\boldsymbol{\alpha}^\top\boldsymbol{x}+\alpha_0) - \lambda\|\boldsymbol{x}-\boldsymbol{x}^n\| &= \sup_{\boldsymbol{w}\in\mathbb{R}^{M_{\mathrm{x}}}} L(\boldsymbol{\alpha}^\top\boldsymbol{w}) - \lambda\|\boldsymbol{w}-\boldsymbol{d}-(\boldsymbol{w}^n-\boldsymbol{d})\| \\
&= \sup_{\boldsymbol{w}\in\mathbb{R}^{M_{\mathrm{x}}}} L(\boldsymbol{\alpha}^\top\boldsymbol{w}) - \lambda\|\boldsymbol{w}-\boldsymbol{w}^n\|
\end{aligned}$$

If $\lambda > 0$, then Lemma 47 of Shafieezadeh-Abadeh et al. (2019) can be directly applied:

$$\sup_{\boldsymbol{w}\in\mathbb{R}^{M_{\mathrm{x}}}} L(\boldsymbol{\alpha}^\top \boldsymbol{w}) - \lambda\|\boldsymbol{w} - \boldsymbol{w}^n\| = \begin{cases} L(\boldsymbol{\alpha}^\top \boldsymbol{w}^n) & \text{if } \mathrm{lip}(L)\cdot\|\boldsymbol{\alpha}\|_* \leq \lambda, \\ +\infty & \text{otherwise,} \end{cases}$$
$$= \begin{cases} L(\boldsymbol{\alpha}^\top \boldsymbol{x}^n + \alpha_0) & \text{if } \mathrm{lip}(L)\cdot\|\boldsymbol{\alpha}\|_* \leq \lambda, \\ +\infty & \text{otherwise.} \end{cases}$$

Note that Lemma 47 of Shafieezadeh-Abadeh et al. (2019) assumes $\lambda > 0$. For the case where $\lambda = 0$, the left-hand side of (64) evaluates to $+\infty$ since $L$ is assumed to be non-constant (*cf.* Section 3.2) and convex. The right-hand side of (64) also evaluates to $+\infty$ since $\mathrm{lip}(L)\cdot\|\boldsymbol{\alpha}\|_* > \lambda$ due to $L$ being non-constant and $\boldsymbol{\alpha} \neq \mathbf{0}$. Hence, the equivalence also extends to the case where $\lambda = 0$.

Now consider the case where $\boldsymbol{\alpha} = \mathbf{0}$. There is no $\boldsymbol{d}$ such that $\boldsymbol{\alpha}^\top\boldsymbol{d} = \alpha_0$ unless $\alpha_0 = 0$. However, when $\boldsymbol{\alpha} = \mathbf{0}$, the left-hand side of (64) has the trivial solution $\boldsymbol{x} = \boldsymbol{x}^n$, and the right-hand side of (64) simplifies to $L(\alpha_0)$ since $0 \leq \lambda$ always holds. Thus, the equivalence still holds, which concludes the proof. □

**Proof of Proposition 11.** The statement can be proven along the lines of the proof of Theorem 14 (ii) by Shafieezadeh-Abadeh et al. (2019) if we leverage Lemma 28 to re-express the embedded maximization over $\boldsymbol{x} \in \mathbb{X}$. Details are omitted for the sake of brevity. □

**Proof of Proposition 12.** For ease of exposition, we adopt the notation introduced after problem (58). By construction of problem (58), $\lambda$ is lower bounded by 0. To see that we can bound $\lambda$ from above as well, we observe that there are optimal solutions for which $\lambda$ does not exceed $\overline{\lambda} = \max\{\overline{\lambda}_1, \overline{\lambda}_2\}$, where

$$\overline{\lambda}_1 = \sup_{\boldsymbol{\beta}\in\mathcal{H}} \mathrm{lip}(L)\cdot\|\boldsymbol{\beta}_{\mathrm{x}}\|_*$$

with the bounded set $\mathcal{H} \subseteq \mathbb{R}^{1+M_{\boldsymbol{x}}+M_{\boldsymbol{z}}}$ containing all admissible hypotheses $\boldsymbol{\beta}$, and

$$\overline{\lambda}_2 = \sup_{\boldsymbol{\beta}\in\mathcal{H}} \max_{n\in[N]} \max_{i\in\mathcal{I}} \max_{\boldsymbol{z}\in\mathbb{Z}} \left\{ \frac{l_{\boldsymbol{\beta}}(\boldsymbol{x}^n, \boldsymbol{z}, iy^n)}{\kappa_{\mathrm{z}} d_{\mathrm{z}}(\boldsymbol{z}, \boldsymbol{z}^n) + \kappa_{\mathrm{y}}\cdot\mathbb{1}[i = -1]} : (\boldsymbol{z}, i) \neq (\boldsymbol{z}^n, 1) \right\}. \tag{65}$$

Indeed, selecting $\lambda \geq \overline{\lambda}_1$ ensures that all hypotheses $\boldsymbol{\beta} \in \mathcal{H}$ are feasible, and selecting $\lambda \geq \overline{\lambda}_2$ implies that for any $\boldsymbol{\beta} \in \mathcal{H}$, all left-hand sides in the first two constraint sets of problem (58) are non-positive, and thus all of these constraints are weakly dominated by the non-negativity constraints on $\boldsymbol{s}$. Note that the numerator in the objective function of (65) is bounded since all suprema and maxima in (65) operate over bounded sets and $l_{\boldsymbol{\beta}}$ is Lipschitz continuous. Moreover, the denominator in the objective function of (65) is bounded from below by a strictly positive quantity since $(\boldsymbol{z}, i) \neq (\boldsymbol{z}^n, 1)$ and $\kappa_{\mathrm{z}}, \kappa_{\mathrm{y}} > 0$. □

**Proof of Corollary 2.** The proof is similar to those of Corollaries 16 and 17 by Shafieezadeh-Abadeh et al. (2019). Details are omitted for the sake of brevity. □

**Proof of Proposition 13.** The statement can be proven along the lines of the proof of Theorem 14 (i) by Shafieezadeh-Abadeh et al. (2019). We omit the details for the sake of brevity. □

**Proof of Proposition 14.** For ease of exposition, we adopt the notation introduced after problem (59). Our proof proceeds in two steps. We first show that without loss of generality, we can impose bounds on each variable $\boldsymbol{q}_{ni}$, $n \in [N]$ and $i \in \mathcal{I}$, and we afterwards show that we can impose non-restrictive bounds on $\lambda$ as well.

To see that each $\boldsymbol{q}_{ni}$ can be bounded, $n \in [N]$ and $i \in \mathcal{I}$, we proceed in two steps. We first argue that ${\boldsymbol{q}_{ni}}^\top(\boldsymbol{d} - \boldsymbol{C}\boldsymbol{x}^n) \geq 0$ for all $n$ and $i$, that is, larger values of $\boldsymbol{q}_{ni}$ weakly increase the left-hand sides in the first two constraint sets of (59). Due to the non-negativity of $f_0$ in $\boldsymbol{s}$ in problem (59), larger values of $\boldsymbol{q}_{ni}$ thus weakly increase the objective function in (59). Non-zero values for $\boldsymbol{q}_{ni}$ can therefore only be optimal if they allow to reduce $\lambda$ via the constraints $\|a_j y^n \cdot \boldsymbol{\beta}_{\mathrm{x}} - t \cdot \boldsymbol{C}^\top \boldsymbol{q}_{ni}\|_* \leq \lambda$. We then derive a bounded set $\mathcal{Q}$ such that $\|a_j y^n \cdot \boldsymbol{\beta}_{\mathrm{x}} - t \cdot \boldsymbol{C}^\top \boldsymbol{q}_{ni}\|_* > \|a_j y^n \cdot \boldsymbol{\beta}_{\mathrm{x}}\|_*$ for all $n \in [N]$, $i = (j,t) \in \mathcal{I}$, all admissible $\boldsymbol{\beta}$ and all $\boldsymbol{q}_{ni} \notin \mathcal{Q}$, that is, the choice $\boldsymbol{q}_{ni} = \boldsymbol{0} \in \mathcal{C}^*$ dominates any feasible choice of $\boldsymbol{q}_{ni}$ outside of $\mathcal{Q}$. In view of the first step, note that $\boldsymbol{d} - \boldsymbol{C}\boldsymbol{x}^n \in \mathcal{C}$ by construction of $\mathbb{X}$ and the fact that $\boldsymbol{x}^n \in \mathbb{X}$. We thus have ${\boldsymbol{q}_{ni}}^\top(\boldsymbol{d} - \boldsymbol{C}\boldsymbol{x}^n) \geq 0$ since $\boldsymbol{q}_{ni} \in \mathcal{C}^*$. As for the second step, consider for each $n$ and $i$ the orthogonal decomposition of the vectors $\boldsymbol{q}_{ni} \in \mathcal{C}^*$ into $\boldsymbol{q}_{ni} = \boldsymbol{q}^0_{ni} + \boldsymbol{q}^+_{ni}$ where $\boldsymbol{q}^0_{ni} \in \mathrm{Null}(\boldsymbol{C}^\top)$ is in the nullspace of $\boldsymbol{C}^\top$ and $\boldsymbol{q}^+_{ni} \in \mathrm{Row}(\boldsymbol{C}^\top)$ is in the row space of $\boldsymbol{C}^\top$. There is a bounded set $\mathcal{Q}^+ \subseteq \mathrm{Row}(\boldsymbol{C}^\top)$ such that $\|a_j y^n \cdot \boldsymbol{\beta}_{\mathrm{x}} - t \cdot \boldsymbol{C}^\top \boldsymbol{q}^+_{ni}\|_* > \|a_j y^n \cdot \boldsymbol{\beta}_{\mathrm{x}}\|_*$ for all $n \in [N]$, $i = (j,t) \in \mathcal{I}$, all admissible $\boldsymbol{\beta}$ and all $\boldsymbol{q}^+_{ni} \notin \mathcal{Q}^+$; note in particular that $\|a_j y^n \cdot \boldsymbol{\beta}_{\mathrm{x}}\|_*$ is bounded due to the assumed boundedness of the hypothesis set and the fact that $\mathcal{I}$ and $\mathbb{Z}$ are finite sets. Similarly, there is a bounded set $\mathcal{Q}^0 \subseteq \mathrm{Null}(\boldsymbol{C}^\top)$ such that for all $\boldsymbol{q}'_{ni} \in \mathrm{Null}(\boldsymbol{C}^\top) \setminus \mathcal{Q}^0$ satisfying $\boldsymbol{q}^+_{ni} + \boldsymbol{q}'_{ni} \in \mathcal{C}^*$ for some $\boldsymbol{q}^+_{ni} \in \mathcal{Q}^+$ there is $\boldsymbol{q}^0_{ni} \in \mathcal{Q}^0$ satisfying $\boldsymbol{q}^+_{ni} + \boldsymbol{q}^0_{ni} \in \mathcal{C}^*$ such that $(\boldsymbol{q}^+_{ni} + \boldsymbol{q}^0_{ni})^\top(\boldsymbol{d} - \boldsymbol{C}\boldsymbol{x}^n) \leq (\boldsymbol{q}^+_{ni} + \boldsymbol{q}'_{ni})^\top(\boldsymbol{d} - \boldsymbol{C}\boldsymbol{x}^n)$, that is, ${\boldsymbol{q}^0_{ni}}^\top \boldsymbol{d} \leq {\boldsymbol{q}'_{ni}}^\top \boldsymbol{d}$, across all $n \in [N]$ and $i = (j,t) \in \mathcal{I}$. Thus, we can without loss of generality restrict the choice of $\boldsymbol{q}_{ni}$ to the bounded set $\mathcal{Q} = \mathcal{C}^* \cap (\mathcal{Q}^0 + \mathcal{Q}^+)$, where the sum is taken in the Minkowski sense.

To see that $\lambda$ can be bounded as well, note that by construction, $\lambda$ is bounded from below by 0. To see that we can bound $\lambda$ from above as well, we observe that there are optimal solutions to problem (59) for which $\lambda$ does not exceed $\overline{\lambda} = \max\{\overline{\lambda}_1, \overline{\lambda}_2\}$, where

$$\overline{\lambda}_1 \;=\; \sup_{\boldsymbol{\beta}\in\mathcal{H}} \max_{n\in[N]} \max_{i=(j,t)\in\mathcal{I}} \max_{\boldsymbol{q}_{ni}\in\mathcal{Q}} \|a_j y^n \cdot \boldsymbol{\beta}_{\mathrm{x}} - t\cdot \boldsymbol{C}^\top \boldsymbol{q}_{ni}\|_*$$

with the bounded set $\mathcal{H} \subseteq \mathbb{R}^{1+M_{\boldsymbol{x}}+M_{\boldsymbol{z}}}$ containing all admissible hypotheses $\boldsymbol{\beta}$, and

$$\overline{\lambda}_2 \;=\; \sup_{\boldsymbol{\beta}\in\mathcal{H}} \max_{n\in[N]} \max_{i=(j,t)\in\mathcal{I}} \max_{\boldsymbol{z}\in\mathbb{Z}} \max_{\boldsymbol{q}_{ni}\in\mathcal{Q}} \left\{ \frac{{\boldsymbol{q}_{ni}}^\top(\boldsymbol{d}-\boldsymbol{C}\boldsymbol{x}^n) + a_j t y^n \cdot (\beta_0 + {\boldsymbol{\beta}_{\mathrm{x}}}^\top \boldsymbol{x}^n + {\boldsymbol{\beta}_{\mathrm{z}}}^\top \boldsymbol{z}) + b_j}{\kappa_{\mathrm{z}} d_{\mathrm{z}}(\boldsymbol{z},\boldsymbol{z}^n) + \kappa_{\mathrm{y}} \cdot \mathbb{1}[t=-1]} \right.$$
$$\left. : (\boldsymbol{z},t) \neq (\boldsymbol{z}^n, 1) \right\}.$$

Indeed, selecting $\lambda \geq \overline{\lambda}_1$ ensures that all hypotheses $\boldsymbol{\beta} \in \mathcal{H}$ are feasible, and selecting $\lambda \geq \overline{\lambda}_2$ implies that for any $\boldsymbol{\beta} \in \mathcal{H}$, all left-hand sides in the first two constraint sets of problem (59) are non-positive, and thus all of these constraints are weakly dominated by the non-negativity constraints on $\boldsymbol{s}$. Note that the numerator in the objective function of $\overline{\lambda}_2$ is bounded since all suprema and maxima operate over bounded sets. Moreover, the denominator in the objective function of $\overline{\lambda}_2$ is bounded from below by a strictly positive quantity since $(\boldsymbol{z},t) \neq (\boldsymbol{z}^n,1)$ and $\kappa_{\mathrm{z}}, \kappa_{\mathrm{y}} > 0$. We thus conclude that variable $\lambda$ in problem (59) can be bounded as well. □

**Proof of Corollary 3.** The proof is similar to that of Corollary 15 by Shafieezadeh-Abadeh et al. (2019). Details are omitted for the sake of brevity. □

The proof of Proposition 15 relies on two lemmas that we will state and prove first.

**Lemma 29.** *The compound norm* $\|[\boldsymbol{\alpha},\nu]\|_{\mathrm{comp}} = \|\boldsymbol{\alpha}\| + \kappa|\nu|$, $\boldsymbol{\alpha} \in \mathbb{R}^{M_{\mathrm{x}}}$, $\nu \in \mathbb{R}$ *and* $\kappa > 0$, *satisfies*

$$\|[\boldsymbol{\alpha},\nu]\|_{\mathrm{comp}^*} = \max\left\{\|\boldsymbol{\alpha}\|_*, \, \frac{|\nu|}{\kappa}\right\}.$$

**Proof of Lemma 29.** By definition of the dual norm, we have that

$$\|[\boldsymbol{\alpha},\nu]\|_{\mathrm{comp}^*} = \left\{ \begin{array}{ll} \underset{(\boldsymbol{x},y)\in\mathbb{R}^{M_{\mathrm{x}}}\times\mathbb{R}}{\text{maximize}} & \boldsymbol{\alpha}^\top \boldsymbol{x} + \nu y \\ \text{subject to} & \|\boldsymbol{x}\| + \kappa|y| \leq 1. \end{array} \right.$$

Note that the optimization problem on the right-hand side satisfies Slater's condition since the feasible region includes the interior point $(\boldsymbol{x},y) = (\boldsymbol{0},0)$. The optimization problem thus has a strong dual. To derive the dual problem, we consider the Lagrange dual function for $\gamma \geq 0$,

$$g(\gamma) = \sup_{(\boldsymbol{x},y)\in\mathbb{R}^{M_\mathrm{x}}\times\mathbb{R}} \quad \boldsymbol{\alpha}^\top \boldsymbol{x} + \nu y - \gamma(\|\boldsymbol{x}\| + \kappa|y| - 1).$$

Note that the maximization is separable over $\boldsymbol{x}$ and $y$. Focusing on the variable $y$, in order for the Lagrange dual function to attain a finite value, we need to have $|\nu| \leq \gamma\kappa$ so that $y^\star = 0$. Under this condition, the Lagrange dual function simplifies to

$$g(\gamma) = \sup_{\boldsymbol{x}\in\mathbb{R}^{M_\mathrm{x}}} \boldsymbol{\alpha}^\top \boldsymbol{x} - \gamma\|\boldsymbol{x}\| + \gamma.$$

We can now apply Lemma 28 with $L$ being the identity to obtain the equivalent reformulation

$$g(\gamma) = \begin{cases} \gamma & \text{if } \|\boldsymbol{\alpha}\|_* \leq \gamma \ , \ |\nu| \leq \gamma\kappa, \\ +\infty & \text{otherwise.} \end{cases}$$

The dual problem therefore is

$$\begin{array}{ll} \underset{\gamma}{\text{minimize}} & \gamma \\ \\ \text{subject to} & |\nu| \leq \gamma\kappa \\ & \|\boldsymbol{\alpha}\|_* \leq \gamma \\ & \gamma \in \mathbb{R}_+, \end{array}$$

which has the optimal solution $\|[\boldsymbol{\alpha},\nu]\|_{\text{comp}^*} = \gamma^\star = \max\{\|\boldsymbol{\alpha}\|_*, \, |\nu|/\kappa\}$. □

**Lemma 30.** *Assume that the loss function $L$ is convex and Lipschitz continuous. For fixed $\boldsymbol{\alpha}, \boldsymbol{x}^n \in \mathbb{R}^{M_\mathrm{x}}$, $\alpha_0, y^n \in \mathbb{R}$, $\kappa > 0$ and $\lambda \in \mathbb{R}_+$, we have*

$$\begin{aligned} &\sup_{(\boldsymbol{x},y)\in\mathbb{R}^{M_\mathrm{x}}\times\mathbb{R}} \quad L(\boldsymbol{\alpha}^\top \boldsymbol{x} - y + \alpha_0) - \lambda\|\boldsymbol{x} - \boldsymbol{x}^n\| - \lambda\kappa|y - y^n| \\ &\qquad\qquad = \begin{cases} L(\boldsymbol{\alpha}^\top \boldsymbol{x}^n - y^n + \alpha_0) & \textit{if } \mathrm{lip}(L)\cdot\max\{\kappa\|\boldsymbol{\alpha}\|_*, 1\} \leq \lambda\kappa, \\ +\infty & \textit{otherwise.} \end{cases} \end{aligned}$$

**Proof of Lemma 30.** Concatenating the variables $\boldsymbol{x}$ and $y$ to $\boldsymbol{w} = [\boldsymbol{x}^\top, y]^\top \in \mathbb{R}^{M_\mathrm{x}+1}$, letting $\boldsymbol{\eta} = [\boldsymbol{\alpha}^\top, -1]^\top$ and $\boldsymbol{w}^n = [(\boldsymbol{x}^n)^\top, y^n]^\top$ and defining the compound norm $\|[\boldsymbol{x}, y]\|_{\text{comp}} = \|\boldsymbol{x}\| +$

$\kappa|y|$, we can write the left-hand side of the equation in the statement of the lemma as

$$\sup_{\boldsymbol{w}\in\mathbb{R}^{M_\mathrm{x}+1}} L(\boldsymbol{\eta}^\top \boldsymbol{w} + \alpha_0) - \lambda\|\boldsymbol{w} - \boldsymbol{w}^n\|_{\mathrm{comp}}.$$

Now we can apply Lemma 28 directly to conclude that

$$\sup_{\boldsymbol{w}\in\mathbb{R}^{M_\mathrm{x}+1}} L(\boldsymbol{\eta}^\top \boldsymbol{w} + \alpha_0) - \lambda\|\boldsymbol{w} - \boldsymbol{w}^n\|_{\mathrm{comp}} = \begin{cases} L(\boldsymbol{\eta}^\top \boldsymbol{w}^n + \alpha_0) & \text{if } \mathrm{lip}(L)\cdot\|\boldsymbol{\eta}\|_{\mathrm{comp}^*} \leq \lambda, \\ +\infty & \text{otherwise.} \end{cases}$$

The statement now follows from Lemma 29, which implies that $\mathrm{lip}(L)\cdot\|\boldsymbol{\eta}\|_{\mathrm{comp}^*} \leq \lambda$ if and only if $\mathrm{lip}(L)\cdot\max\{\|\boldsymbol{\alpha}\|_*, |-1|/\kappa\} \leq \lambda$. □

**Proof of Proposition 15.** The statement can be proven along the lines of the proof of Theorem 4 (ii) by Shafieezadeh-Abadeh et al. (2019) if we leverage Lemma 30 to re-express the embedded maximization over $(\boldsymbol{x}, y) \in \mathbb{R}^{M_\mathrm{x}} \times \mathbb{R}$. Details are omitted for the sake of brevity. □

**Proof of Corollary 4.** The proof is similar to that of Proposition 12. Details are omitted for the sake of brevity. □

**Proof of Corollary 5.** The proof is similar to that of Corollary 5 by Shafieezadeh-Abadeh et al. (2019). Details are omitted for the sake of brevity. □

**Proof of Proposition 16.** The statement can be proven along the lines of the proof of Theorem 4 (i) by Shafieezadeh-Abadeh et al. (2019). We omit the details for the sake of brevity. □

**Proof of Corollary 6.** The proof is similar to that of Proposition 14. Details are omitted for the sake of brevity. □

**Proof of Corollary 7.** The proof is similar to those of Corollaries 6 and 7 by Shafieezadeh-Abadeh et al. (2019). Details are omitted for the sake of brevity. □

# 4 Distributionally and Adversarially Robust Logistic Regression via Intersecting Wasserstein Balls

## Abstract

Adversarially robust optimization (ARO) has emerged as the *de facto* standard for training models that hedge against adversarial attacks in the test stage. While these models are robust against adversarial attacks, they tend to suffer severely from overfitting. To address this issue, some successful methods replace the empirical distribution in the training stage with alternatives including *(i)* a worst-case distribution residing in an ambiguity set, resulting in a distributionally robust (DR) counterpart of ARO; *(ii)* a mixture of the empirical distribution with a distribution induced by an auxiliary (*e.g.*, synthetic, external, out-of-domain) dataset. Inspired by the former, we study the Wasserstein DR counterpart of ARO for logistic regression and show it admits a tractable convex optimization reformulation. Adopting the latter setting, we revise the DR approach by intersecting its ambiguity set with another ambiguity set built using the auxiliary dataset, which offers a significant improvement whenever the Wasserstein distance between the data generating and auxiliary distributions can be estimated. We study the underlying optimization problem, develop efficient solution algorithms, and demonstrate that the proposed method outperforms benchmark approaches on standard datasets.

## 4.1 Introduction

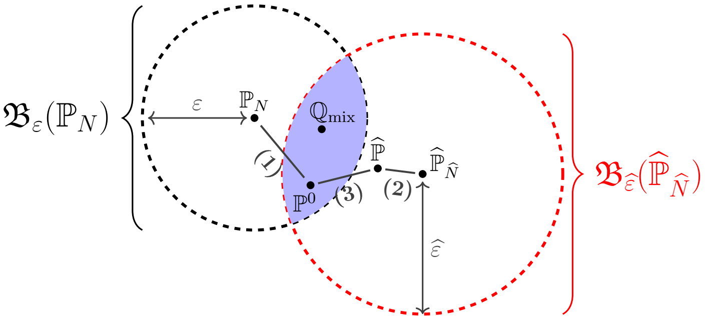
Supervised learning traditionally involves access to a training dataset whose instances are assumed to be independently sampled from a true data-generating distribution (Bishop 2006, Hastie et al. 2009). Optimizing an expected loss for the empirical distribution constructed from such a training set, also known as *empirical risk minimization* (ERM), enjoys several desirable properties in relatively generic settings, including convergence to the true risk minimization problem as the number of training samples increases (Vapnik 1999, Chapter 2). In real-world applications, however, various challenges, such as data scarcity and the existence of adversarial attacks, lead to deteriorated out-of-sample performance for models trained via ERM.

One of the key limitations of ERM, particularly as it is designed to minimize an expected loss for the empirical distribution, emerges from the finite nature of data in practice. This leads ERM to suffer from the 'optimism bias', also known as overfitting (Murphy 2022), or the optimizer's curse (DeMiguel and Nogales 2009, Smith and Winkler 2006), causing deteriorated out-of-sample performance. A popular approach to prevent this phenomenon, *distributionally robust optimization* (DRO; Delage and Ye 2010), optimizes the expected loss for the worst-case distribution residing within a pre-specified ambiguity set.

Another key challenge faced by ERM in practice is adversarial attacks, where an adversary perturbs the observed features during the testing or deployment phase (Szegedy et al. 2014, Goodfellow et al. 2015), also known as evasion corruption at test time (Biggio et al. 2013). For neural networks, the paradigm of *adversarial training* (AT; Madry et al. 2018) is thus designed to provide adversarial robustness by simulating such attacks in the training stage. Several successful variants of AT, specialized to different losses and attacks, have been proposed in the literature to achieve adversarial robustness without significantly reducing performance on training sets (Shafahi et al. 2019, Zhang et al. 2019, Gao et al. 2019, Pang et al. 2022). Some studies (Uesato et al. 2018, Carlini et al. 2019, Wu et al. 2020) investigate the adversarial robustness guarantees of various training algorithms, leading to a research direction focused on heuristic improvements to such models (*e.g.*, Rade and Moosavi-Dezfooli 2022). Our work aligns with another recent line of research (Xing et al. 2022b, Bennouna et al. 2023) on *adversarially robust optimization* (ARO), which constrains ERM to guarantee an *exact*, pre-specified level of adversarial robustness while maximizing training accuracy.

Recently, it has been observed that the two aforementioned notions of robustness can be at odds, as adversarially robust (AR) models suffer from severe overfitting (*robust overfitting*; Raghunathan et al. 2019, Yu et al. 2022, Li and Spratling 2023). Indeed, it is observed that robust overfitting is even more severe than traditional overfitting (Rice et al. 2020). To this end, some works address robust overfitting by revisiting AT algorithms and adding adjustments for better generalization (Chen et al. 2020, Li and Li 2023). In a recent work, Bennouna et al. (2023, Thm 3.2) decompose the error gap of robust overfitting into the statistical error of estimating the true data-generating distribution via the empirical distribution and an adversarial error resulting from the adversarial attacks, hence proposing the simultaneous adoption of DRO and ARO.

In this work, we study logistic regression (LR) for binary classification that is adversarially robust against $\ell_p$-attacks (Croce et al. 2020). To address robust overfitting faced by the adversarially robust LR model, we employ a DRO approach where distributional ambiguity is modeled with the type-1 Wasserstein metric. We base our work on an observation that the worst-case logistic loss under adversarial attacks can be represented as a Lipschitz continuous and convex loss function. This allows us to use existing Wasserstein DRO machinery for Lipschitz losses,



$\mathfrak{B}_{\varepsilon}(\mathbb{P}_N)$
$\varepsilon$
$\mathbb{P}_N$
$\mathbb{Q}_{\text{mix}}$
$\widehat{\mathbb{P}}$
$\widehat{\mathbb{P}}_{\widehat{N}}$
(1)
(3)
(2)
$\mathbb{P}^0$
$\widehat{\varepsilon}$
$\mathfrak{B}_{\widehat{\varepsilon}}(\widehat{\mathbb{P}}_{\widehat{N}})$

Figure 16: *Traditional ARO optimizes the expected adversarial loss over the empirical distribution* $\mathbb{P}_N$ *constructed from* $N$ *i.i.d. samples of the (unknown) true data-generating distribution* $\mathbb{P}^0$*. Replacing* $\mathbb{P}_N$ *with a worst-case distribution in the ball* $\mathfrak{B}_{\varepsilon}(\mathbb{P}_N)$ *gives us its DR counterpart. To reduce the size of this ball, we intersect it with another ball* $\mathfrak{B}_{\widehat{\varepsilon}}(\widehat{P}_{\widehat{N}})$ *while ensuring* $\mathbb{P}^0$ *is still included with high confidence. The latter ball is centered at an empirical distribution* $\widehat{\mathbb{P}}_N$ *constructed from* $\widehat{N}$ *i.i.d. samples of some auxiliary distribution* $\widehat{\mathbb{P}}$*. Recent works using auxiliary data in ARO propose optimizing the expected adversarial loss over a mixture* $\mathbb{Q}_{\text{mix}}$ *of* $\mathbb{P}_N$ *and* $\widehat{P}_{\widehat{N}}$*; we show that this distribution resides in* $\mathfrak{B}_{\varepsilon}(\mathbb{P}_N) \cap \mathfrak{B}_{\widehat{\varepsilon}}(\widehat{\mathbb{P}}_{\widehat{N}})$ *under some conditions.*

and derive an *exact* reformulation of the Wasserstein DR counterpart of adversarially robust LR as a tractable convex problem.

Our main contribution lies in reducing the size of the Wasserstein ambiguity set in the DRO problem mentioned above, in order to create a less conservative problem while preserving the same distributional robustness guarantees. To accomplish this, we draw inspiration from recent work on ARO that leverages auxiliary datasets (*e.g.*, Gowal et al. 2021, Xing et al. 2022b) and revise our DRO problem by intersecting its ambiguity set with another ambiguity set constructed using an auxiliary dataset. Examples of auxiliary data include synthetic data generated from a generative model (*e.g.*, privacy-preserving data release), data in the presence of distributional shifts (*e.g.*, different time periods/regions), noisy data (*e.g.*, measurement errors), or out-of-domain data (*e.g.*, different source); any auxiliary dataset is viable as long as its instances are sampled independently from an underlying data-generating distribution whose Wasserstein distance to the true data-generating distribution is known or can be estimated. Figure 16 illustrates our framework.

The section unfolds as follows. In Section 4.2, we review related literature on DRO and ARO, with a focus on their interactions. We examine the use of auxiliary data in ARO and the intersection of Wasserstein balls in DRO. We discuss open questions for LR to motivate our loss function choice in this work. Section 4.3 gives preliminaries on ERM, ARO, and type-1 Wasserstein DRO. In Section 4.4, we discuss that the adversarial logistic loss can be reformulated as a Lipschitz convex function, enabling the use of Wasserstein DRO machinery for Lipschitz losses. Our main contribution (*cf.* Figure 16) is in Section 4.5, where we provide an explicit reformulation of the distributionally and adversarially robust LR problem over the intersection of two Wasserstein balls, prove the NP-hardness of this problem, and derive a convex relaxation of it. Our work is mainly on *optimization* where we focus on how to solve the underlying problems upon cross-validating Wasserstein ball radii, however, in Section 4.6 we discuss some preliminary statistical approaches to set such radii. We close the section with numerical experiments on standard benchmark datasets in Section 4.7. We borrow the standard notation in DR machine learning, which is elaborated on in our Appendices.

## 4.2 Related Work

**Auxiliary data in ARO.** The use of auxiliary data appears in the ARO literature. In particular, it is shown that additional unlabeled data sampled from the same (Carmon et al. 2019, Xing et al. 2022a) or different (Deng et al. 2021) data-generating distributions could provide adversarial robustness. Sehwag et al. (2022) show that adversarial robustness can be certified even when it is provided for a synthetic dataset as long as the distance between its generator and the true data-generating distribution can be quantified. Gowal et al. (2021) and Xing et al. (2022b) propose optimizing a weighted combination of ARO over empirical and synthetic datasets. We show that the latter approach can be recovered by our model.

**DRO-ARO interactions.** In this work, we optimize ARO against worst-case data-generating distributions residing in an ambiguity set, where the type-1 Wasserstein metric is used for distances since it is arguably the most common choice in machine learning (ML) with Lipschitz losses (Shafieezadeh-Abadeh et al. 2019, Gao 2023). In the literature, it is shown that standard ARO is equivalent to the DRO of the original loss function with a type-$\infty$ Wasserstein metric (Staib and Jegelka 2017, Khim and Loh 2018, Pydi and Jog 2021, Regniez et al. 2022, Frank and Niles-Weed 2024). In other words, in the absence of adversarial attacks, training models adversarially with artificial attacks provide some distributional robustness. Hence, our DR ARO approach can be interpreted as optimizing the logistic loss over the worst-case distribution whose 1-Wasserstein distance is bounded by a pre-specified radius from at least one distribution residing in an $\infty$-Wasserstein ball around the empirical distribution. Conversely, Sinha et al. (2018) discuss that while DRO over Wasserstein balls is intractable for generic losses (*e.g.*, neural networks), its Lagrange relaxation resembles ARO and thus ARO yields a certain degree of (relaxed) distributional robustness (Wu et al. 2020, Bui et al. 2022, Phan et al. 2023).

Such literature suggests that, when there is no concern about the statistical errors caused by using empirical distributions (*e.g.*, in very high-data regimes), one can train DR models to obtain adversarial robustness guarantees. However, as discussed by Bennouna and Van Parys (2022), when statistical errors exist, then we need to be simultaneously robust against adversarial attacks and statistical errors. To the best of our knowledge, there have not been works optimizing a pre-specified level of type-1 Wasserstein distributional robustness (that hedges against overfitting, Kuhn et al. 2019) and adversarial robustness (that hedges against adversarial attacks, Goodfellow et al. 2015) *simultaneously.* To our knowledge, the only approach that considers the exact DR counterpart of ARO is proposed by Bennouna et al. (2023), who model distributional ambiguity with $\varphi$-divergences for neural networks.

**Intersecting ambiguity sets in DRO.** Recent work started to explore the intersection of ambiguity sets for different contexts (Awasthi et al. 2022, Wang et al. 2024) or different metrics (Tanoumand et al. 2023). Our idea of intersecting Wasserstein balls is originated from the "Surround, then Intersect" strategy (Taskesen et al. 2021, §5.2) to train linear regression under sequential domain adaptation in a non-adversarial setting (see the work of Shafahi et al. 2020 and Song et al. 2019 for robustness in domain adaptation/transfer learning). The aforementioned work focuses on the squared loss function with an ambiguity set using the Wasserstein metric developed for the first and second distributional moments. In a recent study, Rychener et al. (2024) generalize most of the previous results and prove that DRO problems over the intersection of two Wasserstein balls admit tractable convex reformulations whenever the loss function is the maximum of concave functions. They also discuss why distributions lying in the intersection of two Wasserstein balls are more natural candidates for the unknown true distribution than those that are Wasserstein barycenters or mixture distributions of the empirical and auxiliary distributions (referred to as heterogeneous data sources; see Example 1, Proposition 2, and Corollary 1).

**Logistic loss in DRO and ARO.** Our choice of LR aligns with the current directions and open questions in the related literature. In the DRO literature, even in the absence of adversarial attacks, the aforementioned work of Taskesen et al. (2021) on the intersection of Wasserstein ambiguity sets is restricted to linear regression. The authors show that this problem admits a tractable convex optimization reformulation, and their proof relies on the properties of the squared loss. Similarly, Rychener et al. (2024) discuss that the logistic loss fails to satisfy the piece-wise concavity assumption and is inherently difficult to optimize over the intersection of Wasserstein balls. We contribute to the DRO literature for adversarial and non-adversarial settings because we show that such a problem would be NP-hard for the logistic loss even without adversarial attacks, and develop specialized approximation techniques. Our problem recovers DR LR (Shafieezadeh-Abadeh et al. 2015, Selvi et al. 2022a) as a special case in the absence of adversarial attacks and auxiliary data. Answering theoretical challenges posed by logistic regression has been useful in answering more general questions in the DRO literature, such as DR LR (Shafieezadeh-Abadeh et al. 2015) leading to DR ML (Shafieezadeh-Abadeh et al. 2019) and mixed-feature DR LR (Selvi et al. 2022a) leading to mixed-feature DR Lipschitz ML (Belbasi et al. 2025). Finally, in the (non-DR) ARO literature, there are recent theory developments on understanding the effect of auxiliary data (*e.g.*, Xing et al. 2022b) specifically for squared and logistic loss functions.

**Single-step adversarial training and single-shot ARO.** Our work proposes a single-shot convex optimization procedure to train logistic models that are both adversarially and distributionally robust. Although the terminology may resemble the recent work on *single-step adversarial training* for neural networks (Wong et al. 2020, Lin et al. 2023, 2024), the two approaches operate differently. Single-step adversarial training generates adversarial perturbations using a single gradient computation at each iteration of iterative model training and improves robustness through updates, with performance typically assessed at intermediate checkpoints. In contrast, our method solves a convex optimization problem once to obtain a model that satisfies both forms of robustness by design. To enable tractability, this approach leverages the convexity and Lipschitz continuity of the loss function under adversarial attacks, which hold for the logistic loss function. While single-step adversarial training applies broadly to general classes of models such as neural networks, our framework offers a complementary, optimization-based perspective in the logistic regression model, where structural properties can be fully exploited.

## 4.3 Preliminaries

We consider a binary classification problem where an instance is modeled as $(\boldsymbol{x}, y) \in \Xi := \mathbb{R}^n \times \{-1, +1\}$ and the labels depend on the features via $\mathrm{Prob}[y \mid \boldsymbol{x}] = [1 + \exp(-y \cdot \boldsymbol{\beta}^\top \boldsymbol{x})]^{-1}$ for some $\boldsymbol{\beta} \in \mathbb{R}^n$; its associated loss is the *logloss* $\ell_{\boldsymbol{\beta}}(\boldsymbol{x}, y) := \log(1 + \exp(-y \cdot \boldsymbol{\beta}^\top \boldsymbol{x}))$.

**Empirical risk minimization.** Let $\mathcal{P}(\Xi)$ denote the set of distributions supported on $\Xi$ and $\mathbb{P}^0 \in \mathcal{P}(\Xi)$ denote the true data-generating distribution. One wants to minimize the expected logloss over $\mathbb{P}^0$, that is

$$\inf_{\boldsymbol{\beta} \in \mathbb{R}^n} \mathbb{E}_{\mathbb{P}^0}[\ell_{\boldsymbol{\beta}}(\boldsymbol{x}, y)]. \tag{RM}$$

In practice, $\mathbb{P}^0$ is hardly ever known, and one resorts to the empirical distribution $\mathbb{P}_N = \frac{1}{N}\sum_{i\in[N]}\delta_{\boldsymbol{\xi}^i}$ where $\boldsymbol{\xi}^i = (\boldsymbol{x}^i, y^i)$, $i \in [N]$, are i.i.d. samples from $\mathbb{P}^0$ and $\delta_{\boldsymbol{\xi}}$ denotes the Dirac distribution supported on $\boldsymbol{\xi}$. The empirical risk minimization (ERM) problem is thus

$$\inf_{\boldsymbol{\beta}\in\mathbb{R}^n} \mathbb{E}_{\mathbb{P}_N}[\ell_{\boldsymbol{\beta}}(\boldsymbol{x}, y)]. \tag{ERM}$$

**Distributionally robust optimization.** To be able to define a distance between distributions, we first define the following feature-label metric on $\Xi$.

**Definition 5.** *The distance between instances $\boldsymbol{\xi} = (\boldsymbol{x}, y) \in \Xi$ and $\boldsymbol{\xi'} = (\boldsymbol{x'}, y') \in \Xi$ for $\kappa \geq 0$ and $q \geq 1$ is*

$$d(\boldsymbol{\xi}, \boldsymbol{\xi'}) = \|\boldsymbol{x} - \boldsymbol{x'}\|_q + \kappa \cdot \mathbb{1}[y \neq y'].$$

Using this metric, we define the Wasserstein distance.

**Definition 6.** *The type-1 Wasserstein distance between distributions $\mathbb{Q}, \mathbb{Q}' \in \mathcal{P}(\Xi)$ is defined as*

$$\mathrm{W}(\mathbb{Q}, \mathbb{Q}') = \inf_{\Pi\in\mathcal{C}(\mathbb{Q},\mathbb{Q}')} \left\{ \int_{\Xi\times\Xi} d(\boldsymbol{\xi}, \boldsymbol{\xi'})\Pi(\mathrm{d}\boldsymbol{\xi}, \mathrm{d}\boldsymbol{\xi'}) \right\},$$

*where $\mathcal{C}(\mathbb{Q}, \mathbb{Q}')$ is the set of couplings of $\mathbb{Q}$ and $\mathbb{Q}'$.*

In finite-data settings, the distance between the true data-generating distribution and the empirical distribution is upper-bounded by some $\epsilon > 0$. The Wasserstein DRO problem is thus defined as

$$\inf_{\boldsymbol{\beta}\in\mathbb{R}^n} \sup_{\mathbb{Q}\in\mathfrak{B}_\varepsilon(\mathbb{P}_N)} \mathbb{E}_{\mathbb{Q}}[\ell_{\boldsymbol{\beta}}(\boldsymbol{x}, y)], \tag{DRO}$$

where $\mathfrak{B}_\varepsilon(\mathbb{P}) := \{\mathbb{Q} \in \mathcal{P}(\Xi) : \mathrm{W}(\mathbb{Q}, \mathbb{P}) \leq \varepsilon\}$ denotes the Wasserstein ball centered at $\mathbb{P} \in \mathcal{P}(\Xi)$ with radius $\varepsilon$. We refer to Mohajerin Esfahani and Kuhn (2018) and Kuhn et al. (2019) for the properties of DRO and estimating $\varepsilon$.

**Adversarially robust optimization.** The goal of adversarial robustness is to provide robustness against adversarial attacks (Goodfellow et al. 2015). An adversarial attack, in the widely studied $\ell_p$-noise setting (Croce et al. 2020), perturbs the features of the test instances $(\boldsymbol{x}, y)$ by adding additive noise $\boldsymbol{z}$ to $\boldsymbol{x}$. The adversary chooses the noise vector $\boldsymbol{z}$, subject to $\|\boldsymbol{z}\|_p \leq \alpha$, so as to maximize the loss $\ell_{\boldsymbol{\beta}}(\boldsymbol{x} + \boldsymbol{z}, y)$ associated with this perturbed test instance.

Therefore, ARO solves the following optimization problem in the training stage to hedge against adversarial perturbations at the test stage:

$$\inf_{\boldsymbol{\beta}\in\mathbb{R}^n} \mathbb{E}_{\mathbb{P}_N}[\sup_{\boldsymbol{z}:\|\boldsymbol{z}\|_p\leq\alpha} \{\ell_{\boldsymbol{\beta}}(\boldsymbol{x}+\boldsymbol{z},y)\}]. \tag{ARO}$$

ARO reduces to ERM when $\alpha = 0$. Note that ARO is identical to feature robust training (Bertsimas et al. 2019b) which is not motivated by adversarial attacks, but by the presence of noisy observations in the training set (Ben-Tal et al. 2009, Gorissen et al. 2015).

**DRO-ARO connection.** A connection between ARO and DRO is noted in the literature (Staib and Jegelka 2017, Proposition 3.1, Khim and Loh 2018, Lemma 22, Pydi and Jog 2021, Lemma 5.1, Regniez et al. 2022, Proposition 2.1, Frank and Niles-Weed 2024, Lemma 3, and Bennouna et al. 2023, §3). Namely, problem ARO is equivalent to a DRO problem

$$\inf_{\boldsymbol{\beta}\in\mathbb{R}^n} \sup_{\mathbb{Q}\in\mathfrak{B}^\infty_\alpha(\mathbb{P}_N)} \mathbb{E}_{\mathbb{Q}}[\ell_{\boldsymbol{\beta}}(\boldsymbol{x},y)], \tag{66}$$

where the ambiguity set $\mathfrak{B}^\infty_\alpha(\mathbb{P}_N)$ is a type-$\infty$ Wasserstein ball (Givens and Shortt 1984) with radius $\alpha$. Hence, in non-adversarial settings, ARO provides robustness with respect to the type-$\infty$ Wasserstein distance. In the case of adversarial attacks, it suffers from robust overfitting as discussed earlier. To address this issue, one straightforward approach is to revisit (66) and replace $\alpha$ with some $\alpha' > \alpha$. This approach, however, does not provide improvements for the out-of-sample performance since *(i)* the type-$\infty$ Wasserstein distance employed in problem (66) uses a metric on the feature space, ignoring labels; *(ii)* type-$\infty$ Wasserstein distances do not provide strong out-of-sample performances in ML (unlike, *e.g.*, the type-1 Wasserstein distance) since the required radii to provide meaningful robustness guarantees are typically too large (Bennouna and Van Parys 2022, §1.2.2, and references therein). We thus study the type-1 Wasserstein counterpart of ARO, which we initiate in the next section.

### 4.4 Distributionally and Adversarially Robust LR

Here we derive the Wasserstein DR counterpart of ARO that will set the ground for our main result in the next section. We impose the following assumption.

**Assumption 4.** *We are given a finite $\varepsilon > 0$ value satisfying* $\mathrm{W}(\mathbb{P}^0, \mathbb{P}_N) \leq \varepsilon$.

The assumption implies that we know an $\varepsilon > 0$ value satisfying $\mathbb{P}^0 \in \mathfrak{B}_\varepsilon(\mathbb{P}_N)$. Typically, however, $\varepsilon$ is either estimated through cross-validation or finite sample statistics, with the assumption then regarded as holding with high confidence (see §4.6 for a review of related results we can borrow). The distributionally and adversarially robust LR problem is thus:

$$\inf_{\boldsymbol{\beta}\in\mathbb{R}^n} \sup_{\mathbb{Q}\in\mathfrak{B}_\varepsilon(\mathbb{P}_N)} \mathbb{E}_{\mathbb{Q}}[\sup_{\boldsymbol{z}:\|\boldsymbol{z}\|_p\leq\alpha} \{\ell_{\boldsymbol{\beta}}(\boldsymbol{x}+\boldsymbol{z},y)\}]. \tag{DR-ARO}$$

By employing a simple duality trick for the inner sup-problem, as commonly applied in robust optimization (Ben-Tal et al. 2009, Bertsimas and den Hertog 2022), we can represent DR-ARO as a standard non-adversarial DRO problem with an updated loss function, which we name the *adversarial loss.*

**Observation 7.** *Problem* DR-ARO *is equivalent to*

$$\inf_{\boldsymbol{\beta}\in\mathbb{R}^n} \sup_{\mathbb{Q}\in\mathfrak{B}_\varepsilon(\mathbb{P}_N)} \mathbb{E}_{\mathbb{Q}}[\ell^\alpha_{\boldsymbol{\beta}}(\boldsymbol{x},y)],$$

*where the* adversarial loss $\ell^\alpha_{\boldsymbol{\beta}}$ *is defined as*

$$\ell^\alpha_{\boldsymbol{\beta}}(\boldsymbol{x},y) := \log(1+\exp(-y\cdot\boldsymbol{\beta}^\top\boldsymbol{x}+\alpha\cdot\|\boldsymbol{\beta}\|_{p^\star})),$$

*for $p^\star$ satisfying $1/p+1/p^\star=1$. The univariate representation $L^\alpha(z) := \log(1+\exp(-z+\alpha\cdot\|\boldsymbol{\beta}\|_{p^\star}))$ of $\ell^\alpha_{\boldsymbol{\beta}}$ is convex and has a Lipschitz modulus of* 1*.*

As a corollary of Observation 7, we can directly employ the techniques proposed by Shafieezadeh-Abadeh et al. (2019) to dualize the inner sup-problem of DR-ARO and obtain a tractable reformulation.

**Corollary 8.** *Problem* DR-ARO *admits the following tractable convex optimization reformulation:*

$$\begin{aligned}
\inf_{\boldsymbol{\beta},\lambda,\boldsymbol{s}} \quad & \varepsilon\lambda + \frac{1}{N}\sum_{i=1}^N s_i \\
\text{s.t.} \quad & \ell^\alpha_{\boldsymbol{\beta}}(\boldsymbol{x}^i,y^i) \leq s_i & \forall i\in[N] \\
& \ell^\alpha_{\boldsymbol{\beta}}(\boldsymbol{x}^i,-y^i) - \lambda\kappa \leq s_i & \forall i\in[N] \\
& \|\boldsymbol{\beta}\|_{q^\star} \leq \lambda \\
& \boldsymbol{\beta}\in\mathbb{R}^n,\ \lambda\geq 0,\ \boldsymbol{s}\in\mathbb{R}^N_+,
\end{aligned}$$

*for $q^\star$ satisfying $1/q+1/q^\star=1$.*

The constraints of this problem are exponential cone representable (derivation is in the appendices) and for $q \in \{1, 2, \infty\}$, the yielding problem can be solved with the exponential cone solver of MOSEK (MOSEK ApS 2023) in polynomial time (Nesterov 2018).

## 4.5 Main Result

In §4.4 we discussed the traditional DRO setting where we have access to an empirical distribution $\mathbb{P}_N$ constructed from $N$ i.i.d. samples of the true data-generating distribution $\mathbb{P}^0$, and we are given (or we estimate) some $\varepsilon$ so that $\mathbb{P}^0 \in \mathfrak{B}_\varepsilon(\mathbb{P}_N)$. Recently in DRO literature, it became a key focus to study the case where we have access to an additional auxiliary empirical distribution $\widehat{\mathbb{P}}_{\widehat{N}}$ constructed from $\widehat{N}$ i.i.d. samples $\widehat{\boldsymbol{\xi}}^j = (\widehat{\boldsymbol{x}}^j, \widehat{y}^j),\ j \in [\widehat{N}]$, of some other distribution $\widehat{\mathbb{P}}$; given the increasing availability of useful auxiliary data in the ARO domain, we explore this direction here. We start with the following assumption.

**Assumption 5.** *We are given finite* $\varepsilon, \widehat{\varepsilon} > 0$ *values satisfying* $\mathrm{W}(\mathbb{P}^0, \mathbb{P}_N) \leq \varepsilon$ *and* $\mathrm{W}(\mathbb{P}^0, \widehat{\mathbb{P}}_{\widehat{N}}) \leq \widehat{\varepsilon}$.

The assumption implies that we know $\varepsilon, \widehat{\varepsilon} > 0$ values satisfying $\mathbb{P}^0 \in \mathfrak{B}_\varepsilon(\mathbb{P}_N) \cap \mathfrak{B}_{\widehat{\varepsilon}}(\widehat{\mathbb{P}}_{\widehat{N}})$. In practice, this assumption is ensured to hold with high confidence by estimating the $\varepsilon$ and $\widehat{\varepsilon}$ values; methods across various domains which we can adopt are reviewed in §4.6. We want to optimize the adversarial loss over the intersection $\mathfrak{B}_\varepsilon(\mathbb{P}_N) \cap \mathfrak{B}_{\widehat{\varepsilon}}(\widehat{\mathbb{P}}_{\widehat{N}})$:

$$\inf_{\boldsymbol{\beta} \in \mathbb{R}^n} \sup_{\mathbb{Q} \in \mathfrak{B}_\varepsilon(\mathbb{P}_N) \cap \mathfrak{B}_{\widehat{\varepsilon}}(\widehat{\mathbb{P}}_{\widehat{N}})} \mathbb{E}_{\mathbb{Q}}[\ell^\alpha_{\boldsymbol{\beta}}(\boldsymbol{x}, y)]. \tag{Inter-ARO}$$

Note that Assumption 5 implies that $\varepsilon$ and $\widehat{\varepsilon}$ guarantee that $\mathfrak{B}_\varepsilon(\mathbb{P}_N) \cap \mathfrak{B}_{\widehat{\varepsilon}}(\widehat{\mathbb{P}}_{\widehat{N}})$ is nonempty, and problem Inter-ARO is thus feasible. This problem is expected to outperform DR-ARO as the ambiguity set is smaller while still including $\mathbb{P}^0$. However, problem Inter-ARO is challenging to solve even in the absence of adversarial attacks ($\alpha = 0$) as we reviewed in §4.2. To address this challenge, we first reformulate Inter-ARO as a semi-infinite optimization problem with finitely many variables.

**Proposition 17.** Inter-ARO *is equivalent to:*

$$\begin{aligned}
\inf_{\substack{\boldsymbol{\beta},\lambda,\widehat{\lambda}\\ \boldsymbol{s},\widehat{\boldsymbol{s}}}} \quad & \varepsilon\lambda + \widehat{\varepsilon}\widehat{\lambda} + \frac{1}{N}\sum_{i=1}^{N} s_i + \frac{1}{\widehat{N}}\sum_{j=1}^{\widehat{N}} \widehat{s}_j \\
\text{s.t.} \quad & \sup_{\boldsymbol{x}\in\mathbb{R}^n} \{\ell^{\alpha}_{\boldsymbol{\beta}}(\boldsymbol{x}, l) - \lambda\|\boldsymbol{x}^i - \boldsymbol{x}\|_q - \widehat{\lambda}\|\widehat{\boldsymbol{x}}^j - \boldsymbol{x}\|_q\} \leq s_i + \frac{\kappa(1-ly^i)}{2}\lambda + \widehat{s}_j + \frac{\kappa(1-l\widehat{y}^j)}{2}\widehat{\lambda} \\
& \qquad\qquad\qquad\qquad\qquad\qquad \forall (i,j,l) \in [N]\times[\widehat{N}]\times\{-1,1\} \\
& \boldsymbol{\beta}\in\mathbb{R}^n,\ \lambda \geq 0,\ \widehat{\lambda}\geq 0,\ \boldsymbol{s}\in\mathbb{R}^N_+,\ \widehat{\boldsymbol{s}}\in\mathbb{R}^{\widehat{N}}_+.
\end{aligned}$$

Even though this problem recovers the tractable problem DR-ARO as $\widehat{\varepsilon}\to\infty$, it is NP-hard in the finite radius settings. We reformulate Inter-ARO as an adjustable robust optimization problem (Ben-Tal et al. 2004, Yanıkoğlu et al. 2019), and borrow tools from this literature to obtain the following result.

**Proposition 18.** Inter-ARO *is equivalent to an adjustable RO problem with $\mathcal{O}(N\cdot\widehat{N})$ two-stage robust constraints, which is NP-hard even when $N = \widehat{N} = 1$.*

The adjustable RO literature has developed a rich arsenal of relaxations that can be leveraged for Inter-ARO. We adopt the 'static relaxation technique' (Bertsimas et al. 2015) to restrict the feasible region of Inter-ARO and obtain a tractable approximation.

**Theorem 11** (main)**.** *The following convex optimization problem is a feasible relaxation of* Inter-ARO*:*

$$\begin{aligned}
\inf_{\substack{\boldsymbol{\beta},\lambda,\widehat{\lambda},\boldsymbol{s},\widehat{\boldsymbol{s}}\\ \boldsymbol{z}^+_{ij},\boldsymbol{z}^-_{ij}}} \quad & \varepsilon\lambda + \widehat{\varepsilon}\widehat{\lambda} + \frac{1}{N}\sum_{i=1}^{N} s_i + \frac{1}{\widehat{N}}\sum_{j=1}^{\widehat{N}} \widehat{s}_j \\
\text{s.t.} \quad & L^{\alpha}(l\cdot\boldsymbol{\beta}^\top\boldsymbol{x}^i + \boldsymbol{z}^{l\top}_{ij}(\widehat{\boldsymbol{x}}^j - \boldsymbol{x}^i)) \leq s_i + \frac{\kappa(1-ly^i)}{2}\lambda + \widehat{s}_j + \frac{\kappa(1-l\widehat{y}^j)}{2}\widehat{\lambda} \\
& \qquad\qquad\qquad\qquad\qquad \forall (i,j,l) \in [N]\times[\widehat{N}]\times\{-1,1\} \\
& \|l\boldsymbol{\beta} - \boldsymbol{z}^l_{ij}\|_{q^\star} \leq \lambda \\
& \qquad\qquad\qquad\qquad\qquad \forall (i,j,l) \in [N]\times[\widehat{N}]\times\{-1,1\} \\
& \|\boldsymbol{z}^l_{ij}\|_{q^\star} \leq \widehat{\lambda} \\
& \qquad\qquad\qquad\qquad\qquad \forall (i,j,l) \in [N]\times[\widehat{N}]\times\{-1,1\} \\
& \boldsymbol{\beta}\in\mathbb{R}^n,\ \lambda \geq 0,\ \widehat{\lambda}\geq 0,\ \boldsymbol{s}\in\mathbb{R}^N_+,\ \widehat{\boldsymbol{s}}\in\mathbb{R}^{\widehat{N}}_+ \\
& \boldsymbol{z}^l_{ij}\in\mathbb{R}^n,\ (i,j,l)\in[N]\times[\widehat{N}]\times\{-1,1\}.
\end{aligned} \tag{Inter-ARO$^\star$}$$

Similarly to DR-ARO, the constraints of Inter-ARO$^\star$ are exponential cone representable (*cf.* appendices).

Recall that for $\widehat{\varepsilon}$ large enough, Inter-ARO reduces to DR-ARO. The following corollary shows that, despite Inter-ARO$^\star$ being a relaxation of Inter-ARO, a similar property holds. That is, "not learning anything from auxiliary data" remains feasible: the static relaxation does not force learning from $\widehat{\mathbb{P}}_{\widehat{N}}$, and it learns from auxiliary data only if the objective improves.

**Corollary 9.** Feasibility of ignoring auxiliary data: *Any feasible solution* $(\boldsymbol{\beta}, \lambda, \boldsymbol{s})$ *of* DR-ARO *can be used to recover a feasible solution* $(\boldsymbol{\beta}, \lambda, \widehat{\lambda}, \boldsymbol{s}, \widehat{\boldsymbol{s}}, \boldsymbol{z}_{ij}^+, \boldsymbol{z}_{ij}^-)$ *for* Inter-ARO$^\star$ *with* $\widehat{\lambda} = 0$, $\widehat{\boldsymbol{s}} = \mathbf{0}$, *and* $\boldsymbol{z}_{ij}^+ = \boldsymbol{z}_{ij}^- = \mathbf{0}$.
Convergence to Inter-ARO*: The optimal value of* Inter-ARO$^\star$ *converges to the optimal value of* Inter-ARO*, with the same set of* $\boldsymbol{\beta}$ *solutions, as* $\widehat{\varepsilon} \to \infty$.

In light of Corollary 9, Appendix 4.D.1 discusses that some simulations in our numerical experiments chose not to incorporate the auxiliary data by setting a sufficiently large $\widehat{\varepsilon}$. We close the section by discussing how Inter-ARO can recover some problems in the DRO and ARO literature. Firstly, recall that Inter-ARO can ignore the auxiliary data once $\widehat{\varepsilon}$ is set large enough, reducing this problem to DR-ARO. Moreover, notice that $\alpha = 0$ reduces $\ell_{\boldsymbol{\beta}}^{\alpha}$ to $\ell_{\boldsymbol{\beta}}$, hence for $\alpha = 0$ and $\widehat{\varepsilon} = \infty$ Inter-ARO recovers the Wasserstein LR model of Shafieezadeh-Abadeh et al. (2015). We next relate Inter-ARO to the problems in the ARO literature that use auxiliary data. The works in this literature (Gowal et al. 2021, Xing et al. 2022b) solve the following

$$\begin{aligned} \inf_{\boldsymbol{\beta} \in \mathbb{R}^n} \frac{1}{N + w\widehat{N}} \Big[ \sum_{i \in [N]} \sup_{\boldsymbol{z}^i \in \mathcal{B}_p(\alpha)} \{\ell_{\boldsymbol{\beta}}(\boldsymbol{x}^i + \boldsymbol{z}^i, y^i)\} + \\ w \sum_{j \in [\widehat{N}]} \sup_{\boldsymbol{z}^j \in \mathcal{B}_p(\alpha)} \{\ell_{\boldsymbol{\beta}}(\widehat{\boldsymbol{x}}^j + \boldsymbol{z}^j, \widehat{y}^j)\} \Big], \end{aligned} \tag{67}$$

for some $w > 0$, where $\mathcal{B}_p(\alpha) := \{\boldsymbol{z} \in \mathbb{R}^n : \|\boldsymbol{z}\|_p \leq \alpha\}$. We observe that (67) resembles a variant of ARO that replaces the empirical distribution $\mathbb{P}_N$ with its mixture with $\widehat{\mathbb{P}}_{\widehat{N}}$:

**Observation 8.** *Problem* (67) *is equivalent to*

$$\inf_{\boldsymbol{\beta} \in \mathbb{R}^n} \quad \mathbb{E}_{\mathbb{Q}_{\text{mix}}}[\ell_{\boldsymbol{\beta}}^{\alpha}(\boldsymbol{x}, y)] \tag{68}$$

*where* $\mathbb{Q}_{\text{mix}} := \lambda \cdot \mathbb{P}_N + (1 - \lambda) \cdot \widehat{\mathbb{P}}_{\widehat{N}}$ *for* $\lambda = \frac{N}{N + w\widehat{N}}$.

We give a condition on $\varepsilon$ and $\widehat{\varepsilon}$ to guarantee that the mixture distribution introduced in

Proposition 8 lives in $\mathfrak{B}_\varepsilon(\mathbb{P}_N) \cap \mathfrak{B}_{\widehat{\varepsilon}}(\widehat{\mathbb{P}}_{\widehat{N}})$, that is, the distribution $\mathbb{Q}_{\text{mix}}$ will be feasible in the sup problem of Inter-ARO.

**Proposition 19.** *For any $\lambda \in (0,1)$ and $\mathbb{Q}_{\text{mix}} = \lambda \cdot \mathbb{P}_N + (1-\lambda) \cdot \widehat{\mathbb{P}}_{\widehat{N}}$, we have $\mathbb{Q}_{\text{mix}} \in \mathfrak{B}_\varepsilon(\mathbb{P}_N) \cap \mathfrak{B}_{\widehat{\varepsilon}}(\widehat{\mathbb{P}}_{\widehat{N}})$ whenever $\varepsilon + \widehat{\varepsilon} \geq \mathrm{W}(\mathbb{P}_N, \widehat{\mathbb{P}}_{\widehat{N}})$ and $\frac{\widehat{\varepsilon}}{\varepsilon} = \frac{\lambda}{1-\lambda}$.*

For $\lambda = \frac{N}{N+\widehat{N}}$, if the intersection $\mathfrak{B}_\varepsilon(\mathbb{P}_N) \cap \mathfrak{B}_{\widehat{\varepsilon}}(\widehat{\mathbb{P}}_{\widehat{N}})$ is nonempty, Proposition 19 implies that a sufficient condition for this intersection to include $\mathbb{Q}_{\text{mix}}$ is $\widehat{\varepsilon}/\varepsilon = N/\widehat{N}$, which is intuitive since the radii of Wasserstein ambiguity sets are chosen inversely proportional to the number of samples (Kuhn et al. 2019, Theorem 18).

## 4.6 Setting Wasserstein Radii

Thus far, we have assumed knowledge of DRO ball radii $\varepsilon$ and $\widehat{\varepsilon}$ that satisfy Assumptions 4 and 5. In this section, we employ Wasserstein finite-sample statistics techniques to estimate these values.

**Setting $\epsilon$ for DR-ARO.** In the following theorem, we present tight characterizations for $\varepsilon$ so that the ball $\mathfrak{B}_\varepsilon(\mathbb{P}_N)$ includes the true distribution $\mathbb{P}^0$ with arbitrarily high confidence. We show that for an $\varepsilon$ chosen in such a manner, DR-ARO is well-defined. The full description of this result is available in our appendices.

**Theorem 12** (abridged collection of results from Fournier and Guillin 2015, Kuhn et al. 2019, Yue et al. 2021)**.** *For light-tailed distribution $\mathbb{P}^0$ and $\varepsilon \geq \mathcal{O}(\frac{\log(\eta^{-1})}{N})^{1/n}$ for $\eta \in (0,1)$, we have:* (i) *$\mathbb{P}^0 \in \mathfrak{B}_\varepsilon(\mathbb{P}_N)$ with $1-\eta$ confidence;* (ii) *DR-ARO overestimates the expected loss for $\mathbb{P}^0$ with $1-\eta$ confidence;* (iii) *DR-ARO is asymptotically consistent $\mathbb{P}^0$-a.s.;* (iv) *worst-case distributions for optimal solutions of DR-ARO are supported on at most $N+1$ outcomes.*

We next derive an analogous result for Inter-ARO.

**Choosing $\epsilon$ and $\widehat{\varepsilon}$ in Inter-ARO.** Recall that Inter-ARO revises DR-ARO by intersecting $\mathfrak{B}_\varepsilon(\mathbb{P}_N)$ with $\mathfrak{B}_{\widehat{\varepsilon}}(\widehat{\mathbb{P}}_{\widehat{N}})$. We need a nonempty intersection for Inter-ARO to be well-defined. A necessary and sufficient condition follows from the triangle inequality $\varepsilon + \widehat{\varepsilon} \geq \mathrm{W}(\mathbb{P}_N, \widehat{\mathbb{P}}_{\widehat{N}})$, where $\mathrm{W}(\mathbb{P}_N, \widehat{\mathbb{P}}_{\widehat{N}})$ can be computed with linear optimization as both distributions are discrete. We also want this intersection to include $\mathbb{P}^0$ with high confidence, in order to satisfy Assumption 5. We next provide a tight characterization for such $\varepsilon, \widehat{\varepsilon}$. The full description of this result is available in our appendices.

**Theorem 13** (abridged)**.** *For light-tailed* $\mathbb{P}^0$ *and* $\widehat{\mathbb{P}}$*, if* $\varepsilon \geq \mathcal{O}(\frac{\log(\eta_1^{-1})}{N})^{1/n}$ *and* $\widehat{\varepsilon} \geq \mathrm{W}(\mathbb{P}^0, \widehat{\mathbb{P}}) + \mathcal{O}(\frac{\log(\eta_2^{-1})}{\widehat{N}})^{1/n}$ *for* $\eta_1, \eta_2 \in (0, 1)$ *with* $\eta := \eta_1 + \eta_2 < 1$*, we have:* (i) $\mathbb{P}^0 \in \mathfrak{B}_{\varepsilon}(\mathbb{P}_N) \cap \mathfrak{B}_{\widehat{\varepsilon}}(\widehat{\mathbb{P}}_{\widehat{N}})$ *with* $1 - \eta$ *confidence;* (ii) Inter-ARO *overestimates true loss with* $1 - \eta$ *confidence.*

**Remark 1.** Inter-ARO *is not asymptotically consistent, given that* $\widehat{N} \to \infty$ *will let* $\widehat{\varepsilon} \to \mathrm{W}(\mathbb{P}^0, \widehat{\mathbb{P}})$ *due to the non-zero constant distance between the true distribution* $\mathbb{P}^0$ *and the auxiliary distribution* $\widehat{\mathbb{P}}$*. Inter-ARO is thus not useful in asymptotic data regimes.*

**Remark 2.** *The assumption that true data-generating distributions are light-tailed is satisfied when* $\Xi$ *is compact, and it is a common assumption for even simple sample average approximation techniques (Mohajerin Esfahani and Kuhn 2018, Assumption 3.3).*

**Knowledge of** $\mathrm{W}(\mathbb{P}^0, \widehat{\mathbb{P}})$**.** In Theorem 13, we use $\mathrm{W}(\mathbb{P}^0, \widehat{\mathbb{P}})$ explicitly. This distance, however, is typically unknown, and a common approach is to cross-validate it[6]. This would be applicable in our setting thanks to Corollary 9, because the relaxation Inter-ARO$^\star$ does not force learning from the auxiliary data unless it is useful, that is, one can seek evidence for the usefulness of the auxiliary data via cross-validation. Moreover, there are several domains where $\mathrm{W}(\mathbb{P}^0, \widehat{\mathbb{P}})$ is known exactly. For some special cases, we can use direct domain knowledge (*e.g.*, the "Uber vs Lyft" example of Taskesen et al. 2021). A recent example comes from learning from multi-source data, where $\mathbb{P}^0$ is named the target distribution and $\widehat{\mathbb{P}}$ is the source distribution (Rychener et al. 2024, §1). Another domain is private data release, where a data holder shares a subset of opt-in data to form $\mathbb{P}_N$, and generates a privacy-preserving synthetic dataset from the rest. The (privately generated) synthetic distribution has a known nonzero Wasserstein distance from the true data-generating distribution (Dwork and Roth 2014, Ullman and Vadhan 2020). See (Rychener et al. 2024, §5) for more scenarios that enable quantifying $\mathrm{W}(\mathbb{P}^0, \widehat{\mathbb{P}})$. Alternatively, one can directly rely on $\mathrm{W}(\mathbb{P}_N, \widehat{\mathbb{P}})$ if it is known, especially when synthetic data generators are trained on the empirical dataset. By employing Wasserstein GANs, which minimize the Wasserstein-1 distance, the distance between the generated distribution and the training distribution is minimized. This ensures that the synthetic distribution remains within a radius of the training distribution (Arjovsky et al. 2017).

## 4.7 Experiments

We conduct a series of experiments, each having a different source of auxiliary data, to test the proposed DR ARO models. We use the following abbreviations, where 'solution' refers to the

[6] In practice, distance between the unknown true and auxiliary data-generating distributions is also cross-validated in the transfer learning and domain adaptation literature (Zhong et al. 2010).

optimal $\boldsymbol{\beta}$ to make decisions:

- `ERM`: Solution of problem ERM (*i.e.*, naiv̈e LR);
- `ARO`: Solution of problem ARO (*i.e.*, adversarially robust LR);
- `ARO+Aux`: Solution of problem (67) (*i.e.*, replacing the empirical distribution of ARO with its mixture with auxiliary data);
- `DRO+ARO`: Solution of DR-ARO (*i.e.*, the Wasserstein DR counterpart of ARO);
- `DRO+ARO+Aux`: Solution of Inter-ARO$^\star$ (*i.e.*, relaxation of Inter-ARO that intersects the ambiguity set of DR-ARO with an auxiliary Wasserstein ball);

All parameters are 5-fold cross-validated from various grids. The Wasserstein radii use the grid $\{10^{-6}, 10^{-5}, 10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 0, 1, 2, 5, 10\}$, which is sufficient to ensure that the rule-of-thumb $\varepsilon = \mathcal{O}(1/\sqrt{N})$ is included around the center of this grid for all experiments conducted. To ensure that the intersections of Wasserstein balls are nonempty, we compute $\mathrm{W}(\mathbb{P}_N, \widehat{\mathbb{P}}_{\widehat{N}})$ once, and discard all combinations $(\varepsilon, \widehat{\varepsilon})$ with $\varepsilon + \widehat{\varepsilon} < \mathrm{W}(\mathbb{P}_N, \widehat{\mathbb{P}}_{\widehat{N}})$. The weight parameter $\omega$ of `ARO+Aux` is cross-validated from the grid $\{10^{-6}, 10^{-5}, 10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 0, 1\}$. We fix the norms defining the feature-label metric and the adversarial attacks to $\ell_1$- and $\ell_2$-norms, respectively. The parameter $\kappa$ (*cf.* Definition 5) is cross-validated from the grid $\{1, \sqrt{n}, n\}$, and since $n$ is the number of features, this grid includes cases where label uncertainty is equivalent to uncertainty of a single feature ($\kappa = 1$), label uncertainty is equivalent to uncertainty of all features combined ($\kappa = n$), and an intermediary case ($\kappa = \sqrt{n}$). The case of ignoring label uncertainty ($\kappa = \infty$) is purposely not included in the grid, since one of the key reasons behind robust overfitting is that while `ARO` is equivalent to a distributionally robust model, the underlying ambiguity is only around the features (*cf.* our discussion in §4.2). All simulated adversarial attacks are worst-case $\ell_p$-attacks that are instance-wise at test time, and the experiments assume we know the strength $\alpha$ and norm $\ell_p$ of the adversarial attacks.

All experiments are conducted in Julia and executed on Intel Xeon 2.66GHz processors with 16GB memory in single-core mode. We use MOSEK's exponential cone optimizer to solve all problems. To interpret the results accurately, recall that `DRO+ARO` and `DRO+ARO+Aux` are the DR models that we propose. Note also that `ERM`, `ARO`, and `DRO+ARO` do not utilize auxiliary data, while `DRO+ARO+Aux` and `ARO+Aux` have access to the same auxiliary datasets across all experiments (*i.e.*, we do not sample different auxiliary distributions for different methods to ensure that our comparisons are made *ceteris paribus*). Moreover, while `ARO` does not have access to auxiliary data, one can interpret `ARO+Aux` as a generalization of `ARO` that also has

| Data | `ERM` | `ARO` | `ARO+Aux` | `DRO+ARO` | `DRO+ARO+Aux` |
|---|---|---|---|---|---|
| absent | 44.02% | 38.82% | 35.95% | 34.22% | **32.64%** |
| anneal | 18.08% | 16.61% | 14.97% | 13.50% | **12.78%** |
| audio | 21.43% | 21.54% | 17.03% | 11.76% | **9.01%** |
| breast-c | 4.74% | 4.93% | 3.87% | 3.06% | **2.52%** |
| contrac | 44.14% | 42.86% | 40.98% | 40.00% | **39.65%** |
| derma | 15.97% | 16.46% | 13.47% | 12.78% | **10.84%** |
| ecoli | 16.30% | 14.67% | 13.26% | 11.11% | **9.78%** |
| spam | 11.35% | 10.23% | 10.16% | 9.83% | **9.81%** |
| spect | 33.75% | 29.69% | 25.78% | 25.47% | **21.56%** |
| p-tumor | 21.84% | 20.81% | 17.35% | 16.18% | **14.78%** |

Table 27: *Out-of-sample errors of UCI experiments with $\ell_2$-attacks of strength $\alpha = 0.05$.*

access to auxiliary datasets since it takes a mixture (with mixture weight $\omega$) of the empirical dataset with the auxiliary dataset. For example, $\omega = 1$ would simply revise `ARO` by appending the empirical dataset with the auxiliary dataset.

### 4.7.1 UCI Datasets (Auxiliary Data is Synthetically Generated)

We compare the out-of-sample error rates of each method on 10 UCI datasets for binary classification (Kelly et al. 2023). For each dataset, we run 10 simulations as follows: *(i)* Select 40% of the data as a test set ($N_{\text{te}} \propto 0.4$); *(ii)* Sample 25% of the remaining to form a training set ($N \propto 0.6 \cdot 0.25$); *(iii)* The rest ($\widehat{N} \propto 0.6 \cdot 0.75$) is used to fit a synthetic generator Gaussian Copula from the SDV package (Patki et al. 2016), which is then used to generate auxiliary data. The mean errors on the test set are reported in Table 27 for $\ell_2$-attacks of strength $\alpha = 0.05$. The best error is always achieved by `DRO+ARO+Aux`, followed by `DRO+ARO`, `DRO+Aux`, `ARO`, `ERM`, respectively. In our appendices, we report similar results for attack strengths $\alpha \in \{0, 0.05, 0.2\}$, and share data preprocessing details and standard deviations of out-of-sample errors.

### 4.7.2 MNIST/EMNIST Datasets (Auxiliary Data is Out-of-Domain)

We use the MNIST digits dataset (LeCun et al. 1998) to classify whether a digit is 1 or 7. For an auxiliary dataset, we use the larger EMNIST digits dataset (Cohen et al. 2017), whose authors summarize that this dataset has additional samples collected from a different group of individuals (high school students). Since EMNIST digits include MNIST digits, we remove the latter from the EMNIST dataset. We simulate the following 25 times: *(i)* Sample 1,000

| Attack | ERM | ARO | ARO+Aux | DRO+ARO | DRO+ARO+Aux |
|---|---|---|---|---|---|
| No attack ($\alpha = 0$) | 1.55% | 1.55% | 1.19% | 0.64% | **0.53%** |
| $\ell_1$ ($\alpha = 68/255$) | 2.17% | 1.84% | 1.33% | 0.66% | **0.57%** |
| $\ell_2$ ($\alpha = 128/255$) | 99.93% | 3.36% | 2.54% | 2.40% | **2.12%** |
| $\ell_\infty$ ($\alpha = 8/255$) | 100.00% | 2.60% | 2.38% | 2.20% | **1.95%** |

Table 28: *Out-of-sample errors of MNIST/EMNIST experiments with various attacks.*

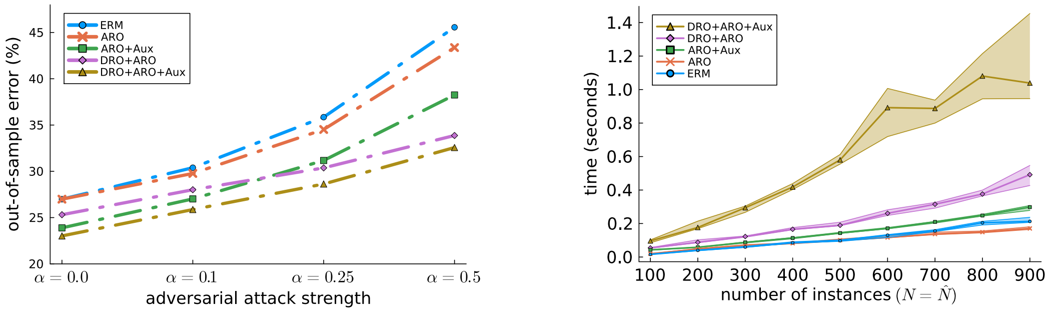


Figure 17: *Out-of-sample errors under varying attack strengths (left) and runtimes under varying numbers of empirical and auxiliary instances (right) of artificial experiments.*

instances from the MNIST dataset as a training set; *(ii)* The remaining instances in the MNIST dataset are our test set; *(iii)* Sample 1,000 instances from the EMNIST dataset as an auxiliary dataset. Table 28 reports the mean out-of-sample errors in various adversarial attack regimes. The results are analogous to the UCI experiments. Additionally, note that in the absence of adversarial attacks ($\alpha = 0$), DRO+ARO coincides with the Wasserstein LR model of Shafieezadeh-Abadeh et al. (2015), and the results thus imply that even without adversarial attacks, we can improve the state-of-the-art DR model by revising its ambiguity set in light of auxiliary data.

### 4.7.3 Artificial Experiments (Auxiliary Data is Perturbed)

We generate empirical and auxiliary datasets by controlling their data-generating distributions (more details in the appendices). We simulate 25 cases, each with $N = 100$ training, $\widehat{N} = 200$ auxiliary, and $N_{\text{te}} = 10,000$ test instances and $n = 100$ features. The performance of benchmark models with varying attacks is available in Figure 17 (left). ERM provides the worst performance, followed by ARO. The relationship between DRO+ARO and ARO+Aux is not monotonic: the former works better in larger attack regimes, conforming to the robust overfitting phenomenon. Finally, Adv+DRO+Aux always performs the best. We conduct a similar simulation for datasets with $n = 100$, and gradually increase $N = \widehat{N}$ to report median (50%±15% quantiles shaded) runtimes of each method (*cf.* Figure 17, right). The fastest methods is `ARO`, followed by `ERM`, `ARO+Aux`, `DRO+ARO`, and `DRO+ARO+Aux`. The slowest is `DRO+ARO+Aux`, but the runtime scales graciously.

## 4.8 Conclusions

We formulate the distributionally robust counterpart of adversarially robust LR. Additionally, we demonstrate how to effectively utilize appropriately curated auxiliary data by intersecting Wasserstein balls. We illustrate the superiority of the proposed approach in terms of out-of-sample performance and confirm its scalability in practical settings.

From a theoretical point of view, it would be natural to extend our work to more loss functions, as is typical for DRO studies stemming from LR. To be able to optimize Inter-ARO$^\star$ for very large-dimensional datasets, an interesting future work is to investigate first-order optimization methods that do not rely on off-the-shelf solvers. We also believe a cutting-plane method tailored for Inter-ARO$^\star$ can also help us scale this problem for large-dimensional problems, since we would avoid monolithically optimizing a problem with $\mathcal{O}(N \cdot \widehat{N})$ exponential cone constraints.

From a practical perspective, the ability to optimize Inter-ARO$^\star$ in high-dimensional settings would also enable fine-tuning the final layer of a pre-trained neural network for binary classification, since this corresponds to logistic regression under a sigmoid activation. In our image recognition experiments, we used the MNIST dataset, as EMNIST served as a natural choice for auxiliary data. Identifying a suitable auxiliary dataset for CIFAR (Krizhevsky et al. 2009) could similarly support new experimental directions.

Finally, recent breakthroughs in foundation models naturally pose the question of whether our ideas in this work apply to these models. For example, Ye et al. (2022) use a pre-trained language model (PLM) to generate synthetic pairs of text sequences and labels which are then used to train downstream models. It would be interesting to adapt our ideas to the text domain to explore robustness in the presence of two PLMs.

## 4.A Notation

Throughout this section, bold lowercase letters denote vectors, while standard lowercase letters are reserved for scalars. A generic data instance is modeled as $\boldsymbol{\xi} = (\boldsymbol{x}, y) \in \Xi := \mathbb{R}^n \times \{-1, +1\}$. For any $p > 0$, $\|\boldsymbol{x}\|_p$ denotes the $p$-norm $(\sum_{i=1}^n |x_i|^p)^{1/p}$ and $\|\boldsymbol{x}\|_{p^\star}$ is its dual norm where $1/p + 1/p^\star = 1$ with the convention of $1/1 + 1/\infty = 1$. The set of probability distributions supported on $\Xi$ is denoted by $\mathcal{P}(\Xi)$. The Dirac measure supported on $\boldsymbol{\xi}$ is denoted by $\delta_{\boldsymbol{\xi}}$. The logloss is defined as $\ell_{\boldsymbol{\beta}}(\boldsymbol{x}, y) = \log(1+\exp(-y \cdot \boldsymbol{\beta}^\top \boldsymbol{x}))$ and its associated univariate loss is $L(z) = \log(1+\exp(-z))$ so that $L(y \cdot \boldsymbol{\beta}^\top \boldsymbol{x}) = \ell_{\boldsymbol{\beta}}(\boldsymbol{x}, y)$. The exponential cone is denoted by $\mathcal{K}_{\exp} = \mathrm{cl}(\{\boldsymbol{\omega} \in \mathbb{R}^3 : \omega_1 \geq \omega_2 \cdot \exp(\omega_3/\omega_2),\ \omega_1 > 0,\ \omega_2 > 0\})$ where cl is the closure operator. The Lipschitz modulus of a univariate function $f$ is defined as $\mathrm{Lip}(f) := \sup_{z,z' \in \mathbb{R}} \{|f(z) - f(z')|/|z - z'| : z \neq z\}$ whereas its effective domain is $\mathrm{dom}(f) = \{z : f(z) < +\infty\}$. For a function $f : \mathbb{R}^n \mapsto \mathbb{R}$, its convex conjugate is $f^*(\boldsymbol{z}) = \sup_{\boldsymbol{x} \in \mathbb{R}^n} \boldsymbol{z}^\top \boldsymbol{x} - f(\boldsymbol{x})$. We reserve $\alpha \geq 0$ for the radii of the norms of adversarial attacks on the features and $\varepsilon \geq 0$ for the radii of distributional ambiguity sets.

## 4.B Missing Proofs

### 4.B.1 Proof of Observation 7

For any $\boldsymbol{\beta} \in \mathbb{R}^n$, with standard robust optimization arguments (Ben-Tal et al. 2009, Bertsimas and den Hertog 2022), we can show that

$$\begin{aligned}
& \sup_{\boldsymbol{z} : \|\boldsymbol{z}\|_p \leq \alpha} \{\ell_{\boldsymbol{\beta}}(\boldsymbol{x} + \boldsymbol{z}, y)\} \\
\iff & \sup_{\boldsymbol{z} : \|\boldsymbol{z}\|_p \leq \alpha} \{\log(1 + \exp(-y \cdot \boldsymbol{\beta}^\top (\boldsymbol{x} + \boldsymbol{z})))\} \\
\iff & \log\left(1 + \exp\left(\sup_{\boldsymbol{z} : \|\boldsymbol{z}\|_p \leq \alpha} \{-y \cdot \boldsymbol{\beta}^\top (\boldsymbol{x} + \boldsymbol{z})\}\right)\right) \\
\iff & \log\left(1 + \exp\left(-y \cdot \boldsymbol{\beta}^\top \boldsymbol{x} + \alpha \cdot \sup_{\boldsymbol{z} : \|\boldsymbol{z}\|_p \leq 1} \{-y \cdot \boldsymbol{\beta}^\top \boldsymbol{z}\}\right)\right) \\
\iff & \log(1 + \exp(-y \cdot \boldsymbol{\beta}^\top \boldsymbol{x} + \alpha \cdot \|-y \cdot \boldsymbol{\beta}\|_{p^\star})) \\
\iff & \log(1 + \exp(-y \cdot \boldsymbol{\beta}^\top \boldsymbol{x} + \alpha \cdot \|\boldsymbol{\beta}\|_{p^\star})),
\end{aligned}$$

where the first step follows from the definition of logloss, the second step follows from the fact that log and exp are increasing functions, the third step takes the constant terms out of the sup problem and exploits the fact that the optimal solution of maximizing a linear function will be at an extreme point of the $\ell_p$-ball, the fourth step uses the definition of dual norm, and finally, the redundant $-y \in \{-1, +1\}$ is omitted from the dual norm. We can therefore define the adversarial loss $\ell^\alpha_{\boldsymbol{\beta}}(\boldsymbol{x}, y) := \log(1+\exp(-y \cdot \boldsymbol{\beta}^\top \boldsymbol{x} + \alpha \cdot \|\boldsymbol{\beta}\|_{p^\star}))$ where $\alpha$ models the strength of the adversary, $\boldsymbol{\beta}$ is the decision vector, and $(\boldsymbol{x}, y)$ is an instance. Replacing $\sup_{\boldsymbol{z}:\|\boldsymbol{z}\|_p \leq \alpha}\{\ell_{\boldsymbol{\beta}}(\boldsymbol{x}+\boldsymbol{z}, y)\}$ in DR-ARO with $\ell^\alpha_{\boldsymbol{\beta}}(\boldsymbol{x}, y)$ concludes the equivalence of the optimization problem.

Furthermore, to see $\mathrm{Lip}(L^\alpha) = 1$, firstly note that since $L^\alpha(z) = \log(1+\exp(-z+\alpha \cdot \|\boldsymbol{\beta}\|_{p^\star}))$ is differentiable everywhere in $z$ and its gradient $L^{\alpha\prime}$ is bounded everywhere, we have that $\mathrm{Lip}(L^{\alpha})$ is equal to $\sup_{z\in\mathbb{R}}\{|L^{\alpha\prime}(z)|\}$. We thus have:

$$L^{\alpha\prime}(z) = \frac{-\exp(-z+\alpha\cdot\|\boldsymbol{\beta}\|_{p^\star})}{1+\exp(-z+\alpha\cdot\|\boldsymbol{\beta}\|_{p^\star})} = \frac{-1}{1+\exp(z-\alpha\cdot\|\boldsymbol{\beta}\|_{p^\star})} \in (-1,0)$$

and $|L^{\alpha\prime}(z)| = [1+\exp(z-\alpha\cdot\|\boldsymbol{\beta}\|_{p^\star})]^{-1} \longrightarrow 1$ as $z \longrightarrow -\infty$. □

### 4.B.2 Proof of Corollary 8

Observation 7 lets us represent DR-ARO as the DR counterpart of empirical minimization of $\ell^{\alpha}_{\boldsymbol{\beta}}$:

$$\begin{array}{lll} \underset{\boldsymbol{\beta}}{\text{minimize}} & \underset{\mathbb{Q}\in\mathfrak{B}_\varepsilon(\mathbb{P}_N)}{\sup} & \mathbb{E}_{\mathbb{Q}}\left[\ell^{\alpha}_{\boldsymbol{\beta}}(\boldsymbol{x},y)\right] \\ \text{subject to} & \boldsymbol{\beta}\in\mathbb{R}^n. & \end{array} \tag{69}$$

Since the univariate loss $L^{\alpha}(z) := \log(1+\exp(-z+\alpha\cdot\|\boldsymbol{\beta}\|_{p^\star}))$ satisfying the identity $L^{\alpha}(\langle y\cdot\boldsymbol{x},\boldsymbol{\beta}\rangle) = \ell^{\alpha}_{\boldsymbol{\beta}}(\boldsymbol{x},y)$ is Lipschitz continuous, Theorem 14 *(ii)* of Shafieezadeh-Abadeh et al. (2019) is immediately applicable. We can therefore rewrite (69) as:

$$\begin{array}{lll} \underset{\boldsymbol{\beta},\ \lambda,\ \boldsymbol{s}}{\text{minimize}} & \lambda\cdot\varepsilon + \dfrac{1}{N}\displaystyle\sum_{i\in[N]} s_i & \\ \text{subject to} & L^{\alpha}(\langle y^i\cdot\boldsymbol{x},\boldsymbol{\beta}\rangle) \leq s_i & \forall i\in[N] \\ & L^{\alpha}(\langle -y^i\cdot\boldsymbol{x},\boldsymbol{\beta}\rangle) - \lambda\cdot\kappa \leq s_i & \forall i\in[N] \\ & \mathrm{Lip}(L^{\alpha})\cdot\|\boldsymbol{\beta}\|_{q^\star} \leq \lambda & \\ & \boldsymbol{\beta}\in\mathbb{R}^n,\ \ \lambda\geq 0,\ \ \boldsymbol{s}\in\mathbb{R}^N. & \end{array}$$

Replacing $\mathrm{Lip}(L^{\alpha}) = 1$ and substituting the definition of $L^{\alpha}$ concludes the proof. □

### 4.B.3 Proof of Proposition 17

We prove Proposition 17 by constructing the optimization problem in its statement. We will thus dualize the inner sup-problem of Inter-ARO for fixed $\boldsymbol{\beta}$. To this end, we present a sequence of reformulations to the inner problem and then exploit strong semi-infinite duality.

By interchanging $\boldsymbol{\xi} = (\boldsymbol{x}, y)$, we first rewrite the inner problem as

$$\begin{array}{ll}
\underset{\mathbb{Q},\Pi,\widehat{\Pi}}{\text{maximize}} & \displaystyle\int_{\boldsymbol{\xi}\in\Xi} \ell_{\boldsymbol{\beta}}^{\alpha}(\boldsymbol{\xi})\mathbb{Q}(\mathrm{d}\boldsymbol{\xi}) \\
\text{subject to} & \displaystyle\int_{\boldsymbol{\xi},\boldsymbol{\xi}'\in\Xi^2} d(\boldsymbol{\xi},\boldsymbol{\xi}')\Pi(\mathrm{d}\boldsymbol{\xi},\mathrm{d}\boldsymbol{\xi}') \leq \varepsilon \\
& \displaystyle\int_{\boldsymbol{\xi}\in\Xi} \Pi(\mathrm{d}\boldsymbol{\xi},\mathrm{d}\boldsymbol{\xi}') = \mathbb{P}_N(\mathrm{d}\boldsymbol{\xi}') \qquad\qquad \forall \boldsymbol{\xi}' \in \Xi \\
& \displaystyle\int_{\boldsymbol{\xi}'\in\Xi} \Pi(\mathrm{d}\boldsymbol{\xi},\mathrm{d}\boldsymbol{\xi}') = \mathbb{Q}(\mathrm{d}\boldsymbol{\xi}) \qquad\qquad \forall \boldsymbol{\xi} \in \Xi \\
& \displaystyle\int_{\boldsymbol{\xi},\boldsymbol{\xi}'\in\Xi^2} d(\boldsymbol{\xi},\boldsymbol{\xi}')\widehat{\Pi}(\mathrm{d}\boldsymbol{\xi},\mathrm{d}\boldsymbol{\xi}') \leq \widehat{\varepsilon} \\
& \displaystyle\int_{\boldsymbol{\xi}\in\Xi} \widehat{\Pi}(\mathrm{d}\boldsymbol{\xi},\mathrm{d}\boldsymbol{\xi}') = \widehat{\mathbb{P}}_{\widehat{N}}(\mathrm{d}\boldsymbol{\xi}') \qquad\qquad \forall \boldsymbol{\xi}' \in \Xi \\
& \displaystyle\int_{\boldsymbol{\xi}'\in\Xi} \widehat{\Pi}(\mathrm{d}\boldsymbol{\xi},\mathrm{d}\boldsymbol{\xi}') = \mathbb{Q}(\mathrm{d}\boldsymbol{\xi}) \qquad\qquad \forall \boldsymbol{\xi} \in \Xi \\
& \mathbb{Q} \in \mathcal{P}(\Xi),\ \Pi \in \mathcal{P}(\Xi^2),\ \widehat{\Pi} \in \mathcal{P}(\Xi^2).
\end{array}$$

Here, the first three constraints specify that $\mathbb{Q}$ and $\mathbb{P}_N$ have a Wasserstein distance bounded by $\varepsilon$ from each other, modeled via their coupling $\Pi$. The latter three constraints similarly specify that $\mathbb{Q}$ and $\widehat{\mathbb{P}}_{\widehat{N}}$ are at most $\widehat{\varepsilon}$ away from each other, modeled via their coupling $\widehat{\Pi}$. As $\mathbb{Q}$ lies in the intersection of two Wasserstein balls in Inter-ARO, the marginal $\mathbb{Q}$ is shared between $\Pi$ and $\widehat{\Pi}$. We can now substitute the third constraint into the objective and the last constraint and obtain:

$$
\begin{aligned}
\underset{\Pi,\widehat{\Pi}}{\text{maximize}} \quad & \int_{\boldsymbol{\xi}\in\Xi} \ell^{\alpha}_{\boldsymbol{\beta}}(\boldsymbol{\xi}) \int_{\boldsymbol{\xi'}\in\Xi} \Pi(\mathrm{d}\boldsymbol{\xi}, \mathrm{d}\boldsymbol{\xi'}) \\
\text{subject to} \quad & \int_{\boldsymbol{\xi},\boldsymbol{\xi'}\in\Xi^2} d(\boldsymbol{\xi},\boldsymbol{\xi'})\Pi(\mathrm{d}\boldsymbol{\xi}, \mathrm{d}\boldsymbol{\xi'}) \leq \varepsilon \\
& \int_{\boldsymbol{\xi}\in\Xi} \Pi(\mathrm{d}\boldsymbol{\xi}, \mathrm{d}\boldsymbol{\xi'}) = \mathbb{P}_N(\mathrm{d}\boldsymbol{\xi'}) \qquad\qquad \forall \boldsymbol{\xi'} \in \Xi \\
& \int_{\boldsymbol{\xi},\boldsymbol{\xi'}\in\Xi^2} d(\boldsymbol{\xi},\boldsymbol{\xi'})\widehat{\Pi}(\mathrm{d}\boldsymbol{\xi}, \mathrm{d}\boldsymbol{\xi'}) \leq \widehat{\varepsilon} \\
& \int_{\boldsymbol{\xi}\in\Xi} \widehat{\Pi}(\mathrm{d}\boldsymbol{\xi}, \mathrm{d}\boldsymbol{\xi'}) = \widehat{\mathbb{P}}_{\widehat{N}}(\mathrm{d}\boldsymbol{\xi'}) \qquad\qquad \forall \boldsymbol{\xi'} \in \Xi \\
& \int_{\boldsymbol{\xi'}\in\Xi} \widehat{\Pi}(\mathrm{d}\boldsymbol{\xi}, \mathrm{d}\boldsymbol{\xi'}) = \int_{\boldsymbol{\xi'}\in\Xi} \Pi(\mathrm{d}\boldsymbol{\xi}, \mathrm{d}\boldsymbol{\xi'}) \quad \forall \boldsymbol{\xi} \in \Xi \\
& \Pi \in \mathcal{P}(\Xi^2),\ \widehat{\Pi} \in \mathcal{P}(\Xi^2).
\end{aligned}
$$

Denoting by $\mathbb{Q}^i(\mathrm{d}\boldsymbol{\xi}) := \Pi(\mathrm{d}\boldsymbol{\xi} \mid \boldsymbol{\xi}^i)$ the conditional distribution of $\Pi$ upon the realization of $\boldsymbol{\xi'} = \boldsymbol{\xi}^i$ and exploiting the fact that $\mathbb{P}_N$ is a discrete distribution supported on the $N$ data points $\{\boldsymbol{\xi}^i\}_{i\in[N]}$, we can use the marginalized representation $\Pi(\mathrm{d}\boldsymbol{\xi}, \mathrm{d}\boldsymbol{\xi'}) = \frac{1}{N}\sum_{i=1}^N \delta_{\boldsymbol{\xi}^i}(\mathrm{d}\boldsymbol{\xi'})\mathbb{Q}^i(\mathrm{d}\boldsymbol{\xi})$. Similarly, we can introduce $\widehat{\mathbb{Q}}^i(\mathrm{d}\boldsymbol{\xi}) := \widehat{\Pi}(\mathrm{d}\boldsymbol{\xi} \mid \widehat{\boldsymbol{\xi}}^i)$ for $\{\widehat{\boldsymbol{\xi}}^i\}_{i\in[\widehat{N}]}$ to exploit the marginalized representation $\widehat{\Pi}(\mathrm{d}\boldsymbol{\xi}, \mathrm{d}\boldsymbol{\xi'}) = \frac{1}{\widehat{N}}\sum_{j=1}^{\widehat{N}} \delta_{\widehat{\boldsymbol{\xi}}^j}(\mathrm{d}\boldsymbol{\xi'})\widehat{\mathbb{Q}}^j(\mathrm{d}\boldsymbol{\xi})$. By using this marginalization representation, we can use the following simplification for the objective function:

$$\int_{\boldsymbol{\xi}\in\Xi} \ell^{\alpha}_{\boldsymbol{\beta}}(\boldsymbol{\xi}) \int_{\boldsymbol{\xi'}\in\Xi} \Pi(\mathrm{d}\boldsymbol{\xi}, \mathrm{d}\boldsymbol{\xi'}) \; = \; \frac{1}{N}\sum_{i=1}^{N} \int_{\boldsymbol{\xi}\in\Xi} \ell^{\alpha}_{\boldsymbol{\beta}}(\boldsymbol{\xi}) \int_{\boldsymbol{\xi'}\in\Xi} \delta_{\boldsymbol{\xi}^i}(\mathrm{d}\boldsymbol{\xi'})\mathbb{Q}^i(\mathrm{d}\boldsymbol{\xi}) \; = \; \frac{1}{N}\sum_{i=1}^{N} \int_{\boldsymbol{\xi}\in\Xi} \ell^{\alpha}_{\boldsymbol{\beta}}(\boldsymbol{\xi})\mathbb{Q}^i(\mathrm{d}\boldsymbol{\xi}).$$

Applying analogous reformulations to the constraints leads to the following reformulation of the inner sup problem of Inter-ARO:

$$\begin{array}{lll}
\underset{\mathbb{Q},\widehat{\mathbb{Q}}}{\text{maximize}} & \displaystyle\frac{1}{N}\sum_{i=1}^{N}\int_{\boldsymbol{\xi}\in\Xi}\ell_{\boldsymbol{\beta}}^{\alpha}(\boldsymbol{\xi})\mathbb{Q}^{i}(\mathrm{d}\boldsymbol{\xi}) & \\
\text{subject to} & \displaystyle\frac{1}{N}\sum_{i=1}^{N}\int_{\boldsymbol{\xi}\in\Xi}d(\boldsymbol{\xi},\boldsymbol{\xi}^{i})\mathbb{Q}^{i}(\mathrm{d}\boldsymbol{\xi}) \leq \varepsilon & \\
& \displaystyle\frac{1}{\widehat{N}}\sum_{j=1}^{\widehat{N}}\int_{\boldsymbol{\xi}\in\Xi}d(\boldsymbol{\xi},\widehat{\boldsymbol{\xi}}^{j})\widehat{\mathbb{Q}}^{j}(\mathrm{d}\boldsymbol{\xi}) \leq \widehat{\varepsilon} & \\
& \displaystyle\frac{1}{N}\sum_{i=1}^{N}\mathbb{Q}^{i}(\mathrm{d}\boldsymbol{\xi}) = \frac{1}{\widehat{N}}\sum_{j=1}^{\widehat{N}}\widehat{\mathbb{Q}}^{j}(\mathrm{d}\boldsymbol{\xi}) & \forall\boldsymbol{\xi}\in\Xi \\
& \mathbb{Q}^{i}\in\mathcal{P}(\Xi),\ \widehat{\mathbb{Q}}^{j}\in\mathcal{P}(\Xi) & \forall i\in[N],\ \forall j\in[\widehat{N}].
\end{array}$$

We now decompose each $\mathbb{Q}^{i}$ into two measures corresponding to $y=\pm 1$, so that $\mathbb{Q}^{i}(\mathrm{d}(\boldsymbol{x},y)) = \mathbb{Q}^{i}_{+1}(\mathrm{d}\boldsymbol{x})$ for $y=+1$ and $\mathbb{Q}^{i}(\mathrm{d}(\boldsymbol{x},y)) = \mathbb{Q}^{i}_{-1}(\mathrm{d}\boldsymbol{x})$ for $y=-1$. We similarly represent each $\widehat{\mathbb{Q}}^{j}$ via $\widehat{\mathbb{Q}}^{j}_{+1}$ and $\widehat{\mathbb{Q}}^{j}_{-1}$ depending on $y$. Note that these new measures are not probability measures as they do not integrate to 1, but non-negative measures supported on $\mathbb{R}^{n}$ (denoted $\in\mathcal{P}_{+}(\mathbb{R}^{n})$).

We get:

$$\begin{array}{lll}
\underset{\mathbb{Q}_{\pm 1}, \widehat{\mathbb{Q}}_{\pm 1}}{\text{maximize}} & \displaystyle \frac{1}{N}\sum_{i=1}^{N}\int_{\boldsymbol{x}\in\mathbb{R}^n} [\ell^{\alpha}_{\boldsymbol{\beta}}(\boldsymbol{x},+1)\mathbb{Q}^i_{+1}(\mathrm{d}\boldsymbol{x}) + \ell^{\alpha}_{\boldsymbol{\beta}}(\boldsymbol{x},-1)\mathbb{Q}^i_{-1}(\mathrm{d}\boldsymbol{x})] & \\
\text{subject to} & \displaystyle \frac{1}{N}\sum_{i=1}^{N}\int_{\boldsymbol{x}\in\mathbb{R}^n} [d((\boldsymbol{x},+1),\boldsymbol{\xi}^i)\mathbb{Q}^i_{+1}(\mathrm{d}\boldsymbol{x}) + d((\boldsymbol{x},-1),\boldsymbol{\xi}^i)\mathbb{Q}^i_{-1}(\mathrm{d}\boldsymbol{x})] \leq \varepsilon & \\
& \displaystyle \frac{1}{\widehat{N}}\sum_{j=1}^{\widehat{N}}\int_{\boldsymbol{x}\in\mathbb{R}^n} [d((\boldsymbol{x},+1),\widehat{\boldsymbol{\xi}}^j)\widehat{\mathbb{Q}}^j_{+1}(\mathrm{d}\boldsymbol{x}) + d((\boldsymbol{x},-1),\widehat{\boldsymbol{\xi}}^j)\widehat{\mathbb{Q}}^j_{-1}(\mathrm{d}\boldsymbol{x})] \leq \widehat{\varepsilon} & \\
& \displaystyle \int_{\boldsymbol{x}\in\mathbb{R}^n} \mathbb{Q}^i_{+1}(\mathrm{d}\boldsymbol{x}) + \mathbb{Q}^i_{-1}(\mathrm{d}\boldsymbol{x}) = 1 & \forall i \in [N] \\
& \displaystyle \int_{\boldsymbol{x}\in\mathbb{R}^n} \widehat{\mathbb{Q}}^j_{+1}(\mathrm{d}\boldsymbol{x}) + \widehat{\mathbb{Q}}^j_{-1}(\mathrm{d}\boldsymbol{x}) = 1 & \forall j \in [\widehat{N}] \\
& \displaystyle \frac{1}{N}\sum_{i=1}^{N}\mathbb{Q}^i_{+1}(\mathrm{d}\boldsymbol{x}) = \frac{1}{\widehat{N}}\sum_{j=1}^{\widehat{N}}\widehat{\mathbb{Q}}^j_{+1}(\mathrm{d}\boldsymbol{x}) & \forall \boldsymbol{x} \in \mathbb{R}^n \\
& \displaystyle \frac{1}{N}\sum_{i=1}^{N}\mathbb{Q}^i_{-1}(\mathrm{d}\boldsymbol{x}) = \frac{1}{\widehat{N}}\sum_{j=1}^{\widehat{N}}\widehat{\mathbb{Q}}^j_{-1}(\mathrm{d}\boldsymbol{x}) & \forall \boldsymbol{x} \in \mathbb{R}^n \\
& \mathbb{Q}^i_{\pm 1} \in \mathcal{P}_+(\mathbb{R}^n), \ \widehat{\mathbb{Q}}^j_{\pm 1} \in \mathcal{P}_+(\mathbb{R}^n) & \forall i \in [N], \ j \in [\widehat{N}].
\end{array}$$

Next, we explicitly write the definition of the metric $d(\cdot,\cdot)$ in the first two constraints as well as use auxiliary measures $\mathbb{A}_{\pm 1} \in \mathcal{P}_+(\mathbb{R}^n)$ to break down the last two equality constraints:

$$
\begin{array}{lll}
\underset{\mathbb{A}_{\pm 1}, \mathbb{Q}_{\pm 1}, \widehat{\mathbb{Q}}_{\pm 1}}{\text{maximize}} & \displaystyle \frac{1}{N}\sum_{i=1}^{N}\int_{\boldsymbol{x}\in\mathbb{R}^n} [\ell^{\alpha}_{\boldsymbol{\beta}}(\boldsymbol{x}, +1)\mathbb{Q}^i_{+1}(\mathrm{d}\boldsymbol{x}) + \ell^{\alpha}_{\boldsymbol{\beta}}(\boldsymbol{x}, -1)\mathbb{Q}^i_{-1}(\mathrm{d}\boldsymbol{x})] & \\
\text{subject to} & \displaystyle \frac{1}{N}\int_{\boldsymbol{x}\in\mathbb{R}^n}\Bigg[\kappa \cdot \sum_{i\in[N]:y^i=-1} \mathbb{Q}^i_{+1}(\mathrm{d}\boldsymbol{x}) + \kappa \cdot \sum_{i\in[N]:y^i=+1} \mathbb{Q}^i_{-1}(\mathrm{d}\boldsymbol{x}) + & \\
& \displaystyle \sum_{i=1}^{N}\|\boldsymbol{x} - \boldsymbol{x}^i\|_q \cdot [\mathbb{Q}^i_{+1}(\mathrm{d}\boldsymbol{x}) + \mathbb{Q}^i_{-1}(\mathrm{d}\boldsymbol{x})]\Bigg] \leq \varepsilon & \\
& \displaystyle \frac{1}{\widehat{N}}\int_{\boldsymbol{x}\in\mathbb{R}^n}\Bigg[\kappa \cdot \sum_{j\in[N]:\widehat{y}^j=-1} \widehat{\mathbb{Q}}^j_{+1}(\mathrm{d}\boldsymbol{x}) + \kappa \cdot \sum_{j\in[N]:\widehat{y}^j=+1} \widehat{\mathbb{Q}}^j_{-1}(\mathrm{d}\boldsymbol{x}) + & \\
& \displaystyle \sum_{j=1}^{\widehat{N}}\|\boldsymbol{x} - \widehat{\boldsymbol{x}}^j\|_q \cdot [\widehat{\mathbb{Q}}^j_{+1}(\mathrm{d}\boldsymbol{x}) + \widehat{\mathbb{Q}}^j_{-1}(\mathrm{d}\boldsymbol{x})]\Bigg] \leq \widehat{\varepsilon} & \\
& \displaystyle \int_{\boldsymbol{x}\in\mathbb{R}^n} \mathbb{Q}^i_{+1}(\mathrm{d}\boldsymbol{x}) + \mathbb{Q}^i_{-1}(\mathrm{d}\boldsymbol{x}) = 1 & \forall i \in [N] \\
& \displaystyle \int_{\boldsymbol{x}\in\mathbb{R}^n} \widehat{\mathbb{Q}}^j_{+1}(\mathrm{d}\boldsymbol{x}) + \widehat{\mathbb{Q}}^j_{-1}(\mathrm{d}\boldsymbol{x}) = 1 & \forall j \in [\widehat{N}] \\
& \displaystyle \frac{1}{N}\sum_{i=1}^{N} \mathbb{Q}^i_{+1}(\mathrm{d}\boldsymbol{x}) = \mathbb{A}_{+1}(\mathrm{d}\boldsymbol{x}) & \forall \boldsymbol{x} \in \mathbb{R}^n \\
& \displaystyle \frac{1}{\widehat{N}}\sum_{j=1}^{\widehat{N}} \widehat{\mathbb{Q}}^j_{+1}(\mathrm{d}\boldsymbol{x}) = \mathbb{A}_{+1}(\mathrm{d}\boldsymbol{x}) & \forall \boldsymbol{x} \in \mathbb{R}^n \\
& \displaystyle \frac{1}{N}\sum_{i=1}^{N} \mathbb{Q}^i_{-1}(\mathrm{d}\boldsymbol{x}) = \mathbb{A}_{-1}(\mathrm{d}\boldsymbol{x}) & \forall \boldsymbol{x} \in \mathbb{R}^n \\
& \displaystyle \frac{1}{\widehat{N}}\sum_{j=1}^{\widehat{N}} \widehat{\mathbb{Q}}^j_{-1}(\mathrm{d}\boldsymbol{x}) = \mathbb{A}_{-1}(\mathrm{d}\boldsymbol{x}) & \forall \boldsymbol{x} \in \mathbb{R}^n \\
& \mathbb{A}_{\pm 1} \in \mathcal{P}_+(\mathbb{R}^n), \ \mathbb{Q}^i_{\pm 1} \in \mathcal{P}_+(\mathbb{R}^n), \ \widehat{\mathbb{Q}}^j_{\pm 1} \in \mathcal{P}_+(\mathbb{R}^n) & \forall i \in [N], \ j \in [\widehat{N}].
\end{array}
$$

The following semi-infinite optimization problem, obtained by standard algebraic duality, is a strong dual to the above problem since $\varepsilon, \widehat{\varepsilon} > 0$ (Shapiro 2001).

$$
\begin{array}{lll}
\underset{\lambda,\widehat{\lambda},\boldsymbol{s},\widehat{\boldsymbol{s}},p_{\pm 1},\widehat{p}_{\pm 1}}{\text{minimize}} & \dfrac{1}{N}\left[N\varepsilon\lambda + \widehat{N}\widehat{\varepsilon}\widehat{\lambda} + \displaystyle\sum_{i=1}^{N} s_i + \sum_{j=1}^{\widehat{N}} \widehat{s}_j\right] & \\
\text{subject to} & \kappa\dfrac{1-y^i}{2}\lambda + \lambda\|\boldsymbol{x}^i - \boldsymbol{x}\|_q + s_i + \dfrac{p_{+1}(x)}{N} \geq \ell^{\alpha}_{\boldsymbol{\beta}}(\boldsymbol{x},+1) & \forall i \in [N],\ \forall \boldsymbol{x} \in \mathbb{R}^n \\
& \kappa\dfrac{1-\widehat{y}^j}{2}\widehat{\lambda} + \widehat{\lambda}\|\widehat{\boldsymbol{x}}^j - \boldsymbol{x}\|_q + \widehat{s}_j + \dfrac{\widehat{p}_{+1}(x)}{\widehat{N}} \geq 0 & \forall j \in [\widehat{N}],\ \forall \boldsymbol{x} \in \mathbb{R}^n \\
& \kappa\dfrac{1+y^i}{2}\lambda + \lambda\|\boldsymbol{x}^i - \boldsymbol{x}\|_q + s_i + \dfrac{p_{-1}(x)}{N} \geq \ell^{\alpha}_{\boldsymbol{\beta}}(\boldsymbol{x},-1) & \forall i \in [N],\ \forall \boldsymbol{x} \in \mathbb{R}^n \\
& \kappa\dfrac{1+\widehat{y}^j}{2}\widehat{\lambda} + \widehat{\lambda}\|\widehat{\boldsymbol{x}}^j - \boldsymbol{x}\|_q + \widehat{s}_j + \dfrac{\widehat{p}_{-1}(x)}{\widehat{N}} \geq 0 & \forall j \in [\widehat{N}],\ \forall \boldsymbol{x} \in \mathbb{R}^n \\
& p_{+1}(\boldsymbol{x}) + \widehat{p}_{+1}(\boldsymbol{x}) \leq 0 & \\
& p_{-1}(\boldsymbol{x}) + \widehat{p}_{-1}(\boldsymbol{x}) \leq 0 & \\
& \lambda \in \mathbb{R}_+,\ \widehat{\lambda} \in \mathbb{R}+,\ \boldsymbol{s} \in \mathbb{R}^N,\ \widehat{\boldsymbol{s}} \in \mathbb{R}^{\widehat{N}} & \\
& p_{\pm 1} : \mathbb{R}^n \mapsto \mathbb{R},\ \widehat{p}_{\pm 1} : \mathbb{R}^n \mapsto \mathbb{R}. &
\end{array}
$$

To eliminate the (function) variables $p_{+1}$ and $\widehat{p}_{+1}$, we first summarize the constraints they appear

$$
\begin{cases}
p_{+1}(\boldsymbol{x}) \geq N \cdot \left[\ell^{\alpha}_{\boldsymbol{\beta}}(\boldsymbol{x},+1) - s_i - \lambda\|\boldsymbol{x}^i - \boldsymbol{x}\|_q - \kappa\dfrac{1-y^i}{2}\lambda\right] & \forall i \in [N],\ \forall \boldsymbol{x} \in \mathbb{R}^n \\
\widehat{p}_{+1}(\boldsymbol{x}) \geq \widehat{N} \cdot \left[-\widehat{s}_j - \widehat{\lambda}\|\widehat{\boldsymbol{x}}^j - \boldsymbol{x}\|_q - \kappa\dfrac{1-\widehat{y}^j}{2}\widehat{\lambda}\right] & \forall j \in [\widehat{N}],\ \forall \boldsymbol{x} \in \mathbb{R}^n \\
p_{+1}(\boldsymbol{x}) + \widehat{p}_{+1}(\boldsymbol{x}) \leq 0 & \forall \boldsymbol{x} \in \mathbb{R}^n,
\end{cases}
$$

and notice that this system is equivalent to the epigraph-based reformulation of the following constraint

$$
\ell^{\alpha}_{\boldsymbol{\beta}}(\boldsymbol{x},+1) - s_i - \lambda\|\boldsymbol{x}^i - \boldsymbol{x}\|_q - \kappa\frac{1-y^i}{2}\lambda + \frac{\widehat{N}}{N} \cdot \left[-\widehat{s}_j - \widehat{\lambda}\|\widehat{\boldsymbol{x}}^j - \boldsymbol{x}\|_q - \kappa\frac{1-\widehat{y}^j}{2}\widehat{\lambda}\right] \leq 0
$$
$$
\forall i \in [N],\ \forall j \in [\widehat{N}],\ \forall \boldsymbol{x} \in \mathbb{R}^n.
$$

We can therefore eliminate $p_{+1}$ and $\widehat{p}_{+1}$. We can also eliminate $p_{-1}$ and $\widehat{p}_{-1}$ since we similarly have:

$$\begin{cases} p_{-1}(\boldsymbol{x}) \geq N \cdot \left[ \ell^{\alpha}_{\boldsymbol{\beta}}(\boldsymbol{x}, -1) - s_i - \lambda \|\boldsymbol{x}^i - \boldsymbol{x}\|_q - \kappa \dfrac{1+y^i}{2} \lambda \right] & \forall i \in [N], \ \forall \boldsymbol{x} \in \mathbb{R}^n \\ \widehat{p}_{-1}(\boldsymbol{x}) \geq \widehat{N} \cdot \left[ -\widehat{s}_j - \widehat{\lambda} \|\widehat{\boldsymbol{x}}^j - \boldsymbol{x}\|_q - \kappa \dfrac{1+\widehat{y}^j}{2} \widehat{\lambda} \right] & \forall j \in [\widehat{N}], \ \forall \boldsymbol{x} \in \mathbb{R}^n \\ p_{-1}(\boldsymbol{x}) + \widehat{p}_{-1}(\boldsymbol{x}) \leq 0 & \forall \boldsymbol{x} \in \mathbb{R}^n \end{cases}$$

$$\iff \ell^{\alpha}_{\boldsymbol{\beta}}(\boldsymbol{x}, -1) - s_i - \lambda \|\boldsymbol{x}^i - \boldsymbol{x}\|_q - \kappa \frac{1+y^i}{2} \lambda + \frac{\widehat{N}}{N} \cdot \left[ -\widehat{s}_j - \widehat{\lambda} \|\widehat{\boldsymbol{x}}^j - \boldsymbol{x}\|_q - \kappa \frac{1+\widehat{y}^j}{2} \widehat{\lambda} \right] \leq 0$$
$$\forall i \in [N], \ \forall j \in [\widehat{N}], \ \forall \boldsymbol{x} \in \mathbb{R}^n.$$

This trick of eliminating $p_{\pm 1}$, $\widehat{p}_{\pm 1}$ is due to the auxiliary distributions $\mathbb{A}_{\pm 1}$ that we introduced; without them, the dual problem is substantially harder to work with. We therefore obtain the following reformulation of the dual problem

$$\begin{array}{ll} \underset{\lambda, \widehat{\lambda}, \boldsymbol{s}, \widehat{\boldsymbol{s}}}{\text{minimize}} & \dfrac{1}{N} \left[ N\varepsilon\lambda + \widehat{N}\widehat{\varepsilon}\widehat{\lambda} + \displaystyle\sum_{i=1}^{N} s_i + \sum_{j=1}^{\widehat{N}} \widehat{s}_j \right] \\ \text{subject to} & \displaystyle\sup_{\boldsymbol{x} \in \mathbb{R}^n} \{ \ell^{\alpha}_{\boldsymbol{\beta}}(\boldsymbol{x}, +1) - \lambda \|\boldsymbol{x}^i - \boldsymbol{x}\|_q - \frac{\widehat{N}}{N} \widehat{\lambda} \|\widehat{\boldsymbol{x}}^j - \boldsymbol{x}\|_q \} \leq \\ & \qquad s_i + \kappa \dfrac{1-y^i}{2} \lambda + \dfrac{\widehat{N}}{N} \cdot \left[ \widehat{s}_j + \kappa \dfrac{1-\widehat{y}^j}{2} \widehat{\lambda} \right] \qquad \forall i \in [N], \ \forall j \in [\widehat{N}] \\ & \displaystyle\sup_{\boldsymbol{x} \in \mathbb{R}^n} \{ \ell^{\alpha}_{\boldsymbol{\beta}}(\boldsymbol{x}, -1) - \lambda \|\boldsymbol{x}^i - \boldsymbol{x}\|_q - \frac{\widehat{N}}{N} \widehat{\lambda} \|\widehat{\boldsymbol{x}}^j - \boldsymbol{x}\|_q \} \leq \\ & \qquad s_i + \kappa \dfrac{1+y^i}{2} \lambda + \dfrac{\widehat{N}}{N} \cdot \left[ \widehat{s}_j + \kappa \dfrac{1+\widehat{y}^j}{2} \widehat{\lambda} \right] \qquad \forall i \in [N], \ \forall j \in [\widehat{N}] \\ & \lambda \geq 0, \ \widehat{\lambda} \geq 0, \ \boldsymbol{s} \in \mathbb{R}^N_+, \ \widehat{\boldsymbol{s}} \in \mathbb{R}^{\widehat{N}}_+, \end{array}$$

where we replaced the $\forall \boldsymbol{x} \in \mathbb{R}^n$ with the worst case realizations by taking the suprema of the constraints over $\boldsymbol{x}$. We also added non-negativity on the definition of $\boldsymbol{s}$ and $\widehat{\boldsymbol{s}}$ which is without loss of generality since this is implied by the first two constraints, which is due to the fact that in the primal reformulation the “integrates to 1” constraints (whose associated dual variables are

$\boldsymbol{s}$ and $\widehat{\boldsymbol{s}}$) can be written as

$$\int_{\boldsymbol{x}\in\mathbb{R}^n} \mathbb{Q}^i_{+1}(\mathrm{d}\boldsymbol{x}) + \mathbb{Q}^i_{-1}(\mathrm{d}\boldsymbol{x}) \leq 1 \quad \forall i \in [N],$$

$$\int_{\boldsymbol{x}\in\mathbb{R}^n} \widehat{\mathbb{Q}}^j_{+1}(\mathrm{d}\boldsymbol{x}) + \widehat{\mathbb{Q}}^j_{-1}(\mathrm{d}\boldsymbol{x}) \leq 1 \quad \forall j \in [\widehat{N}],$$

due to the objective pressure. Relabeling $\frac{\widehat{N}}{N}\widehat{\lambda}$ as $\widehat{\lambda}$ and $\frac{\widehat{N}}{N}\widehat{s}_j$ as $\widehat{s}_j$ simplifies the problem to:

$$\begin{array}{lll}
\underset{\lambda,\widehat{\lambda},\boldsymbol{s},\widehat{\boldsymbol{s}}}{\text{minimize}} & \varepsilon\lambda + \widehat{\varepsilon}\widehat{\lambda} + \dfrac{1}{N}\displaystyle\sum_{i=1}^{N} s_i + \dfrac{1}{\widehat{N}}\sum_{i=1}^{\widehat{N}} \widehat{s}_j & \\
\text{subject to} & \displaystyle\sup_{\boldsymbol{x}\in\mathbb{R}^n} \{\ell^{\alpha}_{\boldsymbol{\beta}}(\boldsymbol{x}, +1) - \lambda\|\boldsymbol{x}^i - \boldsymbol{x}\|_q - \widehat{\lambda}\|\widehat{\boldsymbol{x}}^j - \boldsymbol{x}\|_q\} \leq & \\
& \qquad s_i + \kappa\dfrac{1-y^i}{2}\lambda + \widehat{s}_j + \kappa\dfrac{1-\widehat{y}^j}{2}\widehat{\lambda} & \forall i \in [N],\ \forall j \in [\widehat{N}] \\
& \displaystyle\sup_{\boldsymbol{x}\in\mathbb{R}^n} \{\ell^{\alpha}_{\boldsymbol{\beta}}(\boldsymbol{x}, -1) - \lambda\|\boldsymbol{x}^i - \boldsymbol{x}\|_q - \widehat{\lambda}\|\widehat{\boldsymbol{x}}^j - \boldsymbol{x}\|_q\} \leq & \\
& \qquad s_i + \kappa\dfrac{1+y^i}{2}\lambda + \widehat{s}_j + \kappa\dfrac{1+\widehat{y}^j}{2}\widehat{\lambda} & \forall i \in [N],\ \forall j \in [\widehat{N}] \\
& \lambda \geq 0,\ \widehat{\lambda} \geq 0,\ \boldsymbol{s} \in \mathbb{R}^N_+,\ \widehat{\boldsymbol{s}} \in \mathbb{R}^{\widehat{N}}_+. &
\end{array}$$

Combining all the sup-constraints with the help of an an auxiliary parameter $l \in \{-1, 1\}$ and replacing this problem with the inner problem of Inter-ARO concludes the proof. □

### 4.B.4 Proof of Proposition 18

We first present a technical lemma that will allow us to rewrite a specific type of difference of convex functions (DC) maximization problem that appears in the constraints of Inter-ARO. Rewriting such DC maximization problems is one of the key steps in reformulating Wasserstein DRO problems, and our lemma is inspired from Shafieezadeh-Abadeh et al. (2019, Lemma 47), Shafieezadeh-Abadeh et al. (2023, Theorem 3.8), and Belbasi et al. (2025, Lemma 1) who reformulate maximizing the difference of a convex function and a norm. Our DRO problem Inter-ARO, however, comprises two ambiguity sets, hence the DC term that we investigate will be the difference between a convex function and the sum of *two norms*. This requires a new analysis and we will see that Inter-ARO is NP-hard due to this additional difficulty.

**Lemma 31.** *Suppose that $L : \mathbb{R} \mapsto \mathbb{R}$ is a closed convex function, and $\|\cdot\|_q$ is a norm. For vectors $\boldsymbol{\omega}, \boldsymbol{a}, \widehat{\boldsymbol{a}} \in \mathbb{R}^n$ and scalars $\lambda, \widehat{\lambda} > 0$, we have:*

$$\begin{aligned}
&\sup_{\boldsymbol{x}\in\mathbb{R}^n} \{L(\boldsymbol{\omega}^\top \boldsymbol{x}) - \lambda\|\boldsymbol{a}-\boldsymbol{x}\|_q - \widehat{\lambda}\|\widehat{\boldsymbol{a}}-\boldsymbol{x}\|_q\}\\
= &\sup_{\theta\in\mathrm{dom}(L^*)} \; -L^*(\theta) + \theta\cdot\boldsymbol{\omega}^\top\boldsymbol{a} + \theta\cdot \inf_{\boldsymbol{z}\in\mathbb{R}^n}\{\boldsymbol{z}^\top(\widehat{\boldsymbol{a}}-\boldsymbol{a}) \;:\; |\theta|\cdot\|\boldsymbol{\omega}-\boldsymbol{z}\|_{q^\star} \le \lambda,\; |\theta|\cdot\|\boldsymbol{z}\|_{q^\star} \le \widehat{\lambda}\}
\end{aligned}$$

*Proof.* We denote by $f_{\boldsymbol{\omega}}(\boldsymbol{x}) = \boldsymbol{\omega}^\top\boldsymbol{x}$ and by $g$ the convex function $g(\boldsymbol{x}) = g_1(\boldsymbol{x}) + g_2(\boldsymbol{x})$ where $g_1(\boldsymbol{x}) := \lambda\|\boldsymbol{a}-\boldsymbol{x}\|_q$ and $g_2(\boldsymbol{x}) := \widehat{\lambda}\|\widehat{\boldsymbol{a}}-\boldsymbol{x}\|_q$, and reformulate the sup problem as

$$\sup_{\boldsymbol{x}\in\mathbb{R}^n} L(\boldsymbol{\omega}^\top\boldsymbol{x}) - g(\boldsymbol{x}) \;=\; \sup_{\boldsymbol{x}\in\mathbb{R}^n} (L\circ f_{\boldsymbol{\omega}})(\boldsymbol{x}) - g(\boldsymbol{x}) \;=\; \sup_{\boldsymbol{z}\in\mathbb{R}^n} g^*(\boldsymbol{z}) - (L\circ f_{\boldsymbol{\omega}})^*(\boldsymbol{z}),$$

where the first identity follows from the definition of composition and the second identity employs Toland's duality (Toland 1978) to rewrite difference of convex functions optimization.

By using infimal convolutions (Rockafellar 1997, Theorem 16.4), we can reformulate $g^*$:

$$\begin{aligned}
g^*(\boldsymbol{z}) &= \inf_{\boldsymbol{z}_1,\boldsymbol{z}_2}\{g_1^*(\boldsymbol{z}_1) + g_2^*(\boldsymbol{z}_2) \;:\; \boldsymbol{z}_1+\boldsymbol{z}_2 = \boldsymbol{z}\}\\
&= \inf_{\boldsymbol{z}_1,\boldsymbol{z}_2}\{\boldsymbol{z}_1^\top\boldsymbol{a} + \boldsymbol{z}_2^\top\widehat{\boldsymbol{a}} \;:\; \boldsymbol{z}_1+\boldsymbol{z}_2=\boldsymbol{z},\; \|\boldsymbol{z}_1\|_{q^\star}\le\lambda,\; \|\boldsymbol{z}_2\|_{q^\star}\le\widehat{\lambda}\},
\end{aligned}$$

where the second step uses the definitions of $g_1^*(\boldsymbol{z}_1)$ and $g_2^*(\boldsymbol{z}_2)$. Moreover, we show

$$\begin{aligned}
(L\circ f_{\boldsymbol{\omega}})^*(\boldsymbol{z}) &= \sup_{\boldsymbol{x}\in\mathbb{R}^n} \boldsymbol{z}^\top\boldsymbol{x} - L(\boldsymbol{\omega}^\top\boldsymbol{x})\\
&= \sup_{t\in\mathbb{R},\;\boldsymbol{x}\in\mathbb{R}^n} \{\boldsymbol{z}^\top\boldsymbol{x} - L(t) \;:\; t = \boldsymbol{\omega}^\top\boldsymbol{x}\}\\
&= \inf_{\theta\in\mathbb{R}}\; \sup_{t\in\mathbb{R},\;\boldsymbol{x}\in\mathbb{R}^n} \boldsymbol{z}^\top\boldsymbol{x} - L(t) - \theta\cdot(\boldsymbol{\omega}^\top\boldsymbol{x} - t)\\
&= \inf_{\theta\in\mathbb{R}}\; \sup_{t\in\mathbb{R}}\; \sup_{\boldsymbol{x}\in\mathbb{R}^n} (\boldsymbol{z} - \theta\cdot\boldsymbol{\omega})^\top\boldsymbol{x} - L(t) + \theta\cdot t\\
&= \inf_{\theta\in\mathbb{R}}\; \sup_{t\in\mathbb{R}} \begin{cases} -L(t) + \theta\cdot t & \text{if } \theta\cdot\boldsymbol{\omega} = \boldsymbol{z}\\ +\infty & \text{otherwise.}\end{cases}\\
&= \inf_{\theta\in\mathbb{R}} \begin{cases} L^*(\theta) & \text{if } \theta\cdot\boldsymbol{\omega} = \boldsymbol{z}\\ +\infty & \text{otherwise.}\end{cases}\\
&= \inf_{\theta\in\mathrm{dom}(L^*)} \{L^*(\theta) \;:\; \theta\cdot\boldsymbol{\omega} = \boldsymbol{z}\},
\end{aligned}$$

where the first identity follows from the definition of the convex conjugate, the second identity introduces an additional variable to make this an equality-constrained optimization problem, the third identity takes the Lagrange dual (which is a strong dual since the problem maximizes a concave objective with a single equality constraint), the fourth identity rearranges the expressions, the fifth identity exploits unboundedness of $\boldsymbol{x}$, the sixth identity uses the definition of convex conjugates and the final identity replaces the feasible set $\theta \in \mathbb{R}$ with the domain of $L^\star$ without loss of generality as this is an inf-problem.

Replacing the conjugates allows us to conclude that the maximization problem equals

$$\begin{aligned}
& \sup_{\boldsymbol{z} \in \mathbb{R}^n} g^*(\boldsymbol{z}) + \sup_{\theta \in \mathrm{dom}(L^*)} \{-L^*(\theta) \ : \ \theta \cdot \boldsymbol{\omega} = \boldsymbol{z}\} \\
= & \sup_{\boldsymbol{z} \in \mathbb{R}^n, \ \theta \in \mathrm{dom}(L^*)} \{g^*(\boldsymbol{z}) - L^*(\theta) \ : \ \theta \cdot \boldsymbol{\omega} = \boldsymbol{z}\} \\
= & \sup_{\theta \in \mathrm{dom}(L^*)} g^*(\theta \cdot \boldsymbol{\omega}) - L^*(\theta) \\
= & \sup_{\theta \in \mathrm{dom}(L^*)} - L^*(\theta) + \inf_{\boldsymbol{z}_1, \boldsymbol{z}_2 \in \mathbb{R}^n} \{\boldsymbol{z}_1^\top \boldsymbol{a} + \boldsymbol{z}_2^\top \widehat{\boldsymbol{a}} \ : \ \boldsymbol{z}_1 + \boldsymbol{z}_2 = \theta \cdot \boldsymbol{\omega}, \ \|\boldsymbol{z}_1\|_{q^\star} \leq \lambda, \ \|\boldsymbol{z}_2\|_{q^\star} \leq \widehat{\lambda}\} \\
= & \sup_{\theta \in \mathrm{dom}(L^*)} - L^*(\theta) + \theta \cdot \inf_{\boldsymbol{z}_1, \boldsymbol{z}_2 \in \mathbb{R}^n} \{\boldsymbol{z}_1^\top \boldsymbol{a} + \boldsymbol{z}_2^\top \widehat{\boldsymbol{a}} \ : \ \boldsymbol{z}_1 + \boldsymbol{z}_2 = \boldsymbol{\omega}, \ |\theta| \cdot \|\boldsymbol{z}_1\|_{q^\star} \leq \lambda, \ |\theta| \cdot \|\boldsymbol{z}_2\|_{q^\star} \leq \widehat{\lambda}\} \\
= & \sup_{\theta \in \mathrm{dom}(L^*)} - L^*(\theta) + \theta \cdot \boldsymbol{\omega}^\top \boldsymbol{a} + \theta \cdot \inf_{\boldsymbol{z} \in \mathbb{R}^n} \{\boldsymbol{z}^\top (\widehat{\boldsymbol{a}} - \boldsymbol{a}) \ : \ |\theta| \cdot \|\boldsymbol{\omega} - \boldsymbol{z}\|_{q^\star} \leq \lambda, \ |\theta| \cdot \|\boldsymbol{z}\|_{q^\star} \leq \widehat{\lambda}\}.
\end{aligned}$$

Here, the first identity follows from writing the problem as a single maximization problem, the second identity follows from the equality constraint, the third identity follows from the definition of the conjugate $g^*$, the fourth identity is due to relabeling $\boldsymbol{z}_1 = \theta \cdot \boldsymbol{z}_1$ and $\boldsymbol{z}_2 = \theta \cdot \boldsymbol{z}_2$, and the fifth identity is due to a variable change ($\boldsymbol{z}_1 = \boldsymbol{\omega} - \boldsymbol{z}_2$ relabeled as $\boldsymbol{z}$). □

DC maximization terms similar to the one dealt by Lemma 31 appear on the left-hand side of the constraints of Inter-ARO (*cf.* formulation in Proposition 17). These constraints would admit a tractable reformulation for the case without auxiliary data because the inf-term in the reformulation presented in Lemma 31 does not appear in such cases. To see this, eliminate the second norm (the one associated with auxiliary data) by taking $\widehat{\lambda} = 0$, which will cause the constraint $|\theta| \cdot \|\boldsymbol{z}\|_{q^\star} \leq \widehat{\lambda}$ to force $\boldsymbol{z} = \boldsymbol{0}$, and the alternative formulation will thus be:

$$\begin{cases} \sup\limits_{\theta \in \mathrm{dom}(L^*)} \{-L^*(\theta) + \theta \cdot \boldsymbol{\omega}^\top \boldsymbol{a}\} & \text{if } \sup_{\theta \in \mathrm{dom}(L^*)} \{|\theta|\} \cdot \|\boldsymbol{z}\|_{q^\star} \leq \lambda \\ +\infty & \text{otherwise} \end{cases}$$

$$= \begin{cases} L(\boldsymbol{\omega}^\top \boldsymbol{a}) & \text{if } \mathrm{Lip}(L) \cdot \|\boldsymbol{z}\|_{q^\star} \leq \lambda \\ +\infty & \text{otherwise,} \end{cases}$$

where we used the fact that $L = L^{**}$ and $\sup_{\theta \in \mathrm{dom}(L)} |\theta| = \mathrm{Lip}(L)$ since $L$ is closed convex (Rockafellar 1997, Corollary 13.3.3). Hence, the DC maximization can be represented with a convex function with an additional convex inequality, making the constraints tractable for the case without auxiliary data. For the case with auxiliary data, however, the $\sup_\theta \inf_{\boldsymbol{z}}$ structure makes these constraints equivalent to two-stage robust constraints (with uncertain parameter $\theta$ and adjustable variable $\boldsymbol{z}$), bringing an adjustable robust optimization (Ben-Tal et al. 2004, Yanıkoğlu et al. 2019) perspective to Inter-ARO. By using the univariate representation $\ell^\alpha_{\boldsymbol{\beta}}(\boldsymbol{x}, y) = L^\alpha(y \cdot \boldsymbol{\beta}^\top \boldsymbol{x})$, Inter-ARO can be written as

$$\begin{array}{lll} \underset{\boldsymbol{\beta}, \lambda, \widehat{\lambda}, \boldsymbol{s}, \widehat{\boldsymbol{s}}}{\text{minimize}} & \varepsilon\lambda + \widehat{\varepsilon}\widehat{\lambda} + \dfrac{1}{N}\displaystyle\sum_{j=1}^{N} s_j + \dfrac{1}{\widehat{N}}\displaystyle\sum_{i=1}^{\widehat{N}} \widehat{s}_i & \\ \text{subject to} & \displaystyle\sup_{\boldsymbol{x} \in \mathbb{R}^n} \{L^\alpha(\boldsymbol{\beta}^\top \boldsymbol{x}) - \lambda \|\boldsymbol{x}^i - \boldsymbol{x}\|_q - \widehat{\lambda} \|\widehat{\boldsymbol{x}}^j - \boldsymbol{x}\|_q\} \leq & \\ & \qquad s_i + \kappa \dfrac{1 - y^i}{2} \lambda + \widehat{s}_j + \kappa \dfrac{1 - \widehat{y}^j}{2} \widehat{\lambda} & \forall i \in [N],\ \forall j \in [\widehat{N}] \\ & \displaystyle\sup_{\boldsymbol{x} \in \mathbb{R}^n} \{L^\alpha(-\boldsymbol{\beta}^\top \boldsymbol{x}) - \lambda \|\boldsymbol{x}^i - \boldsymbol{x}\|_q - \widehat{\lambda} \|\widehat{\boldsymbol{x}}^j - \boldsymbol{x}\|_q\} \leq & \\ & \qquad s_i + \kappa \dfrac{1 + y^i}{2} \lambda + \widehat{s}_j + \kappa \dfrac{1 + \widehat{y}^j}{2} \widehat{\lambda} & \forall i \in [N],\ \forall j \in [\widehat{N}] \\ & \boldsymbol{\beta} \in \mathbb{R}^n,\ \lambda \geq 0,\ \widehat{\lambda} \geq 0,\ \boldsymbol{s} \in \mathbb{R}^N_+,\ \widehat{\boldsymbol{s}} \in \mathbb{R}^{\widehat{N}}_+, & \end{array}$$

and applying Lemma 31 to the left-hand side of the constraints gives:

$$
\begin{aligned}
\underset{\boldsymbol{\beta},\lambda,\widehat{\lambda},\boldsymbol{s},\widehat{\boldsymbol{s}}}{\text{minimize}} \quad & \varepsilon\lambda + \widehat{\varepsilon}\widehat{\lambda} + \frac{1}{N}\sum_{j=1}^{N} s_j + \frac{1}{\widehat{N}}\sum_{i=1}^{\widehat{N}} \widehat{s}_i \\
\text{subject to} \quad & \sup_{\theta \in \mathrm{dom}(L^*)} \; - L^{\alpha *}(\theta) + \theta \cdot \boldsymbol{\beta}^\top \boldsymbol{x}^i + \theta \cdot \inf_{\boldsymbol{z} \in \mathbb{R}^n} \{ \boldsymbol{z}^\top (\widehat{\boldsymbol{x}}^j - \boldsymbol{x}^i) \; : \; |\theta| \cdot \|\boldsymbol{\beta} - \boldsymbol{z}\|_{q^\star} \leq \lambda, \; |\theta| \cdot \|\boldsymbol{z}\|_{q^\star} \leq \widehat{\lambda} \} \leq \\
& \qquad s_i + \kappa \frac{1 - y^i}{2}\lambda + \widehat{s}_j + \kappa \frac{1 - \widehat{y}^j}{2}\widehat{\lambda} \quad \forall i \in [N], \; \forall j \in [\widehat{N}] \\
& \sup_{\theta \in \mathrm{dom}(L^*)} \; - L^{\alpha *}(\theta) - \theta \cdot \boldsymbol{\beta}^\top \boldsymbol{x}^i + \theta \cdot \inf_{\boldsymbol{z} \in \mathbb{R}^n} \{ \boldsymbol{z}^\top (\widehat{\boldsymbol{x}}^j - \boldsymbol{x}^i) \; : \; |\theta| \cdot \|-\boldsymbol{\beta} - \boldsymbol{z}\|_{q^\star} \leq \lambda, \; |\theta| \cdot \|\boldsymbol{z}\|_{q^\star} \leq \widehat{\lambda} \} \leq \\
& \qquad s_i + \kappa \frac{1 + y^i}{2}\lambda + \widehat{s}_j + \kappa \frac{1 + \widehat{y}^j}{2}\widehat{\lambda} \quad \forall i \in [N], \; \forall j \in [\widehat{N}] \\
& \boldsymbol{\beta} \in \mathbb{R}^n, \; \lambda \geq 0, \; \widehat{\lambda} \geq 0, \; \boldsymbol{s} \in \mathbb{R}^N_+, \; \widehat{\boldsymbol{s}} \in \mathbb{R}^{\widehat{N}}_+.
\end{aligned}
\tag{70}
$$

Which, equivalently, can be written as the following problem with $2N \cdot \widehat{N}$ two-stage robust constraints:

$$
\begin{aligned}
\underset{\boldsymbol{\beta},\lambda,\widehat{\lambda},\boldsymbol{s},\widehat{\boldsymbol{s}}}{\text{minimize}} \quad & \varepsilon\lambda + \widehat{\varepsilon}\widehat{\lambda} + \frac{1}{N}\sum_{j=1}^{N} s_j + \frac{1}{\widehat{N}}\sum_{i=1}^{\widehat{N}} \widehat{s}_i \\
\text{subject to} \quad & \left[ \forall \theta \in \mathrm{dom}(L^*), \; \exists \boldsymbol{z} \in \mathbb{R}^n \; : \begin{cases} -L^{\alpha *}(\theta) + \theta \cdot \boldsymbol{\beta}^\top \boldsymbol{x}^i + \theta \cdot \boldsymbol{z}^\top (\widehat{\boldsymbol{x}}^j - \boldsymbol{x}^i) \leq s_i + \kappa \dfrac{1 - y^i}{2}\lambda + \widehat{s}_j + \kappa \dfrac{1 - \widehat{y}^j}{2}\widehat{\lambda} \\ |\theta| \cdot \|\boldsymbol{\beta} - \boldsymbol{z}\|_{q^\star} \leq \lambda \\ |\theta| \cdot \|\boldsymbol{z}\|_{q^\star} \leq \widehat{\lambda} \end{cases} \right] \\
& \qquad\qquad\qquad\qquad\qquad \forall i \in [N], \; \forall j \in [\widehat{N}] \\
& \left[ \forall \theta \in \mathrm{dom}(L^*), \; \exists \boldsymbol{z} \in \mathbb{R}^n \; : \begin{cases} -L^{\alpha *}(\theta) - \theta \cdot \boldsymbol{\beta}^\top \boldsymbol{x}^i + \theta \cdot \boldsymbol{z}^\top (\widehat{\boldsymbol{x}}^j - \boldsymbol{x}^i) \leq s_i + \kappa \dfrac{1 + y^i}{2}\lambda + \widehat{s}_j + \kappa \dfrac{1 + \widehat{y}^j}{2}\widehat{\lambda} \\ |\theta| \cdot \|-\boldsymbol{\beta} - \boldsymbol{z}\|_{q^\star} \leq \lambda \\ |\theta| \cdot \|\boldsymbol{z}\|_{q^\star} \leq \widehat{\lambda} \end{cases} \right] \\
& \qquad\qquad\qquad\qquad\qquad \forall i \in [N], \; \forall j \in [\widehat{N}] \\
& \boldsymbol{\beta} \in \mathbb{R}^n, \; \lambda \geq 0, \; \widehat{\lambda} \geq 0, \; \boldsymbol{s} \in \mathbb{R}^N_+, \; \widehat{\boldsymbol{s}} \in \mathbb{R}^{\widehat{N}}_+.
\end{aligned}
\tag{Inter-adjustable}
$$

By using adjustable robust optimization theory, we show that this problem is NP-hard even in the simplest setting. To this end, take $N = \widehat{N} = 1$ as well as $\kappa = 0$; the formulation presented in Proposition 17 reduces to:

$$\begin{array}{ll} \underset{\boldsymbol{\beta},\lambda,\widehat{\lambda},s,\widehat{s}}{\text{minimize}} & \varepsilon\lambda + \widehat{\varepsilon}\widehat{\lambda} + s + \widehat{s} \\ \text{subject to} & \sup_{\boldsymbol{x}\in\mathbb{R}^n} \{\ell^{\alpha}_{\boldsymbol{\beta}}(\boldsymbol{x}, l) - \lambda\|\boldsymbol{x}^1 - \boldsymbol{x}\|_q - \widehat{\lambda}\|\widehat{\boldsymbol{x}}^1 - \boldsymbol{x}\|_q\} \leq s_1 + \widehat{s}_1 \quad \forall l \in \{-1, 1\} \\ & \boldsymbol{\beta} \in \mathbb{R}^n,\ \lambda \geq 0,\ \widehat{\lambda} \geq 0,\ s \geq 0,\ \widehat{s} \geq 0. \end{array}$$

The worst case realization of $l \in \{-1, 1\}$ will always make $\ell^{\alpha}_{\boldsymbol{\beta}}(\boldsymbol{x}, l) = \log(1 + \exp(-l \cdot \boldsymbol{\beta}^\top \boldsymbol{x} + \alpha \cdot \|\boldsymbol{\beta}\|_{p^\star}))$ equal to $\varsigma^{\alpha}_{\boldsymbol{\beta}}(\boldsymbol{x}) = \log(1 + \exp(|l \cdot \boldsymbol{\beta}^\top \boldsymbol{x}| + \alpha \cdot \|\boldsymbol{\beta}\|_{p^\star}))$, where $\varsigma$ inherits similar properties from $\ell$: it is convex in $\boldsymbol{\beta}$ and its univariate representation $S^\alpha$ has the same Lipschitz constant with $L^\alpha$. We can thus represent the above problem as

$$\begin{array}{ll} \underset{\boldsymbol{\beta},\lambda,\widehat{\lambda},s,\widehat{s}}{\text{minimize}} & \varepsilon\lambda + \widehat{\varepsilon}\widehat{\lambda} + s + \widehat{s} \\ \text{subject to} & \sup_{\boldsymbol{x}\in\mathbb{R}^n} \{S^\alpha(\boldsymbol{\beta}^\top \boldsymbol{x}) - \lambda\|\boldsymbol{x}^1 - \boldsymbol{x}\|_q - \widehat{\lambda}\|\widehat{\boldsymbol{x}}^1 - \boldsymbol{x}\|_q\} \leq s + \widehat{s} \\ & \boldsymbol{\beta} \in \mathbb{R}^n,\ \lambda \geq 0,\ \widehat{\lambda} \geq 0,\ s \geq 0,\ \widehat{s} \geq 0. \end{array}$$

Substituting $s + \widehat{s}$ into the objective (due to the objective pressure) allows us to reformulate the above problem as

$$\begin{array}{ll} \underset{\boldsymbol{\beta},\lambda,\widehat{\lambda}}{\text{minimize}} & \varepsilon\lambda + \widehat{\varepsilon}\widehat{\lambda} + \sup_{\boldsymbol{x}\in\mathbb{R}^n} \{S^\alpha(\boldsymbol{\beta}^\top \boldsymbol{x}) - \lambda\|\boldsymbol{x}^1 - \boldsymbol{x}\|_q - \widehat{\lambda}\|\widehat{\boldsymbol{x}}^1 - \boldsymbol{x}\|_q\} \\ \text{subject to} & \boldsymbol{\beta} \in \mathbb{R}^n,\ \lambda \geq 0,\ \widehat{\lambda} \geq 0, \end{array} \tag{71}$$

and an application of Lemma 31 leads us to the following reformulation:

$$\inf_{\substack{\boldsymbol{\beta}\in\mathbb{R}^n \\ \lambda\geq 0,\widehat{\lambda}\geq 0}} \ \sup_{\theta\in\text{dom}(S^*)} \ \inf_{\boldsymbol{z}\in\mathbb{R}^n} \left\{ \varepsilon\lambda + \widehat{\varepsilon}\widehat{\lambda} - S^{\alpha *}(\theta) + \theta \cdot \boldsymbol{\beta}^\top \boldsymbol{x}^1 + \underbrace{\theta \cdot \boldsymbol{z}^\top (\widehat{\boldsymbol{x}}^1 - \boldsymbol{x}^1)}_{(1)} \ : \ \underbrace{|\theta| \cdot \|\boldsymbol{\beta} - \boldsymbol{z}\|_{q^\star} \leq \lambda}_{(2)},\ |\theta| \cdot \|\boldsymbol{z}\|_{q^\star} \leq \widehat{\lambda} \right\}.$$

The objective term (1) has a product of the uncertain parameter $\theta$ and the adjustable variable $\boldsymbol{z}$, and even when (2) is linear such as in the case of $q = 1$ the product of the uncertain parameter with both the decision variable $\boldsymbol{\beta}$ and the adjustable variable $\boldsymbol{z}$ still appear since:

$$|\theta| \cdot \|\boldsymbol{\beta} - \boldsymbol{z}\|_\infty \leq \lambda \iff -\lambda \leq \theta\boldsymbol{\beta} - \theta\boldsymbol{z} \leq \lambda.$$

This reduces problem (71) to a generic two-stage robust optimization problem with random recourse (Subramanyam et al. 2020, Problem 1) which is proven to be NP-hard even if $S^{\alpha *}$ was constant (Guslitser 2002). □

### 4.B.5 Proof of Theorem 11

Consider the reformulation Inter-adjustable of Inter-ARO that we introduced in the proof of Proposition 18. For any $i \in [N]$ and $j \in [\widehat{N}]$, the corresponding constraint in the first group of 'adjustable robust' (∀, ∃) constraints will be:

$$\forall \theta \in \operatorname{dom}(L^*), \exists \boldsymbol{z} \in \mathbb{R}^n : \begin{cases} -L^{\alpha *}(\theta) + \theta \cdot \boldsymbol{\beta}^\top \boldsymbol{x}^i + \theta \cdot \boldsymbol{z}^\top (\widehat{\boldsymbol{x}}^j - \boldsymbol{x}^i) \leq s_i + \kappa \dfrac{1 - y^i}{2} \lambda + \widehat{s}_j + \kappa \dfrac{1 - \widehat{y}^j}{2} \widehat{\lambda} \\ |\theta| \cdot \|\boldsymbol{\beta} - \boldsymbol{z}\|_{q^\star} \leq \lambda \\ |\theta| \cdot \|\boldsymbol{z}\|_{q^\star} \leq \widehat{\lambda}. \end{cases}$$

By changing the order of ∀ and ∃, we obtain:

$$\exists \boldsymbol{z} \in \mathbb{R}^n, \forall \theta \in \operatorname{dom}(L^*) : \begin{cases} -L^{\alpha *}(\theta) + \theta \cdot \boldsymbol{\beta}^\top \boldsymbol{x}^i + \theta \cdot \boldsymbol{z}^\top (\widehat{\boldsymbol{x}}^j - \boldsymbol{x}^i) \leq s_i + \kappa \dfrac{1 - y^i}{2} \lambda + \widehat{s}_j + \kappa \dfrac{1 - \widehat{y}^j}{2} \widehat{\lambda} \\ |\theta| \cdot \|\boldsymbol{\beta} - \boldsymbol{z}\|_{q^\star} \leq \lambda \\ |\theta| \cdot \|\boldsymbol{z}\|_{q^\star} \leq \widehat{\lambda}. \end{cases}$$

Notice that this is a safe approximation, since any fixed $\boldsymbol{z}$ satisfying the latter system is a feasible static solution in the former system, meaning that for every realization of $\theta$ in the first system, the inner $\exists \boldsymbol{z}$ can always 'play' the same $\boldsymbol{z}$ that is feasible in the latter system (hence the latter is named the *static* relaxation, Bertsimas et al. 2015). In the relaxed system, we can drop $\forall \theta$ and keep its worst-case realization instead:

$$\exists \boldsymbol{z} \in \mathbb{R}^n : \begin{cases} \sup_{\theta \in \operatorname{dom}(L^*)} \{-L^{\alpha *}(\theta) + \theta \cdot \boldsymbol{\beta}^\top \boldsymbol{x}^i + \theta \cdot \boldsymbol{z}^\top (\widehat{\boldsymbol{x}}^j - \boldsymbol{x}^i)\} \leq s_i + \kappa \dfrac{1 - y^i}{2} \lambda + \widehat{s}_j + \kappa \dfrac{1 - \widehat{y}^j}{2} \widehat{\lambda} \\ \sup_{\theta \in \operatorname{dom}(L^*)} \{|\theta|\} \cdot \|\boldsymbol{\beta} - \boldsymbol{z}\|_{q^\star} \leq \lambda \\ \sup_{\theta \in \operatorname{dom}(L^*)} \{|\theta|\} \cdot \|\boldsymbol{z}\|_{q^\star} \leq \widehat{\lambda}. \end{cases}$$

The term $\sup_{\theta \in \operatorname{dom}(L^*)} \{-L^{\alpha *}(\theta) + \theta \cdot \boldsymbol{\beta}^\top \boldsymbol{x}^i + \theta \cdot \boldsymbol{z}^\top (\widehat{\boldsymbol{x}}^j - \boldsymbol{x}^i)\}$ is the definition of the biconjugate $L^{\alpha **}(\boldsymbol{\beta}^\top \boldsymbol{x}^i + \boldsymbol{z}^\top (\widehat{\boldsymbol{x}}^j - \boldsymbol{x}^i))$. Since $L^\alpha$ is a closed convex function, we have $L^{\alpha **} = L^\alpha$ (Rockafellar 1997, Corollary 12.2.1). Moreover, $\sup_{\theta \in \operatorname{dom}(L^*)} \{|\theta|\}$ is an alternative representation of the Lipschitz constant of the function $L^\alpha$ (Rockafellar 1997, Corollary 13.3.3), which is equal to 1 as we showed earlier. The adjustable robust constraint thus reduces to:

$$\exists \boldsymbol{z} \in \mathbb{R}^n : \begin{cases} L^\alpha(\boldsymbol{\beta}^\top \boldsymbol{x}^i + \boldsymbol{z}^\top(\widehat{\boldsymbol{x}}^j - \boldsymbol{x}^i)) \leq s_i + \kappa \dfrac{1-y^i}{2}\lambda + \widehat{s}_j + \kappa \dfrac{1-\widehat{y}^j}{2}\widehat{\lambda} \\ \|\boldsymbol{\beta} - \boldsymbol{z}\|_{q^\star} \leq \lambda \\ \|\boldsymbol{z}\|_{q^\star} \leq \widehat{\lambda} \end{cases}$$

as a result of the static relaxation. This relaxed reformulation applies to all $i \in [N]$ and $j \in [\widehat{N}]$ as well as to the second group of adjustable robust constraints analogously. Replacing each constraint of Inter-adjustable with this system concludes the proof. □

### 4.B.6 Proof of Corollary 9

To prove the first statement, take $\widehat{\lambda} = 0$ and observe the constraint $\|\boldsymbol{z}_{ij}^l\|_{q^\star} \leq \widehat{\lambda}$ implies $\boldsymbol{z}_{ij}^l = \boldsymbol{0}$ for all $l \in \{-1, 1\}$, $i \in [N]$, $j \in [\widehat{N}]$. The optimization problem can thus be written without those variables:

$$\begin{aligned} \underset{\boldsymbol{\beta}, \lambda, \boldsymbol{s}, \widehat{\boldsymbol{s}}}{\text{minimize}} \quad & \varepsilon\lambda + \frac{1}{N}\sum_{i=1}^{N} s_i + \frac{1}{\widehat{N}}\sum_{j=1}^{\widehat{N}} \widehat{s}_j \\ \text{subject to} \quad & L^\alpha(l\boldsymbol{\beta}^\top \boldsymbol{x}^i) \leq s_i + \kappa\frac{1 - ly^i}{2}\lambda + \widehat{s}_j \quad \forall l \in \{-1, 1\},\ \forall i \in [N],\ \forall j \in [\widehat{N}] \\ & \|\boldsymbol{\beta}\|_{q^\star} \leq \lambda \\ & \boldsymbol{\beta} \in \mathbb{R}^n,\ \lambda \geq 0,\ \boldsymbol{s} \in \mathbb{R}_+^N,\ \widehat{\boldsymbol{s}} \in \mathbb{R}_+^{\widehat{N}}. \end{aligned}$$

Notice that optimal solutions should satisfy $\widehat{s}_j = \widehat{s}_{j'}$ for all $j, j' \in [N]$. To see this, assume for contradiction that $\exists j, j' \in [N]$ such that $\widehat{s}_j < \widehat{s}_{j'}$. If a constraint indexed with $(l, i, j)$ for arbitrary $l \in \{-1, 1\}$ and $i \in [N]$ is feasible, it means the consraint indexed with $(l, i, j')$ cannot be tight given that these constraints are identical except for the $\widehat{s}_j$ or $\widehat{s}_{j'}$ appearing on the right hand side. Hence, such a solution cannot be optimal as this is a minimization problem, and updating $\widehat{s}_{j'}$ as $\widehat{s}_j$ preserves the feasibility of the problem while decreasing the objective value.

We can thus use a single variable $\tau \in \mathbb{R}_+$ and rewrite the problem as

$$\begin{array}{ll} \underset{\boldsymbol{\beta},\lambda,\boldsymbol{s},\widehat{\boldsymbol{s}}}{\text{minimize}} & \varepsilon\lambda + \dfrac{1}{N}\displaystyle\sum_{i=1}^{N}(s_i + \tau) \\ \text{subject to} & L^{\alpha}(\boldsymbol{\beta}^\top \boldsymbol{x}^i) \leq s_i + \kappa \dfrac{1-y^i}{2}\lambda + \tau \qquad \forall i \in [N] \\ & L^{\alpha}(-\boldsymbol{\beta}^\top \boldsymbol{x}^i) \leq s_i + \kappa \dfrac{1+y^i}{2}\lambda + \tau \quad \forall i \in [N] \\ & \|\boldsymbol{\beta}\|_{q^\star} \leq \lambda \\ & \boldsymbol{\beta} \in \mathbb{R}^n, \ \lambda \geq 0, \ \boldsymbol{s} \in \mathbb{R}^N_+, \ \widehat{\boldsymbol{s}} \in \mathbb{R}^{\widehat{N}}_+, \end{array}$$

where we also eliminated the index $l \in \{-1, 1\}$ by writing the constraints explicitly. Since $s_i$ and $\tau$ both appear as $s_i + \tau$ in this problem, we can use a variable change where we relabel $s_i + \tau$ as $s_i$ (or, equivalently set $\tau = 0$ without any optimality loss). Moreover, the constraints with index $i \in [N]$ are

$$\begin{cases} L^{\alpha}(\boldsymbol{\beta}^\top \boldsymbol{x}^i) \leq s_i + \tau \\ L^{\alpha}(-\boldsymbol{\beta}^\top \boldsymbol{x}^i) \leq s_i + \kappa\lambda + \tau \end{cases} = \begin{cases} L^{\alpha}(y^i \cdot \boldsymbol{\beta}^\top \boldsymbol{x}^i) \leq s_i + \tau \\ L^{\alpha}(-y^i \cdot \boldsymbol{\beta}^\top \boldsymbol{x}^i) \leq s_i + \kappa\lambda + \tau \end{cases}$$

if $y^i = 1$, and similarly they are

$$\begin{cases} L^{\alpha}(\boldsymbol{\beta}^\top \boldsymbol{x}^i) \leq s_i + \kappa\lambda + \tau \\ L^{\alpha}(-\boldsymbol{\beta}^\top \boldsymbol{x}^i) \leq s_i + \tau \end{cases} = \begin{cases} L^{\alpha}(-y^i \cdot \boldsymbol{\beta}^\top \boldsymbol{x}^i) \leq s_i + \kappa\lambda + \tau \\ L^{\alpha}(y^i \cdot \boldsymbol{\beta}^\top \boldsymbol{x}^i) \leq s_i + \tau \end{cases}$$

if $y^i = -1$. Since these are identical, the problem can finally be written as

$$\begin{array}{ll} \underset{\boldsymbol{\beta},\lambda,\boldsymbol{s}}{\text{minimize}} & \varepsilon\lambda + \dfrac{1}{N}\displaystyle\sum_{i=1}^{N} s_i \\ \text{subject to} & \log(1 + \exp(-y^i \cdot \boldsymbol{\beta}^\top \boldsymbol{x}^i + \alpha \cdot \|\boldsymbol{\beta}\|_{p^\star})) \leq s_i \qquad \forall i \in [N] \\ & \log(1 + \exp(y^i \cdot \boldsymbol{\beta}^\top \boldsymbol{x}^i + \alpha \cdot \|\boldsymbol{\beta}\|_{p^\star})) - \lambda\kappa \leq s_i \quad \forall i \in [N] \\ & \|\boldsymbol{\beta}\|_{q^\star} \leq \lambda \\ & \boldsymbol{\beta} \in \mathbb{R}^n, \ \lambda \geq 0, \ \boldsymbol{s} \in \mathbb{R}^N_+, \end{array}$$

where we also used the definition of $L^{\alpha}$. This problem is identical to DR-ARO, which means that feasible solutions of DR-ARO are feasible for Inter-ARO$^\star$ if the additional variables $(\widehat{\lambda}, \widehat{\boldsymbol{s}}, \boldsymbol{z}^l_{ij})$

are set to zero, concluding the first statement of the corollary.

The second statement is immediate since $\widehat{\varepsilon} \to \infty$ forces $\widehat{\lambda} = 0$ due to the term $\widehat{\varepsilon}\widehat{\lambda}$ in the objective of Inter-ARO$^\star$, and this proof shows in such a case Inter-ARO$^\star$ reduces to DR-ARO (which is identical to Inter-ARO when $\varepsilon \to \infty$ by definition). □

### 4.B.7 Proof of Observation 8

By standard linearity arguments and from the definition of $\mathbb{Q}_{\text{mix}}$, we have

$$\begin{aligned}
&\mathbb{E}_{\mathbb{Q}_{\text{mix}}}\left[\sup_{\boldsymbol{z}\in\mathcal{B}_p(\alpha)}\{\ell_{\boldsymbol{\beta}}(\boldsymbol{x}+\boldsymbol{z},y)\}\right]\\
\iff &\int_{(\boldsymbol{x},y)\in\mathbb{R}^n\times\{-1,+1\}}\sup_{\boldsymbol{z}\in\mathcal{B}_p(\alpha)}\{\ell_{\boldsymbol{\beta}}(\boldsymbol{x}+\boldsymbol{z},y)\}\,\mathrm{d}\mathbb{Q}_{\text{mix}}((\boldsymbol{x},y))\\
\iff &\frac{N}{N+w\widehat{N}}\int_{(\boldsymbol{x},y)\in\mathbb{R}^n\times\{-1,+1\}}\sup_{\boldsymbol{z}\in\mathcal{B}_p(\alpha)}\{\ell_{\boldsymbol{\beta}}(\boldsymbol{x}+\boldsymbol{z},y)\}\,\mathrm{d}\mathbb{P}_N((\boldsymbol{x},y))+\\
&\quad\frac{w\widehat{N}}{N+w\widehat{N}}\int_{(\boldsymbol{x},y)\in\mathbb{R}^n\times\{-1,+1\}}\sup_{\boldsymbol{z}\in\mathcal{B}_p(\alpha)}\{\ell_{\boldsymbol{\beta}}(\boldsymbol{x}+\boldsymbol{z},y)\}\,\mathrm{d}\widehat{\mathbb{P}}_{\widehat{N}}((\boldsymbol{x},y))\\
\iff &\frac{N}{N+w\widehat{N}}\cdot\frac{1}{N}\sum_{i\in[N]}\sup_{\boldsymbol{z}^i\in\mathcal{B}_p(\alpha)}\{\ell_{\boldsymbol{\beta}}(\boldsymbol{x}^i+\boldsymbol{z}^i,y^i)\}+\frac{w\widehat{N}}{N+w\widehat{N}}\cdot\frac{1}{\widehat{N}}\sum_{j\in[\widehat{N}]}\sup_{\boldsymbol{z}^j\in\mathcal{B}_p(\alpha)}\{\ell_{\boldsymbol{\beta}}(\widehat{\boldsymbol{x}}^j+\boldsymbol{z}^j,\widehat{y}^j)\}\\
\iff &\frac{1}{N+w\widehat{N}}\left[\sum_{i\in[N]}\sup_{\boldsymbol{z}^i\in\mathcal{B}_p(\alpha)}\{\ell_{\boldsymbol{\beta}}(\boldsymbol{x}^i+\boldsymbol{z}^i,y^i)\}+w\cdot\sum_{j\in[\widehat{N}]}\sup_{\boldsymbol{z}^j\in\mathcal{B}_p(\alpha)}\{\ell_{\boldsymbol{\beta}}(\widehat{\boldsymbol{x}}^j+\boldsymbol{z}^j,\widehat{y}^j)\}\right],
\end{aligned}$$

which coincides with the objective function of (67). Since we have

$$\mathbb{E}_{\mathbb{Q}_{\text{mix}}}\left[\sup_{\boldsymbol{z}\in\mathcal{B}_p(\alpha)}\{\ell_{\boldsymbol{\beta}}(\boldsymbol{x}+\boldsymbol{z},y)\}\right]=\mathbb{E}_{\mathbb{Q}_{\text{mix}}}[\ell^{\alpha}_{\boldsymbol{\beta}}(\boldsymbol{x},y)]$$

we can conclude the proof. □

### 4.B.8 Proof of Proposition 19

We first prove auxiliary results on mixture distributions. To this end, denote by $\mathcal{C}(\mathbb{Q},\mathbb{P})\subseteq\mathcal{P}(\Xi\times\Xi)$ the set of couplings of the distributions $\mathbb{Q}\in\mathcal{P}(\Xi)$ and $\mathbb{P}\in\mathcal{P}(\Xi)$.

**Lemma 32.** *Let $\mathbb{Q},\mathbb{P}^1,\mathbb{P}^2\in\mathcal{P}(\Xi)$ be probability distributions. If $\Pi^1\in\mathcal{C}(\mathbb{Q},\mathbb{P}^1)$ and $\Pi^2\in\mathcal{C}(\mathbb{Q},\mathbb{P}^2)$, then, $\lambda\cdot\Pi^1+(1-\lambda)\cdot\Pi^2\in\mathcal{C}(\mathbb{Q},\lambda\cdot\mathbb{P}^1+(1-\lambda)\cdot\mathbb{P}^2)$ for all $\lambda\in(0,1)$.*

*Proof.* Let $\Pi = \lambda \cdot \Pi^1 + (1-\lambda) \cdot \Pi^2$ and $\mathbb{P} = \lambda \cdot \mathbb{P}^1 + (1-\lambda) \cdot \mathbb{P}^2$. To have $\Pi \in \mathcal{C}(\mathbb{Q}, \mathbb{P})$ we need $\Pi(\mathrm{d}\boldsymbol{\xi}, \Xi) = \mathbb{Q}(\mathrm{d}\boldsymbol{\xi})$ and $\Pi(\Xi, \mathrm{d}\boldsymbol{\xi'}) = \mathbb{P}(\mathrm{d}\boldsymbol{\xi'})$. To this end, observe that

$$\begin{aligned}\Pi(\mathrm{d}\boldsymbol{\xi}, \Xi) &= \lambda \cdot \Pi^1(\mathrm{d}\boldsymbol{\xi}, \Xi) + (1-\lambda) \cdot \Pi^2(\mathrm{d}\boldsymbol{\xi}, \Xi) \\ &= \lambda \cdot \mathbb{Q} + (1-\lambda) \cdot \mathbb{Q} \;=\; \mathbb{Q}\end{aligned}$$

where the second identity uses the fact that $\Pi^1 \in \mathcal{C}(\mathbb{Q}, \mathbb{P}^1)$. Similarly, we can show:

$$\begin{aligned}\Pi(\Xi, \mathrm{d}\boldsymbol{\xi}) &= \lambda \cdot \Pi^1(\Xi, \mathrm{d}\boldsymbol{\xi}) + (1-\lambda) \cdot \Pi^2(\Xi, \mathrm{d}\boldsymbol{\xi}) \\ &= \lambda \cdot \mathbb{P}^1 + (1-\lambda) \cdot \mathbb{P}^2 \;=\; \mathbb{P},\end{aligned}$$

which concludes the proof. □

We further prove the following intermediary result.

**Lemma 33.** *Let $\mathbb{Q}, \mathbb{P}^1, \mathbb{P}^2 \in \mathcal{P}(\Xi)$ and $\mathbb{P} = \lambda \cdot \mathbb{P}^1 + (1-\lambda) \cdot \mathbb{P}^2$ for some $\lambda \in (0,1)$. We have:*

$$\mathrm{W}(\mathbb{Q}, \mathbb{P}) \leq \lambda \cdot \mathrm{W}(\mathbb{Q}, \mathbb{P}^1) + (1-\lambda) \cdot \mathrm{W}(\mathbb{Q}, \mathbb{P}^2).$$

*Proof.* The Wasserstein distance between $\mathbb{Q}, \mathbb{Q}' \in \mathcal{P}(\Xi)$ can be written as:

$$\mathrm{W}(\mathbb{Q}, \mathbb{Q}') = \min_{\Pi \in \mathcal{C}(\mathbb{Q}, \mathbb{Q}')} \left\{ \int_{\Xi \times \Xi} d(\boldsymbol{\xi}, \boldsymbol{\xi'}) \Pi(\mathrm{d}\boldsymbol{\xi}, \mathrm{d}\boldsymbol{\xi'}) \right\},$$

and since $d$ is a feature-label metric (*cf.* Definition 5) the minimum is well-defined (Villani 2009, Theorem 4.1). We name the optimal solutions to the above problem the *optimal couplings*. Let $\Pi^1$ be an optimal coupling of $\mathrm{W}(\mathbb{Q}, \mathbb{P}^1)$ and let $\Pi^2$ be an optimal coupling of $\mathrm{W}(\mathbb{Q}, \mathbb{P}^2)$ and define $\Pi^c = \lambda \cdot \Pi^1 + (1-\lambda) \cdot \Pi^2$. We have

$$\begin{aligned}\mathrm{W}(\mathbb{Q}, \mathbb{P}) &= \min_{\Pi \in \mathcal{C}(\mathbb{Q}, \mathbb{P})} \left\{ \int_{\Xi \times \Xi} d(\boldsymbol{\xi}, \boldsymbol{\xi'}) \Pi(\mathrm{d}\boldsymbol{\xi}, \mathrm{d}\boldsymbol{\xi'}) \right\} \\ &\leq \int_{\Xi \times \Xi} d(\boldsymbol{\xi}, \boldsymbol{\xi'}) \Pi^c(\mathrm{d}\boldsymbol{\xi}, \mathrm{d}\boldsymbol{\xi'}) \\ &= \lambda \cdot \int_{\Xi \times \Xi} d(\boldsymbol{\xi}, \boldsymbol{\xi'}) \Pi^1(\mathrm{d}\boldsymbol{\xi}, \mathrm{d}\boldsymbol{\xi'}) + (1-\lambda) \cdot \int_{\Xi \times \Xi} d(\boldsymbol{\xi}, \boldsymbol{\xi'}) \Pi^2(\mathrm{d}\boldsymbol{\xi}, \mathrm{d}\boldsymbol{\xi'}) \\ &= \lambda \cdot \mathrm{W}(\mathbb{Q}, \mathbb{P}^1) + (1-\lambda) \cdot \mathrm{W}(\mathbb{Q}, \mathbb{P}^2),\end{aligned}$$

where the first identity uses the definition of the Wasserstein metric, the inequality is due to

Lemma 32 as $\Pi^c$ is a feasible coupling (not necessarily optimal), the equality that follows uses the definition of $\Pi^c$ and the linearity of integrals, and the final identity uses the fact that $\Pi^1$ and $\Pi^2$ were constructed to be the optimal couplings. □

We now prove the proposition (we refer to $\mathbb{Q}_{\text{mix}}$ in the statement of this lemma simply as $\mathbb{Q}$). To prove $\mathbb{Q} \in \mathfrak{B}_\varepsilon(\mathbb{P}_N) \cap \mathfrak{B}_{\widehat{\varepsilon}}(\widehat{\mathbb{P}}_{\widehat{N}})$, it is sufficient to show that $\mathrm{W}(\mathbb{P}_N, \mathbb{Q}) \leq \varepsilon$ and $\mathrm{W}(\widehat{\mathbb{P}}_{\widehat{N}}, \mathbb{Q}) \leq \widehat{\varepsilon}$ jointly hold. By using Lemma 33, we can derive the following inequalities:

$$\begin{aligned}
\mathrm{W}(\mathbb{P}_N, \mathbb{Q}) &\leq \lambda \cdot \underbrace{\mathrm{W}(\mathbb{P}_N, \mathbb{P}_N)}_{=0} + (1-\lambda) \cdot \mathrm{W}(\mathbb{P}_N, \widehat{\mathbb{P}}_{\widehat{N}}) \\
\mathrm{W}(\widehat{\mathbb{P}}_{\widehat{N}}, \mathbb{Q}) &\leq \lambda \cdot \mathrm{W}(\mathbb{P}_N, \widehat{\mathbb{P}}_{\widehat{N}}) + (1-\lambda) \cdot \underbrace{\mathrm{W}(\widehat{\mathbb{P}}_{\widehat{N}}, \widehat{\mathbb{P}}_{\widehat{N}})}_{=0}.
\end{aligned}$$

Therefore, sufficient conditions on $\mathrm{W}(\mathbb{P}_N, \mathbb{Q}) \leq \varepsilon$ and $\mathrm{W}(\widehat{\mathbb{P}}_{\widehat{N}}, \mathbb{Q}) \leq \widehat{\varepsilon}$ would be:

$$\begin{cases} (1-\lambda) \cdot \mathrm{W}(\mathbb{P}_N, \widehat{\mathbb{P}}_{\widehat{N}}) \leq \varepsilon \\ \lambda \cdot \mathrm{W}(\mathbb{P}_N, \widehat{\mathbb{P}}_{\widehat{N}}) \leq \widehat{\varepsilon}. \end{cases}$$

Moreover, given that $\varepsilon + \widehat{\varepsilon} \geq \mathrm{W}(\mathbb{P}_N, \widehat{\mathbb{P}}_{\widehat{N}})$, the sufficient conditions further simplify to

$$\begin{cases} (1-\lambda) \cdot \widehat{\varepsilon} \leq \lambda \cdot \varepsilon \\ \lambda \cdot \varepsilon \leq (1-\lambda) \cdot \widehat{\varepsilon}. \end{cases} \iff \lambda \cdot \varepsilon = (1-\lambda) \cdot \widehat{\varepsilon},$$

which is implied when $\dfrac{\lambda}{1-\lambda} = \dfrac{\widehat{\varepsilon}}{\varepsilon}$, concluding the proof. □

### 4.B.9 Proof of Theorem 12

Since each result in the statement of this theorem is abridged, we will present these results sequentially as separate results. We review the existing literature to characterize $\mathfrak{B}_\varepsilon(\mathbb{P}_N)$, in a similar fashion with the results presented in (Selvi et al. 2022a, Appendix A) for the logistic loss, by revising them to the adversarial loss whenever necessary. The $N$-fold product distribution of $\mathbb{P}^0$ from which the training set $\mathbb{P}_N$ is constructed is denoted below by $[\mathbb{P}^0]^N$.

**Theorem 14.** *Assume there exist $a > 1$ and $A > 0$ such that $\mathbb{E}_{\mathbb{P}^0}[\exp(\|\boldsymbol{\xi}\|^a)] \leq A$ for a norm $\|\cdot\|$ on $\mathbb{R}^n$. Then, there are constants $c_1, c_2 > 0$ that only depend on $\mathbb{P}^0$ through $a$, $A$, and $n$,* *such that* $[\mathbb{P}^0]^N(\mathbb{P}^0 \in \mathfrak{B}_\varepsilon(\mathbb{P}_N)) \geq 1 - \eta$ *holds for any confidence level* $\eta \in (0,1)$ *as long as the Wasserstein ball radius satisfies the following optimal characterization*

$$\varepsilon \geq \begin{cases} \left(\dfrac{\log(c_1/\eta)}{c_2 \cdot N}\right)^{1/\max\{n,2\}} & \text{if } N \geq \dfrac{\log(c_1/\eta)}{c_2} \\ \left(\dfrac{\log(c_1/\eta)}{c_2 \cdot N}\right)^{1/a} & \text{otherwise.} \end{cases}$$

*Proof.* The statement follows from Theorem 18 of Kuhn et al. (2019). The presented decay rate $\mathcal{O}(N^{-1/n})$ of $\varepsilon$ as $N$ increases is optimal (Fournier and Guillin 2015). □

Now that we gave a confidence for the unknown radius $\varepsilon$ satisfying $\mathbb{P}^0 \in \mathfrak{B}_\varepsilon(\mathbb{P}_N)$, we analyze the underlying optimization problems. Most of the theory is well-established for logistic loss function, and in the following we show that similar results follow for the adversarial loss function. For convenience, we state DR-ARO again by using the adversarial loss function as defined in Observation 7:

$$\begin{array}{lll} \underset{\boldsymbol{\beta}}{\text{minimize}} & \sup\limits_{\mathbb{Q} \in \mathfrak{B}_\varepsilon(\mathbb{P}_N)} & \mathbb{E}_{\mathbb{Q}}[\ell^\alpha_{\boldsymbol{\beta}}(\boldsymbol{x}, y)] \\ \text{subject to} & \boldsymbol{\beta} \in \mathbb{R}^n. & \end{array} \tag{DR-ARO}$$

**Theorem 15.** *If the assumptions of Theorem 14 are satisfied and* $\varepsilon$ *is chosen as in the statement of Theorem 14, then*

$$[\mathbb{P}^0]^N \left( \mathbb{E}_{\mathbb{P}^0}[\ell^\alpha_{\boldsymbol{\beta}^\star}(\boldsymbol{x}, y)] \leq \sup_{\mathbb{Q} \in \mathfrak{B}_\varepsilon(\mathbb{P}_N)} \mathbb{E}_{\mathbb{Q}}[\ell^\alpha_{\boldsymbol{\beta}^\star}(\boldsymbol{x}, y)] \right) \geq 1 - \eta$$

*holds for all* $\eta \in (0,1)$ *and all optimizers* $\boldsymbol{\beta}^\star$ *of DR-ARO.*

*Proof.* The statement follows from Theorem 19 of Kuhn et al. (2019) given that $\ell^\alpha_{\boldsymbol{\beta}}$ is a finite-valued continuous loss function. □

Theorem 15 states that the optimal value of DR-ARO overestimates the true loss with arbitrarily high confidence $1 - \eta$. Despite the desired overestimation of the true loss, we show that DR-ARO is still asymptotically consistent if we restrict the set of admissible $\boldsymbol{\beta}$ to a bounded set[7].

[7] Note that, this is without loss of generality given that we can normalize the decision boundary of linear classifiers.

**Theorem 16.** *If we restrict the hypotheses $\boldsymbol{\beta}$ to a bounded set $\mathcal{H} \subseteq \mathbb{R}^n$, and parameterize $\varepsilon$ as $\varepsilon_N$ to show its dependency to the sample size, then, under the assumptions of Theorem 14, we have*

$$\sup_{\mathbb{Q} \in \mathfrak{B}_{\varepsilon_N}(\mathbb{P}_N)} \mathbb{E}_{\mathbb{Q}}[\ell^{\alpha}_{\boldsymbol{\beta}^\star}(\boldsymbol{x}, y)] \underset{N \to \infty}{\longrightarrow} \mathbb{E}_{\mathbb{P}^0}[\ell^{\alpha}_{\boldsymbol{\beta}^\star}(\boldsymbol{x}, y)] \quad \mathbb{P}^0\text{-almost surely,}$$

*whenever $\varepsilon_N$ is set as specified in Theorem 14 along with its finite-sample confidence $\eta_N$, and they satisfy $\sum_{N \in \mathbb{N}} \eta_N < \infty$ and $\lim_{N \to \infty} \varepsilon_N = 0$.*

*Proof.* If we show that there exists $\boldsymbol{\xi}^0 \in \Xi$ and $C > 0$ such that $\ell^{\alpha}_{\boldsymbol{\beta}}(\boldsymbol{x}, y) \leq C(1 + d(\boldsymbol{\xi}, \boldsymbol{\xi}^0))$ holds for all $\boldsymbol{\beta} \in \mathcal{H}$ and $\boldsymbol{\xi} \in \Xi$ (that is, the adversarial loss satisfies a growth condition), the statement will follow immediately from Theorem 20 of (Kuhn et al. 2019).

To see that the growth condition is satisfied, we first substitute the definition of $\ell^{\alpha}_{\boldsymbol{\beta}}$ and $d$ explicitly, and note that we would like to show there exists $\boldsymbol{\xi}^0 \in \Xi$ and $C > 0$ such that

$$\log(1 + \exp(-y \cdot \boldsymbol{\beta}^\top \boldsymbol{x} + \alpha \cdot \|\boldsymbol{\beta}\|_{p^\star})) \leq C(1 + \|\boldsymbol{x} - \boldsymbol{x}^0\|_q + \kappa \cdot \mathbb{1}[y \neq y^0])$$

holds for all $\boldsymbol{\beta} \in \mathcal{H}$ and $\boldsymbol{\xi} \in \Xi$. We take $\boldsymbol{\xi}^0 = (\mathbf{0}, y^0)$ and show that the right-hand side of the inequality can be lower bounded as:

$$\begin{aligned} C(1 + \|\boldsymbol{x} - \boldsymbol{x}^0\|_q + \kappa \cdot \mathbb{1}[y \neq y^0]) &= C(1 + \|\boldsymbol{x}\|_q + \kappa \cdot \mathbb{1}[y \neq y^0]) \\ &\geq C(1 + \|\boldsymbol{x}\|_q). \end{aligned}$$

Moreover, the left-hand side of the inequality can be upper bounded for any $\boldsymbol{\beta} \in \mathcal{H} \subseteq [-M, M]^n$ (for some $M > 0$) and $\boldsymbol{\xi} = (\boldsymbol{x}, y) \in \Xi$ as:

$$\begin{aligned} \log(1 + \exp(-y \cdot \boldsymbol{\beta}^\top \boldsymbol{x} + \alpha \cdot \|\boldsymbol{\beta}\|_{p^\star})) &\leq \log(1 + \exp(|\boldsymbol{\beta}^\top \boldsymbol{x}| + \alpha \cdot \|\boldsymbol{\beta}\|_{p^\star})) \\ &\leq \log(2 \cdot \exp(|\boldsymbol{\beta}^\top \boldsymbol{x}| + \alpha \cdot \|\boldsymbol{\beta}\|_{p^\star})) \\ &= \log(2) + |\boldsymbol{\beta}^\top \boldsymbol{x}| + \alpha \cdot \|\boldsymbol{\beta}\|_{p^\star} \\ &\leq \log(2) + \sup_{\boldsymbol{\beta} \in [-M, M]^n} \{|\boldsymbol{\beta}^\top \boldsymbol{x}|\} + \alpha \cdot \sup_{\boldsymbol{\beta} \in [-M, M]^n} \{\|\boldsymbol{\beta}\|_{p^\star}\} \\ &= \log(2) + M \cdot \|\boldsymbol{x}\|_1 + M \cdot \alpha \\ &\leq \log(2) + M \cdot n^{(q-1)/q} \cdot \|\boldsymbol{x}\|_1 + M \cdot \alpha \end{aligned}$$

where the final inequality uses Hölder's inequality to bound the 1-norm with the $q$-norm. Thus, it suffices to show that we have

$$\log(2) + M \cdot n^{(q-1)/q} \cdot \|\boldsymbol{x}\|_1 + M \cdot \alpha \leq C(1 + \|\boldsymbol{x}\|_q) \quad \forall \boldsymbol{\xi} \in \Xi,$$

which is satisfied for any $C \geq \max\{\log(2) + M \cdot \alpha,\ M \cdot n^{(q-1)/q}\}$. This completes the proof by showing the growth condition is satisfied. □

So far, we reviewed tight characterizations for $\varepsilon$ so that the ball $\mathfrak{B}_\varepsilon(\mathbb{P}_N)$ includes the true distribution $\mathbb{P}^0$ with arbitrarily high confidence, proved that the DRO problem DR-ARO overestimates the true loss, while converging to the true problem asymptotically as the confidence $1-\eta$ increases and the radius $\varepsilon$ decreases simultaneously. Finally, we discuss that for optimal solutions $\boldsymbol{\beta}^\star$ to DR-ARO, there are worst case distributions $\mathbb{Q}^\star \in \mathfrak{B}_\varepsilon(\mathbb{P}_N)$ of nature's problem that are supported on at most $N+1$ outcomes.

**Theorem 17.** *If we restrict the hypotheses $\boldsymbol{\beta}$ to a bounded set $\mathcal{H} \subseteq \mathbb{R}^n$, then there are distributions $\mathbb{Q}^\star \in \mathfrak{B}_\varepsilon(\mathbb{P}_N)$ that are supported on at most $N+1$ outcomes and satisfy:*

$$\mathbb{E}_{\mathbb{Q}^\star}[\ell^\alpha_{\boldsymbol{\beta}}(\boldsymbol{x}, y)] \;=\; \sup_{\mathbb{Q} \in \mathfrak{B}_\varepsilon(\mathbb{P}_N)} \mathbb{E}_{\mathbb{Q}}[\ell^\alpha_{\boldsymbol{\beta}}(\boldsymbol{x}, y)].$$

*Proof.* The proof follows from (Yue et al. 2021). □

See the proof of Selvi et al. (2022a, Theorem 8) and the discussion that follows for insights and further analysis on these results presented.

### 4.B.10 Proof of Theorem 13

Firstly, since $\widehat{\mathbb{P}}_{\widehat{N}}$ is constructed from i.i.d. samples of $\widehat{\mathbb{P}}$, we can overestimate the distance $\widehat{\varepsilon}_1 = \mathrm{W}(\widehat{\mathbb{P}}_{\widehat{N}}, \widehat{\mathbb{P}})$ analogously by applying Theorem 14, *mutatis mutandis.* This leads us to the following result where the joint (independent) $N$-fold product distribution of $\mathbb{P}^0$ and the $\widehat{N}$-fold product distribution of $\widehat{\mathbb{P}}$ is denoted below by $[\mathbb{P}^0 \times \widehat{\mathbb{P}}]^{N \times \widehat{N}}$.

**Theorem 18.** *Assume that there exist $a > 1$ and $A > 0$ such that $\mathbb{E}_{\mathbb{P}^0}[\exp(\|\boldsymbol{\xi}\|^a)] \leq A$, and there exist $\widehat{a} > 1$ and $\widehat{A} > 0$ such that $\mathbb{E}_{\widehat{\mathbb{P}}}[\exp(\|\boldsymbol{\xi}\|^{\widehat{a}})] \leq \widehat{A}$ for a norm $\|\cdot\|$ on $\mathbb{R}^n$. Then, there are constants $c_1, c_2 > 0$ that only depends on $\mathbb{P}^0$ through $a$, $A$, and $n$, and constants $\widehat{c}_1, \widehat{c}_2 > 0$ that only depends on $\widehat{\mathbb{P}}$ through $\widehat{a}$, $\widehat{A}$, and $n$ such that $[\mathbb{P}^0 \times \widehat{\mathbb{P}}]^{N \times \widehat{N}}(\mathbb{P}^0 \in \mathfrak{B}_\varepsilon(\mathbb{P}_N) \cap \mathfrak{B}_{\widehat{\varepsilon}}(\widehat{\mathbb{P}}_{\widehat{N}})) \geq 1 - \eta$ holds for any confidence level $\eta \in (0,1)$ as long as the Wasserstein ball radii satisfy the following* *characterization*

$$\varepsilon \geq \begin{cases} \left(\dfrac{\log(c_1/\eta_1)}{c_2 \cdot N}\right)^{1/\max\{n,2\}} & \textit{if } N \geq \dfrac{\log(c_1/\eta_1)}{c_2} \\ \left(\dfrac{\log(c_1/\eta_1)}{c_2 \cdot N}\right)^{1/a} & \textit{otherwise} \end{cases}$$

$$\widehat{\varepsilon} \geq \mathrm{W}(\mathbb{P}^0, \widehat{\mathbb{P}}) + \begin{cases} \left(\dfrac{\log(\widehat{c}_1/\eta_2)}{\widehat{c}_2 \cdot \widehat{N}}\right)^{1/\max\{n,2\}} & \textit{if } \widehat{N} \geq \dfrac{\log(\widehat{c}_1/\eta_2)}{\widehat{c}_2} \\ \left(\dfrac{\log(\widehat{c}_1/\eta_2)}{\widehat{c}_2 \cdot \widehat{N}}\right)^{1/\widehat{a}} & \textit{otherwise} \end{cases}$$

*for some* $\eta_1, \eta_2 > 0$ *satisfying* $\eta_1 + \eta_2 = \eta$.

*Proof.* It immediately follows from Theorem 14 that $[\mathbb{P}^0]^N(\mathbb{P}^0 \in \mathfrak{B}_\varepsilon(\mathbb{P}_N)) \geq 1 - \eta_1$ holds. If we take $\widehat{\varepsilon}_1 > 0$ as

$$\widehat{\varepsilon}_1 \geq \begin{cases} \left(\dfrac{\log(\widehat{c}_1/\eta_2)}{\widehat{c}_2 \cdot \widehat{N}}\right)^{1/\max\{n,2\}} & \text{if } \widehat{N} \geq \dfrac{\log(\widehat{c}_1/\eta_2)}{\widehat{c}_2} \\ \left(\dfrac{\log(\widehat{c}_1/\eta_2)}{\widehat{c}_2 \cdot \widehat{N}}\right)^{1/\widehat{a}} & \text{otherwise,} \end{cases}$$

then, we similarly have $[\widehat{\mathbb{P}}]^{\widehat{N}}(\widehat{\mathbb{P}} \in \mathfrak{B}_{\widehat{\varepsilon}_1}(\widehat{\mathbb{P}}_{\widehat{N}})) \geq 1 - \eta_2$. Since the following implication follows from the triangle inequality:

$$\widehat{\mathbb{P}} \in \mathfrak{B}_{\widehat{\varepsilon}_1}(\widehat{\mathbb{P}}_{\widehat{N}}) \implies \mathbb{P}^0 \in \mathfrak{B}_{\widehat{\varepsilon}_1 + \mathrm{W}(\mathbb{P}^0, \widehat{\mathbb{P}})}(\widehat{\mathbb{P}}_{\widehat{N}}),$$

we have that $[\widehat{\mathbb{P}}]^{\widehat{N}}(\mathbb{P}^0 \in \mathfrak{B}_\varepsilon(\widehat{\mathbb{P}}_{\widehat{N}})) \geq 1 - \eta_2$. These results, along with the facts that $\widehat{\mathbb{P}}_{\widehat{N}}$ and $\mathbb{P}_N$ are independently sampled from their true distributions, imply:

$$\begin{aligned} & [\mathbb{P}^0 \times \widehat{\mathbb{P}}]^{N \times \widehat{N}}(\mathbb{P}^0 \notin \mathfrak{B}_\varepsilon(\mathbb{P}_N) \vee \mathbb{P}^0 \notin \mathfrak{B}_{\widehat{\varepsilon}}(\widehat{\mathbb{P}}_{\widehat{N}})) \\ \leq & [\mathbb{P}^0 \times \widehat{\mathbb{P}}]^{N \times \widehat{N}}(\mathbb{P}^0 \notin \mathfrak{B}_\varepsilon(\mathbb{P}_N)) + [\mathbb{P}^0 \times \widehat{\mathbb{P}}]^{N \times \widehat{N}}(\mathbb{P}^0 \notin \mathfrak{B}_{\widehat{\varepsilon}}(\widehat{\mathbb{P}}_{\widehat{N}})) \\ = & [\mathbb{P}^0]^N(\mathbb{P}^0 \notin \mathfrak{B}_\varepsilon(\mathbb{P}_N)) + [\widehat{\mathbb{P}}]^{\widehat{N}}(\mathbb{P}^0 \notin \mathfrak{B}_{\widehat{\varepsilon}}(\widehat{\mathbb{P}}_{\widehat{N}})) < \eta_1 + \eta_2 \end{aligned}$$

implying the desired result $[\mathbb{P}^0 \times \widehat{\mathbb{P}}]^{N \times \widehat{N}}(\mathbb{P}^0 \in \mathfrak{B}_\varepsilon(\mathbb{P}_N) \cap \mathfrak{B}_{\widehat{\varepsilon}}(\widehat{\mathbb{P}}_{\widehat{N}})) \geq 1 - \eta$. □

The second statement immediately follows under the assumptions of Theorem 18: Inter-ARO overestimates the true loss analogously as Theorem 15 with an identical proof.

## 4.C Exponential Conic Reformulation of DR-ARO

For any $i \in [N]$, the constraints of DR-ARO are

$$\begin{cases} \log(1 + \exp(-y^i \cdot \boldsymbol{\beta}^\top \boldsymbol{x}^i + \alpha \cdot \|\boldsymbol{\beta}\|_{p^\star})) \leq s_i \\ \log(1 + \exp(y^i \cdot \boldsymbol{\beta}^\top \boldsymbol{x}^i + \alpha \cdot \|\boldsymbol{\beta}\|_{p^\star})) - \lambda \cdot \kappa \leq s_i, \end{cases}$$

which, by using an auxiliary variable $u$, can be written as

$$\begin{cases} \log(1 + \exp(-y^i \cdot \boldsymbol{\beta}^\top \boldsymbol{x}^i + u)) \leq s_i \\ \log(1 + \exp(y^i \cdot \boldsymbol{\beta}^\top \boldsymbol{x}^i + u)) - \lambda \cdot \kappa \leq s_i \\ \alpha \cdot \|\boldsymbol{\beta}\|_{p^\star} \leq u. \end{cases}$$

Following the conic modeling guidelines of MOSEK ApS (2023), for new variables $v_i^+, w_i^+ \in \mathbb{R}$, the first constraint can be written as

$$\left\{ v_i^+ + w_i^+ \leq 1,\ (v_i^+, 1, [-u + y^i \cdot \boldsymbol{\beta}^\top \boldsymbol{x}^i) - s_i] \in \mathcal{K}_{\text{exp}},\ (w_i^+, 1, -s_i) \in \mathcal{K}_{\text{exp}}, \right.$$

by using the definition of the exponential cone $\mathcal{K}_{\text{exp}}$. Similarly, for new variables $v_i^-, w_i^- \in \mathbb{R}$, the second constraint can be written as

$$\left\{ v_i^- + w_i^- \leq 1,\ (v_i^-, 1, [-u - y^i \cdot \boldsymbol{\beta}^\top \boldsymbol{x}^i] - s_i - \lambda \cdot \kappa) \in \mathcal{K}_{\text{exp}},\ (w_i^-, 1, -s_i - \lambda \cdot \kappa) \in \mathcal{K}_{\text{exp}}. \right.$$

Applying this for all $i \in [N]$ concludes that the following is the conic formulation of DR-ARO:

$$\begin{array}{lll}
\underset{\substack{\boldsymbol{\beta},\ \lambda,\ \boldsymbol{s},\ u \\ \boldsymbol{v}^+, \boldsymbol{w}^+, \boldsymbol{v}^-, \boldsymbol{w}^-}}{\text{minimize}} & \lambda \cdot \varepsilon + \dfrac{1}{N} \displaystyle\sum_{i \in [N]} s_i & \\
\text{subject to} & v_i^+ + w_i^+ \leq 1 & \forall i \in [N] \\
& (v_i^+, 1, [-u + y^i \cdot \boldsymbol{\beta}^\top \boldsymbol{x}^i] - s_i) \in \mathcal{K}_{\text{exp}},\ (w_i^+, 1, -s_i) \in \mathcal{K}_{\text{exp}} & \forall i \in [N] \\
& v_i^- + w_i^- \leq 1 & \forall i \in [N] \\
& (v_i^-, 1, [-u - y^i \cdot \boldsymbol{\beta}^\top \boldsymbol{x}^i] - s_i - \lambda \cdot \kappa) \in \mathcal{K}_{\text{exp}},\ (w_i^-, 1, -s_i - \lambda \cdot \kappa) \in \mathcal{K}_{\text{exp}} & \forall i \in [N] \\
& \alpha \cdot \|\boldsymbol{\beta}\|_{p^\star} \leq u & \\
& \|\boldsymbol{\beta}\|_{q^\star} \leq \lambda & \\
& \boldsymbol{\beta} \in \mathbb{R}^n,\ \lambda \geq 0,\ \boldsymbol{s} \in \mathbb{R}^N,\ u \in \mathbb{R},\ \boldsymbol{v}^+, \boldsymbol{w}^+, \boldsymbol{v}^-, \boldsymbol{w}^- \in \mathbb{R}^N. &
\end{array}$$

## 4.D Further Details on Numerical Experiments

### 4.D.1 UCI Experiments

**Preprocessing UCI datasets.** We experiment on 10 UCI datasets (Kelly et al. 2023) (*cf.* Table 29). We use Python 3 for preprocessing these datasets. Classification problems with more than two classes are converted to binary classification problems (most frequent class/others). For all datasets, numerical features are standardized, the ordinal categorical features are left as they are, and the nominal categorical features are processed via one-hot encoding. As mentioned in the main section, we obtain auxiliary (synthetic) datasets via SDV, which is also implemented in Python 3.

**Detailed misclassification results on the UCI datasets.** Table 30 contains detailed results on the out-of-sample error rates of each method on 10 UCI datasets for classification. All parameters are 5-fold cross-validated: Wasserstein radii from the grid $\{10^{-6}, 10^{-5}, 10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 0, 1, 2, 5, 10\}$, $\kappa$ from the grid $\{1, \sqrt{n}, n\}$ the weight parameter of `ARO+Aux` from grid $\{10^{-6}, 10^{-5}, 10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 0, 1\}$. We fix the norm defining the feature-label metric to the $\ell_1$-norm, and test $\ell_2$-attacks, but other choices with analogous results are also implemented.

Finally, we demonstrate that our theory, especially `DRO+ARO+Aux`, contributes to the DRO literature even without adversarial attacks. In this case of $\alpha = 0$, `ERM` and `ARO` would be equivalent, and `DRO+ARO` would reduce to the traditional DR LR model (Shafieezadeh-Abadeh et al.

| DataSet | $N$ | $\widehat{N}$ | $N_{\text{te}}$ | $n$ |
|---|---|---|---|---|
| absent | 111 | 333 | 296 | 74 |
| annealing | 134 | 404 | 360 | 41 |
| audiology | 33 | 102 | 91 | 102 |
| breast-cancer | 102 | 307 | 274 | 90 |
| contraceptive | 220 | 663 | 590 | 23 |
| dermatology | 53 | 161 | 144 | 99 |
| ecoli | 50 | 151 | 135 | 9 |
| spambase | 690 | 2,070 | 1,841 | 58 |
| spect | 24 | 72 | 64 | 23 |
| prim-tumor | 50 | 153 | 136 | 32 |

Table 29: *Size of the UCI datasets.*

| Data | $\alpha$ | ERM | ARO | ARO+Aux | DRO+ARO | DRO+ARO+Aux |
|---|---|---|---|---|---|---|
| absent | 0.05 | 44.02% (± 2.89) | 38.82% (± 2.86) | 35.95% (± 3.78) | 34.22% (± 2.70) | **32.64%** (± 2.54) |
| | 0.20 | 73.65% (± 4.14) | 51.49% (± 3.39) | 49.56% (± 3.80) | 45.61% (± 2.32) | **44.90%** (± 2.30) |
| annealing | 0.05 | 18.08% (± 1.89) | 16.61% (± 2.16) | 14.97% (± 1.39) | 13.50% (± 2.98) | **12.78%** (± 2.78) |
| | 0.20 | 37.31% (± 3.92) | 23.08% (± 2.82) | 21.30% (± 1.93) | 20.70% (± 1.32) | **19.53%** (± 1.42) |
| audiology | 0.05 | 21.43% (± 3.64) | 21.54% (± 3.92) | 17.03% (± 2.90) | 11.76% (± 3.28) | **9.01%** (± 3.54) |
| | 0.20 | 37.91% (± 6.78) | 29.34% (± 5.89) | 20.44% (± 2.75) | 20.00% (± 3.01) | **17.91%** (± 3.28) |
| breast-cancer | 0.05 | 4.74% (± 1.26) | 4.93% (± 1.75) | 3.87% (± 1.17) | 3.06% (± 0.79) | **2.52%** (± 0.50) |
| | 0.20 | 9.93% (± 1.73) | 8.14% (± 2.01) | 6.09% (± 1.79) | 5.04% (± 1.11) | **4.67%** (± 0.99) |
| contraceptive | 0.05 | 44.14% (± 2.80) | 42.86% (± 2.59) | 40.98% (± 0.95) | 40.00% (± 1.33) | **39.65%** (± 1.15) |
| | 0.20 | 66.19% (± 5.97) | 43.49% (± 2.24) | **42.71%** (± 1.47) | **42.71%** (± 1.47) | **42.71%** (± 1.47) |
| dermatology | 0.05 | 15.97% (± 2.64) | 16.46% (± 1.67) | 13.47% (± 1.97) | 12.78% (± 1.61) | **10.84%** (± 1.24) |
| | 0.20 | 30.07% (± 4.24) | 28.54% (± 3.25) | 21.53% (± 2.17) | 22.64% (± 2.15) | **20.21%** (± 1.58) |
| ecoli | 0.05 | 16.30% (± 4.42) | 14.67% (± 5.13) | 13.26% (± 3.07) | 11.11% (± 5.52) | **9.78%** (± 2.61) |
| | 0.20 | 51.41% (± 3.37) | 42.67% (± 2.91) | 41.85% (± 2.95) | 39.70% (± 2.68) | **38.89%** (± 2.57) |
| spambase | 0.05 | 11.35% (± 0.77) | 10.23% (± 0.54) | 10.16% (± 0.56) | 9.83% (± 0.37) | **9.81%** (± 0.38) |
| | 0.20 | 27.32% (± 2.11) | 15.83% (± 0.77) | 15.70% (± 0.76) | 15.67% (± 0.72) | **15.50%** (± 0.68) |
| spect | 0.05 | 33.75% (± 5.17) | 29.69% (± 5.46) | 25.78% (± 3.06) | 25.47% (± 3.38) | **21.56%** (± 2.74) |
| | 0.20 | 54.22% (± 9.88) | 37.5% (± 3.53) | 35.16% (± 2.47) | 33.75% (± 2.68) | **30.16%** (± 3.61) |
| prim-tumor | 0.05 | 21.84% (± 4.55) | 20.81% (± 3.97) | 17.35% (± 3.59) | 16.18% (± 3.83) | **14.78%** (± 2.89) |
| | 0.20 | 34.19% (± 6.17) | 25.37% (± 4.58) | 21.62% (± 3.45) | 21.84% (± 3.34) | **19.63%** (± 2.71) |

Table 30: *Mean (± std) out-of-sample errors of UCI datasets, each with 10 simulations. Results for adversarial ($\ell_2$-)attack strengths $\alpha = 0.05$ and $\alpha = 0.2$ are shared.*

2015). ARO+Aux would be interpreted as revising the empirical distribution of ERM to a mixture (mixture weight cross-validated) of the empirical and auxiliary distributions. DRO+ARO+Aux, on the other hand, can be interpreted as DRO over a carefully reduced ambiguity set (intersection of the empirical and auxiliary Wasserstein balls). The results are in Table 31. Analogous results follow as before (that is, DRO+ARO+Aux is the 'winning' approach, DRO+ARO and ARO+Aux alternate for the 'second' approach), with the exception of the dataset *contraceptive*, where ARO+Aux

| Data | `ERM` | `ARO` | `ARO+Aux` | `DRO+ARO` | `DRO+ARO+Aux` |
|---|---|---|---|---|---|
| absent | 36.28% | 36.28% | 31.86% | 28.31% | **27.74%** |
| annealing | 10.61% | 10.61% | 7.64% | **7.14%** | **7.14%** |
| audiology | 14.94% | 14.94% | 12.97% | 10.11% | **7.69%** |
| breast-cancer | 6.64% | 6.64% | 5.22% | 2.55% | **2.15%** |
| contraceptive | 35.00% | 35.00% | **33.75%** | 34.56% | 33.85% |
| dermatology | 16.04% | 16.04% | 11.60% | 9.93% | **8.06%** |
| ecoli | 6.74% | 6.74% | 4.96% | 5.19% | **4.37%** |
| spambase | 8.95% | 8.95% | 8.52% | 8.34% | **8.16%** |
| spect | 30.74% | 30.74% | 24.69% | 22.35% | **18.75%** |
| prim-tumor | 22.79% | 22.79% | 17.28% | 15.07% | **13.97%** |

Table 31: *Mean out-of-sample errors of UCI experiments without adversarial attacks.*

outperforms others.

**Cross-validated Wasserstein radii.** In Corollary 9, we discussed the feasibility of ignoring the auxiliary data when its data-generating distribution is distant from the true data-generating distribution. To examine whether large values of $\widehat{\varepsilon}$ are selected via cross-validation in our experiments, we investigated their histograms and observed that this indeed occurs frequently. For example, as shown in Table 31, when there are no adversarial attacks, `DRO+ARO` and `DRO+ARO+Aux` yield identical errors on the 'annealing' dataset. This is because the two models become equivalent: `DRO+ARO+Aux` selects a large $\widehat{\varepsilon}$ (either 5 or 10), which ensures that the intersection of the Wasserstein balls reduces to the empirical distribution. With $\alpha = 0.05$, `DRO+ARO+Aux` selects a large $\widehat{\varepsilon}$ in 7 out of 10 simulations, while in the remaining 3 it selects a smaller value, resulting in a nontrivial intersection and, in turn, improved performance over `DRO+ARO`. We conclude that our method selects a nontrivial $\widehat{\varepsilon}$ only when there is evidence that the auxiliary data is *useful*, that is, when its data-generating distribution is sufficiently close to the true one. For instance, on the 'absent' dataset, we always have $\widehat{\varepsilon} \in \{10^{-2}, 10^{-1}\}$. We also revised our numerical experiments by modifying the Gaussian copula synthesizer to enforce uniform marginals (which results in significant information loss), and in this setting, our models consistently selected large $\widehat{\varepsilon}$ values.

**Different training/auxiliary dataset ratio.** Recall that in the UCI experiments, we sampled 15% of the original dataset as the training set, and used 45% to generate a synthetic auxiliary dataset. This setup was chosen to simulate scenarios with limited training data, where overfitting is a particular concern. If we reverse these proportions and use 45% of the original dataset for training and 15% for generating auxiliary data instead, we find that the best-performing method remains `DRO+ARO+Aux`, and the relative ranking among the models remains

| Attack | ERM | ARO | ARO+Aux | DRO++ | DRO+ARO | DRO+ARO+Aux |
|---|---|---|---|---|---|---|
| No attack ($\alpha = 0$) | 1.55% | 1.55% | 1.19% | 0.72% | 0.64% | **0.53%** |
| $\ell_1$ ($\alpha = 68/255$) | 2.17% | 1.84% | 1.33% | 1.40% | 0.66% | **0.57%** |
| $\ell_2$ ($\alpha = 128/255$) | 99.93% | 3.36% | 2.54% | 2.72% | 2.40% | **2.12%** |
| $\ell_\infty$ ($\alpha = 8/255$) | 100.00% | 2.60% | 2.38% | 2.31% | 2.20% | **1.95%** |

Table 32: *Additional MNIST/EMNIST Benchmark.*

qualitatively similar. One notable difference is that the improvement of the auxiliary-data-oblivious DRO+ARO method over the non-DR ARO+Aux model becomes less significant, and is even reversed in the case of $\alpha = 0.05$ on the 'ecoli' dataset: ERM (14.59%), ARO (11.41%), ARO+Aux (9.04%), DRO+ARO (9.41%), DRO+ARO+Aux (8.89%). As expected, since more data is drawn from the true data-generating distribution, all methods exhibit improved out-of-sample performance compared to the original setup. However, in this case, DRO+ARO no longer outperforms ARO+Aux.

### 4.D.2 MNIST/EMNIST Experiments

The setting in the MNIST/EMNIST experiments is similar to that in the UCI experiments. However, for auxiliary data, we use the EMNIST dataset which we accessed via the MLDatasets package of Julia.

Moreover, in §4.2 we reviewed the literature showing that when statistical error is not a concern, that is, when optimizing over the empirical distribution cannot cause overfitting (*e.g.*, in high-data regimes), then adversarial training is equivalent to a type-$\infty$ Wasserstein DRO problem with radius $\varepsilon = \alpha$. Hence, a natural question is whether increasing the value of $\varepsilon$ further also provides distributional robustness. To this end, in Table 32, we revise Table 28 and add an additional benchmark DRO++. Here, we take $\varepsilon = \alpha + \varepsilon'$, and cross-validate $\varepsilon'$ from the same grid that we cross-validated $\varepsilon$ for methods DRO+ARO and DRO+ARO+Aux. We observe that DRO++ does not improve over DRO+ARO, which is expected given that type-$\infty$ Wasserstein DRO does not provide better generalizations, unlike type-1 Wasserstein DRO (*cf.* the discussion at the end of §4.3). Yet, this method improves over ARO in all cases, and even over ARO+Aux in the no-attack or $\ell_\infty$-attack settings.

### 4.D.3 Artificial Experiments

**Data generation.** We sample a 'true' $\boldsymbol{\beta}$ from a unit $\ell_2$-ball, and generate data as summarized in Algorithm 11. Such a dataset generation gives $N$ instances from the same true data-generating distribution. In order to obtain $\hat{N}$ auxiliary dataset instances, we perturb the probabilities $p^i$ with standard random normal noise which is equivalent to sampling i.i.d. from a *perturbed* distribution. Testing is always done on true data, that is, the test set is sampled according to Algorithm 11.

**Algorithm 11:** Data from a ground truth logistic classifier

```
Input: set of feature vectors $\boldsymbol{x}^i,\ i \in [N]$; vector $\boldsymbol{\beta}$
Output: $(\boldsymbol{x}^i, y^i),\ i \in [N]$
for $i \in \{1, \ldots, N\}$ do
    Find the probability $p^i = \left[1 + \exp(-\boldsymbol{\beta}^\top \boldsymbol{x}^i)\right]^{-1}$;
    Sample $u = \mathcal{U}(0, 1)$;
    if $p^i \geq u$ then
        $y^i = +1$;
    else
        $y^i = -1$;
```

**Strength of the attack and importance of auxiliary data.** In the main section we discussed how the strength of an attack determines whether using auxiliary data in ARO (`ARO+Aux`) or considering distributional ambiguity (`DRO+ARO`) is more effective, and observed that unifying them to obtain `DRO+ARO+Aux` yields the best results in all attack regimes. Now we focus on the methods that rely on auxiliary data, namely `ARO+Aux` and `DRO+ARO+Aux` and explore the importance of auxiliary data $\widehat{\mathbb{P}}_{\widehat{N}}$ in comparison to its empirical counterpart $\mathbb{P}_N$. Table 33 shows the average values of $w$ for problem (67) obtained via cross-validation. We see that the greater the attack strength is the more we should use the auxiliary data in `ARO+Aux`. The same relationship holds for the average of $\varepsilon/\widehat{\varepsilon}$ obtained via cross-validation in Inter-ARO, which means that the relative size of the Wasserstein ball built around the empirical distribution gets larger compared to the same ball around the auxiliary data, that is, ambiguity around the auxiliary data is smaller than the ambiguity around the empirical data. We highlight as a possible future research direction exploring when a larger attack *per se* implies the intersection will move towards the auxiliary data distribution.

**More results on scalability.** We further simulate 25 cases with an $\ell_2$-attack strength of $\alpha = 0.2$, $N = 200$ instances in the training dataset, $\widehat{N} = 200$ instances in the auxiliary dataset, and we vary the number of features $n$. We report the median (50% ± 15% quantiles shaded) runtimes of each method in Figure 18. The fastest methods are `ERM` and `ARO` among which the faster one depends on $n$ (as the adversarial loss includes a regularizer of $\boldsymbol{\beta}$), followed by `ARO+Aux`,

| Attack | `ARO+Aux` (cross-validated $w$) | `DRO+ARO+Aux` (cross-validated $\varepsilon/\widehat{\varepsilon}$) |
|---|---|---|
| $\alpha = 0$ | 0.002 | 0.0120 |
| $\alpha = 0.1$ | 0.046 | 0.172 |
| $\alpha = 0.25$ | 0.086 | 0.232 |
| $\alpha = 0.5$ | 0.290 | 0.241 |

Table 33: *Mean $w$ in problem (67) and $\varepsilon/\widehat{\varepsilon}$ in problem Inter-ARO across 25 simulations of cross-validating $\omega$, $\varepsilon$, and $\widehat{\varepsilon}$.*

`DRO+ARO`, and `DRO+ARO+Aux`, respectively. `DRO+ARO+Aux` is the slowest, which is expected given that `DRO+ARO` is its special for large $\widehat{\varepsilon}$. The runtime however scales graciously.

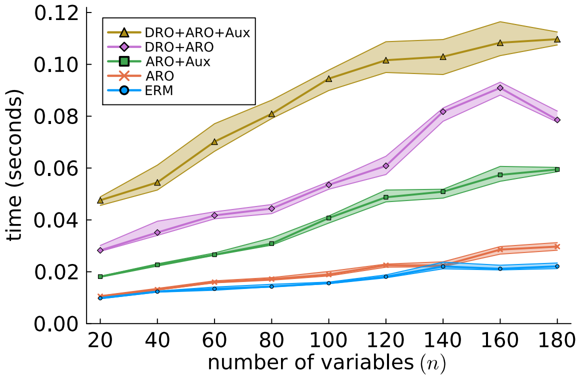


Figure 18: *Runtimes under a varying number of features in the artificially generated empirical and auxiliary datasets.*

Finally, we focus further on `DRO+ARO+Aux` which solves problem Inter-ARO with $\mathcal{O}(n \cdot N \cdot \widehat{N})$ variables and exponential cone constraints. For $n = 1,000$ and $N = \widehat{N} = 10,000$, we observe that the runtimes vary between 134 to 232 seconds across 25 simulations.

# Chapter III

# Nonconvex Optimization

In the previous two chapters, I encountered NP-hard problems in their respective domains and tackled them using tools from stochastic and robust optimization. As I find these tools valuable, I have also worked on extending them by developing efficient approximation methods with strong performance guarantees. This chapter presents my contributions in this area.

In this chapter, Section 5 is based on the following work (Selvi et al. 2022b):

**Aras Selvi, Aharon Ben-Tal, Ruud Brekelmans, Dick den Hertog** (2022). Convex maximization via adjustable robust optimization. **INFORMS Journal on Computing**.
- 2021 INFORMS Computing Society Student Paper Award (First Place)

**Note:** An earlier version of this work appeared in my master's thesis at Tilburg University, where I focused on convex maximization subject to linear constraints. At the beginning of my doctoral studies, I broadened the scope to include nonlinear constraints and conducted additional experiments. I include the present version in this thesis because one of its extensions (see Section 5.A) builds directly on the work in Chapter 6, which I developed entirely during my doctoral studies. I therefore view Chapters 5 and 6 as complementary components of my research on nonconvex optimization.

Section 6 is based on the following work (Selvi et al. 2023):

**Aras Selvi, Dick den Hertog, Wolfram Wiesemann** (2023). A reformulation-linearization technique for optimization over simplices. **Mathematical Programming**.

# 5 Convex Maximization via Adjustable Robust Optimization

## Abstract

Maximizing a convex function over convex constraints is an NP-hard problem in general. We prove that such a problem can be reformulated as an adjustable robust optimization (ARO) problem where each adjustable variable corresponds to a unique constraint of the original problem. We use ARO techniques to obtain approximate solutions to the convex maximization problem. In order to demonstrate the complete approximation scheme, we distinguish the case where we have just one nonlinear constraint and the case where we have multiple linear constraints. Concerning the first case, we give three examples where one can analytically eliminate the adjustable variable and approximately solve the resulting static robust optimization problem efficiently. More specifically, we show that the norm constrained log-sum-exp (geometric) maximization problem can be approximated by (convex) exponential cone optimization techniques. Concerning the second case of multiple linear constraints, the equivalent ARO problem can be represented as an adjustable robust linear optimization (ARLO) problem. Using linear decision rules then returns a safe approximation of the constraints. The resulting problem is a convex optimization problem, and solving this problem gives an upper bound on the global optimum value of the original problem. By using the optimal linear decision rule, we obtain a lower bound solution as well. We derive the approximation problems explicitly for quadratic maximization, geometric maximization, and sum-of-max-linear-terms maximization problems with multiple linear constraints. Numerical experiments show that, contrary to the state-of-the-art solvers, we can approximate large-scale problems swiftly with tight bounds. In several cases, we have equal upper and lower bounds, which concludes that we have global optimality guarantees in these cases.

## 5.1 Introduction

We propose a new approximation method for the convex maximization problem:

$$\begin{aligned} \max_{x\in\mathbb{R}^n} \quad & f(Ax+b) \\ \text{s.t.} \quad & x \in U, \end{aligned}$$

where $U \subset \mathbb{R}^n$ is a compact set defined by convex constraints and $f : \mathbb{R}^m \mapsto \mathbb{R}$ is a closed convex function with an arbitrary linear input $Ax + b$ for $A \in \mathbb{R}^{m\times n}$, and $b \in \mathbb{R}^m$. The seminal paper of Tuy (1964) is regarded as the first approach to solve convex maximization problems, where $U$ is a polyhedron.

There are many real-life problems that can be reformulated as convex maximization problems. Rebennack et al. (2009) show that the fixed charge network flow problem can be formulated as a convex maximization problem. Moreover, Zwart (1974) shows that two important problem classes are equivalent to convex maximization problems, namely cost minimization problems with the cost function being subject to economies of scale, and linear optimization problems that involve 'yes' or 'no' decisions (binary variables). Many machine learning (ML) problems can be formulated as convex maximization problems, for instance Mangasarian (1996) shows that fundamental problems in ML, misclassification minimization and feature selection, are equivalent to convex maximization problems. The same author more recently proposes the use of absolute value inequalities for classifying unlabeled data, which results in a problem of minimizing a concave function on a polyhedral set (Mangasarian 2015). Other important examples from data science are variants of principal component analysis (PCA). Zass and Shashua (2007) show that the nonnegative PCA problem and the sparse PCA problem are convex quadratic maximization problems. Additionally, a popular approach to solve difference of convex functions (DC) programs is the convex-concave method, which iteratively solves convex minimization and convex maximization problems (assuming the constraints are convex) (Lipp and Boyd 2016). DC programming is being used to solve many problems in machine learning, data science, biology, security and transportation (Le Thi and Pham Dinh 2018). Also many problems arising in graph theory can be formulated as convex maximization problems, a well known example is the MAX-CUT problem (Goemans and Williamson 1994). A lot of variations of integer linear and integer quadratic optimization problems over polyhedra can be written as convex maximization problems (Benson 1995). Convex maximization naturally appears in robust optimization when finding the worst-case scenario of a constraint which is a convex function of the uncertain parameter. A similar problem appears when one applies the adversarial approach (Bienstock and Özbay 2008) to solve a robust convex optimization problem. In this approach, at the step of adding worst-case uncertainty realization to the discrete uncertainty set, one needs to maximize convex functions.

A local or global solution of the convex maximization problem is necessarily at an extreme point of the feasible region (Rockafellar 1997), hence there exist many methods to solve convex maximization problems by searching for extreme point solutions, but this approach is itself very hard. It is shown that the convex maximization problem is NP-hard in very simple cases (e.g., quadratic maximization over a hypercube), and even verifying local optimality is NP-hard (Pardalos and Schnitger 1988). Hence, there are many papers to approximate the convex maximization problem (Benson 1995). The survey paper (Pardalos and Rosen 1986) collects such works until the 1980s. Most of the proposed methods use linear underestimator functions, which are derived by the so-called convex envelopes. These algorithms have a disadvantage, namely, the size of the sub problems grows in every new iteration, which makes them impractical in general. Moreover, the proposed methods in the literature are designed only for some specific cases (e.g., (Zwart 1974)). All of the well-accepted methods to solve the most studied convex maximization problem, quadratic maximization, are based on cutting plane methods, iterative numerical methods such as the element methods, and techniques of branches and borders based on the decomposition of the feasible set as summarized in (Audet et al. 2005, Andrianova et al. 2016). These papers also indicate that such methods do not provide solutions in reasonable time for practical problems.

Convex maximization is frequently being studied in the scope of DC programming in recent optimization research. As summarized by Lipp and Boyd (2016), the early approaches reformulated the DC programming problems as convex maximization problems (Tuy 1986, Tuy and Horst 1988, Horst et al. 1991). One can see how the methods in convex maximization are adopted for DC programming literature in the work of Horst and Thoai (1999). Lipp and Boyd (2016) provide a thorough literature review in convex maximization, and state that the literature to solve such problems mostly relies on branch and bound or cutting plane methods which are very slow in practice. In this respect, in order to be able to cope with large DC programming problems, Lipp and Boyd (2016) propose a heuristic algorithm to find a decent local solution of the convex maximization problem.

In this work a new method to approximately solve the convex maximization problem is presented. The method starts by reformulating the convex maximization problem as an adjustable robust optimization (ARO) problem. The ARO problem has a number of adjustable variables equal to the number of the original constraints. The adjustable variables appear nonlinearly, hence the ARO problem is still a hard problem. Therefore, we apply approximation methods used in the ARO literature in order to approximate the ARO problem. This way we derive convex optimization problems which provide upper and lower bounds of the original convex maximization problem.

The rest of the section is organized as follows. In Section 5.2, we present our main theorem for single constrained convex maximization, and show how to reformulate the convex maximization problem as an ARO problem. We exploit the relationship between equivalent formulations to show how to obtain a solution in the original problem by using the solution of the ARO problem. We give three cases of convex maximization over a single norm constraint, show how to analytically eliminate the adjustable variable, and (approximately) solve the resulting static robust optimization problem. The approximate solution of the ARO problem gives an upper bound to the convex maximization problem, and by using this solution we show how to obtain a lower bound for the original problem. In Section 5.3, we derive the ARO reformulation of a general convex maximization problem with arbitrary convex constraints, and show that one way to approximately solve this ARO problem is by relaxing adjustable variables to static variables. We specifically investigate the case of convex maximization over multiple linear constraints. It is shown that the ARO reformulation of this case is an adjustable robust linear optimization (ARLO) problem. Adoption of linear decision rules for the adjustable variables enables one to derive efficient upper and lower bound approximation problems. Explicit approximations are obtained for convex quadratic, geometric, and sum-of-max-linear-terms maximization problems. The numerical experiments follow in Section 5.4, which illustrate that our approximation problems can be solved significantly faster than the state-of-the-art optimization solvers, and provide tight optimality gaps in most of the cases. In cases where the upper bound is equal to the lower bound, a guarantee of global maximizer is obtained. We conclude the work in Section 5.5 by discussing our findings and giving future research directions.

## 5.2 ARO Reformulation of Single-Constrained Convex Maximization

In this section we show how to reformulate a single convex constrained convex maximization problem as an equivalent ARO problem, and how to (approximately) solve it. In the next section we will generalize these results to problems with multiple constraints.

Let $f : \mathbb{R}^m \mapsto \mathbb{R}$ and $g : \mathbb{R}^n \mapsto \mathbb{R}$ be closed convex functions. We assume that $\exists \bar{x} \in \mathbb{R}^n : g(\bar{x}) < \rho$ for scalar $\rho \in \mathbb{R}$. Moreover, let $A \in \mathbb{R}^{m \times n}$ and $b \in \mathbb{R}^m$. We consider the following convex maximization problem with a single convex constraint:

$$\begin{aligned} \max_{x \in \mathbb{R}^n} \quad & f(Ax + b) \\ \text{s.t.} \quad & g(x) \leq \rho. \end{aligned} \tag{72}$$

The feasible set is denoted by $U$, i.e., $U = \{x \in \mathbb{R}^n : \; g(x) \leq \rho\}$. Let $f^*$ denote the convex conjugate of $f$. The perspective of $f$ is defined as $z_0 f(\frac{z}{z_0})$ for $z_0 \geq 0$, and our understanding for

$0f(\frac{z}{0})$ is the recession function $\lim_{z_0 \downarrow 0} z_0 f(\frac{z}{z_0})$ (Rockafellar 1997) for the rest of the section. The following theorem shows that problem (72) is equivalent to an ARO problem where $w \in \operatorname{dom} f^*$ is the uncertain parameter and $\lambda \geq 0$ is the adjustable variable.

**Theorem 19.** *The optimal objective value of problem* (72) *is equal to the optimal objective value of the following adjustable robust optimization problem:*

$$\begin{aligned} \inf_{\tau \in \mathbb{R}} \quad & \tau \\ \text{s.t.} \quad & \forall w \in \operatorname{dom} f^*,\ \exists \lambda \geq 0 : \quad \lambda \rho + \lambda g^*\left(\frac{A^\top w}{\lambda}\right) + b^\top w - f^*(w) \leq \tau. \end{aligned} \tag{73}$$

*The solution $\bar{x}$ that attains this value satisfies $\bar{x} \in \arg\sup\left\{(A^\top \bar{w})^\top x \ : \ x \in U\right\}$ with $\bar{w} \in \operatorname{dom} f^*$ being the parameter realization where the constraint of* (73) *is tight at optimality.*

*Proof.* Problem (72) can be written as:

$$\begin{aligned} \inf_{\tau \in \mathbb{R}} \quad & \tau \\ \text{s.t.} \quad & f(Ax+b) \leq \tau, \quad \forall x \in U. \end{aligned} \tag{74}$$

Since $f$ is a closed convex function, we have:

$$f(z) = f^{**}(z) = \sup_{w \in \operatorname{dom} f^*} \{z^\top w - f^*(w)\},$$

where $f^{**}$ is the biconjugate of $f$. Hence, the constraint of problem (74) becomes:

$$\begin{aligned} \forall x \in U : \ f(Ax+b) \leq \tau \iff & \forall x \in U : \sup_{w \in \operatorname{dom} f^*} \{(Ax+b)^T w - f^*(w)\} \leq \tau \\ \iff & \sup_{x \in U} \left\{ \sup_{w \in \operatorname{dom} f^*} \{(A^\top w)^\top x + b^\top w - f^*(w)\} \right\} \leq \tau \qquad \text{(75a)} \\ \iff & \sup_{w \in \operatorname{dom} f^*} \left\{ \sup_{x \in U} \left\{(A^\top w)^\top x\right\} + b^\top w - f^*(w) \right\} \leq \tau, \qquad \text{(75b)} \end{aligned}$$

where in step (75a) we exploit the fact that if a constraint holds for the worst-case then it holds for any case, and in step (75b) we change the order of the supremum operators. We replace the inner problem (linear maximization over a convex set) with its Lagrangian dual problem and obtain:

$$\begin{aligned}
\sup_x \left\{ (A^\top w)^\top x : \ g(x) \le \rho \right\} &= \inf_{\lambda \ge 0} \left\{ \sup_x \{ (A^\top w)^\top x - \lambda g(x) \} + \lambda\rho \right\} \\
&= \inf_{\lambda > 0} \left\{ \lambda \sup_x \left\{ \frac{(A^\top w)^\top}{\lambda} x - g(x) \right\} + \lambda\rho \right\} && (76a) \\
&= \inf_{\lambda \ge 0} \left\{ \lambda g^* \left( \frac{A^\top w}{\lambda} \right) + \lambda\rho \right\}. && (76b)
\end{aligned}$$

We substitute (76b) into (75b) to conclude:

$$\begin{aligned}
& \forall x \in U : \ f(Ax + b) \le \tau \\
\iff & \sup_{w \in \operatorname{dom} f^*} \left\{ \inf_{\lambda \ge 0} \left\{ \lambda\rho + \lambda g^* \left( \frac{A^\top w}{\lambda} \right) \right\} + b^\top w - f^*(w) \right\} \le \tau && (77a) \\
\iff & \forall w \in \operatorname{dom} f^*, \ \exists \lambda \ge 0 : \ \lambda\rho + \lambda g^* \left( \frac{A^\top w}{\lambda} \right) + b^\top w - f^*(w) \le \tau. && (77b)
\end{aligned}$$

Minimizing $\tau$ over (77b) gives problem (73). The optimal solution $\bar{x}$ of problem (72) can be retrieved from the optimal solution of ARO problem (73). Firstly, notice that $\bar{x}$ solves the outer problem of (75a) and the inner problem of (75b). Let $(\bar{\lambda}, \bar{w})$ denote the solution of the optimization problem (77a), where $\bar{\lambda}$ is a function of $\bar{w}$ due to the inner problem. By the equivalence introduced above, $\bar{w}$ solves problem (75b), hence we can retrieve $\bar{x}$ from the inner problem of (75b), i.e.,:

$$\bar{x} \in \arg\sup \left\{ (A^\top \bar{w})^\top x \ : \ g(x) \le \rho \right\}, \tag{78}$$

which is a convex optimization problem. □

**Remark 3.** *If $g$ is differentiable and the inverse of its gradient ($\nabla^{-1} g(\cdot)$) exists, then we can analytically obtain the solution as $\bar{x} = \nabla^{-1} g \left( \frac{A^\top \bar{w}}{\lambda} \right)$ which follows from (76a).* ■

In the ARO reformulation (73) the adjustable variable $\lambda$ appears nonlinearly, hence this is also a difficult problem. However, there is only one adjustable variable, and it appears in a single (semi-infinite) constraint. In the following, we consider three cases where one can derive explicit expressions for $\lambda$ (e.g., the analytic worst case for $\lambda$, as a function of $w$), which circumvents the nonlinearity. In these examples we do not show how to derive the convex conjugate of various $g(x)$ functions, but these can be found in, e.g., (Boyd and Vandenberghe 2004).

**Corollary 10** (2-norm Constraint)**.** *Let $g(x) := \|x-a\|_2$ in problem (72), where $a \in \mathbb{R}^n$ is a parameter. Then, the global optimum value $\tau^*$ of this problem is*

$$\tau^* = \sup_{w \in \operatorname{dom} f^*} \rho \left\|A^\top w\right\|_2 + a^\top A^\top w + b^\top w - f^*(w). \tag{79}$$

*Furthermore, if the domain of $f^*$ consists of linear inequalities, an upper bound value of $\tau^*$ can be found by solving a convex relaxation of (79). A corresponding lower bound solution of problem (72) is $x^* = (A^\top w^*)\dfrac{\rho}{||A^\top w^*||_2} + a$, where $w^* \in \operatorname{dom} f^*$ is the upper bound solution.*

*Proof.* Let the feasible set of problem (72) be defined as:

$$U_1 = \left\{x \in \mathbb{R}^n : \; g(x) = \frac{1}{2}\|x-a\|_2^2 \leq \frac{1}{2}\rho^2\right\}.$$

The conjugate of the squared norm is $g^*(z) = \frac{1}{2}\|z\|_2^2 + z^\top a$, hence, minimizing $\tau$ over constraint (77) is minimizing $\tau$ subject to:

$$\begin{aligned}
&\sup_{w \in \operatorname{dom} f^*} \left\{\inf_{\lambda \geq 0}\left\{\frac{1}{2}\rho^2\lambda + \lambda g^*\left(\frac{A^\top w}{\lambda}\right)\right\} + b^\top w - f^*(w)\right\} \leq \tau \\
\iff &\sup_{w \in \operatorname{dom} f^*} \left\{\min_{\lambda \geq 0}\left\{\frac{1}{2}\rho^2\lambda + \frac{1}{2}\lambda\left\|\frac{A^\top w}{\lambda}\right\|_2^2\right\} + a^\top A^\top w + b^\top w - f^*(w)\right\} \leq \tau \\
\iff &\sup_{w \in \operatorname{dom} f^*} \left\{\rho\left\|A^\top w\right\|_2 + a^\top A^\top w + b^\top w - f^*(w)\right\} \leq \tau,
\end{aligned}$$

where the last step holds since the inner minimization problem is a convex problem with the optimal decision rule $\bar{\lambda} = \rho^{-1}\left\|A^\top w\right\|_2$. Therefore, for the feasible set $U_1$, problem (72) is equivalent to the following optimization problem:

$$\tau^* = \sup_{w \in \operatorname{dom} f^*} \rho \left\|A^\top w\right\|_2 + a^\top A^\top w + b^\top w - f^*(w). \tag{80}$$

Notice that at the original problem (72) with $U = U_1$, the constraint is convex and the convexity of the objective function $f(Ax+b)$ makes the problem non-convex, whereas in problem (80) maximizing $-f^*(w)$ is fine and the 2-norm makes the problem non-convex.

Although we conclude that problem (80) is non-convex, there exist strong approximation methods, and an example is when the domain of $f^*$ consists of linear inequalities, i.e.,

$$\operatorname{dom} f^* = \{w \;:\; \alpha_i^\top w \leq \beta_i, \; i = 1, \ldots, d\},$$

for $\alpha_i \in \mathbb{R}^m,\ \beta_i \in \mathbb{R}$. In this setting, the only difficult part is maximizing the sum of a convex function and a concave function over linear constraints. To be able to approximate problem (80) with a convex problem we use the reformulation-linearization technique (RLT), whose details can be found in (Sherali and Adams 2013). We also tighten the RLT relaxation by using a 'positive semi-definite cut' (Sherali and Fraticelli 2002) and obtain (derivation is in Appendix 5.A of the supplements):

$$\begin{array}{lll} \sup\limits_{V \in \mathbb{S}^{m\times m},\ w \in \mathbb{R}^m} & \rho\sqrt{\mathrm{tr}(A^\top V A)} + a^\top A^\top w + b^\top w - f^*(w) & \\ \text{s.t.} & \alpha_i^\top w - \beta_i \leq 0, & i = 1, \dots, d \\ & \alpha_i^\top V \alpha_j - (\beta_i \alpha_j + \beta_j \alpha_i)^\top w + \beta_i \beta_j \geq 0, & i \leq j = 1, \dots, d \\ & \begin{pmatrix} V & w \\ w^\top & 1 \end{pmatrix} \succeq 0, & \end{array} \tag{81}$$

where $\mathrm{tr}(\cdot)$ denotes the trace operator. For instance, suppose $f$ is the log-sum-exp function. Since $f^*(w)$ is the negative entropy of $w$ in its domain (standard $m$-dimensional simplex), problem (81) is an exponential-cone representable problem. It can further be proved that for this case the semi-definite constraint is redundant, and the optimal value of $V$ is the diagonal matrix $\mathrm{Diag}(w)$ (Selvi et al. 2023). So the only variable is $w \in \mathbb{R}^m$ and we can solve problem (81) with the exponential cone solver of MOSEK (MOSEK ApS 2019).

Going back to the general setting of $f$, it is straightforward to see that problem (81) upper bounds the optimal objective value of problem (80) since an optimal solution $\bar{w}$ in problem (80) is feasible in problem (81) by taking $V = \bar{w}\bar{w}^\top$. We can also obtain a lower bound on problem (80) by using the upper bound solution. Solving $\bar{x} \in \arg\sup\left\{(A^\top \bar{w})^\top x \ :\ x \in U_1\right\}$, the optimal $\bar{x}$ value can be recovered by $\bar{x} = \nabla^{-1} g\left(\frac{A^\top \bar{w}}{\lambda}\right) = (A^\top \bar{w})\dfrac{\rho}{||A^\top \bar{w}||_2} + a$. Since we approximate $\bar{w}$ of problem (80) with $w^*$ in problem (81), a lower bounding approximation of the optimal solution $\bar{x}$ is $x^* = (A^\top w^*)\dfrac{\rho}{||A^\top w^*||_2} + a$. Moreover, the constraint of the original convex maximization problem is indeed tight for $x^*$, i.e., $\frac{1}{2}||x^* - a||_2^2 = \frac{1}{2}\rho^2$. This shows us that this method gives us an extreme point lower bound solution, which is desirable as otherwise $x^*$ cannot even be locally optimal, i.e., every local and global solution of problem (72) is at an extreme point of the feasible set. □

Numerical experiments of Corollary 10 are in Section 5.4.1

**Corollary 11** (Box Constraints). *Let $g(x) := \|x-a\|_\infty$ in problem (72), where $a \in \mathbb{R}^n$ is a parameter. Then, the global optimum value $\tau^*$ of this problem is*

$$\tau^* = \sup_{w \in \operatorname{dom} f^*} \rho \left\| A^\top w \right\|_1 + a^\top A^\top w + b^\top w - f^*(w), \tag{82}$$

*which can be solved via mixed-integer convex optimization. Furthermore, the optimal solution $\bar{x}$ of problem (72) that attains $\tau^*$ can be found by solving the linear optimization problem $\bar{x} \in$ $\arg\sup\{(A^\top \bar{w})^\top x \ : \ \|x-a\|_\infty \leq \rho\}$, where $\bar{w}$ is the solution of (82).*

*Proof.* The feasible region defined by box constraints is:

$$U_2 = \left\{x \in \mathbb{R}^n : \ g(x) = \|x-a\|_\infty \leq \rho\right\}.$$

Although box constraints are actually a collection of multiple constraints, by using the ∞-norm we can apply our theorem for a single constraint. Using the conjugate

$$g^*(z) = \begin{cases} z^\top a & \text{if } \|z\|_1 \leq 1 \\ \infty & \text{otherwise,} \end{cases}$$

in (77), we obtain that problem (72) is equivalent to minimizing $\tau$ over:

$$\begin{aligned} &\sup_{w \in \operatorname{dom} f^*} \left\{ \min_{\lambda \geq 0} \left\{ \lambda\rho + a^\top A^\top w : \ \left\| A^\top w \right\|_1 \leq \lambda \right\} + b^\top w - f^*(w) \right\} \leq \tau \\ \iff &\sup_{w \in \operatorname{dom} f^*} \left\{ \rho \left\| A^\top w \right\|_1 + a^\top A^\top w + b^\top w - f^*(w) \right\} \leq \tau, \end{aligned}$$

since the minimizer of the inner problem is $\bar{\lambda} = \left\| A^\top w \right\|_1$. This also shows that the piecewise linear decision rule (LDR) is optimal for this problem, as $\bar{\lambda}$ has the structure of a piecewise LDR. We can conclude that if the uncertainty set is $U_2$, the optimal value of problem (72) is given by:

$$\tau^* = \sup_{w \in \operatorname{dom} f^*} \rho \left\| A^\top w \right\|_1 + a^\top A^\top w + b^\top w - f^*(w). \tag{83}$$

Notice that (83) is still a hard problem due to the convexity of $\rho \left\| A^\top w \right\|_1$. However, this 1-norm can be represented by linear terms using extra binary variables for absolute values (see, e.g., (Löfberg 2016)). There are many efficient solvers which can solve the resulting mixed-integer convex optimization problem. For example, if $f$ is a log-sum-exp function, then $f^*$ is the negative entropy (with linear domain constraints), hence problem (83) will be a mixed-integer exponential cone representable problem and MOSEK can efficiently solve these type of problems.

From Theorem 19 it follows that the optimal solution $\bar{x}$ of the original problem that attains the value $\tau^*$ can be retrieved by solving $\bar{x} \in \arg\sup \left\{ (A^\top \bar{w})^\top x \;:\; x \in U_2 \right\}$ (as we cannot derive a closed-form solution due to the $\infty$-norm). □

Numerical experiments of Corollary 11 can be found in Section 5.4.2.

**Corollary 12** ($p$-norm Constraint)**.** *Let* $g(x) := \|x-a\|_p$ *in problem* (72)*, where* $p \geq 1$ *and* $a \in \mathbb{R}^n$ *is a parameter. Then, the global optimum value* $\tau^*$ *of this problem is*

$$\tau^* = \sup_{w \in \operatorname{dom} f^*} \rho \left\| A^\top w \right\|_q + a^\top A^\top w + b^\top w - f^*(w), \tag{84}$$

*where* $q$ *is given by* $1/q + 1/p = 1$*. The solution* $\bar{x}$ *that attains this value in problem* (72) *is obtained by solving* $\bar{x} \in \arg\sup\{(A^\top \bar{w})^\top x \;:\; \|x-a\|_p \leq \rho\}$*, where* $\bar{w}$ *is the solution of* (84)*.*

*Proof.* The dual norm of the $p$-norm is the $q$-norm, where $q$ is given by $\frac{1}{q} + \frac{1}{p} = 1$. The rest of this proof is analogous to the proof of Corollary 11, hence we can conclude that the optimal objective value of problem (72) is $\tau^* = \sup_{w \in \operatorname{dom} f^*} \rho \left\| A^\top w \right\|_q + a^\top A^\top w + b^\top w - f^*(w)$. The solution $\bar{x}$ can be found by similar arguments as in Corollary 11. □

Corollary 12 generalizes the previous two corollaries, i.e., $p = 2, \infty$. Solving problem (84) is in general difficult, however the non-convexity of this problem is not due to $f^*(w)$, but the dual norm of the original $g(x)$. This may in turn significantly simplify solving the convex maximization problem. In Corollary 10 we showed cases where we can efficiently approximate problem (84) for $p = 2$ and obtain the corresponding lower bound solution, and in Corollary 11 we showed how to globally solve problem (84) by mixed-integer convex optimization, along with the corresponding solution of problem 72 for $p = \infty$.

## 5.3 ARO Reformulation of Multiple-Constrained Convex Maximization

In this section, we first extend Theorem 19 to convex maximization problems with multiple convex constraints. Then, we give tractable upper and lower bound relaxation problems for the case of multiple linear constraints.

### 5.3.1 Convex Maximization over Multiple Convex Constraints

Let $f : \mathbb{R}^m \mapsto \mathbb{R}$ be a closed convex objective function, and $g_j : \mathbb{R}^n \mapsto \mathbb{R}, j = 1, \dots, q$ be closed convex functions. We assume that $\exists \bar{x}_j \in \mathbb{R}^n : \; g_j(\bar{x}_j) < \rho_j$ for $j = 1, \dots, q$. Similarly to the previous setting, let $A \in \mathbb{R}^{m \times n}$ and $b \in \mathbb{R}^m$. We consider the following convex maximization problem with multiple convex constraints:

$$\begin{aligned} \max_{x \in \mathbb{R}^n} \quad & f(Ax + b) \\ \text{s.t.} \quad & g_j(x) \le \rho_j, \quad j = 1, \dots, q. \end{aligned} \tag{85}$$

Let $U = \{x \in \mathbb{R}^n : \; g_i(x) \le \rho_i, \; i = 1, \dots, q\}$ denote the feasible set of problem (85).

**Theorem 20.** *The optimal objective value of problem* (85) *is equal to the optimal objective value of the following adjustable robust optimization problem:*

$$\begin{aligned} & \inf_{\tau} \; \tau \\ \text{s.t.}\, \forall w \in \operatorname{dom} f^*, \exists \lambda \in \mathbb{R}^q_+, z \in \mathbb{R}^{n \times q} : & \begin{cases} \sum_{j=1}^q \lambda_j \rho_j + \sum_{i=1}^q \lambda_i g_i^* \left( \dfrac{z_i}{\lambda_i} \right) + b^\top w - f^*(w) \le \tau \\ \sum_{j=1}^q z_j = A^\top w, \end{cases} \end{aligned} \tag{86}$$

*where $z_j$ denotes the $j$-th column of $z$, and $\mathbb{R}_+$ is the set of nonnegative real numbers. The optimal solution $\bar{x}$ that attains this value satisfies $\bar{x} \in \arg\sup \left\{ (A^\top \bar{w})^\top x \; : \; x \in U \right\}$ with $\bar{w} \in \operatorname{dom} f^*$ being the parameter realization where the first constraint of* (86) *is tight at the optimal solution.*

*Proof.* From the proof of Theorem 19 we see that problem (85) can be written as the minimization of $\tau$ over:

$$\sup_{w \in \operatorname{dom} f^*} \left\{ \sup_{x \in U} \left\{ (A^\top w)^\top x \right\} + b^\top w - f^*(w) \right\} \le \tau. \tag{87}$$

Consider the inner problem in constraint (87). Taking its Lagrangian dual problem gives:

$$\begin{aligned} & \sup_{x \in U} \left\{ (A^\top w)^\top x \right\} \\ = & \inf_{\lambda \in \mathbb{R}^q_+} \left\{ \sup_{x \in \mathbb{R}^n} \left\{ (A^\top w)^\top x - \sum_{i=1}^q \lambda_i g_i(x) \right\} + \sum_{j=1}^q \lambda_j \rho_j \right\} \end{aligned}$$

$$= \inf_{\lambda \in \mathbb{R}^q_+} \left\{ \left( \sum_{i=1}^{q} \lambda_i g_i \right)^* (A^\top w) + \sum_{j=1}^{q} \lambda_j \rho_j \right\}. \tag{88}$$

To simplify $\left( \sum_{i=1}^{q} \lambda_i g_i \right)^*$, we use the fact that the conjugate of sum of convex functions can be written as the *infimal convolution* of these functions (Rockafellar 1997). This equivalence gives:

$$\begin{aligned}
\left( \sum_{i=1}^{q} \lambda_i g_i \right)^* (A^\top w) &= \inf_{z_1, \ldots, z_q \in \mathbb{R}^n} \left\{ \sum_{i=1}^{q} (\lambda_i g_i)^*(z_i) : \sum_{j=1}^{q} z_j = A^\top w \right\} \\
&= \inf_{z_1, \ldots, z_q \in \mathbb{R}^n} \left\{ \sum_{i=1}^{q} \lambda_i (g_i)^* \left( \frac{z_i}{\lambda_i} \right) : \sum_{j=1}^{q} z_j = A^\top w \right\},
\end{aligned}$$

where the last step holds since $\lambda_i g_i^* \left( \frac{z_i}{\lambda_i} \right) = \lambda_i \sup_x \left\{ \frac{z_i^\top x}{\lambda_i} - g_i(x) \right\} = (\lambda_i g_i)^*(z_i)$. Therefore, (88) becomes:

$$\inf_{\lambda \in \mathbb{R}^q_+} \left\{ \sum_{j=1}^{q} \lambda_j \rho_j + \inf_{z_1, \ldots, z_q \in \mathbb{R}^n} \left\{ \sum_{i=1}^{q} \lambda_i g_i^* \left( \frac{z_i}{\lambda_i} \right) : \sum_{j=1}^{q} z_j = A^\top w \right\} \right\},$$

and so constraint (87) reduces to:

$$\sup_{w \in \operatorname{dom} f^*} \left\{ \inf_{\substack{z \in \mathbb{R}^{n \times q} \\ \lambda \in \mathbb{R}^q_+}} \left\{ \sum_{j=1}^{q} \lambda_j \rho_j + \sum_{i=1}^{q} \lambda_i g_i^* \left( \frac{z_i}{\lambda_i} \right) + b^\top w - f^*(w) \ : \ \sum_{j=1}^{q} z_j = A^\top w \right\} \right\} \leq \tau. \tag{89}$$

Minimizing $\tau$ over (89) gives us us the global optimum value of problem (85). The optimal solution $\bar{x}$ of the original problem can be retrieved by solving $\bar{x} \in \arg\sup \left\{ (A^\top \bar{w})^\top x \ : \ x \in U \right\}$ where $\bar{w} \in \operatorname{dom} f^*$ maximizes the left-hand side of (89) (the reasoning is identical to Theorem 19, i.e., such $\bar{w}$ maximizes (87) equivalently). □

One way to approximate the ARO problem (86) is to use static variables for the adjustable variables. Assume without loss of generality that in problem (85) we have $x \in \mathbb{R}^n_+$. It is possible to verify that this problem is equivalent to problem (86) with the last constraint being $\sum_{j=1}^{q} z_j \geq A^\top w$ (instead of equality). By relaxing the adjustable variables to static variables, we obtain a tractable approximation of the problem. Hence, the following convex optimization problem returns an upper bound to the original problem:

$$\begin{aligned}\inf_{\tau\in\mathbb{R},\ \lambda\in\mathbb{R}^q_+,\ z\in\mathbb{R}^{n\times q}} \quad & \tau \\ \text{s.t.} \quad & \textstyle\sum_{j=1}^q \lambda_j\rho_j + \sum_{i=1}^q \lambda_i g_i^*\left(\frac{z_i}{\lambda_i}\right) + \sup\limits_{w\in\operatorname{dom} f^*} \left\{b^\top w - f^*(w)\right\} \le \tau \\ & \sup_{w\in\operatorname{dom} f^*} \left\{A_i^\top w\right\} \le \textstyle\sum_{j=1}^q z_{ij}, \qquad i=1,\dots,n.\end{aligned} \tag{90}$$

Moreover, we know that solving $\arg\sup\limits_{x\in\mathbb{R}^n_+} \left\{(A^\top\bar{w})^\top x - \sum_{i=1}^q \bar{\lambda}_i g_i(x)\right\}$ gives us the global optimum solution, where $(\bar{\lambda}, \bar{w})$ solves the optimization problem of (89). Since we used a safe-approximation for this problem, we can use the $\lambda$ value which solves problem (90) as an approximation of $\bar{\lambda}$, and try $n+1$ many scenarios for $w$ solving each supremization problem in (90), in order to obtain a lower bound $\bar{x}$. Such a lower bound approach is thoroughly described in Section 5.3.2 where we approximate the problem of maximizing a convex function over linear constraints.

When the constraints are linear, problem (86) is linear in the adjustable variables, hence this deserves a separate treatment which is done in the next subsection.

### 5.3.2 Convex Maximization over a Polyhedron

In this subsection, we consider problem (85) in which the feasible set is a polyhedron with a nonempty interior. Application of Theorem 20 yields an attractive adjustable robust linear optimization reformulation, for which efficient approximations exist in the literature. We first illustrate how the problem can be approximated with lower and upper bounds. Then, we consider special cases of the objective function $f$: quadratic, log-sum-exp (geometric), and sum-of-max-terms.

Formally, we work on the following problem for $D\in\mathbb{R}^{q\times n}$ and $d\in\mathbb{R}^n$:

$$\begin{aligned}\max_{x\in\mathbb{R}^n_+} \quad & f(Ax+b) \\ \text{s.t.} \quad & Dx\le d,\end{aligned} \tag{91}$$

which is a special case of problem (85) with

$$U=\{x\in\mathbb{R}^n_+ :\ Dx\le d\}. \tag{92}$$

Let $D_{(j)}$ denote the $j$-th row of $D$. Then, the $j$-th constraint of $U$ is given by

$$g_j(x) \leq \rho_j \iff D_{(j)}x \leq d_j.$$

By Theorem 20, the optimal objective values of problem (91) and the following problem are equal:

$$\inf_{\tau\in\mathbb{R}} \tau \text{ s. t.} \forall w \in \operatorname{dom} f^*, \exists \lambda \in \mathbb{R}^q_+, z \in \mathbb{R}^{n\times q} : \begin{cases} d^\top \lambda + b^\top w - f^*(w) \leq \tau \\ z_i \leq D_{(i)}\lambda_i, \quad i = 1, \dots, q \\ \sum_{j=1}^q z_j = A^\top w, \end{cases}$$

Constraints $\sum_{j=1}^q z_j = A^\top w$ and $z_i \leq D_{(i)}\lambda_i$, $i = 1, \dots, q$ together can be written as $A^\top w \leq D^\top \lambda$. Therefore, problem (91) has the same optimal objective value as the ARO problem:

$$\inf_{\tau\in\mathbb{R}} \tau \text{ s. t.} \forall w \in \operatorname{dom} f^*, \exists \lambda \in \mathbb{R}^q : \begin{cases} d^\top \lambda + b^\top w - f^*(w) \leq \tau \\ D^\top \lambda \geq A^\top w \\ \lambda \geq \mathbf{0}. \end{cases} \tag{93}$$

Notice that the final problem is a linear ARO problem with fixed recourse (linearity is obtained by lifting the $-f^*(w)$ term to the uncertainty set). There are many possible methods one can use to solve such a problem, for example, one can solve this problem to optimality by eliminating the adjustable variables via Fourier-Motzkin Elimination for ARO (Zhen et al. 2018), which is efficiently applicable for small-sized problems. We refer to Yanıkoğlu et al. (2019) for a survey of alternative methods to solve this linear ARO problem. In the remaining of this section we show how to derive tractable problems to find upper and lower bounds on the optimal objective value of problem (91).

In Appendix 5.B of the supplements it is shown that by using linear decision rules, one obtains a (tractable) safe approximation of the constraints of ARO problem (93) as:

$$\inf_{u\in\mathbb{R}^q, V\in\mathbb{R}^{q\times m}, r\in\mathbb{R}^q, \tau\in\mathbb{R}} \tau \begin{cases} d^\top(u + Vw + rw_0) + b^\top w + w_0 - \tau \leq 0, & \forall \begin{pmatrix} w_0 & w^\top \end{pmatrix}^\top \in W & \text{(94a)} \\ -D^\top(u + Vw + rw_0) + A^\top w \leq \mathbf{0}, & \forall \begin{pmatrix} w_0 & w^\top \end{pmatrix}^\top \in W & \text{(94b)} \\ -(u + Vw + rw_0) \leq \mathbf{0} & \forall \begin{pmatrix} w_0 & w^\top \end{pmatrix}^\top \in W, & \text{(94c)} \end{cases}$$

with $W = \left\{(w_0 \; w^\top)^\top \in \mathbb{R}^{m+1} : \; w_0 + f^*(w) \leq 0\right\}$, which is used to prove the following result.

**Theorem 21.** *The optimal objective value of the following problem is an upper bound to the optimal objective value of problem (91):*

$$\inf_{u\in\mathbb{R}^q, V\in\mathbb{R}^{q\times m}, r\in\mathbb{R}^q, \tau\in\mathbb{R}} \tau \;\text{ s.t.} \left\{\begin{array}{rcr} d^\top u + \left(1+d^\top r\right) f\left(\dfrac{V^\top d + b}{1+d^\top r}\right) \leq \tau & & \\ & & (95a) \\ 1 + d^\top r \geq 0 & & \\ -D_i^\top u + \left(-D_i^\top r\right) f\left(\dfrac{A_i - V^\top D_i}{-D_i^\top r}\right) \leq 0 & & \\ & i = 1, \ldots, n & (95b) \\ -D_i^\top r \geq 0 & & \\ -u_i + (-r_i) f\left(\dfrac{V_{(i)}^\top}{r_i}\right) \leq 0 & & \\ & i = 1, \ldots, q & (95c) \\ -r_i \geq 0. & & \end{array}\right.$$

*Here, $V_{(i)}$ stands for the $i$-th row of $V$ where $A_i$, $D_i$ are the $i$-th columns of $A$, $D$.*

**Remark 4.** *Problem (95) is a convex optimization problem whose complexity depends on the perspective of $f$. If $f$ is positively homogeneous, the perspective function is $z_0 f(\frac{z}{z_0}) = f(z)$ which in turn makes problem (95) easier (without variable $r$).* ■

**Remark 5.** *When problem (95) is hard to solve, the adversarial approach (Bienstock and Özbay 2008) could be a valuable alternative, which avoids directly solving this problem. This approach takes the equivalent safe approximation problem (94), replaces $W$ with a finite set $\overline{W}$, and solves the resulting linear optimization problem (LP). Then, the $\left(w_0 \quad w^\top\right)^\top \in W$ values that violate each of the constraints the most are added to set $\overline{W}$. Iterating this procedure until there is no violation guarantees optimality at termination. Obviously, this approach does not use perspective functions, but if the number of LPs solved is large, then this also becomes a hard problem.* ■

The solution of the upper bound problem can be used to obtain a (potentially good) lower bound for problem (91) by using what was proposed for two-stage fixed-recourse robust constraints by Hadjiyiannis et al. (2011) and extended by Zhen et al. (2017). Theorem 20 states that the global optimum solution of problem (91) can be obtained by solving $\bar{x} \in \arg\sup\{(A^\top \bar{w})^\top x \; : \; x \in U\}$ where $\bar{w} \in \operatorname{dom} f^*$ is the scenario where the first constraint of (93) is tight. However, finding $\bar{w}$ is as hard as solving (93) (see the proof of Theorem 20), while for any $w \in \operatorname{dom} f^*$, solving $\arg\sup\{(A^\top w)^\top x \; : \; x \in U\}$ gives us a feasible (lower bound) solution of problem (91).

Hence, we generate a finite set $\overline{W}$, which will be used to obtain good lower bound solutions, i.e., for each $w \in \overline{W}$ we solve:

$$x \in \arg\sup_{x \in U} \{(A^\top w)^\top x\}, \tag{96}$$

and the best solution is the one giving the best objective $f(Ax+b)$. The set $\overline{W}$ can be generated in multiple ways, and we use our upper bound solution for this purpose, namely $(\hat{u}, \hat{r}, \hat{V}, \hat{\tau})$ standing for optimal $(u, r, V, \tau)$ values in (95). We plug this solution in the safe approximation (94) of the ARO problem (93) that is equivalent to the original convex maximization problem. Then, we generate elements of $\overline{W}$ as the worst-case parameter realizations in each constraint of (94), independently. For example, the first element of $\overline{W}$ is obtained from the first constraint (94a) as:

$$\begin{aligned}
\overline{W}^1 &\in \arg\sup_{w \in \operatorname{dom} f^*} \{ \sup_{w_0 \le -f^*(w)} \quad d^\top(\hat{u} + \hat{V}w + \hat{r}w_0) + b^\top w + w_0 - \hat{\tau} \} \\
&= \arg\sup_{w \in \operatorname{dom} f^*} \{ \sup_{w_0 \le -f^*(w)} \quad (1 + d^\top \hat{r}) w_0 + (d^\top \hat{V} + b^\top) w + d^\top \hat{u} - \hat{\tau} \} \\
&= \arg\sup_{w \in \operatorname{dom} f^*} \left\{ -(1 + d^\top \hat{r}) f^*(w) + (d^\top \hat{V} + b^\top) w + d^\top \hat{u} - \hat{\tau} \right\},
\end{aligned}$$

We thus construct $\overline{W}$ as $\overline{W} = \overline{W}^1 \cup [\bigcup_{i=1}^n \overline{W}_i^2] \cup [\bigcup_{i=1}^q \overline{W}_i^3]$ with:

$$\overline{W}^1 \in \arg\sup_{w \in \operatorname{dom} f^*} \left\{ -(1 + d^\top r) f^*(w) + (d^\top \hat{V} + b^\top) w + d^\top \hat{u} - \hat{\tau} \right\}, \tag{97}$$

$$\overline{W}_i^2 \in \arg\sup_{w \in \operatorname{dom} f^*} \left\{ (D_i^\top r) f^*(w) + (A_i^\top - D_i^\top \hat{V}) w - D_i^\top \hat{u} \right\}, \qquad i = 1, \ldots, n \tag{98}$$

$$\overline{W}_i^3 \in \arg\sup_{w \in \operatorname{dom} f^*} \left\{ -\hat{u}_i - \hat{V}_i w + \hat{r}_i f^*(w) \right\}, \qquad i = 1, \ldots, q. \tag{99}$$

In the above notation $\overline{W}$ is a finite set of $n+q+1$ many $w$ scenarios. With this approach, we aim to represent $\bar{w}$ (the worst-case parameter realization of the original ARO problem) as accurately as possible by plugging the optimal LDR from the upper bound relaxation problem (95) and then collecting the worst-case parameter realization in each constraint of the safe approximation. We emphasize that the solution of the upper bound relaxation is directly being used in the process of obtaining a lower bound. Hence, the upper bound problem is not only being solved to get an upper bound value, but more importantly also to obtain a good solution to the original problem. The upper/lower bound approximation scheme is summarized in Appendix 5.C of the supplements.

We next derive the approximations for specific problems, namely convex quadratic maximization, convex log-sum-exp (geometric) maximization, and convex sum-of-max-linear-terms maximization. The results are shared in Tables 34, 35, and 36. Complete derivations can be found in Appendix 5.D of the supplements. In summary, we see that the upper bound approximation of the convex quadratic maximization problem is found by solving a second-order cone optimization problem, and from the solution of this problem the lower bound scenarios can be collected analytically. For geometric maximization, the upper bound problem is a convex exponential cone optimization problem, and the lower bound scenarios can also be collected by solving multiple exponential cone optimization problems. Finally, for sum-of-max-linear-terms maximization, the upper bound problem is a linear optimization problem, and the lower bound scenarios can be found analytically. All of the original problems are known to be very hard problems, while the approximation problems are mainstream convex optimization problems and there exist many powerful solvers to solve such problems. In the numerical experiments, we use Mosek as a conic optimization solver and CPLEX (IBM ILOG CPLEX 2014) as a linear optimization solver.

| Problem | Parameters and Assumptions | | Convex Maximization Formulation |
|---|---|---|---|
| Quadratic Maximization | $g(x) = x^\top Qx + \ell^\top x$,<br>$Q \succeq 0$,<br>$\ell \in \mathbb{R}^n$, | convex quadratic funtion<br>positive semi-definite matrix<br>linear coefficients vector | $\max_{x\in\mathbb{R}^n_+} \; g(x) = x^\top Qx + \ell^\top x$<br>s. t. $Dx \le d$ |
| Geometric Maximization | $f(z) = \log\left(\sum_{i=1}^m \exp(z_i)\right)$,<br>$A \in \mathbb{R}^{m\times n}$,<br>$b \in \mathbb{R}^m$, | log-sum-exp function<br>linear coefficients matrix<br>linear constants vector | $\max_{x\in\mathbb{R}^n_+} \; f(Ax+b) = \log\left(\sum_{i=1}^m \exp(A_{(i)}x + b_i)\right)$<br>s. t. $Dx \le d$ |
| Sum-of-Max-Linear-Terms Optimization | $f(z) = \sum_{k=1}^K \max_{j\in\mathcal{I}_k} \{z_j\}$,<br>$K > 0$,<br>$\mathcal{I}_k \subseteq \{1,\ldots,m\}$,<br>$\mathcal{I}_k \cap \mathcal{I}_\ell = \emptyset$,<br>$\cup_{k=1}^K \mathcal{I}_k = \{1,\ldots,m\}$,<br>$A \in \mathbb{R}^{m\times n}$,<br>$b \in \mathbb{R}^m$, | sum-of-max-terms function<br>number of max-terms<br>indexes for $k$-th max term<br>for $k \ne \ell$, w.l.o.g.<br>w.l.o.g.<br>linear coefficients matrix<br>linear constants vector | $\max_{x\in\mathbb{R}^n_+} \; f(Ax+b) = \sum_{k=1}^K \max_{j\in\mathcal{I}_k}\{A_{(j)}x + b_j\}$<br>s. t. $Dx \le d$<br>Note: By using logical programming one can reformulate as mixed integer optimization |

Table 34: *Quadratic, Geometric, and Sum-of-Max-Linear-Terms Maximization problems. In each case we introduce the problem setting, and share the main convex maximization problem that we are interested in solving. All problems have the same linear constraints $Dx \le d$ for $D \in \mathbb{R}^{q\times n}$, $d \in \mathbb{R}^q$.*

| Problem | Upper Bound Problem | Problem Type and Variables |
|---|---|---|
| Quadratic Maximization | inf $\tau$<br>s. t. $d^\top u + \bar{v}^\top d - (1+\tau)/2 + \left\lVert \begin{pmatrix} \tilde{V}^\top d \\ \hat{v}^\top d + (1-\tau)/2 \end{pmatrix} \right\rVert_2 \le 0$<br>$-D_i^\top u + \frac{\ell_i}{2} - \bar{v}^\top D_i + \left\lVert \begin{pmatrix} L_i - \tilde{V}^\top D_i \\ \ell_i/2 - \hat{v}^\top D_i \end{pmatrix} \right\rVert_2 \le 0, \quad i = 1,\ldots,n$<br>$-u_i - \bar{v}_i + \left\lVert \begin{pmatrix} -\tilde{V}_{(i)}^\top \\ -\hat{v}_i \end{pmatrix} \right\rVert_2 \le 0, \quad i = 1,\ldots,q$ | Second-order Cone Optimization Problem<br>$\tau \in \mathbb{R}, u \in \mathbb{R}^q, \bar{v} \in \mathbb{R}^q, \hat{v} \in \mathbb{R}^q, \tilde{V} \in \mathbb{R}^{q\times m}$,<br>$L$ is obtained by decomposition $Q = L^\top L$ |
| Geometric Maximization | inf $\tau$<br>s. t. $1 + d^\mathsf{T} r \ge \sum_{j=1}^m z_j^{(1)}$<br>$\left(z_j^{(1)}, 1 + d^\mathsf{T} r, \left(V_j^\top d + b_j - \tau + d^\mathsf{T} u\right)\right) \in \mathcal{K}_{\exp}, \; j = 1,\ldots,m,$<br>$-D_i^\mathsf{T} r \ge \sum_{j=1}^m z_{ij}^{(2)}, \quad i = 1,\ldots,n$<br>$\left(z_{ij}^{(2)}, -D_i^\mathsf{T} r, (A_{i,(j)} - V_j^\mathsf{T} D_i - D_i^T u)\right) \in \mathcal{K}_{\exp}, \; j = 1,\ldots,m, \quad i = 1,\ldots,n$<br>$-r_i \ge \sum_{j=1}^m z_{ij}^{(3)}, \quad i = 1,\ldots,q$<br>$\left(z_{ij}^{(3)}, -r_i, (-V_{j,(i)} - u_i)\right) \in \mathcal{K}_{\exp}, \quad j = 1,\ldots,m, \quad i = 1,\ldots,q$ | Exponential Cone Optimization Problem<br>$r \in \mathbb{R}^q, u \in \mathbb{R}^q, V \in \mathbb{R}^{q\times m}, \tau \in \mathbb{R}$,<br>$z^{(1)} \in \mathbb{R}^m, z^{(2)} \in \mathbb{R}^{n\times m}, z^{(3)} \in \mathbb{R}^{q\times m}$,<br>$\mathcal{K}_{\exp}$ denotes the exponential cone |
| Sum-of-Max-Linear-Terms Optimization | inf $\tau$<br>s. t. $d^\top u + \sum_{k=1}^K \max_{j\in\mathcal{I}_k}\{V_j^\top d + b_j\} \le \tau$<br>$-D_i^\top u + \sum_{k=1}^K \max_{j\in\mathcal{I}_k}\{A_{i,(j)} - V_j^\top D_i\} \le 0, \quad i = 1,\ldots,n$<br>$-u_i + \sum_{k=1}^K \max_{j\in\mathcal{I}_k}\{V_{j,(i)}\} \le 0, \quad i = 1,\ldots,q$ | Linear Optimization Problem<br>$u \in \mathbb{R}^q, V \in \mathbb{R}^{q\times m}, \tau \in \mathbb{R}$,<br>for linearity use auxiliary variables |

Table 35: *For each of the three problems, we show the upper bound approximation problems. These problems are special forms of upper bound problem (95) of Theorem 21.*

## 5.4 Numerical Experiments

In this section we present numerical experiments to support the theory developed in this work. We use YALMIP through MATLAB 2018b (MATLAB 2018) to call various solvers, and report the solver times below (we do not reflect the problem formulation times in YALMIP). Numerical experiments are obtained by using a standard personal computer with an 8-th Generation

| Problem | Lower Bound Scenarios | Problem Type |
|---|---|---|
| Quadratic Maximization | $\overline{W}^1 = \begin{bmatrix} h\left(\begin{pmatrix} \tilde{V}^\top d \\ d^\top \hat{v} + (1-\tau)/2 \end{pmatrix}\right) \\ 1 \end{bmatrix}$<br>$\overline{W}_i^2 = \begin{bmatrix} h\left(\begin{pmatrix} L_i - \tilde{V}^\top D_i \\ \ell_i/2 - D_i^\top \hat{v} \end{pmatrix}\right) \\ 1 \end{bmatrix}, \qquad i = 1, \dots, n$<br>$\overline{W}_i^3 = \begin{bmatrix} h\left(\begin{pmatrix} -\tilde{V}_{(i)}^\top \\ -\hat{v}_i \end{pmatrix}\right) \\ 1 \end{bmatrix}, \qquad i = 1, \dots, q,$<br>where $h(\cdot)$ normalizes its input. | Analytic Scenarios |
| Geometric Maximization | $\overline{W}^1 \in \operatorname*{arg\,sup}_{w,t \in \mathbb{R}^m} \ (1 + d^\top r)(\sum_{j=1}^m t_j) + (d^\top V + b^\top)w + d^\top u - \tau$<br>$\text{s.t.} \ (1, w_j, t_j) \in \mathcal{K}_{\text{exp}}, \quad j = 1, \dots, m$<br>$\sum_{j=1}^m w_j = 1$<br>$\overline{W}_i^2 \in \operatorname*{arg\,sup}_{w,t \in \mathbb{R}^m} \ (-D_i^\top r)(\sum_{j=1}^m t_j) + (A_i^\top - D_i^\top V)w - D_i^\top u$<br>$\text{s.t.} \ (1, w_j, t_j) \in \mathcal{K}_{\text{exp}}, \quad j = 1, \dots, m \qquad i = 1, \dots, n$<br>$\sum_{j=1}^m w_j = 1$<br>$\overline{W}_i^3 \in \operatorname*{arg\,sup}_{w,t \in \mathbb{R}^m} \ (-r_i)(\sum_{j=1}^m t_j) + (-V_{(i)})w - u_i$<br>$\text{s.t.} \ (1, w_j, t_j) \in \mathcal{K}_{\text{exp}}, \quad j = 1, \dots, m \qquad i = 1, \dots, q$<br>$\sum_{j=1}^m w_j = 1$ | Each scenario collection is an exponential cone problem |
| Sum-of-Max-Linear-Terms Optimization | $\overline{W}^1 \in \operatorname*{arg\,sup}_{w} \ d^\top (u + Vw) + b^\top w - \tau$<br>$\text{s.t.} \ w_j \geq 0, \quad j = 1, \dots, m$<br>$\sum_{j \in \mathcal{I}_k} w_j = 1, \quad k = 1, \dots, K$<br>$\overline{W}_i^2 \in \operatorname*{arg\,sup}_{w} \ A_i^\top w - D_i^\top (u + Vw)$<br>$\text{s.t.} \ w_j \geq 0, \quad j = 1, \dots, m \qquad i = 1, \dots, n$<br>$\sum_{j \in \mathcal{I}_k} w_j = 1, \quad k = 1, \dots, K$<br>$\overline{W}_i^3 \in \operatorname*{arg\,sup}_{w} \ -u_i - V_{(i)} w$<br>$\text{s.t.} \ w_j \geq 0, \quad j = 1, \dots, m \qquad i = 1, \dots, q$<br>$\sum_{j \in \mathcal{I}_k} w_j = 1, \quad k = 1, \dots, K$ | Each optimization problem can be split to $K$ optimization problems that are analytically solvable |

Table 36: *For each of the three problems, we share the lower bound scenario collection steps. As discussed before, these scenarios will help us to find a lower bound solution by solving linear optimization problems corresponding to each scenario.*

Intel(R) Core(TM) i7-8750H processor. The details of the data and codes of these experiments are shared in Appendix 5.G of the supplements.

### 5.4.1 Log-Sum-Exp Maximization over a 2-norm Constraint

Consider problem (72) with $g(x) := ||x - a||_2$. In Section 5.2 under Corollary 10 we discussed that for $f(z) = \log(\sum_{i=1}^m \exp(z_i))$, the upper bound approximation problem is exponential-cone representable, and we can analytically obtain the lower bound solution immediately. The approximation problem (81) is solved by MOSEK.

To benchmark our approximation, we used general-purpose nonlinear optimization solvers Artelys Knitro (Byrd et al. 2006), IPOPT (Wächter and Biegler 2006), and BMIBNB of YALMIP (Löfberg 2004). Knitro appeared to yield the best result, so we report Knitro as a benchmark. Although Knitro does not guarantee global optimality it can find the global optimum value faster and more often compared to the other solvers we tried. We also compare our approximation with the global optimization solver BARON. The results can be found in Table 37.

| Problem | Knitro | | Knitro Multi-Start | | | BARON | | **Approximation** | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | value | time | value | start | time | value | time | upper | lower | time |
| # 1 $(n = 5, m = 5)$ | 9.7890 | 0.05 | - | - | - | 9.7890 | 0.40 | 10.0186 | 9.7816 | 0.22 |
| # 2 $(n = 20, m = 15)$ | 54.0329 | 0.18 | 55.8913 | 2 | 0.43 | 25.0731 | 1,800.00 | 56.0521 | 55.7287 | 0.22 |
| # 3 $(n = 100, m = 120)$ | 241.1606 | 1.61 | - | - | - | NA | 1,800.00 | 241.1606 | 241.1606 | 0.23 |
| # 4 $(n = 20, m = 40)$ | 156.6875 | 0.40 | 179.1224 | 5 | 5.52 | NA | 1,800.00 | 179.1224 | 179.1224 | 0.24 |
| # 5 $(n = 50, m = 100)$ | 324.8785 | 1.17 | 370.9066 | 3 | 6.55 | NA | 1,800.00 | 370.9066 | 370.9066 | 0.27 |
| # 6 $(n = 100, m = 100)$ | 428.7200 | 2.37 | 472.0475 | 2 | 8.86 | NA | 1,800.00 | 472.0475 | 472.0475 | 0.25 |
| # 7 $(n = 200, m = 30)$ | 551.3160 | 0.68 | 570.8467 | 20 | 18.45 | NA | 1,800.00 | 570.8467 | 570.8467 | 0.23 |
| # 8 $(n = 400, m = 80)$ | 570.7248 | 3.51 | 601.6886 | 4 | 14.35 | NA | 1,800.00 | 601.6886 | 601.6886 | 0.25 |
| # 9 $(n = 50, m = 20)$ | error | 0.62 | - | - | - | error | 1,800.00 | 890.7722 | 890.7722 | 0.24 |
| # 10 $(n = 10,000, m = 100)$ | NA | 1,800.00 | - | - | - | - | - | 456.6373 | 455.5054 | 0.29 |
| # 11 $(n = 1,000, m = 1,000)$ | 156.4414 | 1311.00 | - | - | - | - | - | 178.4665 | 141.9482 | 0.59 |
| # 12 $(n = 2,000, m = 700)$ | 238.4521 | 1,800.00 | - | - | - | - | - | 325.0580 | 311.7564 | 0.51 |

Table 37: *Comparison of our approximation method with Knitro and BARON for the* 2*-norm constrained log-sum-exp maximization problem. The first column gives the problem numbers, and gives the dimension of these problems (recall the decision variable $x$ is an $n$-dimensional vector and the function $f(Ax + b)$ takes $m$-dimensional vector input, hence $A$ is a matrix of size $m \times n$). The 'value' column stands for the best objective value computed by the corresponding solver, and 'time' stands for how many seconds it took for the solver to compute this. 'start' means the minimum number of starting points for Knitro to find the corresponding value. The column 'upper' is our upper bound approximation, and 'lower' is the lower bound approximation. We allow 30-minutes for solvers, hence 'NA' means that the solver cannot find any feasible point in the time limit, otherwise the best solution found is written. The entry 'error' means that there are numerical errors and the solver terminates.*

These results show that BARON can only find the global optimum value (and guarantee optimality) in the first problem, which is the smallest one. For the other problems BARON is not usable, which is mainly due to the fact that the log-sum-exp function is highly nonlinear as well as evaluating it is a hard task due to the exponential-terms. Moreover, we see that our approximation method (except Problems 1 and 2) is faster than Knitro without multi-start which is designed to find a local optimum. For Problems 4-7 Knitro cannot find the global optimum without using the multi-start option (which needs manual tuning) and there is no guarantee of global optimality at the end, while our method solves the problem with a global optimality guarantee (upper bounds are equal to lower bounds). For Problem 1 and Problem 2 we cannot find the global optimum value, which shows our method does not necessarily find the global optimum in all of the cases. Problem 9 is a problem where the exponential summands in the logarithm operator get very large and the computers accept these terms as infinity, so the solvers cannot compute any solution, and they do not automatically scale the problem. However, since our upper bound approximation does not compute the log-sum-exp function directly, it does not suffer from such numerical issues. To compute a lower bound we need function evaluations and to address the numerical issues we automatically scale the problem. Problem 10 has $10,000$ variables, so none of the solvers can find an initial solution within 1,800 seconds, whereas our method finds the upper and lower bounds within 0.29 seconds. Problem 11 has $m = 1,000$ and $n = 1,000$. This problem has a larger approximation gap, which may be a result of the increasing dimension (recall that the proposed approximation problem has $m$ variables). Finally, Problem 12 has a large dimension in terms of $m$ and $n$. Knitro returns a feasible solution within 1,800 seconds but it is not a local optimum (large $n$ increases the dimension of the main problem). We also have a gap between the upper and lower bound approximations in this problem.

### 5.4.2 Log-Sum-Exp Maximization over an $\infty$-norm Constraint

We consider problem (72) with $g(x) := ||x - a||_\infty$. In Corollary 11 we discussed that we can find the global optimum value of problem (72) by solving a mixed-integer convex optimization problem. In light of our findings, when $f$ is a log-sum-exp function, we can represent this problem as a mixed-integer exponential-cone optimization problem and MOSEK can solve these types of problems efficiently. We use the same solvers as in the previous subsection as benchmarks. In this problem, we find the exact global optimum value, i.e., we do not approximate this value. The solutions that attain the global optimum values are also efficiently retrieved, as problem (78) is maximizing a linear function over box constraints. The results are presented in Table 38.

| Problem | Knitro | | Knitro Multi-Start | | | BARON | | **Exact** | |
|---|---|---|---|---|---|---|---|---|---|
| | value | time | value | start | time | value | time | value | time |
| # 1 $(n = 5, m = 5)$ | 9.10 | 0.03 | 15.00 | 2 | 0.04 | 15.00 | 0.06 | 15.00 | 0.28 |
| # 2 $(n = 20, m = 5)$ | 46.34 | 0.04 | 54.50 | 3 | 0.11 | 23.50 | 1,800.00 | 54.50 | 0.39 |
| # 3 $(n = 50, m = 20)$ | NA | 0.20 | - | - | - | NA | 1,800.00 | 1,723.30 | 48.00 |
| # 4 $(n = 180, m = 20)$ | 54.39 | 0.50 | 57.70 | 15 | 6.17 | NA | 1,800.00 | 57.70 | 149.80 |
| # 5 $(n = 30, m = 300)$ | 35.76 | 0.50 | 39.62 | 30 | 33.91 | NA | 1,800.00 | 39.62 | 18.37 |

Table 38: *Comparison of our exact method with Knitro and BARON for $\infty$-norm constrained log-sum-exp maximization. The meanings of common columns are the same as Table 37, and similarly time is in seconds. Here, 'Exact' section stands for our method which computes exactly the global optimum value as shown in Corollary 11.*

In general, our method's computation time is higher compared to the previous problem, mainly because we solve a mixed-integer exponential cone optimization problem. However, contrary to Knitro (even with multi-start), our method provides a global optimality guarantee. Similarly to the previous example, BARON is not scalable for larger problems.

### 5.4.3 Convex Quadratic Maximization over Linear Constraints

We consider the first problem type from Table 34, e.g., maximizing a convex quadratic function with respect to linear constraints. The state-of-the-art solver for convex quadratic maximization is CPLEX (version 12.6 onwards), which uses a branch-and-bound method based on McCormick relaxations and SDP cuts. The algorithm terminates at a global optimum, but as for any branch and bound algorithm, this may take an exponential number of steps (Boyd and Mattingley 2007). On the one hand, our upper bound approximation method solves a second-order cone optimization problem, so we use MOSEK for this purpose. On the other hand, our lower bound approximation method solves multiple linear optimization problems after collecting lower bound scenarios analytically, hence we use CPLEX for this purpose. The results are given in Table 39.

| Problem | CPLEX | | | Upper Bound | | Lower Bound | |
|---|---|---|---|---|---|---|---|
| | value | time | restricted | value | time | value | time |
| # 1 $(n = 20, q = 10)$ | 394.75 | 0.08 | 394.75 | 702.05 | 0.43 | 394.75 | 0.16 |
| # 2 $(n = 20, q = 10)$ | 884.75 | 0.10 | 884.75 | 1,192.10 | 0.43 | 884.75 | 0.15 |
| # 3 $(n = 10, q = 15)$ | 4,674.70 | 0.15 | 4,674.70 | 4,753.10 | 0.01 | 4,674.70 | 0.17 |
| # 4 $(n = 50, q = 62)$ | 175,710.00 | 0.81 | 175,710.00 | 177,640.00 | 0.29 | 175,710.00 | 0.02 |
| # 5 $(n = 100, q = 130)$ | 692,610.00 | 16.50 | 692,610.00 | 707,520.00 | 2.30 | 692,610.00 | 1.40 |
| # 6 $(n = 200, q = 240)$ | 6,020,800.00 | 466.00 | 6,020,800.00 | 6,100,200.00 | 25.00 | 6,020,800.00 | 3.40 |
| # 7 $(n = 240, q = 280)$ | 7,303.00 | 3,600.00 | 7,303.00 | 1,961,700.00 | 31.21 | 1,855,700.00 | 4.18 |

Table 39: *Comparison of our approximation method and CPLEX with multiple linear constrained convex quadratic maximization. Here the column names that are used in the previous tables have the same meaning, and time is similarly in seconds. Recall that q is the number of (linear) constraints. The 'Upper Bound' section is our proposed upper bound, and 'Lower Bound' is the proposed lower bound. In 'restricted' column of CPLEX, we are giving a time limit of the total time needed by our approximation method (upper + lower) and record the best CPLEX can find. Note that when the 'restricted' case finds the same value as its 'value' (e.g., the unlimited time case), it means CPLEX finds the global optimum solution but cannot guarantee global optimality yet.*

Problems 1-3 are small-sized problems where we see that CPLEX finds the global optimum faster than our approximation method's total time. In Problems 4-5 CPLEX finds the global optimum slower than our approximation, while with a time limit of the time our approximation takes CPLEX still finds the global optimum, but cannot guarantee global optimality. A similar argument holds for Problem 6, but the run-time of CPLEX increases significantly. Finally in Problem 7, CPLEX cannot find the global optimum solution in an hour (3,600 seconds). In general, we see that our lower bound values are all equal to the global optimum value (except for Problem 7) with a reasonable run-time, however we cannot guarantee global optimality or give tight upper bounds, unlike in our other experiments.

### 5.4.4 Log-Sum-Exp Maximization over Linear Constraints

We consider the second problem type from Table 34, where we are interested in maximizing a log-sum-exp function over linear constraints. We use the state-of-the-art general purpose global optimization solver BARON as the main solver. We also use Knitro as a local optimization solver. Since our upper and lower bound problems are exponential cone representable problems, we use MOSEK solver for solving this problem. The results can be found in Table 40.

| Problem | Knitro | | Knitro Multi-Start | | | Baron | | Upper Bound | | Lower Bound | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | value | time | value | start | time | value | time | value | time | value | time |
| # 1 (size= 10) | 16.8763 | 0.05 | 35.2008 | 3 | 0.10 | 35.2008 | 0.32 | 35.2008 | 0.05 | 35.2008 | 0.01 |
| # 1 (size= 40) | 248.7589 | 0.29 | - | - | - | 248.7589 | 0.36 | 248.7589 | 1.13 | 248.7589 | 0.01 |
| # 1 (size= 60) | 282.6632 | 0.48 | 386.0733 | 5 | 1.95 | 384.6207 | 114.38* | 386.0733 | 2.95 | 386.0733 | 0.01 |
| # 1 (size= 100) | 667.0963 | 2.19 | 676.8081 | 20 | 37.77 | 547.8524 | 10.94* | 676.8080 | 9.98 | 676.8080 | 0.01 |
| # 2 (size= 20) | 64.5087 | 0.10 | 64.8926 | 10 | 0.43 | 64.6196 | 0.28* | 65.2268 | 0.37 | 64.8926 | 0.30 |
| # 3 (size= 50) | 102.5824 | 0.32 | 145.3380 | 50 | 11.24 | 144.1676 | 0.64* | 145.3380 | 9.73 | 145.3380 | 0.01 |
| # 4 (size= 100) | 166.8399 | 1.85 | 176.1074 | 50 | 77.05 | 166.8399 | 3.17* | 176.1074 | 241.57 | 176.1074 | 0.01 |
| # 5 (size= 10) | 34.8293 | 0.05 | 45.0356 | 3 | 0.10 | 45.0356 | 1,000.00 | 45.0356 | 0.18 | 45.0356 | 0.01 |
| # 6 (size= 30) | 74.8013 | 0.17 | 76.0362 | 15 | 1.86 | 76.0362 | 0.72* | 76.1283 | 14.71 | 76.0362 | 4.37 |

Table 40: *Comparison of our approximation method with Knitro and BARON. Size of the problem means $n = q = m$ and these are equal to the given size, i.e., the number of variables, number of constraints, and input dimension of the log-sum-exp function are equal. A star (*) next to the solution-time of BARON means the solver encountered numerical issues and it returns the best solution in the reported time.*

Except for Problems 2 and 6, we find global optimum upper and lower bounds, enabling us to guarantee global optimality. For Problems 2 and 6 we obtain tight upper bounds, and our lower bounds are globally optimal. Our upper bound approximation problem works fast, however it is an exponential cone problem with $2q + qm$ variables in total. Depending on the size of the problem, our approximation may get slower as shown in Problem 4. However, in such large-scale problems Knitro also needs a lot of multiple starting points, and as it cannot find an upper bound to the solution obtained, it cannot guarantee optimality. Knitro needs multi-start to find the global optimum in all of the problems except for Problem 1 with $n = 40$. Also, it should be noted that BARON cannot guarantee global optimality in several cases. This is because BARON has numerical difficulties in these problems.

### 5.4.5 Sum-of-Max-Linear-Terms Maximization over Linear Constraints

Finally, we consider the third problem type from Table 34. For this problem, we solve our approximation problems (which are all linear problems) with CPLEX. We also compare our approximations with directly solving this convex maximization problem via GUROBI version 9.0 (Gurobi Optimization 2018). Although GUROBI does not solve sum-of-max-linear-terms maximization problems explicitly, YALMIP gives the mixed integer linear optimization problem reformulation of this problem successfully (by applying logical programming). The reason we chose GUROBI as a benchmark solver is because it gives a better performance than CPLEX in our numerical experiments. Therefore, we compare our linear approximation problems to the performance of GUROBI on mixed integer optimization to solve the problem to global optimality. Recall that mixed integer optimization is NP-hard (Nemhauser and Wolsey 1988). The results are given in Table 41.

Our method does not necessarily yield a global optimum contrary to solving the mixed-integer optimization reformulation of the problem, however, in large problems our method converges to a lower bound solution very swiftly. On the other hand, GUROBI suffers severely from the curse of dimensionality in mixed integer optimization. Therefore, we see that in bigger problems our lower bound quality is better than the best solution found by GUROBI, while in small-sized problems the solver converges to global optima very quickly. Our method particularly stands out in terms of the speed. We find the global optimum (and certify optimality) in Problems 1,2,7,8. Note that in Problem 1 the global optimality of our approximation is not a coincidence because in this problem we have $K = 1$, and we know linear decision rules are optimal in this case (Ardestani-Jaafari and Delage 2016). Moreover, GUROBI cannot find the global optimum in the allowed 1,000 seconds for Problems 4,5,6,8,9,10,12,13. Except for Problems 3,4,5, our lower bound solutions are better than the ones found by GUROBI, or equal with a faster computation time. In Problem 3, GUROBI finds a better value than our lower bound in the same time (time restricted case). In the largest problems, namely 6 and 13, the best solutions found by GUROBI in 1,000 seconds are considerably lower than our lower bound solutions, where our method computes the upper and lower bounds within 150 and 11 seconds, respectively.

| Problem | GUROBI | | | Upper Bound | | Lower Bound | |
|---|---|---|---|---|---|---|---|
| | value | time | restricted | value | time | value | time |
| # 1 $(n = 5, \lvert\mathcal{I}_k\rvert = 5, K = 1)$ | 23.29 | 0.01 | 23.29 | 23.29 | 0.01 | 23.29 | 0.07 |
| # 2 $(n = 5, \lvert\mathcal{I}_k\rvert = 5, K = 10)$ | 233.94 | 0.10 | 233.94 | 233.94 | 0.01 | 233.94 | 0.08 |
| # 3 $(n = 20, \lvert\mathcal{I}_k\rvert = 10, K = 10)$ | 1,081.60 | 7.18 | 1,055.20 | 1,169.40 | 0.20 | 1,053.10 | 0.62 |
| # 4 $(n = 30, \lvert\mathcal{I}_k\rvert = 20, K = 20)$ | 4,105.10 | 1,000.00* | 3,970.80 | 4,499.10 | 0.72 | 3,976.00 | 1.71 |
| # 5 $(n = 100, \lvert\mathcal{I}_k\rvert = 40, K = 30)$ | 22,224.00 | 1,000.00* | 21,129.90 | 27,939.00 | 16.69 | 21,751.00 | 4.46 |
| # 6 $(n = 200, \lvert\mathcal{I}_k\rvert = 50, K = 50)$ | 54,213.00 | 1,000.00* | 54,213.00 | 83,466.00 | 146.22 | 61,794.00 | 4.46 |
| # 7 $(n = 10, \lvert\mathcal{I}_k\rvert = 5, K = 2)$ | 113.71 | 0.08 | 113.71 | 113.71 | 0.01 | 113.71 | 0.15 |
| # 8 $(n = 10, \lvert\mathcal{I}_k\rvert = 50, K = 50)$ | 479.50 | 1,000.00* | 479.50 | 3,052.90 | 1.37 | 3,052.90 | 0.27 |
| # 9 $(n = 30, \lvert\mathcal{I}_k\rvert = 50, K = 50)$ | 1,877.30 | 1,000.00* | 539.10 | 2,894.60 | 42.38 | 2,783.50 | 1.27 |
| # 10 $(n = 50, \lvert\mathcal{I}_k\rvert = 60, K = 60)$ | 1,673.70 | 1,000.00* | 1,673.70 | 3,323.50 | 356.23 | 2,997.40 | 0.57 |
| # 11 $(n = 20, \lvert\mathcal{I}_k\rvert = 10, K = 10)$ | 3,002.40 | 93.85 | 2,401.10 | 3,032.00 | 0.47 | 3,002.40 | 0.55 |
| # 12 $(n = 20, \lvert\mathcal{I}_k\rvert = 50, K = 10)$ | 3,114.20 | 1,000.00* | 96.80 | 3,452.20 | 0.43 | 3,349.00 | 0.39 |
| # 13$(n = 20, \lvert\mathcal{I}_k\rvert = 100, K = 50)$ | 1,091.80 | 1,000.00* | 489.60 | 17,568.00 | 10.06 | 17,062.00 | 1.13 |

Table 41: *Comparison of our approximation method with GUROBI for multiple linear constrained sum-of-max-linear-terms maximization. The column descriptions are the same as the previous tables. The size of the problem is defined by n (number of variables), $\mathcal{I}_k$ (number of elements in the set of each max-term), and K (number of max-terms). GUROBI solves each problem's mixed integer linear optimization reformulation, hence for large sized problems this can take a vast amount of time. Therefore we give 1,000 seconds time limits for each run of GUROBI. The 'value' column of GUROBI shows the value computed by GUROBI within 1,000 seconds, thus if the time is 1,000*, it means GUROBI cannot compute the global optimum solution within the time limit and returns the best solution found. The 'restricted' column gives the best value GUROBI can compute within a limitation of the time that it takes for our method to find the upper and lower bound values.*

## 5.5 Conclusions

Maximizing a convex function over convex constraints is known to be a hard problem even in its simplest cases. One can either try to solve the problem globally, or aim to obtain a good lower bound solution. We show that, as the size of the convex maximization problem gets larger, the local optimization solvers return solutions with large global optimality gaps, and in many cases the global optimization solvers cannot terminate.

In this work, we show how to use adjustable robust optimization techniques to solve the convex maximization problem. The main idea is to transfer the difficulty of the main non-convex problem to the nonlinearity of an equivalent (convex) ARO problem. Exploiting the rich ARO literature gives us strong methods to tightly approximate the convex maximization problem efficiently. More specifically, we show convex maximization problems whose ARO reformulations can be simplified by eliminating the adjustable variables, or problems where the adjustable variables can be restricted by using decision rules. The subsequent applications of robust optimization techniques approximate the resulting problem. Furthermore, we explicitly derive tractable upper and lower bound approximation problems of some well-known convex maximization problems. By using similar techniques, we also show a class of convex maximization problems which can be reformulated as mixed-integer convex optimization problems; although the latter is still NP-hard, the existence of powerful solvers makes these reformulations more practical to solve.

Since we approximate the equivalent ARO problem of the convex maximization problem, the gap between the upper and lower bounds we propose can be explained by the ARO theory. For instance, in the problems where linear decision rules are used to restrict the adjustable variables, the gap of the approximations can be explained by the performance of the linear decision rules. This theoretical bridge allows us to guarantee global optimality of the proposed approximation in some (easy) convex maximization problems whose ARO reformulations admit linear decision rules optimally.

The numerical experiments demonstrate the efficiency and strength of the proposed approximation. To be more specific, maximizing the log-sum-exp function over a ball constraint is approximated in less than half a second, where most of the problems are approximated without any gap. If the constraints are instead box constraints, the problem can be represented as a mixed-integer exponential cone optimization problem, and we can solve this problem to global optimality while the global optimization solvers fail to solve the original problem in most cases. We explicitly work with the maximization of convex quadratic, log-sum-exp, and sum-of-max-terms functions over linear constraints. Regarding the first case, we show that almost optimal lower bounds can be obtained very swiftly even for large problems where global optimization solvers cannot find a solution better than the starting point. For the second case, we find global lower and upper bounds in almost all problems, with the exceptions being approximated tightly, even for problems where the solvers face numerical difficulties due to the nature of geometric maximization. For the third case, as both of our upper and lower bound approximations are obtained via linear optimization, we find tight approximations swiftly, while the solvers are successful in small-sized problems but fail termination in large problems.

There are two directions for future work: application and theory. For the former one, the proposed approximation methods can be applied to one of the numerous real-life convex maximization problems. Applying these approximations in the convex maximization step of the convex-concave method can be useful for DC programming. Another implementation can be integrating our methodology with the global optimization solvers, where the solvers can first obtain a good lower bound by using our solution (along with the corresponding upper bound), and then iterate to find a globally optimal solution. As for theory, one can use piecewise-linear decision rules or nonlinear decision rules to restrict the adjustable variables in the proposed ARO reformulation. We do not give theoretical guarantees of the approximation gaps, hence developing such guarantees is essential. Furthermore, one can develop an algorithm by dividing the feasible region of the convex maximization problem into parts and solve our approximations in each part, which may give tighter bounds.

## 5.A RLT Relaxation of Problem (79)

Suppose we have a symmetric matrix variable $V \in \mathbb{S}^{m\times m}$ such that $V = ww^\top$. The following can be applied to replace the convex term $||A^\top w||_2$:

$$||A^\top w||_2 = \sqrt{w^\top AA^\top w} = \sqrt{\text{tr}\,(w^\top AA^\top w)} = \sqrt{\text{tr}\,(A^\top ww^\top A)}.$$

Therefore, we use the following concave reformulation of $||A^\top w||_2$:

$$||A^\top w||_2 = \sqrt{\text{tr}\,(A^\top VA)}$$

This concave reformulation is of course based on the assumption $V = ww^\top$. Moreover, we use the main idea of RLT and multiply each of the original constraints $\alpha_i^\top w - \beta_i \leq 0,\ \alpha_j^\top w - \beta_j \leq 0$ to obtain:

$$\begin{aligned} & (\alpha_i^\top w - \beta_i)(\alpha_j^\top w - \beta_j) \geq 0 \\ \iff & \alpha_i^\top ww^\top \alpha_j - (\beta_i\alpha_j + \beta_j\alpha_i)^\top w + \beta_i\beta_j \geq 0 \qquad (100) \\ \iff & \alpha_i^\top V\alpha_j - (\beta_i\alpha_j + \beta_j\alpha_i)^\top w + \beta_i\beta_j \geq 0. \qquad (101) \end{aligned}$$

Although $V = ww^\top$ is assumed, it is a non-convex constraint, so we relax it as:

$$V \succeq ww^\top \iff \begin{pmatrix} V & w \\ w^\top & 1 \end{pmatrix} \succeq 0. \qquad (102)$$

Thus, problem (79) is relaxed by the following convex optimization problem:

$$\begin{array}{lll}\sup\limits_{V\in\mathbb{S}^{m\times m},\ w\in\mathbb{R}^m} & \rho\sqrt{\operatorname{tr}(A^\top VA)}+a^\top A^\top w+b^\top w-f^*(w) & \\ \text{s.t.} & \alpha_i^\top w-\beta_i\leq 0, & i=1,\dots,d \\ & \alpha_i^\top V\alpha_j-(\beta_i\alpha_j+\beta_j\alpha_i)^\top w+\beta_i\beta_j\geq 0, & i\leq j=1,\dots,d \\ & \begin{pmatrix} V & w \\ w^\top & 1\end{pmatrix}\succeq 0, & \end{array} \tag{103}$$

which concludes the proof.

### 5.B Proof of Theorem 21

We showed that problem (91) can be represented as ARO problem (93). As shown by Roos et al. (2018), we can lift the nonlinear term $f^*(w)$ to the uncertainty set by introducing an auxiliary uncertain parameter $w_0$. Hence, the set of constraints of the ARO problem is equivalent to:

$$\forall\begin{pmatrix} w_0 \\ w\end{pmatrix}\in W,\ \exists\lambda\in\mathbb{R}^q:\begin{cases} d^\top\lambda+b^\top w+w_0\leq\tau & \text{(104a)} \\ D^\top\lambda\geq A^\top w & \text{(104b)} \\ \lambda\geq\mathbf{0}, & \text{(104c)}\end{cases}$$

where we define the new uncertainty set as

$$W=\left\{\begin{pmatrix} w_0 \\ w\end{pmatrix}\in\mathbb{R}^{m+1}:\ w_0+f^*(w)\leq 0\right\}. \tag{105}$$

A safe approximation of the constraint set is obtained by using a linear decision rule for the adjustable variable:

$$\lambda=u+Vw+rw_0,$$

where $u\in\mathbb{R}^q, V\in\mathbb{R}^{q\times m}$ and $r\in\mathbb{R}^q$. Substituting this LDR in (104a) leads to

$$\begin{aligned} & d^\top\lambda+b^\top w+w_0\leq\tau && \forall\begin{pmatrix} w_0 & w^\top\end{pmatrix}^\top\in W \\ \iff & d^\top(u+Vw+rw_0)+b^\top w+w_0\leq\tau && \forall\begin{pmatrix} w_0 & w^\top\end{pmatrix}^\top\in W\end{aligned}$$

$$\iff d^\top u + \begin{pmatrix} w_0 \\ w \end{pmatrix}^\top \begin{pmatrix} 1 + d^\top r \\ V^\top d + b \end{pmatrix} \leq \tau \qquad\qquad \forall \begin{pmatrix} w_0 & w^\top \end{pmatrix}^\top \in W$$

$$\iff d^\top u + \delta^*\left( \begin{pmatrix} 1 + d^T r \\ V^T d + b \end{pmatrix} \middle| W \right) \leq \tau. \tag{106}$$

To be able to find the tractable robust counterpart of (106), we derive the support function of the new uncertainty set $W$, which is

$$\begin{aligned} \delta^*\left( \begin{pmatrix} z_0 \\ z \end{pmatrix} \middle| W \right) &= \sup_{(w_0, w)^\top \in W} \{ z_0 w_0 + z^\top w \} \\ &= \begin{cases} \sup_{w \in \mathbb{R}^m} \{ z^\top w - z_0 f^*(w) \} & \text{if } z_0 > 0 \\ \sup_{w \in \operatorname{dom} f^*} \{ z^\top w \} & \text{if } z_0 = 0 \\ +\infty & \text{otherwise} \end{cases} \\ &= \begin{cases} z_0 f\left(\frac{z}{z_0}\right) & \text{if } z_0 \geq 0 \\ +\infty & \text{otherwise.} \end{cases} \end{aligned} \tag{107}$$

Above, for the case of $z_0 = 0$, we use the property $\sup_{w \in \operatorname{dom} f^*} \{ z^\top w \} = \delta^*(z | \operatorname{dom} f^*) = \lim_{z_0 \downarrow 0} z_0 f\left(\frac{z}{z_0}\right)$, and the result follows since for $z_0 = 0$ we have the understanding of $\lim_{z_0 \downarrow 0} z_0 f\left(\frac{z}{z_0}\right)$ for the perspective (Rockafellar 1997). By substituting (107) into (106) we obtain:

$$d^\top u + \delta^*\left( \begin{pmatrix} 1 + d^T r \\ V^T d + b \end{pmatrix} \middle| W \right) \leq \tau$$

$$\iff \begin{cases} d^\top u + (1 + d^\top r) f\left( \dfrac{V^\top d + b}{1 + d^\top r} \right) \leq \tau \\ 1 + d^\top r \geq 0. \end{cases}$$

Hence, by using LDRs (104a) becomes exactly (95a).

Following the same steps for (104b) yields us to (95b):

$$
\begin{aligned}
& D^\top \lambda \geq A^\top w && \forall \begin{pmatrix} w_0 & w^\top \end{pmatrix}^\top \in W \\
\iff \quad & D_i^\top \lambda \geq A_i^\top w && \forall \begin{pmatrix} w_0 & w^\top \end{pmatrix}^\top \in W,\ i = 1, \ldots, n \\
\iff \quad & \begin{cases} -D_i^\top u + (-D_i^\top r) f \left( \dfrac{A_i - V^\top D_i}{-D_i^\top r} \right) \leq 0 \\ -D_i^\top r \geq 0 \end{cases} && i = 1, \ldots, n.
\end{aligned}
$$

Similarly (104c) becomes (95c):

$$
\begin{aligned}
& \lambda \geq 0 \\
\iff \quad & -u_i - V_{(i)} w - r_i w_0 \leq 0 && \forall \begin{pmatrix} w_0 & w^\top \end{pmatrix}^\top \in W,\ i = 1, \ldots, q \\
\iff \quad & \begin{cases} -u_i + (-r_i) f \left( \frac{-V_{(i)}^\top}{-r_i} \right) \\ -r_i \geq 0 \end{cases} && i = 1, \ldots, q.
\end{aligned}
$$

As we use an LDR for the adjustable variable, the optimal objective value of (95) is an upper bound to (91).

## 5.C Upper and Lower Bound Approximation of Problem (91)

We summarize the process of finding upper and lower bounds on the global optimum objective value of problem (91).

**Algorithm 12:** *Obtaining upper and lower bounds for problem* (91)

**input :** $f$, $A$, $b$, $U$

**output:** Upper bound value $\hat{\tau}$, lower bound solution $x^*$ with value $f(Ax^* + b)$

1. Obtain the upper bound solution by solving (95), i.e., $(\hat{u}, \hat{V}, \hat{r}, \hat{\tau}) \in$

$$\operatorname*{arg\,inf}_{u\in\mathbb{R}^q, V\in\mathbb{R}^{q\times m}, r\in\mathbb{R}^q, \tau\in\mathbb{R}} \tau \ \text{ s.t.} \begin{cases} d^\top u + \left(1 + d^{\mathsf{T}} r\right) f\left(\dfrac{V^\top d + b}{1 + d^\top r}\right) \le \tau & \\ 1 + d^{\mathsf{T}} r \ge 0 & \\ -D_i^{\mathsf{T}} u + \left(-D_i^\top r\right) f\left(\dfrac{A_i - V^\top D_i}{-D_i^\top r}\right) \le 0 & i = 1, \dots, n \\ -D_i^{\mathsf{T}} r \ge 0 & \\ -u_i + (-r_i) f\left(\dfrac{V_{(i)}^\top}{r_i}\right) \le 0 & i = 1, \dots, q \\ -r_i \ge 0, & \end{cases}$$

   or alternatively via the adversarial approach. $\hat{\tau}$ is an upper bound value.

2. Generate a finite set of (potential) worst-case ARO scenarios by plugging the optimal LDR back in the safe approximation of the original ARO, and by collecting worst-case scenario of each constraint, i.e., $\overline{W} = \overline{W}^1 \cup [\bigcup_{i=1}^n \overline{W}_i^2] \cup [\bigcup_{i=1}^q \overline{W}_i^3]$ with:

$$\begin{aligned} &\overline{W}^1 \in \operatorname*{arg\,sup}_{w\in\operatorname{dom} f^*} \left\{-(1 + d^\top r) f^*(w) + (d^\top \hat{V} + b^\top) w + d^\top \hat{u} - \hat{\tau}\right\}, && \\ &\overline{W}_i^2 \in \operatorname*{arg\,sup}_{w\in\operatorname{dom} f^*} \left\{(D_i^\top r) f^*(w) + (A_i^\top - D_i^\top \hat{V}) w - D_i^\top \hat{u}\right\}, && i = 1, \dots, n \\ &\overline{W}_i^3 \in \operatorname*{arg\,sup}_{w\in\operatorname{dom} f^*} \left\{-\hat{u}_i - \hat{V}_i w + \hat{r}_i f^*(w)\right\}, && i = 1, \dots, q. \end{aligned}$$

3. For all $w \in \overline{W}$, solve the linear optimization problem:

$$x \in \operatorname*{arg\,sup}_{x\in U} \{(A^\top w)^\top x\},$$

   and return $x$ that achieves the highest $f(Ax + b)$ as the lower bound solution of problem (91).

## 5.D Complete Derivation of Specific Problems in Section 5.3.2

### 5.D.1 Quadratic Optimization

Here we consider problem (91) when the objective function is a convex quadratic function. For the problem of maximizing a convex quadratic function over a polyhedron, we can find an upper bound by solving a second-order cone optimization problem, and we can find a lower bound by solving a linear optimization problem.

Consider the convex quadratic function $g : \mathbb{R}^n \mapsto \mathbb{R}$ defined by:

$$g(x) = x^\top Qx + \ell^\top x,$$

where $\ell \in \mathbb{R}^n$ and $Q$ is a symmetric positive semi-definite (psd) matrix. Maximizing this function over a polyhedral set can be written as the robust optimization problem:

$$\begin{aligned} \inf \quad & \tau \\ \text{s.t.} \quad & x^\top Qx + \ell^\top x \leq \tau, \qquad \forall x \in U, \end{aligned} \tag{109}$$

where $U = \{x \in \mathbb{R}^n_+ : \ Dx \leq d\}$ for $D \in \mathbb{R}^{q \times n}, \ d \in \mathbb{R}^q$. We use the conic representation of the constraints of problem (109):

$$\left\| \begin{pmatrix} (1 + \ell^\top x - \tau)/2 \\ Lx \end{pmatrix} \right\|_2 - (1 - \ell^\top x + \tau)/2 \leq 0,$$

where $L$ is the psd decomposition $Q = L^\top L$. Therefore, the constraint of problem (109) can be written as a robust conic constraint. Define $f : \mathbb{R}^{m+1} \times \mathbb{R} \mapsto \mathbb{R}$ by:

$$f \begin{pmatrix} z \\ \tilde{z} \end{pmatrix} = \left\| z \right\|_2 + \tilde{z}, \tag{110}$$

with $z \in \mathbb{R}^{m+1}$ and $\tilde{z} \in \mathbb{R}$. It can be verified that $f$ is positively homogeneous and that the conjugate of this function for $w \in \mathbb{R}^{m+1}, \ \tilde{w} \in \mathbb{R}$ is:

$$f^* \begin{pmatrix} w \\ \tilde{w} \end{pmatrix} = \begin{cases} 0 & \text{if } \tilde{w} = 1 \text{ and } ||w||_2 \leq 1 \\ +\infty & \text{otherwise.} \end{cases}$$

Defining

$$A = \begin{bmatrix} L \\ \ell^\top/2 \\ \ell^\top/2 \end{bmatrix}, \quad b = \begin{bmatrix} \mathbf{0} \\ (1-\tau)/2 \\ (-1-\tau)/2 \end{bmatrix}, \tag{111}$$

it follows that:

$$f\left(Ax+b\right) = f\left(\begin{pmatrix} Lx \\ (1+\ell^\top x - \tau)/2 \\ (\ell^\top x - 1 - \tau)/2 \end{pmatrix}\right) = \left\| \begin{pmatrix} Lx \\ (1+\ell^\top x - \tau)/2 \end{pmatrix} \right\|_2 - (1 - \ell^\top x + \tau)/2.$$

Hence, the constraint of problem (109) is equivalent to $f\left(Ax+b\right) \le 0$. Therefore, problem (109) can be rewritten as:

$$\begin{aligned} \inf \quad & \tau \\ \text{s. t.} \quad & f\left(\begin{pmatrix} Lx \\ (1+\ell^\top x - \tau)/2 \\ (\ell^\top x - 1 - \tau)/2 \end{pmatrix}\right) \le 0, \quad \forall x \in U. \end{aligned} \tag{112}$$

An upper bound of this problem can now be obtained by applying Theorem 21 and exploiting the positive homogeneity of $f$ (see Appendix 5.E). The upper bound is the optimal value of the problem:

$$\begin{aligned} \inf \quad & \tau \\ \text{s. t.} \quad & d^\top u + \bar{v}^\top d - (1+\tau)/2 + \left\| \begin{pmatrix} \tilde{V}^\top d \\ \hat{v}^\top d + (1-\tau)/2 \end{pmatrix} \right\|_2 \le 0 \\ & -D_i^\top u + \frac{\ell_i}{2} - \bar{v}^\top D_i + \left\| \begin{pmatrix} L_i - \tilde{V}^\top D_i \\ \ell_i/2 - \hat{v}^\top D_i \end{pmatrix} \right\|_2 \le 0, \qquad i = 1, \ldots, n \\ & -u_i - \bar{v}_i + \left\| \begin{pmatrix} -\tilde{V}_{(i)}^\top \\ -\hat{v}_i \end{pmatrix} \right\|_2 \le 0, \qquad i = 1, \ldots, q, \end{aligned} \tag{113}$$

in which the variables are $\tau \in \mathbb{R}, u \in \mathbb{R}^q, \bar{v} \in \mathbb{R}^q, \hat{v} \in \mathbb{R}^q, \tilde{V} \in \mathbb{R}^{q \times m}$.

In order to compute a lower bound, we use the optimal solution $(\tau, u, \bar{v}, \hat{v}, \tilde{V})$ to problem (113) by obtaining a collection of worst case scenarios $\overline{W}$ from (97), (98), and (99). These problems can be solved analytically as explained in Appendix 5.F. This yields the scenarios:

$$
\begin{aligned}
\overline{W}^1 &= \begin{bmatrix} h\left(\begin{pmatrix} \tilde{V}^\top d \\ d^\top \hat{v} + (1-\tau)/2 \end{pmatrix}\right) \\ 1 \end{bmatrix} & & \\
\overline{W}_i^2 &= \begin{bmatrix} h\left(\begin{pmatrix} L_i - \tilde{V}^\top D_i \\ \ell_i/2 - D_i^\top \hat{v} \end{pmatrix}\right) \\ 1 \end{bmatrix} & & i = 1, \ldots, n \\
\overline{W}_i^3 &= \begin{bmatrix} h\left(\begin{pmatrix} -\tilde{V}_{(i)}^\top \\ -\hat{v}_i \end{pmatrix}\right) \\ 1 \end{bmatrix} & & i = 1, \ldots, q,
\end{aligned}
\tag{114}
$$

where $h(a) = a/||a||_2$ normalizes its input. Using these worst-case scenarios, the candidate solutions $\bar{x}^{(j)}$ are obtained by solving (96), and we can substitute them in the main objective function as $f(A\bar{x}^{(j)} + b)$ to find the best lower bound.

### 5.D.2 Geometric Optimization

Geometric Optimization (GO) is a class of optimization problems originally introduced by Duffin (1967). A practical tutorial can be found in the work of Boyd et al. (2007). Even though it can have many representations, we focus on the GO variant where the objective is maximizing the convex log-sum-exp objective. The log-sum-exp function $f : \mathbb{R}^m \mapsto \mathbb{R}$ is defined as

$$
f(z) = \log\left(\sum_{i=1}^{m} \exp(z_i)\right), \tag{115}
$$

and we are interested in solving problems of the following type

$$
\begin{aligned}
\max_{x} \quad & f(Ax+b) = \log\left(\textstyle\sum_{i=1}^{m} \exp(A_{(i)}x + b_i)\right) \\
\text{s.t.} \quad & x \in U,
\end{aligned}
\tag{116}
$$

where $U = \{x \in \mathbb{R}_+^n \ : \ Dx \leq d\}$. This problem may appear in robust geometric optimization problems. If one applies the adversarial approach for such problems, in the step of adding worst-case uncertainty realization to the discrete uncertainty set, one will need to maximize the convex geometric function.

The conjugate $f^* : \mathbb{R}^m \mapsto \mathbb{R}$ of the log-sum-exp function (115) is

$$f^*(w) = \begin{cases} \sum_{i=1}^m w_i \log(w_i) & \text{if } w \in \mathbb{R}^m_+ \text{ and } \sum_{i=1}^m w_i = 1 \\ \infty & \text{otherwise.} \end{cases} \tag{117}$$

We observe that $f^*(w)$ is the negative-entropy function of $w$ on its domain, which is a standard $m$-dimensional simplex. It is well known that the negative entropy is a strictly convex function. Next we show that the upper bound and lower bound approximation problems (of problem (116)) are exponential cone representable. This allows one to use the power of today's conic programming solvers, e.g., Mosek's exponential cone optimization solver. We start by introducing the exponential cone, which is the following convex subset of $\mathbb{R}^3$:

$$\mathcal{K}_{\text{exp}} = \{(x_1, x_2, x_3) : x_1 \geq x_2 \exp(x_3/x_2), x_2 > 0\} \cup \{(x_1, 0, x_3) : x_1 \geq 0, x_3 < 0\}.$$

So, the exponential cone is the closure of the set of points which satisfy $x_1 \geq x_2 \exp(x_3/x_2)$, $x_1, x_2 > 0$. In the next corollary we show that the upper bound problem (95) is exponential cone representable.

**Corollary 13** (Upper Bound Approximation)**.** *Upper bound problem* (95) *is exponential cone representable with the following problem with variables* $r \in \mathbb{R}^q, u \in \mathbb{R}^q, V \in \mathbb{R}^{q\times m}, \tau \in \mathbb{R}, z^{(1)} \in \mathbb{R}^m, z^{(2)} \in \mathbb{R}^{n\times m}, z^{(3)} \in \mathbb{R}^{q\times m}$*:*

$$\inf_\tau \ \ s.t. \begin{cases} 1 + d^\top r \geq \sum_{j=1}^m z_j^{(1)}, \\ (\, z_j^{(1)},\ 1 + d^\top r,\ V_j^\top d + b_j - \tau + d^\top u\,) \in \mathcal{K}_{\text{exp}}, \quad j = 1, \ldots, m & (118a) \\ -D_i^\top r \geq \sum_{j=1}^m z_{ij}^{(2)}, \\ (\, z_{ij}^{(2)},\ -D_i^\top r,\ A_{i,(j)} - V_j^\top D_i - D_i^\top u\,) \in \mathcal{K}_{\text{exp}}, \quad j = 1, \ldots, m,\ i = 1, \ldots, n & (118b) \\ -r_i \geq \sum_{j=1}^m z_{ij}^{(3)}, \\ (\, z_{ij}^{(3)},\ -r_i,\ -V_{j,(i)} - u_i\,) \in \mathcal{K}_{\text{exp}}, \quad j = 1, \ldots, m,\ i = 1, \ldots, q & (118c) \end{cases}$$

■

*Proof.* Roos et al. (2018) show that if a function is conically representable, so is its perspective in the same cone. Log-sum-exp is an exponential cone representable function (MOSEK ApS 2023), and we show how to represent a convex inequality system of its perspective with exponential cones. Consider the following set of constraints:

$$\begin{cases} t \geq x_0 \log(\exp(x_1/x_0) + \ldots + \exp(x_m/x_0)) \\ x_0 > 0. \end{cases} \tag{119}$$

By using the proof in (Roos et al. 2018), we can write the following equivalent constraint set:

$$\begin{cases} x_0 \geq \sum_{j=1}^{m} z_j \\ (z_j, x_0, (x_j - t)) \in \mathcal{K}_{\text{exp}}, \quad j = 1, \ldots, m. \end{cases} \tag{120}$$

Since constraints of type (119) appear in the upper bound approximation problem (95), we can use the equivalent representation (120) in each of the constraints to obtain problem (118). □

The lower bound can also be obtained by solving exponential cone programs. The worst-case scenario collection is obtained by solving (97), (98), (99). Here $(r, u, V, r, \tau)$ are all parameters taken from the solution of the upper bound problem. This set of problems can be formulated as exponential conic problems, which is shown in the next corollary.

**Corollary 14** (Lower Bound Scenarios). *The problems* (97), (98), *and* (99) *can be written as the following exponential cone problems:*

$$\begin{array}{llll}
(97): & \underset{w,t}{\arg\sup} & \left\{(1 + d^\top r)(\sum_{j=1}^{m} t_j) + (d^\top V + b^\top)w + d^\top u - \tau\right\} & \\
& \text{s. t.} & (1, w_j, t_j) \in \mathcal{K}_{\text{exp}}, \quad j = 1, \ldots, m & \\
& & \sum_{j=1}^{m} w_j = 1, & \\
(98): & \underset{w,t}{\arg\sup} & \left\{(-D_i^\top r)(\sum_{j=1}^{m} t_j) + (A_i^\top - D_i^\top V)w - D_i^\top u\right\} & \\
& \text{s. t.} & (1, w_j, t_j) \in \mathcal{K}_{\text{exp}}, \quad j = 1, \ldots, m & i = 1, \ldots, n \\
& & \sum_{j=1}^{m} w_j = 1, & \\
(99): & \underset{w,t}{\arg\sup} & \left\{(-r_i)(\sum_{j=1}^{m} t_j) + (-V_{(i)})w - u_i\right\} & \\
& \text{s. t.} & (1, w_j, t_j) \in \mathcal{K}_{\text{exp}}, \quad j = 1, \ldots, m & i = 1, \ldots, q \\
& & \sum_{j=1}^{m} w_j = 1. &
\end{array}$$

*Proof.* Serrano (2015) shows that the negative entropy function is exponential conically representable, and the following problems are equivalent (here $w \in \mathbb{R}^m$):

$$\begin{array}{lll} \sup_w & \{c_0(-\sum_{i=1}^m w_i \log(w_i))\} & \\ \text{s.t.} & w_i \geq 0, \qquad i = 1, \ldots, m. & \end{array} = \begin{array}{lll} \sup_{w,t} & \{c_0 \sum_{i=1}^m t_i\} & \\ \text{s.t.} & (1, w_i, t_i) \in \mathcal{K}_{\exp}, \quad i = 1, \ldots, m. & \end{array} \tag{121}$$

The result now follows by substitution of the conjugate (117) in (97), (98), and (99), respectively and then applying equivalence (121). □

### 5.D.3 Sum-of-Max-Linear-Terms Optimization

Formally, the sum-of-max-terms function $f : \mathbb{R}^m \mapsto \mathbb{R}$ is written as

$$f(z) = \sum_{k=1}^{K} \max_{j \in \mathcal{I}_k} \{z_j\}, \tag{122}$$

where the set $\mathcal{I}_k \subseteq \{1, \ldots, m\}$ for each $k \in \{1, \ldots, K\}$. Moreover, we can assume $\mathcal{I}_k \cap \mathcal{I}_\ell = \emptyset$ for any $k \neq \ell$ and $\cup_{k=1}^K = \{1, \ldots, m\}$ without loss of generality, since otherwise we can add components to $z$ to make this statement hold. The sum-of-max-linear-terms function we cover at this section is represented as

$$f(Ax + b) = \sum_{k=1}^{K} \max_{j \in \mathcal{I}_k} \{A_{(j)}x + b_j\},$$

which is a convex and positively homogeneous function. The main convex maximization problem we are interested in is maximizing $f(Ax + b)$ over $U = \{x \in \mathbb{R}^n_+ : \ Dx \leq d\}$, formally:

$$\begin{array}{ll} \max\limits_{x} & \sum_{k=1}^{K} \max\limits_{j \in \mathcal{I}_k} \{A_{(j)}x + b_j\} \\ \text{s.t.} & x \in U. \end{array} \tag{123}$$

This problem naturally arises when one applies the adversarial approach to robust optimization problems with uncertain sum-of-max-linear-terms constraints.

The conjugate of sum-of-max-linear-terms (122) is given by Roos et al. (2018) as:

$$f^*(w) = \begin{cases} 0 & \text{if } w_i \geq 0 \ \forall i = 1, \ldots, m, \ \sum_{j \in \mathcal{I}_k} w_j = 1 \\ \infty & \text{otherwise.} \end{cases}$$

The formulation of the upper bound approximation for maximizing sum-of-max-linear-terms function over a polyhedron can be greatly simplified. This is due to the fact that sum-of-max-linear-terms function is a positively homogeneous function as well as the trick of introducing auxiliary variables which give us a linear optimization problem in return.

Since the function is a positively homogeneous function, we can write the upper bound approximation problem (95) as:

$$\begin{array}{lll} \inf\limits_{u\in\mathbb{R}^q, V\in\mathbb{R}^{q\times m}, \tau\in\mathbb{R}} & \tau & \\ \text{s. t.} & d^\top u + \sum_{k=1}^K \max\limits_{j\in\mathcal{I}_k}\{V_j^\top d + b_j\} & \leq \tau \\ & -D_i^\top u + \sum_{k=1}^K \max\limits_{j\in\mathcal{I}_k}\{A_{i,(j)} - V_j^\top D_i\} & \leq 0 \quad i=1,\ldots,n \\ & -u_i - \sum_{k=1}^K \max\limits_{j\in\mathcal{I}_k}\{V_{j,(i)}\} & \leq 0 \quad i=1,\ldots,q. \end{array} \tag{124}$$

Problem (124) can be reformulated as a linear optimization problem by using auxiliary variables. The worst-case scenarios for computing the lower bound are obtained from problems (97), (98), and (99), which simplify to:

$$(97): \operatorname*{arg\,sup}_{w\in\operatorname{dom} f^*} \left\{d^\top(u+Vw) + b^\top w - \tau\right\}, \tag{125}$$

$$(98): \operatorname*{arg\,sup}_{w\in\operatorname{dom} f^*} \left\{A_i^\top w - D_i^\top(u+Vw)\right\}, \qquad i=1,\ldots,n \tag{126}$$

$$(99): \operatorname*{arg\,sup}_{w\in\operatorname{dom} f^*} \left\{-u_i - V_{(i)}w\right\}, \qquad i=1,\ldots,q, \tag{127}$$

where we do not have the conjugate terms since $f$ is a homogeneous function so its conjugate takes value 0. Therefore, the worst-case scenarios of each constraint are given by the following problems:

- For (125):

$$\begin{array}{lll} \sup_w & d^\top(u+Vw) + b^\top w - \tau & \\ \text{s. t.} & w_j \geq 0, & j=1,\ldots,m \\ & \sum_{j\in\mathcal{I}_k} w_j = 1, & k=1,\ldots,K. \end{array} \tag{128}$$

  Recalling the only variable here is $w$, this is a linear optimization problem. Moreover, since we have $\mathcal{I}_k \cap \mathcal{I}_{k'} = \emptyset$ for $k \neq k'$, we can separate this problem to $K$ independent optimization problems, where each problem $k$ is:

$$\begin{aligned} c_k = \sup_{w\geq 0} \quad & d^\top (u + V_{\{j\}} y) + b_{\{j\}}^\top y \\ \text{s. t.} \quad & \textstyle\sum_{i=1}^{|\mathcal{I}_k|} y_i = 1. \end{aligned} \tag{129}$$

Here $y \in \mathbb{R}^{|\mathcal{I}_k|}$ is the $w$ components corresponding to the $k$-th term in the sum-of-max-linear-terms function definition. Similarly, $V_{\{j\}}$, $b_{\{j\}} \in \mathbb{R}^{|\mathcal{I}_k|}$ are the components of $V, b$ corresponding to the $k$-th term. Notice that problem (129) is a linear optimization problem over a simplex. The optimal value will have $y_i = 1$ for some $i$ and $y_{i'} = 0$ for all $i' \neq i$. Therefore, the solution is

$$c_k = \max_{i=1,\dots,|\mathcal{I}_k|} \{d^\top (u + V_{\{j\},i})\} + b_{\{j\},i},$$

where $V_{\{j\},i}, b_{\{j\},i}$ represent the $i$-th columns of $V_{\{j\}}$ and $b_{\{j\}}$, respectively. Hence, the optimal value of (128) is given by $-\tau + \sum_{k=1}^K c_k$. The arg max value can be retrieved easily by detecting which $y_i$ variables took value 1; there will be exactly $K$ ones in the result and the rest will be zeros.

- For (126), for all $i = 1, \dots, n$:

$$\begin{aligned} \sup\nolimits_w \quad & A_i^\top w - D_i^\top (u + Vw) & \\ \text{s. t.} \quad & w_j \geq 0, & j = 1, \dots, m \\ & \textstyle\sum_{j \in \mathcal{I}_k} w_j = 1, & k = 1, \dots, K. \end{aligned}$$

Similarly, this problem can be separated to $K$ independent linear optimization problems over simplices. The optimal solution can be found analytically.

- For (127), for all $i = 1, \dots, q$:

$$\begin{aligned} \sup\nolimits_w \quad & -u_i - V_{(i)} w & \\ \text{s. t.} \quad & w_j \geq 0, & j = 1, \dots, m \\ & \textstyle\sum_{j \in \mathcal{I}_k} w_j = 1, & k = 1, \dots, K. \end{aligned}$$

This problem can be solved analytically once again, concluding that all of the worst-case scenario finding procedure can be solved analytically.

### 5.E Upper Bound Approximation of Quadratic Maximization via SOCO

We follow Theorem 21 to apply the upper bound approximation for problem (112). Because $f$ is a positively homogeneous function, the upper bound problem (95) reduces to the following problem for the variables $u \in \mathbb{R}^q, V \in \mathbb{R}^{q\times(m+2)}, \tau \in \mathbb{R}$:

$$\begin{aligned}
\inf \quad & \tau \\
\text{s. t.} \quad & d^\top u + f\left(V^\top d + b\right) \le 0 \\
& -D_i^\top u + f\left(A_i - V^\top D_i\right) \le 0, \quad i = 1,\dots,n \\
& -u_i + f\left(-V_{(i)}^\top\right) \le 0, \qquad\quad\; i = 1,\dots,q.
\end{aligned} \tag{130}$$

Since $V$ has $m+2$ columns, we represent it as:

$$V = \begin{bmatrix} \tilde{V} & \hat{v} & \bar{v} \end{bmatrix} \quad \text{where} \quad \tilde{V} \in \mathbb{R}^{q\times m}, \hat{v} \in \mathbb{R}^q, \bar{v} \in \mathbb{R}^q.$$

Constraints of problem (130) can be simplified to respectively:

$$\begin{cases}
\quad d^\top u + f\left(V^\top d + b\right) \le 0 \\
= \quad d^\top u + f\left(\begin{pmatrix} \tilde{V}^\top d \\ \hat{v}^\top d + 1/2 - \tau/2 \\ \bar{v}^\top d - 1/2 - \tau/2 \end{pmatrix}\right) \le 0 \\
= \quad d^\top u + \bar{v}^\top d - \dfrac{1}{2} - \dfrac{\tau}{2} + \left\| \begin{pmatrix} \tilde{V}^\top d \\ \hat{v}^\top d + \dfrac{1}{2} - \dfrac{\tau}{2} \end{pmatrix} \right\|_2 \le 0
\end{cases}$$

$$\begin{cases}
\quad -D_i^\top u + f\left(A_i - V^\top D_i\right) \le 0 \\
= \quad -D_i^\top u + f\left(\begin{pmatrix} L_i - \tilde{V}^\top D_i \\ \ell_i/2 - \hat{v}^\top D_i \\ \ell_i/2 - \bar{v}^\top D_i \end{pmatrix}\right) \le 0 \\
= \quad -D_i^\top u + \dfrac{\ell_i}{2} - \bar{v}^\top D_i + \left\| \begin{pmatrix} L_i - \tilde{V}^\top D_i \\ \ell_i/2 - \hat{v}^\top D_i \end{pmatrix} \right\|_2 \le 0
\end{cases}$$

$$\begin{cases} & -u_i + f\left(-V_{(i)}^\top\right) \le 0 \\ = & -u_i + f\left(\begin{pmatrix} -\tilde{V}_{(i)}^\top \\ -\hat{v}_i \\ -\bar{v}_i \end{pmatrix}\right) \le 0 \\ = & -u_i - \bar{v}_i + \left\| \begin{pmatrix} -\tilde{V}_{(i)}^\top \\ -\hat{v}_i \end{pmatrix} \right\|_2 \le 0. \end{cases}$$

Thus the upper bound approximation problem can be represented as the following second-order conic program:

$$\begin{aligned} \inf \quad & \tau \\ \text{s.t.} \quad & d^\top u + \bar{v}^\top d - (1+\tau)/2 + \left\| \begin{pmatrix} \tilde{V}^\top d \\ \hat{v}^\top d + (1-\tau)/2 \end{pmatrix} \right\|_2 \le 0 \\ & -D_i^\top u + \frac{\ell_i}{2} - \bar{v}^\top D_i + \left\| \begin{pmatrix} L_i - \tilde{V}^\top D_i \\ \ell_i/2 - \hat{v}^\top D_i \end{pmatrix} \right\|_2 \le 0, \qquad i = 1, \dots, n \\ & -u_i - \bar{v}_i + \left\| \begin{pmatrix} -\tilde{V}_{(i)}^\top \\ -\hat{v}_i \end{pmatrix} \right\|_2 \le 0, \qquad i = 1, \dots, q. \end{aligned} \tag{131}$$

## 5.F Lower Bound Scenarios for Quadratic Maximization

In the light of problems (97), (98), (99), the worst-case scenarios are collected by:

$$(97): \quad \underset{(w,\tilde{w}) \in \operatorname{dom} f^*}{\arg\sup} \left\{ -(1 + d^\top r) f^* \begin{pmatrix} w \\ \tilde{w} \end{pmatrix} + (d^\top V + b^\top) \begin{pmatrix} w \\ \tilde{w} \end{pmatrix} + d^\top u \right\}, \tag{132}$$

$$(98): \quad \underset{(w,\tilde{w}) \in \operatorname{dom} f^*}{\arg\sup} \left\{ (D_i^\top r) f^* \begin{pmatrix} w \\ \tilde{w} \end{pmatrix} + (A_i^\top - D_i^\top V) \begin{pmatrix} w \\ \tilde{w} \end{pmatrix} - D_i^\top u \right\}, \qquad i = 1, \dots, n \tag{133}$$

$$(99): \quad \underset{(w,\tilde{w}) \in \operatorname{dom} f^*}{\arg\sup} \left\{ -u_i - V_{(i)} \begin{pmatrix} w \\ \tilde{w} \end{pmatrix} + r_i f^* \begin{pmatrix} w \\ \tilde{w} \end{pmatrix} \right\}, \qquad i = 1, \dots, q. \tag{134}$$

We already showed the convex conjugate of $f$ takes value 0 in its domain. Recalling

$$A = \begin{bmatrix} L \\ \ell^\top/2 \\ \ell^\top/2 \end{bmatrix}, \quad b = \begin{bmatrix} \mathbf{0} \\ (1-\tau)/2 \\ (-1-\tau)/2 \end{bmatrix}, V = \begin{bmatrix} \tilde{V} & \hat{v} & \bar{v} \end{bmatrix},$$

we can rewrite problem (132) as:

$$\begin{aligned} \sup_{w\in\mathbb{R}^{m+1},\tilde{w}\in\mathbb{R}} \quad & \left(d^\top \begin{bmatrix}\tilde{V} & \hat{v} & \bar{v}\end{bmatrix} + \begin{bmatrix}\mathbf{0} & (1-\tau)/2 & (-1-\tau)/2\end{bmatrix}\right)\begin{pmatrix} w \\ \tilde{w}\end{pmatrix} + d^\top u \\ \text{s.t.} \quad & \tilde{w} = 1 \\ & ||w||_2 \leq 1. \end{aligned}$$

By using $\tilde{w} = 1$, we can eliminate $\tilde{w}$ from the problem. Moreover, $w$ only appears in a linear term, so we can change $||w||_2 \leq 1$ constraint to be $||w||_2 = 1$ instead, i.e., $w$ is a unit vector. So the problem becomes finding the value of:

$$\sup_{w:||w||_2=1} \left\{ \begin{bmatrix} \tilde{V}^\top d \\ \hat{v}^\top d + (1-\tau)/2 \end{bmatrix}^\top w \right\} + d^\top u + d^\top \bar{v} - (1+\tau)/2.$$

Hence we need to maximize a linear function over the unit ball, which can be solved analytically. This yields the objective value:

$$\left\| \begin{pmatrix} \bar{V}^\top d \\ d^\top \hat{v} + (1-\tau)/2 \end{pmatrix} \right\|_2 + d^\top u + d^\top \bar{v} - (1+\tau)/2, \tag{135}$$

and the maximizer is:

$$\overline{W}_1 = \begin{bmatrix} h\left(\begin{pmatrix} \bar{V}^\top d \\ d^\top \hat{v} + (1-\tau)/2 \end{pmatrix}\right) \\ 1 \end{bmatrix},$$

where $h(a) = a/||a||$ normalizes its input. Notice that the last element 1 comes since $\tilde{w} = 1$ is in the domain of convex conjugate. The worst-case of constraints (133) and (134) are obtained via similar calculations.

## 5.G Data of Problems in Numerical Experiments

Here we explain the construction of the test data. The exact problem data, solutions, and the codes are available for download on GitHub at `https://github.com/selvi-aras/convex-max`.

### 5.G.1 Experiments of Section 5.4.1

**Problem 1** The problem data is:

$$A = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ 1 & 0 & 0 & 0 & 1 \\ 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 & 1 \\ 0 & 1 & 0 & 0 & 0 \end{bmatrix}, \quad b = \begin{bmatrix} -1 \\ 0 \\ 1 \\ 1 \\ 1 \end{bmatrix}, \quad a = \begin{bmatrix} 0 \\ 2 \\ 1 \\ 1 \\ 0 \end{bmatrix}, \quad \rho = 3.$$

For the next problems we generate larger problems by uniform random sampling (denoted simply as $\sim$). Remember that $A \in \mathbb{R}^{m\times n}$, $b \in \mathbb{R}^m$, $a \in \mathbb{R}^n$. We let $A_{ij}$ denote the elements of $A$.

**Problem 2:** $m = 15,\ n = 20,\ A_{ij} \sim \{0,1\},\ b_i \sim [-5,5],\ a_j \sim [0,3], \rho = 8$

**Problem 3:** $m = 120,\ n = 100,\ A_{ij} \sim \{0,1\},\ b_i \sim [-5,5],\ a_j \sim [0,4], \rho = 14$

**Problem 4:** $m = 40,\ n = 20,\ A_{ij} \sim \{-4,-3,\ldots,3,4\},\ b_i \sim [-5,5],\ a_j \sim [0,4], \rho = 10$

**Problem 5:** $m = 100,\ n = 50,\ A_{ij} \sim [-5,5],\ b_i \sim [-2,2],\ a_j \sim [-4,4], \rho = 12$

**Problem 6:** $m = 100,\ n = 100,\ A_{ij} \sim [-4,4],\ b_i \sim [-3,3],\ a_j \sim [-4,4], \rho = 15$

**Problem 7:** $m = 30,\ n = 200,\ A_{ij} \sim [-4,2],\ b_i \sim [-1,1],\ a_j \sim [-3,3], \rho = 16$

**Problem 8:** $m = 80,\ n = 400,\ A_{ij} \sim [-2,1],\ b_i \sim [-\frac{1}{2},\frac{1}{2}],\ a_j \sim [-1,1], \rho = 12$

**Problem 9:** $m = 20,\ n = 50,\ A_{ij} \sim [0,8],\ b_i \sim [-1,1],\ a_j \sim [0,4], \rho = 14$

**Problem 10:** $m = 100,\ n = 10{,}000,\ A_{ij} \sim [-\frac{1}{2},\frac{1}{2}],\ b_i \sim [-\frac{1}{4},\frac{1}{4}],\ a_j \sim [-\frac{1}{2},\frac{1}{2}], \rho = 15$

**Problem 11:** $m = 1{,}000,\ n = 1{,}000,\ A_{ij} \sim [-\frac{1}{2},\frac{1}{2}],\ b_i \sim [-\frac{1}{4},\frac{1}{4}],\ a_j \sim [-\frac{1}{2},\frac{1}{2}], \rho = 18$

**Problem 12:** $m = 700,\ n = 2{,}000,\ A_{ij} \sim [-\frac{1}{2},\frac{1}{2}],\ b_i \sim [-\frac{1}{4},\frac{1}{4}],\ a_j \sim [-\frac{1}{2},\frac{1}{2}], \rho = 24.$

### 5.G.2 Experiments of Section 5.4.2

**Problem 1:** $A = \begin{bmatrix} -1 & -1 & 1 & -1 & 1 \\ 1 & -1 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 & -1 \\ 0 & 0 & -1 & -1 & -1 \\ 0 & 0 & 0 & 0 & -1 \end{bmatrix}, \ b = \mathbf{0}_{5\times 1}, a = \mathbf{0}_{5\times 1}, \rho = 3$

**Problem 2:** $m = 5,\ n = 20, A_{ij} \sim [-1,1],\ b_i \sim [-2,2],\ a_j \sim [0,1], \rho = 5$

**Problem 3:** $m = 20,\ n = 50, A_{ij} \sim [-10,10],\ b_i \sim [-3,3],\ a_j \sim [-2,2], \rho = 6$

**Problem 4:** $m = 20,\ n = 180, A_{ij}{\sim}[-1,0.5],\ b_i = 0,\ a_j \sim [0,1], \rho = 1$

**Problem 5:** $m = 300,\ n = 30, A_{ij}{\sim}[-1,1],\ b_i = 0,\ a_j \sim [0,1], \rho = 2$

### 5.G.3 Experiments of Section 5.4.3

**Problem 1 (Enkhbat et al. 2006)** In this example, we solve:

$$\begin{aligned} \max_{x\in\mathbb{R}^{20}_{+}} \quad & \tfrac{1}{2}\textstyle\sum_{i=1}^{20}(x_i-2)^2 \\ \text{s.t.} \quad & Dx \le d, \end{aligned}$$

where

$$D^\top = \begin{bmatrix} -3 & 7 & 0 & -5 & 1 & 1 & 0 & 2 & -1 & 1 \\ 7 & 0 & -5 & 1 & 1 & 0 & 2 & -1 & -1 & 1 \\ 0 & -5 & 1 & 1 & 0 & 2 & -1 & -1 & -9 & 1 \\ -5 & 1 & 1 & 0 & 2 & -1 & -1 & -9 & 3 & 1 \\ 1 & 1 & 0 & 2 & -1 & -1 & -9 & 3 & 5 & 1 \\ 1 & 0 & 2 & -1 & -1 & -9 & 3 & 5 & 0 & 1 \\ 0 & 2 & -1 & -1 & -9 & 3 & 5 & 0 & 0 & 1 \\ 2 & -1 & -1 & -9 & 3 & 5 & 0 & 0 & 1 & 1 \\ -1 & -1 & -9 & 3 & 5 & 0 & 0 & 1 & 7 & 1 \\ -1 & -9 & 3 & 5 & 0 & 0 & 1 & 7 & -7 & 1 \\ -9 & 3 & 5 & 0 & 0 & 1 & 7 & -7 & -4 & 1 \\ 3 & 5 & 0 & 0 & 1 & 7 & -7 & -4 & -6 & 1 \\ 5 & 0 & 0 & 1 & 7 & -7 & -4 & -6 & -3 & 1 \\ 0 & 0 & 1 & 7 & -7 & -4 & -6 & -3 & 7 & 1 \\ 0 & 1 & 7 & -7 & -4 & -6 & -3 & 7 & 0 & 1 \\ 1 & 7 & -7 & -4 & -6 & -3 & 7 & 0 & -5 & 1 \\ 7 & -7 & -4 & -6 & -3 & 7 & 0 & -5 & 1 & 1 \\ -7 & -4 & -6 & -3 & 7 & 0 & -5 & 1 & 1 & 1 \\ -4 & -6 & -3 & 7 & 0 & -5 & 1 & 1 & 0 & 1 \\ -6 & -3 & 7 & 0 & -5 & 1 & 1 & 0 & 2 & 1 \end{bmatrix}, d = \begin{bmatrix} -5 \\ 2 \\ -1 \\ -3 \\ 5 \\ 4 \\ -1 \\ 0 \\ 9 \\ 40. \end{bmatrix} \tag{136}$$

**Problem 2** Same as problem 1, but the objective function is $\frac{1}{2}\sum_{i=1}^{20}(x_i+5)^2$.

Problems 3-7 use the following problem:

$$\begin{aligned} \max_{x\in\mathbb{R}^n_+} \quad & x^\top L^\top L x \\ \text{s. t.} \quad & Dx \leq d \\ & x \leq x_u, \end{aligned}$$

where $L \in \mathbb{R}^{m\times n}$ is a matrix generated randomly where all the entries are sampled uniformly from $0-1$. Moreover, let $D$ be a similar matrix with all $0-1$ random coefficients (except the last problem uses $0-2$ random coefficients) and $d$ to have integer entries uniformly distributed in range $[d_l, d_u]$. Denote $q$ to be the number of total constraints.

**Problem 3** $x_u = 5, d_l = 20, d_u = 60, n = 10, q = 15$

**Problem 4** $x_u = 3, d_l = 30, d_u = 60, n = 50, q = 62$

**Problem 5** $x_u = 2, d_l = 80, d_u = 120, n = 100, q = 130$

**Problem 6** $x_u = 2, d_l = 160, d_u = 240, n = 200, q = 240$

**Problem 7** $x_u = 1, d_l = 150, d_u = 300, n = 240, q = 280.$

### 5.G.4 Experiments of Section 5.4.4

**Problem 1** considers the following problem:

$$\begin{aligned} \max_{x\in\mathbb{R}^n_+} \quad & \log\left(\textstyle\sum_{i=1}^m \exp(A_{(i)}x)\right) \\ \text{s. t.} \quad & -\frac{i}{n} \leq x_i \leq \frac{n}{i}, \quad i = 1, \ldots, n, \end{aligned}$$

where $A_{ij} \sim [-3, 3]$. In the numerical experiments $n$ will vary.

Problems 2-4 consider the following problem:

$$\begin{aligned} \max_{x\in\mathbb{R}^n_+} \quad & \log\left(\textstyle\sum_{i=1}^m \exp(A_{(i)}x + b_i)\right) \\ \text{s. t.} \quad & Dx \leq d. \end{aligned}$$

**Problem 2** $n = q = m = 20, A_{ij} \sim [-3, 3],\ b_i \sim [-2, 2],\ D_{i,j} \sim [0, 1],\ d_i \sim [10, 30]$

**Problem 3** $n = q = m = 50, A_{ij} \sim [-3, 3],\ b_i \sim [-1, 1],\ D_{i,j} \sim [0, 1],\ d_i \sim [20, 60]$

**Problem 4** $n = q = m = 100, A_{ij} \sim [-3, 3],\ b_i \sim [-1, 1],\ D_{i,j} \sim [0, 1],\ d_i \sim [25, 75]$

Problems 5-6 consider:

$$\begin{aligned} \max_{x\in\mathbb{R}^n_+} \quad & \log\left(\textstyle\sum_{i=1}^m \exp(A_{(i)}x+b_i)\right) \\ \text{s. t.} \quad & x_i \le c, \quad i=1,\dots,n \\ & x_i + x_j \le u_{ij}, \quad i,j=1,\dots,n,\ i\ne j, \end{aligned}$$

**Problem 5** $n=m=10,\ A_{ij}\sim[-3,3],\ b_i\sim[-1,1],\ u_{ij}\sim[5,15], c=8$

**Problem 6** $n=m=30,\ A_{ij}\sim[-3,3],\ b_i\sim[-1,1],\ u_{ij}\sim[4,12], c=6.$

### 5.G.5 Experiments of Section 5.4.5

For the easiness of bookkeeping, we generate problem with every max-term having the same number of elements, i.e., $|\mathcal{I}_k| = |\mathcal{I}_{k'}|\ \forall k,k'\in\{1,\dots,K\}$.

Problems 1-6 are defined by:

$$\begin{aligned} \max_{x\in\mathbb{R}^n_+} \quad & \textstyle\sum_{k=1}^K \max_{j\in\mathcal{I}_k}\{A_{(j)}x\} \\ \text{s. t.} \quad & -\frac{i}{n} \le x_i \le \frac{n}{i}, \quad i=1,\dots,n, \end{aligned}$$

where:

**Problem 1** $n=5,\ A_{ij}\sim[-5,5],\ |\mathcal{I}_k|=5,\ K=1$

**Problem 2** $n=5,\ A_{ij}\sim[-5,5],\ |\mathcal{I}_k|=5,\ K=10$

**Problem 3** $n=20,\ A_{ij}\sim[-5,5],\ |\mathcal{I}_k|=10,\ K=10$

**Problem 4** $n=30,\ A_{ij}\sim[-5,5],\ |\mathcal{I}_k|=20,\ K=20$

**Problem 5** $n=100,\ A_{ij}\sim[-5,5],\ |\mathcal{I}_k|=40,\ K=30$

**Problem 6** $n=200,\ A_{ij}\sim[-4,4],\ |\mathcal{I}_k|=50,\ K=50.$

Problems 7-10 are defined by:

$$\begin{aligned} \max_{x\in\mathbb{R}^n_+} \quad & \textstyle\sum_{k=1}^K \max_{j\in\mathcal{I}_k}\{A_{(j)}x+b_j\} \\ \text{s. t.} \quad & Dx \le d, \end{aligned}$$

**Problem 7** $n=10,\ A_{ij}\sim[-5,5],\ b_j\sim[-10,10],\ D_{ij}\sim[0,1],\ d_i\sim[5,15],\ |\mathcal{I}_k|=5,\ K=2$

**Problem 8** $n=10,\ A_{ij}\sim[-5,5],\ b_j\sim[-10,10],\ D_{ij}\sim[0,1],\ d_i\sim[5,15],\ |\mathcal{I}_k|=50,\ K=50$

**Problem 9** $n=30,\ A_{ij}\sim[-5,5],\ b_j\sim[-10,10],\ D_{ij}\sim[0,1],\ d_i\sim[5,15],\ |\mathcal{I}_k|=50,\ K=50$

**Problem 10** $n = 50,\ A_{ij} \sim [-5, 5],\ b_j \sim [-10, 10],\ D_{ij} \sim [0, 1],\ d_i \sim [5, 15],\ |\mathcal{I}_k| = 60, K = 60$ .

Problems 10-13 consider the same problem as above, but $D$ and $d$ are as given in (136) with $n = 20$. The objective function varies as:

**Problem 11** $A_{ij} \sim [-5, 10],\ b_j \sim [-10, 10],\ |\mathcal{I}_k| = 10,\ K = 10$

**Problem 12** $A_{ij} \sim [-5, 10],\ b_j \sim [-10, 10],\ |\mathcal{I}_k| = 50,\ K = 10$

**Problem 13** $A_{ij} \sim [-5, 10],\ b_j \sim [-10, 10],\ |\mathcal{I}_k| = 100,\ K = 50$ .

# 6 A Reformulation-Linearization Technique for Optimization over Simplices

## Abstract

We study non-convex optimization problems over simplices. We show that for a large class of objective functions, the convex approximation obtained from the Reformulation-Linearization Technique (RLT) admits optimal solutions that exhibit a sparsity pattern. This characteristic of the optimal solutions allows us to conclude that *(i)* a linear matrix inequality constraint, which is often added to tighten the relaxation, is vacuously satisfied and can thus be omitted, and *(ii)* the number of decision variables in the RLT relaxation can be reduced from $\mathcal{O}(n^2)$ to $\mathcal{O}(n)$. Taken together, both observations allow us to reduce computation times by up to several orders of magnitude. Our results can be specialized to indefinite quadratic optimization problems over simplices and extended to non-convex optimization problems over the Cartesian product of two simplices as well as specific classes of polyhedral and non-convex feasible regions. Our numerical experiments illustrate the promising performance of the proposed framework.

## 6.1 Introduction

In this work, we study non-convex optimization problems of the form

$$\begin{array}{ll} \sup\limits_{\boldsymbol{x}} & f(\boldsymbol{x}) + g(\boldsymbol{x}) \\ \text{s.t.} & \boldsymbol{A}\boldsymbol{x} \leq \boldsymbol{b} \\ & \boldsymbol{x} \in \mathbb{R}^n, \end{array} \tag{137}$$

where $f : \mathbb{R}^n \mapsto \mathbb{R}$ is a generic function, $g : \mathbb{R}^n \mapsto \mathbb{R}$ is concave, $\boldsymbol{A} \in \mathbb{R}^{m \times n}$ and $\boldsymbol{b} \in \mathbb{R}^m$. Since $f$ is not necessarily concave, problem (137) is a hard optimization problem even if $P = NP$ (Nesterov 2004, Theorem 1). In the special case where $f$ is convex, problem (137) recovers the class of DC (difference-of-convex-functions) optimization problems over a polyhedron Horst and Thoai (1999). Significant efforts have been devoted to solving problem (137) exactly (most commonly via branch-and-bound techniques) or approximately (often via convex approximations). For both tasks, the Reformulation-Linearization Technique (RLT) can be used to obtain tight yet readily solvable convex relaxations of (137).

Originally, RLT has been introduced to equivalently reformulate binary quadratic optimization problems as mixed-binary linear optimization problems Adams and Sherali (1986). To this end, each linear constraint in the original problem is multiplied with each binary decision variable to generate implied quadratic inequalities. These inequalities are subsequently linearized through the introduction of auxiliary decision variables whose values coincide with the generated quadratic terms. This idea is reminiscent of the McCormick envelopes McCormick (1976), which relax bilinear expressions by introducing implied inequalities that are subsequently linearized. RLT has been extended to (continuous) polynomial optimization problems Sherali and Tuncbilek (1992), where implied inequalities are generated from multiplying and subsequently linearizing existing bound constraints.

In this work, we consider a variant of RLT—the Reformulation-Convexification Technique Sherali and Tuncbilek (1995)—which applies to linearly constrained optimization problems that maximize a non-concave objective function. This RLT variant (which we henceforth simply call ‘RLT’ for ease of exposition) replaces the non-concave function $f$ in problem (137) with an auxiliary function $f' : \mathbb{R}^{n\times n} \times \mathbb{R}^n \mapsto \mathbb{R}$ that is concave over the lifted domain $(\boldsymbol{X}, \boldsymbol{x}) \in \mathbb{S}^n \times \mathbb{R}^n$ and that satisfies $f'(\boldsymbol{X}, \boldsymbol{x}) = f(\boldsymbol{x})$ whenever $\boldsymbol{X} = \boldsymbol{x}\boldsymbol{x}^\top$. For the special case where $f(\boldsymbol{x}) = \boldsymbol{x}^\top \boldsymbol{P}\boldsymbol{x}$ for an indefinite symmetric matrix $\boldsymbol{P} \in \mathbb{S}^n$, for example, we can choose $f'(\boldsymbol{X}, \boldsymbol{x}) = \langle \boldsymbol{P}, \boldsymbol{X}\rangle$. RLT then augments problem (137) with the decision matrix $\boldsymbol{X} \in \mathbb{S}^n$ and the constraints

$$\boldsymbol{a}_i^\top \boldsymbol{X}\boldsymbol{a}_j - (b_i\boldsymbol{a}_j + b_j\boldsymbol{a}_i)^\top \boldsymbol{x} + b_ib_j \geq 0 \qquad \forall i, j = 1, \ldots, m, \tag{138}$$

where $\boldsymbol{a}_i^\top$ denotes the $i$-th row of the matrix $\boldsymbol{A}$. The constraints (138) are justified by the fact that the pairwise multiplications $(\boldsymbol{a}_i^\top \boldsymbol{x} - b_i)(\boldsymbol{a}_j^\top \boldsymbol{x} - b_j)$ of the constraints in problem (137) have to be non-negative, and those multiplications coincide with the constraints (138) whenever $\boldsymbol{X} = \boldsymbol{x}\boldsymbol{x}^\top$. To obtain a convex relaxation of problem (137), the non-convex constraint $\boldsymbol{X} = \boldsymbol{x}\boldsymbol{x}^\top$ is either removed (which we henceforth refer to as ‘classical RLT’, see Sherali and Adams (2013)) or relaxed to the linear matrix inequality (LMI) constraint $\boldsymbol{X} \succeq \boldsymbol{x}\boldsymbol{x}^\top$ (henceforth referred to as RLT/SDP, see Anstreicher (2009), Bao et al. (2011), Sherali and Fraticelli (2002)). Even though the matrix $\boldsymbol{X}$ linearizes quadratic terms, we emphasize that the problems we are considering are not restricted to quadratic programs since $f$ may be a generic nonlinear function.

RLT and its extensions have been exceptionally successful in providing tight approximations to indefinite quadratic Sherali and Fraticelli (2002), polynomial Sherali and Tuncbilek (1992) and generic non-convex optimization problems Zhen et al. (2021), Liberti and Pantelides (2006), and RLT is routinely implemented in state-of-the-art optimization software, including ANTIGONE Misener and Floudas (2014), CPLEX IBM ILOG CPLEX (2014), GLoMIQO Misener and Floudas (2013) and GUROBI Gurobi Optimization (2018).

In this work, we assume that the constraints of problem (137) describe an $n$-dimensional simplex. Under this assumption, we show that for a large class of functions $f$ that admit a monotone lifting (which includes, among others, various transformations of quadratic functions as well as the negative entropy), the RLT relaxation of problem (137) admits an optimal solution $(\boldsymbol{X}^\star, \boldsymbol{x}^\star)$ that satisfies $\boldsymbol{X}^\star = \mathrm{diag}(\boldsymbol{x}^\star)$. This has two important consequences. Firstly, we show that when the feasible region of problem (137) is a simplex, $\boldsymbol{X}^\star = \mathrm{Diag}(\boldsymbol{x}^\star)$ satisfies $\boldsymbol{X}^\star \succeq \boldsymbol{x}^\star \boldsymbol{x}^{\star\top}$, that is, the RLT and RLT/SDP relaxations are equivalent, and the computationally expensive LMI constraint $\boldsymbol{X} \succeq \boldsymbol{x}\boldsymbol{x}^\top$ can be omitted in RLT/SDP. Secondly, we do not need to introduce the decision matrix $\boldsymbol{X} \in \mathbb{S}^n$ in the RLT relaxation, which amounts to a dramatic reduction in the size of the resulting relaxation. We also discuss how our result can be extended to instances of problem (137) over the Cartesian product of two simplices, a generic polyhedron, or a non-convex feasible region as well as an indefinite quadratic objective function.

Indefinite quadratic optimization over simplices (also known as standard quadratic optimization) has a long history, and it has found applications, among others, in mean/variance portfolio selection and the determination of the maximal cliques on a node-weighted graph Bomze (1998). More generally, non-convex polynomial optimization problems over simplices have been proposed for the global optimization of neural networks Beliakov and Abraham (2002), portfolio optimization using the expected shortfall risk measure Bertsimas et al. (2004) and the computation of the Lebesgue constant for polynomial interpolation over a simplex Hesthaven (1998); see De Klerk et al. (2008) for a general discussion. Simplicial decompositions of non-convex optimization problems are also studied extensively in the global optimization literature Horst et al. (2000).

The remainder of this section proceeds as follows. We analyze the RLT relaxations of simplex instances of problem (137) in Section 6.2 and report on numerical experiments in Section 6.3, respectively. Appendix 6.A extends our findings to well-structured optimization problems over the Cartesian product of two simplices, specific classes of polyhedral and non-convex feasible regions, as well as indefinite quadratic objective functions. Appendix 6.B, finally, contains additional numerical experiments.

**Notation.** We denote by $\mathbb{R}^n$ ($\mathbb{R}^n_+$) the (non-negative orthant of the) $n$-dimensional Euclidean space and by $\mathbb{Q}$ the set of rational numbers. The cone of (positive semidefinite) symmetric matrices in $\mathbb{R}^{n\times n}$ is denoted by $\mathbb{S}^n$ ($\mathbb{S}^n_+$). Bold lower and upper case letters denote vectors and matrices, respectively, while standard lower case letters are reserved for scalars. We denote the $i$-th component of a vector $\boldsymbol{x}$ by $x_i$, the $(i, j)$-th element of a matrix $\boldsymbol{A}$ by $A_{ij}$ and the $i$-th row of a matrix $\boldsymbol{A}$ by $\boldsymbol{a}_i^\top$. We write $\boldsymbol{X} \succeq \boldsymbol{Y}$ to indicate that $\boldsymbol{X} - \boldsymbol{Y}$ is positive semidefinite. The trace operator is denoted by $\operatorname{tr}(\cdot)$, and the trace inner product between two symmetric matrices is given by $\langle \cdot, \cdot \rangle$. Finally, $\operatorname{Diag}(\boldsymbol{x})$ is a diagonal matrix whose diagonal elements coincide with the components of the vector $\boldsymbol{x}$.

## 6.2 RLT and RLT/SDP over Simplices

This section studies instances of problem (137) where the constraints $\boldsymbol{A}\boldsymbol{x} \leq \boldsymbol{b}$ describe the $n$-dimensional probability simplex:

$$\begin{array}{ll} \sup\limits_{\boldsymbol{x}} & f(\boldsymbol{x}) + g(\boldsymbol{x}) \\ \text{s.t.} & \displaystyle\sum_{i=1}^{n} x_i = 1 \\ & \boldsymbol{x} \in \mathbb{R}^n_+. \end{array} \tag{139}$$

Assuming that the feasible region describes a probability simplex, as opposed to any other full-dimensional simplex in $\mathbb{R}^n$, does not restrict generality. Indeed, we can always redefine the objective function as $f(\boldsymbol{x}) \leftarrow f(\boldsymbol{T}\boldsymbol{x})$ and $g(\boldsymbol{x}) \leftarrow g(\boldsymbol{T}\boldsymbol{x})$ for the invertible matrix $\boldsymbol{T} \in \mathbb{R}^{n \times n}$ that has as columns the extreme points of the simplex to be considered. The pairwise products between the constraints $x_i \geq 0$, $i = 1, \ldots, n$, and $\sum_{i=1}^n x_i = 1$ result in the RLT constraints

$$\boldsymbol{X} \geq \boldsymbol{0}, \qquad \sum_{j=1}^{n} X_{ij} = \sum_{j=1}^{n} X_{ji} = x_i \quad \forall i = 1, \ldots, n;$$

here we omit the constraint $\sum_{i=1}^n \sum_{j=1}^n X_{ij} = 1$ as it is implied by the above constraints and the fact that $\sum_{i=1}^n x_i = 1$. Thus, the RLT relaxation of problem (139) can be written as

$$\begin{array}{lll} \sup\limits_{\boldsymbol{X}, \boldsymbol{x}} & f'(\boldsymbol{X}, \boldsymbol{x}) + g(\boldsymbol{x}) & \\ \text{s.t.} & \displaystyle\sum_{j=1}^{n} X_{ij} = \sum_{j=1}^{n} X_{ji} = x_i & \forall i = 1, \ldots, n \\ & \displaystyle\sum_{i=1}^{n} x_i = 1 & \\ & \boldsymbol{X} \geq \boldsymbol{0}, \quad \boldsymbol{X} \in \mathbb{S}^n, \quad \boldsymbol{x} \in \mathbb{R}^n_+, & \end{array} \tag{140}$$

where the auxiliary function $f'$ has to be suitably chosen, while the RLT/SDP relaxation contains the additional LMI constraint $\boldsymbol{X} \succeq \boldsymbol{x}\boldsymbol{x}^\top$.

We now define a condition which ensures that the RLT relaxation (140) of problem (139) admits an optimal solution $(\boldsymbol{X}^\star, \boldsymbol{x}^\star)$ with $\boldsymbol{X}^\star = \operatorname{diag}(\boldsymbol{x}^\star)$.

**Definition 7.** *We say that $f : \mathbb{R}^n_+ \mapsto \mathbb{R}$ has a* monotone lifting *if there is a concave function $f' : \mathbb{S}^n \times \mathbb{R}^n_+ \mapsto \mathbb{R}$ such that $f'(\boldsymbol{X}, \boldsymbol{x}) = f(\boldsymbol{x})$ whenever $\boldsymbol{X} = \boldsymbol{x}\boldsymbol{x}^\top$, as well as $f'(\boldsymbol{X}', \boldsymbol{x}) \geq f'(\boldsymbol{X}, \boldsymbol{x})$ for all $(\boldsymbol{X}, \boldsymbol{x}) \in \mathbb{S}^n \times \mathbb{R}^n_+$ and all $\boldsymbol{X}' \in \mathbb{S}^n$ satisfying $\boldsymbol{X}' \succeq \boldsymbol{X}$.*

The requirement in Definition 7 that $f'(\boldsymbol{X}, \boldsymbol{x}) = f(\boldsymbol{x})$ whenever $\boldsymbol{X} = \boldsymbol{x}\boldsymbol{x}^\top$ is needed for the correctness of the RLT relaxation. The concavity of $f'$ is required for the RLT relaxation to be a convex optimization problem. The assumption that $f'(\boldsymbol{X}', \boldsymbol{x}) \geq f'(\boldsymbol{X}, \boldsymbol{x})$ for all $(\boldsymbol{X}, \boldsymbol{x}) \in \mathbb{S}^n \times \mathbb{R}^n_+$ and all $\boldsymbol{X}' \in \mathbb{S}^n$ satisfying $\boldsymbol{X}' \succeq \boldsymbol{X}$, finally, will allow us to deduce an optimal solution for $\boldsymbol{X}$ based on the value of $\boldsymbol{x}$. Indeed, we will see below in Theorem 22 that the RLT relaxation (140) of an instance of problem (139) admits optimal solutions $(\boldsymbol{X}^\star, \boldsymbol{x}^\star)$ satisfying $\boldsymbol{X}^\star = \operatorname{Diag}(\boldsymbol{x}^\star)$ whenever the auxiliary function $f'$ in (140) is a monotone lifting of the function $f$ in (139). Intuitively speaking, Definition 7 enables us to weakly improve any solution $(\boldsymbol{X}, \boldsymbol{x})$ satisfying $\boldsymbol{X} \neq \operatorname{diag}(\boldsymbol{x})$ by iteratively moving off-diagonal elements of $\boldsymbol{X}$ to the diagonal. Before presenting the formal result, we provide some examples of functions $f$ that admit monotone liftings.

**Proposition 20.** *The following function classes have monotone liftings:*

1. ***Generalized linearithmic functions:*** *$f(\boldsymbol{x}) = \sum_{\ell=1}^L (\boldsymbol{t}_\ell^\top \boldsymbol{x} + t_\ell) \cdot h_\ell(\boldsymbol{t}_\ell^\top \boldsymbol{x} + t_\ell)$ with* (i) *$\boldsymbol{t}_\ell \in \mathbb{R}^n_+$, $t_\ell \in \mathbb{R}_+$ and $h_\ell : \mathbb{R} \mapsto \mathbb{R}$ concave and non-decreasing, or* (ii) *$\boldsymbol{t}_\ell \in \mathbb{R}^n$, $t_\ell \in \mathbb{R}$ and $h_\ell : \mathbb{R} \mapsto \mathbb{R}$ affine and non-decreasing.*

2. ***Linear combinations:*** *$f(\boldsymbol{x}) = \sum_{\ell=1}^L t_\ell \cdot f_\ell(\boldsymbol{x})$ with $t_\ell \in \mathbb{R}_+$, where each $f_\ell : \mathbb{R}^n \mapsto \mathbb{R}$ has a monotone lifting.*

3. ***Concave compositions:*** *$h(\boldsymbol{x}) = g(f(\boldsymbol{x}))$ for $f : \mathbb{R}^n_+ \mapsto \mathbb{R}$ with a monotone lifting as well as a concave and non-decreasing $g : \mathbb{R} \mapsto \mathbb{R}$.*

4. ***Linear pre-compositions:*** *$h(\boldsymbol{x}) = f(\boldsymbol{T}\boldsymbol{x})$ for $f : \mathbb{R}^p_+ \mapsto \mathbb{R}$ with a monotone lifting as well as $\boldsymbol{T} \in \mathbb{R}^{p \times n}$.*

5. ***Pointwise minima:*** *$h(\boldsymbol{x}) = \min\{f_1(\boldsymbol{x}), \ldots, f_L(\boldsymbol{x})\}$ where each $f_\ell : \mathbb{R}^n_+ \mapsto \mathbb{R}$ has a monotone lifting.*

*Proof.* In view of case *(i)* of the first statement, we choose

$$f'(\boldsymbol{X}, \boldsymbol{x}) \;=\; \sum_{\ell=1}^L (\boldsymbol{t}_\ell^\top \boldsymbol{x} + t_\ell) \cdot h_\ell\left(\frac{\boldsymbol{t}_\ell^\top \boldsymbol{X} \boldsymbol{t}_\ell + 2 t_\ell \boldsymbol{t}_\ell^\top \boldsymbol{x} + t_\ell^2}{\boldsymbol{t}_\ell^\top \boldsymbol{x} + t_\ell}\right),$$

which is concave in $(\boldsymbol{X}, \boldsymbol{x})$ since it constitutes the sum of perspectives of concave functions (Boyd and Vandenberghe 2004, §3.2.2 and §3.2.6). Whenever $\boldsymbol{X} = \boldsymbol{x}\boldsymbol{x}^\top$, we have

$$\boldsymbol{t}_\ell^\top \boldsymbol{X} \boldsymbol{t}_\ell + 2t_\ell \boldsymbol{t}_\ell^\top \boldsymbol{x} + t_\ell^2 \;=\; \boldsymbol{t}_\ell^\top \boldsymbol{x}\boldsymbol{x}^\top \boldsymbol{t}_\ell + 2t_\ell \boldsymbol{t}_\ell^\top \boldsymbol{x} + t_\ell^2 \;=\; (\boldsymbol{t}_\ell^\top \boldsymbol{x} + t_\ell)^2,$$

and thus the standard limit convention for perspective functions implies that $f'(\boldsymbol{X}, \boldsymbol{x}) = f(\boldsymbol{x})$ for all $\boldsymbol{x} \in \mathbb{R}^n_+$. Moreover, for any $\boldsymbol{x} \in \mathbb{R}^n_+$, we have

$$\boldsymbol{t}_\ell^\top \boldsymbol{X}' \boldsymbol{t}_\ell \;\geq\; \boldsymbol{t}_\ell^\top \boldsymbol{X} \boldsymbol{t}_\ell \qquad \forall \boldsymbol{X}, \boldsymbol{X}' \in \mathbb{S}^n \,:\, \boldsymbol{X}' \succeq \boldsymbol{X},$$

where the inequality holds since $\boldsymbol{X}' - \boldsymbol{X} \succeq 0$. We conclude that

$$h_\ell \left( \frac{\boldsymbol{t}_\ell^\top \boldsymbol{X}' \boldsymbol{t}_\ell + 2t_\ell \boldsymbol{t}_\ell^\top \boldsymbol{x} + t_\ell^2}{\boldsymbol{t}_\ell^\top \boldsymbol{x} + t_\ell} \right) \;\geq\; h_\ell \left( \frac{\boldsymbol{t}_\ell^\top \boldsymbol{X} \boldsymbol{t}_\ell + 2t_\ell \boldsymbol{t}_\ell^\top \boldsymbol{x} + t_\ell^2}{\boldsymbol{t}_\ell^\top \boldsymbol{x} + t_\ell} \right)$$

as $2t_\ell \boldsymbol{t}_\ell^\top \boldsymbol{x} + t_\ell^2 \geq 0$ and $\boldsymbol{t}_\ell^\top \boldsymbol{x} + t_\ell \geq 0$ due to the non-negativity of $\boldsymbol{t}_\ell$, $t_\ell$ and $\boldsymbol{x}$, which implies that $f'(\boldsymbol{X}', \boldsymbol{x}) \geq f'(\boldsymbol{X}, \boldsymbol{x})$ as desired.

One readily verifies that in the special case where each $h_\ell$ is affine, the concavity of $f'$, the agreement of $f'$ with $f$ when $\boldsymbol{X} = \boldsymbol{x}\boldsymbol{x}^\top$ and the monotonicity of $f'$ with respect to $\boldsymbol{X}' \succeq \boldsymbol{X}$ continue to hold even when $\boldsymbol{t}_\ell$ and/or $t_\ell$ fail to be non-negative. This establishes case *(ii)* of the first statement.

As for the second statement, let $f'_\ell : \mathbb{S}^n \times \mathbb{R}^n_+ \mapsto \mathbb{R}$ be monotone liftings of $f_\ell$, $\ell = 1, \ldots, L$. We claim that $f'(\boldsymbol{X}, \boldsymbol{x}) = \sum_{\ell=1}^L t_\ell \cdot f'_\ell(\boldsymbol{X}, \boldsymbol{x})$ is a monotone lifting of $f$. Indeed, one readily verifies that $f'$ inherits concavity in $(\boldsymbol{X}, \boldsymbol{x})$ and agreement with $f$ when $\boldsymbol{X} = \boldsymbol{x}\boldsymbol{x}^\top$ from its constituent functions $f'_\ell$. Moreover, since $f'_\ell(\boldsymbol{X}', \boldsymbol{x}) \geq f'_\ell(\boldsymbol{X}, \boldsymbol{x})$ for all $\boldsymbol{X}, \boldsymbol{X}' \in \mathbb{S}^n$ with $\boldsymbol{X}' \succeq \boldsymbol{X}$, $\ell = 1, \ldots, L$, we have $f'(\boldsymbol{X}', \boldsymbol{x}) \geq f'(\boldsymbol{X}, \boldsymbol{x})$ as well.

In view of the third statement, let $f'$ be a monotone lifting of $f$. We claim that in this case, $h'(\boldsymbol{X}, \boldsymbol{x}) = g(f'(\boldsymbol{X}, \boldsymbol{x}))$ is a monotone lifting of $h$. Indeed, $h'$ is a non-decreasing concave transformation of a concave function and is thus concave (Boyd and Vandenberghe 2004, §3.2.5). Moreover, since $f'(\boldsymbol{X}, \boldsymbol{x}) = f(\boldsymbol{x})$ for $\boldsymbol{X} = \boldsymbol{x}\boldsymbol{x}^\top$, we have

$$h'(\boldsymbol{X}, \boldsymbol{x}) \;=\; g(f'(\boldsymbol{X}, \boldsymbol{x})) \;=\; g(f(\boldsymbol{x})) \;=\; h(\boldsymbol{x})$$

whenever $\boldsymbol{X} = \boldsymbol{x}\boldsymbol{x}^\top$. Finally, the monotonicity of $g$ implies that

$$h'(\boldsymbol{X}', \boldsymbol{x}) \;=\; g(f'(\boldsymbol{X}', \boldsymbol{x})) \;\geq\; g(f'(\boldsymbol{X}, \boldsymbol{x})) \;=\; h'(\boldsymbol{X}, \boldsymbol{x})$$

for all $\boldsymbol{X}, \boldsymbol{X}' \in \mathbb{S}^n$ with $\boldsymbol{X}' \succeq \boldsymbol{X}$.

For the fourth statement, we set $h'(\boldsymbol{X}, \boldsymbol{x}) = f'(\boldsymbol{T}\boldsymbol{X}\boldsymbol{T}^\top, \boldsymbol{T}\boldsymbol{x})$, where $f'$ is a monotone lifting of $f$. The function $h'$ is concave since it constitutes a composition of a concave function with a linear function (Boyd and Vandenberghe 2004, §3.2.2). Moreover, for any $\boldsymbol{x} \in \mathbb{R}^n_+$ and $\boldsymbol{X} = \boldsymbol{x}\boldsymbol{x}^\top$, we have

$$h'(\boldsymbol{X}, \boldsymbol{x}) \;=\; f'(\boldsymbol{T}\boldsymbol{X}\boldsymbol{T}^\top, \boldsymbol{T}\boldsymbol{x}) \;=\; f(\boldsymbol{T}\boldsymbol{x}),$$

where the second identity holds since $\boldsymbol{T}\boldsymbol{X}\boldsymbol{T}^\top = (\boldsymbol{T}\boldsymbol{x})(\boldsymbol{T}\boldsymbol{x})^\top$ whenever $\boldsymbol{X} = \boldsymbol{x}\boldsymbol{x}^\top$ as well as $f'(\boldsymbol{X}, \boldsymbol{x}) = f(\boldsymbol{x})$ for $\boldsymbol{X} = \boldsymbol{x}\boldsymbol{x}^\top$. To see that $h'(\boldsymbol{X}', \boldsymbol{x}) \geq h'(\boldsymbol{X}, \boldsymbol{x})$ for all $\boldsymbol{x} \in \mathbb{R}^n_+$ and all $\boldsymbol{X}, \boldsymbol{X}' \in \mathbb{S}^n$ satisfying $\boldsymbol{X}' \succeq \boldsymbol{X}$, we note that

$$h'(\boldsymbol{X}', \boldsymbol{x}) \;=\; f'(\boldsymbol{T}\boldsymbol{X}'\boldsymbol{T}^\top, \boldsymbol{T}\boldsymbol{x}) \;\geq\; f'(\boldsymbol{T}\boldsymbol{X}\boldsymbol{T}^\top, \boldsymbol{T}\boldsymbol{x}) \;=\; h'(\boldsymbol{X}, \boldsymbol{x}),$$

where the inequality follows from the fact that

$$\boldsymbol{X}' \succeq \boldsymbol{X} \quad\Longrightarrow\quad \boldsymbol{T}(\boldsymbol{X}' - \boldsymbol{X})\boldsymbol{T}^\top \succeq \boldsymbol{0} \quad\Longrightarrow\quad \boldsymbol{T}\boldsymbol{X}'\boldsymbol{T}^\top \succeq \boldsymbol{T}\boldsymbol{X}\boldsymbol{T}^\top$$

and the assumption that $f'$ is a monotone lifting.

For the last statement, we set $h'(\boldsymbol{X}, \boldsymbol{x}) = \min\{f'_1(\boldsymbol{X}, \boldsymbol{x}), \ldots, f'_L(\boldsymbol{X}, \boldsymbol{x})\}$, where $f'_\ell : \mathbb{S}^n \times \mathbb{R}^n_+ \mapsto \mathbb{R}$ is a monotone lifting of $f_\ell$ for all $\ell = 1, \ldots, L$. The function $h'$ is concave as it is a minimum of concave functions (Boyd and Vandenberghe 2004, §3.2.3). Moreover, for any $\boldsymbol{x} \in \mathbb{R}^n_+$ and $\boldsymbol{X} = \boldsymbol{x}\boldsymbol{x}^\top$, we have

$$h'(\boldsymbol{X}, \boldsymbol{x}) \;=\; \min\{f'_1(\boldsymbol{X}, \boldsymbol{x}), \ldots, f'_L(\boldsymbol{X}, \boldsymbol{x})\} \;=\; \min\{f_1(\boldsymbol{x}), \ldots, f_L(\boldsymbol{x})\} \;=\; f(\boldsymbol{x}),$$

since each $f'_\ell$ is a monotone lifting of $f_\ell$. Similarly, for any $\boldsymbol{x} \in \mathbb{R}^n_+$ and any $\boldsymbol{X}, \boldsymbol{X}' \in \mathbb{S}^n$ satisfying $\boldsymbol{X}' \succeq \boldsymbol{X}$, we have

$$\begin{aligned} h'(\boldsymbol{X}', \boldsymbol{x}) \;&=\; \min\{f'_1(\boldsymbol{X}', \boldsymbol{x}), \ldots, f'_L(\boldsymbol{X}', \boldsymbol{x})\} \\ &\geq\; \min\{f'_1(\boldsymbol{X}, \boldsymbol{x}), \ldots, f'_L(\boldsymbol{X}, \boldsymbol{x})\} \;=\; h'(\boldsymbol{X}, \boldsymbol{x}), \end{aligned}$$

where the inequality again follows from the fact that each $f'_\ell$ is a monotone lifting of $f_\ell$. This concludes the proof. □

Through an iterative application of its rules, Proposition 20 allows us to construct a rich family of functions that admit monotone liftings. We next list several examples that are of particular interest.

**Corollary 15.** *The functions listed below have monotone liftings.*

1. ***Convex quadratic functions:*** *$f(\boldsymbol{x}) = \boldsymbol{x}^\top \boldsymbol{Q}\boldsymbol{x} + \boldsymbol{q}^\top \boldsymbol{x} + q$ with $\boldsymbol{Q} \in \mathbb{S}^n_+$.*
2. ***Conic quadratic functions:*** *$f(\boldsymbol{x}) = \|\boldsymbol{F}\boldsymbol{x}\|_2 + \boldsymbol{f}^\top \boldsymbol{x} + f$, where $\boldsymbol{F} \in \mathbb{R}^{k\times n}$, $\boldsymbol{f} \in \mathbb{R}^n$ and $f \in \mathbb{R}$.*
3. ***Negative entropy:*** *$f(\boldsymbol{x}) = \sum_{i=1}^n c_i \cdot x_i \ln x_i$ with $c_i \in \mathbb{R}_+$.*
4. ***Power functions:*** *$f(x) = x^a$ with $a \in [1, 2]$ and $a \in \mathbb{Q}$.*

*Proof.* In view of the first statement, let $\boldsymbol{Q} = \boldsymbol{L}^\top \boldsymbol{L}$ for $\boldsymbol{L} \in \mathbb{R}^{n\times n}$, where $\boldsymbol{L}$ can be computed from a Cholesky decomposition. Identifying $\boldsymbol{t}_\ell^\top$ with the $\ell$-th row of $\boldsymbol{L}$ and setting $t_\ell = 0$, $\ell = 1, \ldots, n$, we then obtain

$$\begin{aligned} f(\boldsymbol{x}) \;&=\; (\boldsymbol{L}\boldsymbol{x})^\top (\boldsymbol{L}\boldsymbol{x}) \;+\; \boldsymbol{q}^\top \boldsymbol{x} + q \\ &=\; \sum_{\ell=1}^n (\boldsymbol{t}_\ell^\top \boldsymbol{x})^2 \;+\; \boldsymbol{q}^\top \boldsymbol{x} + q \\ &=\; \sum_{\ell=1}^n (\boldsymbol{t}_\ell^\top \boldsymbol{x} + t_\ell) \cdot h_\ell(\boldsymbol{t}_\ell^\top \boldsymbol{x} + t_\ell) \;+\; \boldsymbol{q}^\top \boldsymbol{x} + q, \end{aligned}$$

where $h_\ell : \mathbb{R} \mapsto \mathbb{R}$ is the identity function, $\ell = 1, \ldots, n$. The first expression on the right-hand side satisfies the conditions of the first statement of Proposition 20 and thus admits a monotone lifting. The remaining term $g(\boldsymbol{x}) = \boldsymbol{q}^\top \boldsymbol{x} + q$ admits the trivial lifting $g'(\boldsymbol{X}, \boldsymbol{x}) = \boldsymbol{q}^\top \boldsymbol{x} + q$, and the second statement of Proposition 20 thus implies that the function $f$ has a monotone lifting as well.

As for the second statement, we note that

$$f(\boldsymbol{x}) \;=\; \sqrt{\boldsymbol{x}^\top \boldsymbol{F}^\top \boldsymbol{F}\boldsymbol{x}} \;+\; \boldsymbol{f}^\top \boldsymbol{x} + f.$$

Since $\boldsymbol{F}^\top \boldsymbol{F} \succeq \boldsymbol{0}$ by construction, the term $\boldsymbol{x}^\top \boldsymbol{F}^\top \boldsymbol{F}\boldsymbol{x}$ has a monotone lifting due to the first statement of this corollary. Moreover, since $x \mapsto \sqrt{x}$ is non-decreasing and concave, the third statement of Proposition 20 implies that the expression $\sqrt{\boldsymbol{x}^\top \boldsymbol{F}^\top \boldsymbol{F}\boldsymbol{x}}$ admits a monotone lifting. The remaining term $g(\boldsymbol{x}) = \boldsymbol{f}^\top \boldsymbol{x} + f$ again admits the trivial lifting $g'(\boldsymbol{X}, \boldsymbol{x}) = \boldsymbol{f}^\top \boldsymbol{x} + f$, and the second statement of Proposition 20 thus implies that the function $f$ has a monotone lifting as well.

In view of the third statement, we first note that each term $x_i \ln x_i$ has a monotone lifting if we choose $\boldsymbol{t}_i = \mathbf{e}_i$, where $\mathbf{e}_i$ denotes the $i$-th canonical basis vector in $\mathbb{R}^n$, and $t_i = 0$ in the first statement of Proposition 20. Since $f$ constitutes a weighted sum of these terms, the existence of its monotone lifting then follows from the second statement of Proposition 20.

As for the last statement, we note that $f(x) = x \cdot h(x)$ with $h(x) = x^{a-1}$. Since $h$ is concave and non-decreasing, the first statement of Proposition 20 implies that $f$ has a monotone lifting.

□

Any indefinite quadratic function can be represented as the sum of a convex quadratic and a concave quadratic function Fampa et al. (2017), Park (2016). Thus, if problem (139) optimizes an indefinite quadratic function over a simplex (*i.e.*, if it is a standard quadratic optimization problem), then we can redefine its objective function as a sum of a convex quadratic and a concave quadratic function and subsequently apply the first statement in Proposition 20 to the convex part of the objective function.

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

**Theorem 22.** *If the function $f$ in problem (139) has a monotone lifting $f'$, then the corresponding RLT relaxation (140) has an optimal solution $(\boldsymbol{X}^\star, \boldsymbol{x}^\star)$ satisfying $\boldsymbol{X}^\star = \mathrm{Diag}(\boldsymbol{x}^\star)$.*

*Proof.* The RLT relaxation (140) maximizes the concave and, *a fortiori*, continuous function $f'(\boldsymbol{X}, \boldsymbol{x}) + g(\boldsymbol{x})$ over a compact feasible region. The Weierstrass theorem thus guarantees that the optimal value of problem (140) is attained.

Let $(\boldsymbol{X}^\star, \boldsymbol{x}^\star)$ be an optimal solution to the RLT relaxation (140). If $\boldsymbol{X}^\star = \mathrm{Diag}(\boldsymbol{x}^\star)$, then there is nothing to prove. If $\boldsymbol{X}^\star \neq \mathrm{Diag}(\boldsymbol{x}^\star)$, on the other hand, then there is $i, j \in \{1, \ldots, n\}$, $i \neq j$, such that $X^\star_{ij} = X^\star_{ji} > 0$. Define $\boldsymbol{X}' \in \mathbb{S}^n$ as $\boldsymbol{X}' = \boldsymbol{X}^\star + \boldsymbol{T}$, where $T_{ij} = T_{ji} = -X^\star_{ij}$, $T_{ii} = T_{jj} = X^\star_{ij}$ and $T_{kl} = 0$ for all other components $k, l$. Note that $\boldsymbol{T} \succeq \mathbf{0}$ since $\boldsymbol{z}^\top \boldsymbol{T} \boldsymbol{z} = X^\star_{ij}(z_i - z_j)^2 \geq 0$ for all $\boldsymbol{z} \in \mathbb{R}^n$. We thus have $\boldsymbol{X}' = \boldsymbol{X}^\star + \boldsymbol{T} \succeq \boldsymbol{X}^\star$, which implies that $f'(\boldsymbol{X}', \boldsymbol{x}^\star) \geq f'(\boldsymbol{X}^\star, \boldsymbol{x}^\star)$ since $f'$ is a monotone lifting of $f$. In addition, the row and column sums of $\boldsymbol{X}^\star$ and $\boldsymbol{X}'$ coincide by construction, and thus $(\boldsymbol{X}', \boldsymbol{x}^\star)$ is also feasible in the RLT relaxation (140).

By construction, the matrix $\boldsymbol{X}'$ contains two non-zero off-diagonal elements less than the matrix $\boldsymbol{X}^\star$. An iterative application of the argument from the previous paragraph eventually results in an optimal diagonal matrix $\boldsymbol{X}'$, which by the constraints of the RLT relaxation (140) must coincide with $\mathrm{Diag}(\boldsymbol{x}^\star)$. This proves the statement of the theorem.

□

Theorem 22 allows us to replace the $n \times n$ decision matrix $\boldsymbol{X}$ in the RLT relaxation (140) of problem (139) with $\mathrm{Diag}(\boldsymbol{x})$ and thus significantly reduce the size of the optimization problem. Our numerical results (*cf.* Section 6.3) indicate that this can in turn result in dramatic savings in solution time. Another important consequence of Theorem 22 is given next.

**Corollary 16.** *If the function $f$ in problem* (139) *has a monotone lifting $f'$, then the optimal value of the corresponding RLT relaxation* (140) *coincides with the optimal value of the corresponding RLT/SDP relaxation.*

*Proof.* Recall that the RLT/SDP relaxation of problem (139) is equivalent to the RLT relaxation (140), except for the additional constraint that $\boldsymbol{X} \succeq \boldsymbol{x}\boldsymbol{x}^\top$. According to Theorem 22, it thus suffices to show that $\mathrm{Diag}(\boldsymbol{x}^\star) \succeq \boldsymbol{x}^\star \boldsymbol{x}^{\star\top}$ for the optimal solution $(\boldsymbol{X}^\star, \boldsymbol{x}^\star)$ considered in the theorem's statement.

Note that the constraints of the RLT relaxation (140) imply that $\boldsymbol{x}^\star \geq \boldsymbol{0}$ and $\sum_{i=1}^n x_i^\star = 1$. For any vector $\boldsymbol{y} \in \mathbb{R}^n$, we can thus construct a random variable $\tilde{Y}$ that attains the value $y_i$ with probability $x_i^\star$, $i = 1, \ldots, n$. We then have

$$\boldsymbol{y}^\top \mathrm{Diag}(\boldsymbol{x}^\star)\, \boldsymbol{y} \;=\; \mathbb{E}\big[\tilde{Y}^2\big] \;\geq\; \mathbb{E}\big[\tilde{Y}\big]^2 \;=\; \boldsymbol{y}^\top \big[\boldsymbol{x}^\star \boldsymbol{x}^{\star\top}\big]\, \boldsymbol{y},$$

since $\mathbb{V}\mathrm{ar}\big[\tilde{Y}\big] = \mathbb{E}\big[\tilde{Y}^2\big] - \mathbb{E}\big[\tilde{Y}\big]^2 \geq 0$. We thus conclude that $\mathrm{Diag}(\boldsymbol{x}^\star) - \boldsymbol{x}^\star \boldsymbol{x}^{\star\top} \succeq \boldsymbol{0}$, that is, the optimal solution $(\boldsymbol{X}^\star, \boldsymbol{x}^\star)$ considered by Theorem 22 vacuously satisfies the LMI constraint of the RLT/SDP relaxation.

□

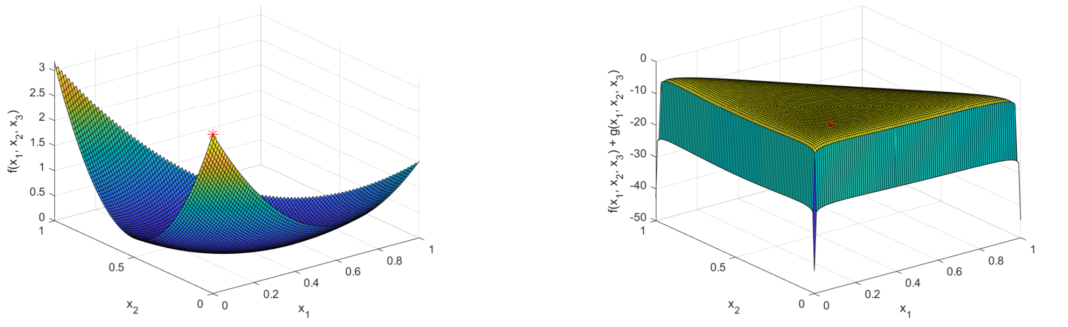
Corollary 16 shows that whenever $f$ has a monotone lifting, the RLT/SDP reformulation offers no advantage over the RLT relaxation (140) of problem (139).

### 6.3 Numerical Experiments

We compare our RLT formulation against standard RLT and RLT/SDP implementations on non-convex optimization problems over simplices. All experiments are run on an 8-th Generation Intel(R) Core(TM) i7-8750H processor using MATLAB 2018b MATLAB (2018), YALMIP R20200930 Löfberg (2004) and MOSEK 9.2.28 MOSEK ApS (2019).

We consider instances of problem (139) whose objective functions satisfy

$$f(\boldsymbol{x}) \;=\; \left\| \boldsymbol{D}\boldsymbol{Q}(\boldsymbol{x} - \frac{1}{n} \cdot \boldsymbol{1}) \right\|^2 \quad \text{and} \quad g(\boldsymbol{x}) \;=\; \frac{1}{n} \sum_{i=1}^n \ln(x_i),$$



(a) Plot of $f(\boldsymbol{x})$ over the simplex

(b) Plot of $f(\boldsymbol{x}) + g(\boldsymbol{x})$ over the simplex

Figure 19: *Example non-convex optimization instance for $n = 3$. The convex quadratic function $f$ is minimized at the center of the simplex and maximized at a vertex. The addition of the concave barrier function $g$ ensures that the overall maximum is attained in the interior of the simplex.*

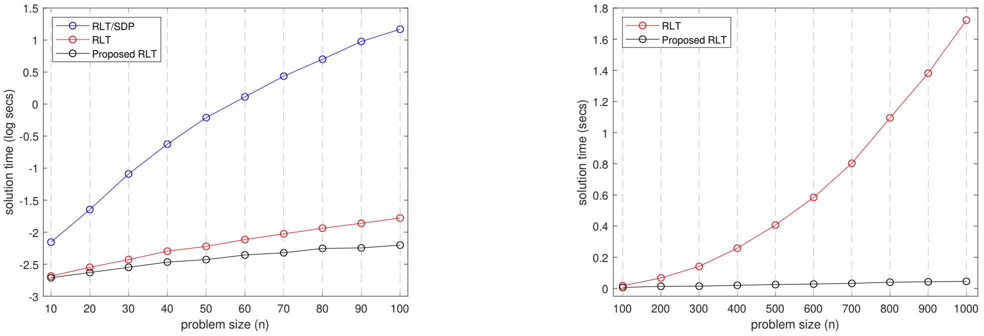


Figure 20: *Median solution times (in $\log_{10}$ secs, left, and secs, right) of our RLT formulation ('Proposed RLT') and the standard RLT and RLT/SDP formulations over 25 non-convex simplicial optimization instances.*

where $\boldsymbol{D} \in \mathbb{S}^n$ is a diagonal scaling matrix whose diagonal elements are chosen uniformly at random from the interval $[0, 10]$, $\boldsymbol{Q} \in \mathbb{R}^{n \times n}$ is a uniformly sampled rotation matrix Mezzadri (2006), and $\mathbf{1} \in \mathbb{R}^n$ is the vector of all ones (*cf.* Figure 19).

It follows from our discussion in Section 6.2 that the optimal values of the RLT and RLT/SDP relaxations coincide for the test instances considered in this section, and there are always optimal solutions $(\boldsymbol{X}^\star, \boldsymbol{x}^\star)$ satisfying $\boldsymbol{X}^\star = \text{Diag}(\boldsymbol{x}^\star)$. Figure 20 compares the runtimes of our RLT formulation, which replaces the matrix $\boldsymbol{X}$ with $\text{Diag}(\boldsymbol{x})$, with those of the standard RLT and RLT/SDP formulations. As expected, our RLT formulation substantially outperforms both alternatives.

## 6.A Theoretical Extensions

We extend our findings to instances of problem (137) whose feasible regions constitute the Cartesian product of two simplices (Appendix 6.A.1) and specific classes of bounded polyhedra (Appendix 6.A.2) and non-convex sets (Appendix 6.A.3), as well as to quadratic optimization problems whose objective functions do not directly admit monotone liftings (Appendix 6.A.4).

### 6.A.1 Cartesian Product of Two Simplices

Consider the following extension of problem (139),

$$\begin{aligned}\sup_{\boldsymbol{x},\boldsymbol{y}} \quad & f(\boldsymbol{x},\boldsymbol{y}) + g(\boldsymbol{x},\boldsymbol{y}) \\ \text{s.t.} \quad & \sum_{i=1}^{n_1} x_i = 1, \;\; \sum_{j=1}^{n_2} y_j = 1 \\ & \boldsymbol{x} \in \mathbb{R}_+^{n_1}, \boldsymbol{y} \in \mathbb{R}_+^{n_2},\end{aligned} \tag{141}$$

which optimizes the sum of a generic function $f$ and a (jointly) concave function $g$ over the Cartesian product of two simplices. The standard RLT reformulation for this problem introduces the $(n_1+n_2)^2$ auxiliary decision variables $\begin{pmatrix} \boldsymbol{X} & \boldsymbol{Z} \\ \boldsymbol{Z}^\top & \boldsymbol{Y} \end{pmatrix} \in \mathbb{S}^{n_1+n_2}$ as well as the following additional constraints:

$$\begin{array}{lll} \sum_{j=1}^{n_1} X_{ij} \;=\; \sum_{j=1}^{n_1} X_{ji} \;=\; x_i & & \forall i = 1,\ldots,n_1 \\ \sum_{j=1}^{n_2} Y_{ij} \;=\; \sum_{j=1}^{n_2} Y_{ji} \;=\; y_i & & \forall i = 1,\ldots,n_2 \\ \sum_{j=1}^{n_2} Z_{ij} \;=\; x_i, \;\; \sum_{j=1}^{n_1} Z_{jk} \;=\; y_k & & \forall i = 1,\ldots,n_1, \; \forall k = 1,\ldots,n_2 \\ X_{ij}, Y_{kl}, Z_{ik} \geq 0 & & \forall i,j = 1,\ldots,n_1, \; \forall k,l = 1,\ldots,n_2. \end{array} \tag{142}$$

Using similar arguments as in Section 6.2, we now show that a significant number of decision variables can be removed from the RLT relaxation if function $f$ in problem (141) has a monotone lifting $f'$.

**Theorem 23.** *If the function $f$ in problem (141) has a monotone lifting $f'$, then the corresponding RLT relaxation has an optimal solution $(\boldsymbol{X}^\star, \boldsymbol{Y}^\star, \boldsymbol{Z}^\star, \boldsymbol{x}^\star, \boldsymbol{y}^\star)$ satisfying $\boldsymbol{X}^\star = \mathrm{Diag}(\boldsymbol{x}^\star)$ and $\boldsymbol{Y}^\star = \mathrm{Diag}(\boldsymbol{y}^\star)$.*

*Proof.* Fix any optimal solution $(\boldsymbol{X}^\star, \boldsymbol{Y}^\star, \boldsymbol{Z}^\star, \boldsymbol{x}^\star, \boldsymbol{y}^\star)$ to problem (141). The statement follows if we apply the arguments of the proof of Theorem 22 to the blocks $\boldsymbol{X}^\star$ and $\boldsymbol{Y}^\star$ of the matrix $\begin{pmatrix} \boldsymbol{X}^\star & \boldsymbol{Z}^\star \\ \boldsymbol{Z}^{\star\top} & \boldsymbol{Y}^\star \end{pmatrix}$. □ □

Note that in Theorem 23 we cannot apply the same arguments to the blocks $\boldsymbol{Z}^\star$ and $\boldsymbol{Z}^{\star\top}$ since they do not lie on the diagonal of the main matrix. We can furthermore show that, as in the case of a single simplex, the RLT and RLT/SDP relaxations are equally tight for problem (141).

**Corollary 17.** *If the function $f$ in problem* (141) *has a monotone lifting $f'$, then the optimal value of the corresponding RLT relaxation coincides with the optimal value of the corresponding RLT/SDP relaxation.*

*Proof.* Given an optimal solution $(\boldsymbol{X}^\star, \boldsymbol{Y}^\star, \boldsymbol{Z}^\star, \boldsymbol{x}^\star, \boldsymbol{y}^\star)$ to problem (141) that satisfies $\boldsymbol{X}^\star = \mathrm{Diag}(\boldsymbol{x}^\star)$ and $\boldsymbol{Y}^\star = \mathrm{Diag}(\boldsymbol{y}^\star)$, as justified by Theorem 23, the statement of the corollary follows if we show that

$$\begin{pmatrix} \mathrm{Diag}(\boldsymbol{x}^\star) & \boldsymbol{Z}^\star \\ \boldsymbol{Z}^{\star\top} & \mathrm{Diag}(\boldsymbol{y}^\star) \end{pmatrix} \succeq \begin{pmatrix} \boldsymbol{x}^\star \\ \boldsymbol{y}^\star \end{pmatrix} \begin{pmatrix} \boldsymbol{x}^\star \\ \boldsymbol{y}^\star \end{pmatrix}^\top .$$

For any $\boldsymbol{a} \in \mathbb{R}^{n_1}$ and $\boldsymbol{b} \in \mathbb{R}^{n_2}$, we have that

$$\begin{aligned} &\begin{pmatrix} \boldsymbol{a} \\ \boldsymbol{b} \end{pmatrix}^\top \begin{pmatrix} \mathrm{Diag}(\boldsymbol{x}^\star) - \boldsymbol{x}^\star\boldsymbol{x}^{\star\top} & \boldsymbol{Z}^\star - \boldsymbol{x}^\star\boldsymbol{y}^{\star\top} \\ \boldsymbol{Z}^{\star\top} - \boldsymbol{y}^\star\boldsymbol{x}^{\star\top} & \mathrm{Diag}(\boldsymbol{y}^\star) - \boldsymbol{y}^\star\boldsymbol{y}^{\star\top} \end{pmatrix} \begin{pmatrix} \boldsymbol{a} \\ \boldsymbol{b} \end{pmatrix} \\ &= \boldsymbol{a}^\top \left(\mathrm{Diag}(\boldsymbol{x}^\star) - \boldsymbol{x}^\star\boldsymbol{x}^{\star\top}\right) \boldsymbol{a} \; + \; \boldsymbol{b}^\top \left(\mathrm{Diag}(\boldsymbol{y}^\star) - \boldsymbol{y}^\star\boldsymbol{y}^{\star\top}\right) \boldsymbol{b} \\ &\qquad + \; 2\boldsymbol{a}^\top \left(\boldsymbol{Z}^\star - \boldsymbol{x}^\star\boldsymbol{y}^{\star\top}\right) \boldsymbol{b} \qquad \geq \; 0. \end{aligned} \tag{143}$$

The last inequality follows from similar arguments as in the proof of Corollary 16, which show that

$$\boldsymbol{a}^\top \left(\mathrm{Diag}(\boldsymbol{x}^\star) - \boldsymbol{x}^\star\boldsymbol{x}^{\star\top}\right) \boldsymbol{a} \; = \; \mathbb{V}\mathrm{ar}\big[\tilde{A}\big], \;\; \boldsymbol{b}^\top \left(\mathrm{Diag}(\boldsymbol{y}^\star) - \boldsymbol{y}^\star\boldsymbol{y}^{\star\top}\right) \boldsymbol{b} \; = \; \mathbb{V}\mathrm{ar}\big[\tilde{B}\big]$$

for the random variables $\tilde{A}$ and $\tilde{B}$ that attain the values $a_i$ and $b_j$ with probabilities $x_i^\star$ and $y_j^\star$, $i = 1, \ldots, n_1$ and $j = 1, \ldots, n_2$, respectively. Likewise, we observe that

$$\boldsymbol{a}^\top \left(\boldsymbol{Z}^\star - \boldsymbol{x}^\star\boldsymbol{y}^{\star\top}\right) \boldsymbol{b} \; = \; \mathbb{E}\big[\tilde{A}\tilde{B}\big] - \mathbb{E}\big[\tilde{A}\big]\mathbb{E}\big[\tilde{B}\big],$$

where we assume that the random variable $\tilde{A}\tilde{B}$ attains the values $a_i b_j$ with probability $Z^\star_{ij}$, $i = 1,\ldots,n_1$ and $j = 1,\ldots,n_2$. Note that this joint probability distribution is consistent with our marginal distributions specified above since the RLT constraints (142) guarantee that $Z^\star_{ij} \geq 0$, $\sum_{k=1}^{n_2} Z^\star_{ik} = x^\star_i$ and $\sum_{k=1}^{n_1} Z^\star_{kj} = y^\star_j$. The previous arguments imply that the sum on the left-hand side of the inequality in (143) evaluates to

$$\mathbb{V}\mathrm{ar}\big[\tilde{A}\big] \;+\; \mathbb{V}\mathrm{ar}\big[\tilde{B}\big] \;+\; 2\mathbb{C}\mathrm{ov}\big[\tilde{A},\, \tilde{B}\big] \;=\; \mathbb{V}\mathrm{ar}\big[\tilde{A} + \tilde{B}\big] \;\geq\; 0,$$

which concludes the proof. □ □

We emphasize that one can readily construct counterexamples which show that our results in this section do *not* extend to three or more simplices.

### 6.A.2 Linear Constraints

Consider a generic instance of problem (137) whose feasible region is bounded, and let the columns of the matrix $\boldsymbol{V} \in \mathbb{R}^{n\times p}$ denote the $p$ extreme points of the feasible region. Problem (137) is equivalent to

$$\begin{array}{ll} \sup\limits_{\boldsymbol{x}'} & f(\boldsymbol{V}\boldsymbol{x}') + g(\boldsymbol{V}\boldsymbol{x}') \\ \text{s.t.} & \sum\limits_{i=1}^{p} x'_i = 1 \\ & \boldsymbol{x}' \in \mathbb{R}^p_+, \end{array} \tag{144}$$

which is an instance of problem (139) studied in Section 6.2. The fourth statement of Proposition 20 implies that the objective component $f$ in problem (144) has a monotone lifting whenever the component $f$ in the original problem (137) has one. Note that the number $p$ of decision variables in problem (144) is typically exponential in the number $m$ of constraints in the original problem (137). Notable exceptions exist, however, such as bijective transformations of the 1-norm ball, $\{\boldsymbol{T}\boldsymbol{x} \,:\, \boldsymbol{x} \in \mathbb{R}^n,\ \ \|\boldsymbol{x}\|_1 \leq 1\}$ with $\boldsymbol{T} \in \mathbb{R}^{n\times n}$ invertible, which have $p = 2n$ extreme points, as well as the unit simplex, $\{\boldsymbol{x} \in \mathbb{R}^n_+ \,:\, \sum_i x_i \leq 1\}$, which has $p = n + 1$ extreme points.

The RLT relaxation of the original problem (137) is typically strictly weaker than the corresponding RLT/SDP relaxation. Corollary 16 shows, however, that both relaxations are equally tight in the lifted problem (144), and Theorem 22 allows us to replace the decision matrix $\boldsymbol{X}' \in \mathbb{S}^p$ in the lifted problem with $\mathrm{Diag}(\boldsymbol{x}')$. We next compare the tightness of the RLT/SDP relaxation of the original problem (137) with the RLT relaxation of the lifted problem (144).

**Theorem 24.** *The RLT relaxation of the lifted problem* (144) *is at least as tight as the RLT/SDP relaxation of the original problem* (137).

*Proof.* The statement of the theorem follows if for any feasible solution $(\boldsymbol{X}', \boldsymbol{x}')$ of the RLT relaxation of the lifted problem (144) we can construct a feasible solution $(\boldsymbol{X}, \boldsymbol{x})$ of the RLT/SDP relaxation of the original problem (137) that attains a weakly larger objective value. To this end, fix any feasible solution $(\boldsymbol{X}', \boldsymbol{x}')$ of the RLT relaxation of problem (144) and set $(\boldsymbol{X}, \boldsymbol{x}) = (\boldsymbol{V}\,\mathrm{Diag}(\boldsymbol{x}')\,\boldsymbol{V}^\top, \boldsymbol{V}\boldsymbol{x}')$. Intuitively speaking, this choice of $(\boldsymbol{X}, \boldsymbol{x})$ interprets $\boldsymbol{x}'$ as the convex weights of the vertices $\boldsymbol{V} = [\boldsymbol{V}_1 \ldots \boldsymbol{V}_p]$ of the feasible region of problem (137) and therefore sets $\boldsymbol{x} = \boldsymbol{V}\boldsymbol{x}'$. Moreover, note that $\mathrm{diag}(\boldsymbol{x}')$ is an 'optimal' representation of $\boldsymbol{x}'\boldsymbol{x}'^\top$ in the RLT relaxation of problem (144). Since $\boldsymbol{x}'\boldsymbol{x}'^\top$ corresponds to $\boldsymbol{V}\boldsymbol{x}'\boldsymbol{x}'^\top\boldsymbol{V}^\top$ in problem (137), we thus set $\boldsymbol{X} = \boldsymbol{V}\mathrm{diag}(\boldsymbol{x}')\boldsymbol{V}^\top$.

We first note that the objective value of $(\boldsymbol{X}, \boldsymbol{x})$ in the relaxation of (137) is at least as large as the objective value of $(\boldsymbol{X}', \boldsymbol{x}')$ in the relaxation of (144):

$$\begin{aligned} f'(\boldsymbol{X}, \boldsymbol{x}) + g(\boldsymbol{x}) &= f'(\boldsymbol{V}\,\mathrm{Diag}(\boldsymbol{x}')\,\boldsymbol{V}^\top, \boldsymbol{V}\boldsymbol{x}') + g(\boldsymbol{V}\boldsymbol{x}') \\ &\geq f'(\boldsymbol{V}\boldsymbol{X}'\,\boldsymbol{V}^\top, \boldsymbol{V}\boldsymbol{x}') + g(\boldsymbol{V}\boldsymbol{x}'). \end{aligned}$$

Here, the left-hand side represents the objective value of $(\boldsymbol{X}, \boldsymbol{x})$ in the RLT/SDP relaxation of problem (137) and the right-hand side represents the objective value of $(\boldsymbol{X}', \boldsymbol{x}')$ in the RLT relaxation of problem (144) if we adopt the monotone lifting proposed in the proof of statement 4 of Proposition 20. The inequality holds since $\mathrm{Diag}(\boldsymbol{x}') \succeq \boldsymbol{X}'$, which can be shown using similar arguments as in the proof of Theorem 22.

To see that $(\boldsymbol{X}, \boldsymbol{x})$ is feasible for the RLT/SDP relaxation of problem (137), we first note that

$$\boldsymbol{A}\boldsymbol{x} \;=\; \boldsymbol{A}\boldsymbol{V}\boldsymbol{x}' \;\leq \boldsymbol{b},$$

since $\boldsymbol{V}\boldsymbol{x}'$ is a convex combination of the vertices of the polyhedron $\{\boldsymbol{x} \in \mathbb{R}^n : \boldsymbol{A}\boldsymbol{x} \leq \boldsymbol{b}\}$. Moreover, we have

$$\boldsymbol{X} \;=\; \boldsymbol{V}\,\mathrm{Diag}(\boldsymbol{x}')\,\boldsymbol{V}^\top \;\succeq\; \boldsymbol{V}\boldsymbol{x}'\boldsymbol{x}'^\top\,\boldsymbol{V}^\top \;=\; \boldsymbol{x}\boldsymbol{x}^\top,$$

where the inequality holds since $\mathrm{Diag}(\boldsymbol{x}') \succeq \boldsymbol{x}'\boldsymbol{x}'^\top$ due to similar arguments as in the proof of Corollary 16. In the remainder of the proof, we show that the RLT constraints (138) hold as well. To this end, we note that

$$\begin{aligned}
& \boldsymbol{a}_i^\top \boldsymbol{X} \boldsymbol{a}_j - (b_i \boldsymbol{a}_j + b_j \boldsymbol{a}_i)^\top \boldsymbol{x} + b_i b_j \\
= \;& \boldsymbol{a}_i^\top \boldsymbol{V} \operatorname{Diag}(\boldsymbol{x}') \, \boldsymbol{V}^\top \boldsymbol{a}_j - (b_i \boldsymbol{a}_j + b_j \boldsymbol{a}_i)^\top \boldsymbol{V} \boldsymbol{x}' + b_i b_j \\
= \;& \boldsymbol{a}_i^\top \left( \sum_{\ell=1}^{p} x'_\ell \boldsymbol{V}_\ell \boldsymbol{V}_\ell^\top \right) \boldsymbol{a}_j - (b_i \boldsymbol{a}_j + b_j \boldsymbol{a}_i)^\top \left( \sum_{\ell=1}^{p} x'_\ell \boldsymbol{V}_\ell \right) + b_i b_j, \qquad (145)
\end{aligned}$$

where $\boldsymbol{V}_\ell$ denotes the $\ell$-th column of $\boldsymbol{V}$. To see that (145) is indeed non-negative, we distinguish between four cases based on the values of $b_i$ and $b_j$.

Case 1: $b_i, b_j = 0$. In this case, the expression (145) simplifies to $\sum_{\ell=1}^{p} x'_\ell \cdot (\boldsymbol{a}_i^\top \boldsymbol{V}_\ell) \, (\boldsymbol{a}_j^\top \boldsymbol{V}_\ell)$, which constitutes a sum of non-negative terms since $\boldsymbol{x}' \geq \boldsymbol{0}$ as well as $\boldsymbol{a}_i^\top \boldsymbol{V}_\ell \leq b_i = 0$ and $\boldsymbol{a}_j^\top \boldsymbol{V}_\ell \leq b_j = 0$.

Case 2: $b_i \neq 0, b_j = 0$ or $b_i = 0, b_j \neq 0$. We assume that $b_i \neq 0, b_j = 0$; the other case follows by symmetry. Assume further that $b_i > 0$; the case where $b_i < 0$ can be shown similarly. Dividing the expression (145) by $b_i$ and removing the terms that contain $b_j$ yields

$$\sum_{\ell=1}^{p} x'_\ell \cdot \left( \frac{\boldsymbol{a}_i^\top \boldsymbol{V}_\ell}{b_i} - 1 \right) \cdot \boldsymbol{V}_\ell^\top \boldsymbol{a}_j,$$

and this expression constitutes a sum of non-negative terms since $x'_\ell \geq 0$ multiplies the product of two non-positive terms: We have $\boldsymbol{a}_i^\top \boldsymbol{V}_\ell / b_i \leq 1$ since $\boldsymbol{a}_i^\top \boldsymbol{V}_\ell \leq b_i$, and we have $\boldsymbol{V}_\ell^\top \boldsymbol{a}_j \leq b_j = 0$.

Case 3: $b_i, b_j > 0$ or $b_i, b_j < 0$. We assume that $b_i, b_j > 0$; the other case follows similarly. Dividing the expression (145) by $b_i b_j > 0$ yields

$$\sum_{\ell=1}^{p} x'_\ell \, (\alpha_\ell \beta_\ell - \alpha_\ell - \beta_\ell) + 1 \quad \text{with } \alpha_\ell = \frac{\boldsymbol{a}_i^\top \boldsymbol{V}_\ell}{b_i} \text{ and } \beta_\ell = \frac{\boldsymbol{a}_j^\top \boldsymbol{V}_\ell}{b_j}, \qquad (146)$$

where $\alpha_\ell, \beta_\ell \leq 1$ since $\boldsymbol{a}_i^\top \boldsymbol{V}_\ell \leq b_i$ and $\boldsymbol{a}_j^\top \boldsymbol{V}_\ell \leq b_j$. Since

$$\min \left\{ \alpha\beta - \alpha - \beta \, : \, \alpha, \beta \leq 1, \;\; \alpha, \beta \in \mathbb{R} \right\} \; = \; -1,$$

each multiplier of $x'_\ell$ in (146) is bounded from below by $-1$, which implies that the sum involving $x'_\ell$ is bounded from below by $-1$, and thus the overall expression (146) is non-negative as desired.

Case 4: $b_i > 0, b_j < 0$ or $b_i < 0, b_j > 0$. We assume that $b_i > 0, b_j < 0$; the other case follows by symmetry. Dividing (145) by $b_i b_j < 0$ yields (146), which now needs to be non-*positive*. Note that $\alpha_\ell \leq 1$ while $\beta_\ell \geq 1$, since $b_j < 0$. The statement now follows from an argument analogous to the previous case as

$$\max \{\alpha\beta - \alpha - \beta \; : \; \alpha \leq 1, \;\; \beta \geq 1 \;\; \alpha, \beta \in \mathbb{R}\} \; = \; -1,$$

which implies that the overall expression (146) is non-positive as desired. □ □

For problems with a moderate number $p$ of vertices, the RLT relaxation of the lifted problem (144), which involves $p$ non-negative decision variables and a single constraint, might be easier to solve than the RLT/SDP relaxation of the original problem (137), which involves $\mathcal{O}(n^2)$ decision variables, $\mathcal{O}(m^2)$ constraints as well as a restriction to the semidefinite cone. In addition, as proven in Theorem 24, the RLT relaxation of the lifted problem is always at least as tight as the standard RLT/SDP relaxation.

### 6.A.3 Nonlinear Constraints

We now study the following generalization of problem (139):

$$\begin{array}{ll} \sup\limits_{\boldsymbol{x}} & f(\boldsymbol{x}) + g(\boldsymbol{x}) \\ \text{s. t.} & \displaystyle\sum_{i=1}^n x_i = 1 \\ & f_i(\boldsymbol{x}) + g_i(\boldsymbol{x}) \geq 0 \qquad \forall i = 1, \dots, m \\ & \boldsymbol{x} \in \mathbb{R}^n_+. \end{array} \tag{147}$$

Here, $f : \mathbb{R}^n \mapsto \mathbb{R}$ as well as $f_i : \mathbb{R}^n \mapsto \mathbb{R}$ are assumed to admit monotone liftings, and $g : \mathbb{R}^n \mapsto \mathbb{R}$ as well as $g_i : \mathbb{R}^n \mapsto \mathbb{R}$ are concave, $i = 1, \dots m$. If we replace both $f$ and $f_i : \mathbb{R}^n \mapsto \mathbb{R}$ with their respective monotone liftings $f'$ and $f'_i$, $i = 1, \dots m$, then one can readily verify that the RLT relaxation

$$\begin{array}{lll} \sup\limits_{\boldsymbol{X}, \boldsymbol{x}} & f'(\boldsymbol{X}, \boldsymbol{x}) + g(\boldsymbol{x}) & \\ \text{s. t.} & \displaystyle\sum_{j=1}^n X_{ij} = \sum_{j=1}^n X_{ji} = x_i & \forall i = 1, \dots, n \\ & \displaystyle\sum_{i=1}^n x_i = 1 & \\ & f'_i(\boldsymbol{X}, \boldsymbol{x}) + g_i(\boldsymbol{x}) \geq 0 & \forall i = 1, \dots, m \\ & \boldsymbol{X} \geq \boldsymbol{0}, \;\; \boldsymbol{X} \in \mathbb{S}^n, \;\; \boldsymbol{x} \in \mathbb{R}^n_+ & \end{array} \tag{148}$$

is optimized by a solution $(\boldsymbol{X}^\star, \boldsymbol{x}^\star)$ that satisfies $\boldsymbol{X}^\star = \mathrm{Diag}(\boldsymbol{x}^\star)$ as well as $\boldsymbol{X}^\star \succeq \boldsymbol{x}^\star \boldsymbol{x}^{\star\top}$. The definition of monotone liftings implies that (148) is a convex optimization problem.

A special case of problem (147) arises when the constraint functions $f_i$, $i = 1, \ldots, m$, are absent and when $f$ and $g_i$, $i = 1, \ldots, m$, depend on separate parts of the decision vector $\boldsymbol{x}$, that is, if problem (147) can be written as

$$\begin{array}{ll} \sup\limits_{\boldsymbol{x}, \boldsymbol{y}} & f(\boldsymbol{x}) + g(\boldsymbol{x}, \boldsymbol{y}) \\ \text{s.t.} & \displaystyle\sum_{i=1}^{n_1} x_i = 1 \\ & \boldsymbol{x} \in \mathbb{R}^{n_1}_+, \quad \boldsymbol{y} \in \mathcal{Y}, \end{array}$$

where $\mathcal{Y} \subseteq \mathbb{R}^{n_2}$ denotes the feasible region for the decision vector $\boldsymbol{y}$. Omitting the RLT constraints that involve cross-products of the constraints involving $\boldsymbol{x}$ and the constraints involving $\boldsymbol{y}$, Theorem 22 and Corollary 16 imply that the RLT relaxation

$$\begin{array}{ll} \sup\limits_{\boldsymbol{X}, \boldsymbol{x}, \boldsymbol{y}} & f'(\boldsymbol{X}, \boldsymbol{x}) + g(\boldsymbol{x}, \boldsymbol{y}) \\ \text{s.t.} & \displaystyle\sum_{j=1}^{n_1} X_{ij} = \sum_{j=1}^{n_1} X_{ji} = x_i \qquad \forall i = 1, \ldots, n_1 \\ & \displaystyle\sum_{i=1}^{n_1} x_i = 1 \\ & \boldsymbol{X} \geq \boldsymbol{0}, \quad \boldsymbol{X} \in \mathbb{S}^{n_1}, \quad \boldsymbol{x} \in \mathbb{R}^{n_1}_+, \quad \boldsymbol{y} \in \mathcal{Y} \end{array}$$

has an optimal solution $(\boldsymbol{X}^\star, \boldsymbol{x}^\star, \boldsymbol{y}^\star)$ satisfying $\boldsymbol{X}^\star = \mathrm{Diag}(\boldsymbol{x}^\star)$.

### 6.A.4 Standard Quadratic Optimization

A standard quadratic optimization problem maximizes a (usually non-convex) quadratic function $\varphi(\boldsymbol{x}) = \boldsymbol{x}^\top \boldsymbol{Q} \boldsymbol{x} + \boldsymbol{q}^\top \boldsymbol{x} + q$, $\boldsymbol{Q} \in \mathbb{S}^n$, $\boldsymbol{q} \in \mathbb{R}^n$ and $q \in \mathbb{R}$, over the probability simplex. Since $\boldsymbol{Q} \not\preceq \boldsymbol{0}$ in general, our results from Section 6.2 are not directly applicable. By decomposing $\boldsymbol{Q}$ into $\boldsymbol{Q} = \boldsymbol{Q}^+ - \boldsymbol{Q}^-$ such that $\boldsymbol{Q}^+, \boldsymbol{Q}^- \succeq \boldsymbol{0}$, however, we obtain an instance of problem (139) where $f(\boldsymbol{x}) = \boldsymbol{x}^\top \boldsymbol{Q}^+ \boldsymbol{x} + \boldsymbol{q}^\top \boldsymbol{x} + q$ and $g(\boldsymbol{x}) = -\boldsymbol{x}^\top \boldsymbol{Q}^- \boldsymbol{x}$. The first statement of Corollary 15 then allows us to apply Theorem 22 and Corollary 16 to the reformulated standard quadratic optimization problem. It is worth noting that different decomposition schemes could lead to different RLT relaxations of varying tightness. For a review of decomposition schemes, we refer to Fampa et al. (2017), Park (2016).

Instead of decomposing the objective function of the standard quadratic optimization problem and utilizing the results from Section 6.2, one can alternatively apply the RLT or RLT/SDP relaxation directly to the original standard quadratic optimization problem. Our numerical results indicate that for the eigenvalue-based matrix decomposition, the RLT/SDP relaxation outperforms our formulation in terms of tightness, whereas the RLT relaxation and our formulation are in general incomparable, that is, either formulation can be superior for a given instance. In terms of runtime, on the other hand, our formulation outperforms the RLT and RLT/SDP relaxations. This is not surprising as our formulation optimizes over $n$ decision variables, whereas the RLT and RLT/SDP relaxations involve $\mathcal{O}(n^2)$ decision variables due to the presence of the decision matrix $\boldsymbol{X}$.

## 6.B Additional Numerical Experiments

We compare our RLT formulation against standard RLT and RLT/SDP implementations on non-convex optimization problems over polyhedra (Appendix 6.B.1) as well as on indefinite quadratic optimization problems over simplices (Appendix 6.B.2).

### 6.B.1 Non-Convex Optimization over Polyhedra

We consider instances of problem (144) where

$$f(\boldsymbol{x}) \;=\; \|\boldsymbol{D}\boldsymbol{Q}\boldsymbol{x}\|^2 \quad \text{and} \quad g(\boldsymbol{x}) \;=\; \sum_{i=1}^{n} \ln(1 - x_i^2).$$

Here, $\boldsymbol{D} \in \mathbb{S}^n$ and $\boldsymbol{Q} \in \mathbb{R}^{n \times n}$ are generated as in Section 6.3, and the feasible region (prior to its lifting) is the hypercube $[-1, 1]^n$ in $\mathbb{R}^n$. Following our discussion in Appendix 6.A.2, our RLT reformulation operates on the lifted space $\mathbb{R}^{2^n}$, where the feasible region is described by a probability simplex whose vertices correspond to the vertices of the hypercube, whereas the standard RLT and RLT/SDP reformulations operate directly on the formulation (137) that involves of $2n$ halfspaces in $\mathbb{R}^n$.

Figure 21 reports the optimality gaps and solution times for instances with $n = 1, \ldots, 10$ decision variables. As expected from Theorem 24, our RLT formulation outperforms both RLT and RLT/SDP in terms of the objective value. Interestingly, the outperformance over RLT is substantial and grows with the dimension $n$. On the other hand, since the number of decision variables in our RLT reformulation grows exponentially in $n$, our reformulation is only viable for small problem instances with up to $n = 10$ decision variables.

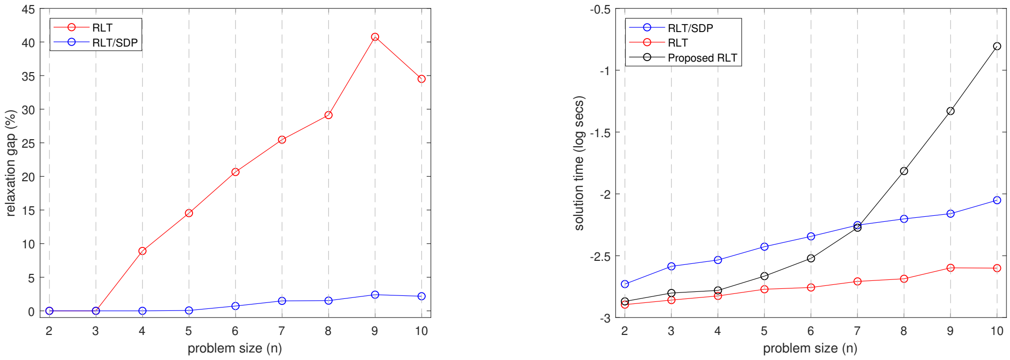


Figure 21: *Median relaxation gaps (left) and solution times (right, in* $\log_{10}$ *secs) of our proposed RLT formulation as well as the standard RLT and RLT/SDP formulations over 25 non-convex hypercubic optimization instances. The relaxation gaps are recorded as* $100\% \cdot (z - z^\star)/\max\{1, |z^\star|\}$*, where* $z^\star$ *is the optimal value of the proposed RLT relaxation, and* $z$ *refers to the optimal value of the standard RLT or RLT/SDP relaxations.*

With its exponential number of vertices, the hypercubic feasible region of our previous experiment constitutes the least favourable setting for our proposed RLT formulation. We next study instances of problem (144) where $k < n$ of the decision variables (hereafter $\boldsymbol{y}$) reside in a hypercube, whereas the remaining $n - k$ decision variables (hereafter $\boldsymbol{x}$) are restricted to a simplex. In this case, the feasible region of the original problem is described by $n + k + 2$ halfspaces in $\mathbb{R}^n$, whereas the feasible region of the lifted problem constitutes a simplex with $2^k \cdot (n - k)$ vertices. The objective function is described by

$$f(\boldsymbol{x}, \boldsymbol{y}) = \left\| \boldsymbol{DQ} \begin{pmatrix} \boldsymbol{x} - \frac{1}{n-k} \cdot \mathbf{1} \\ \boldsymbol{y} \end{pmatrix} \right\|^2, \quad g(\boldsymbol{x}, \boldsymbol{y}) \;=\; \frac{1}{n-k} \sum_{i=1}^{n-k} \ln(x_i) + \sum_{i=1}^{k} \ln(1 - y_i^2),$$

where $\boldsymbol{D} \in \mathbb{S}^n$ and $\boldsymbol{Q} \in \mathbb{R}^{n \times n}$ are generated as before. Figure 22 reports the runtimes of our proposed RLT reformulation as well as the standard RLT and RLT/SDP formulations for problem instances with $k = 3$ and $n = 10, 20, \ldots, 150$ decision variables. Our RLT reformulation significantly outperforms the RLT/SDP formulation in terms of runtimes, and it also improves upon the standard RLT formulation for $n \geq 60$. We note that in terms of the relaxation gaps, our RLT reformulation also outperforms both standard RLT (by about 2%) and RLT/SDP (by about 0.5%), which is in accordance with Theorem 24. Since the differences are small (around 1%), however, we do not illustrate them in the graph.

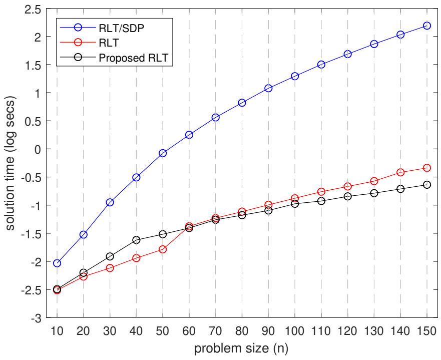
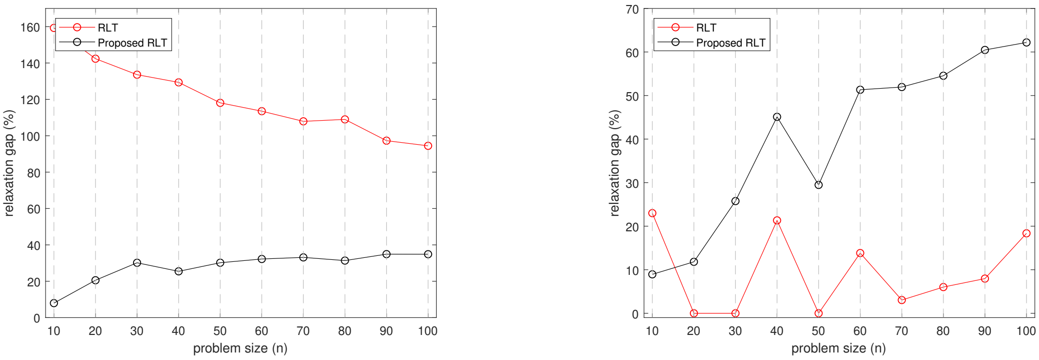


Figure 22: *Median solution times (in $\log_{10}$ secs) of our RLT formulation ('Proposed RLT') as well as the standard RLT and RLT/SDP formulations over 25 non-convex optimization instances whose feasible region emerges from the Cartesian product of a simplex and a hypercube.*



(a) Type 1 problems

(b) Type 2 problems

Figure 23: *Median relaxation gaps of our proposed RLT formulation ('Proposed RLT') and the standard RLT formulation over 25 standard quadratic optimization instances. The relaxation gaps are recorded as* $100\% \cdot (z - z^\star)/\max\{1, |z^\star|\}$*, where* $z^\star$ *is the optimal value of the RLT/SDP relaxation and* $z$ *refers to the optimal value of our RLT formulation or the standard RLT formulation.*

### 6.B.2 Standard Quadratic Optimization

In our final experiment, we maximize an indefinite quadratic function $\varphi(\boldsymbol{x}) = \boldsymbol{x}^\top \boldsymbol{Q} \boldsymbol{x}$ over the simplex in $\mathbb{R}^n$. To this end, we select $\boldsymbol{Q} = \boldsymbol{V}\boldsymbol{D}\boldsymbol{V}^\top$, where $\boldsymbol{D} \in \mathbb{S}^n$ is a diagonal scaling matrix whose diagonal elements are sampled uniformly at random from the interval $[-7.5, 2.5]$ (type 1) or $[-5, 5]$ (type 2), and $\boldsymbol{V} \in \mathbb{R}^{n \times n}$ is a uniformly sampled rotation matrix Mezzadri (2006).

Following our discussion in Appendix 6.A.4, our RLT formulation decomposes the function $\varphi$ into a convex part $f(\boldsymbol{x}) = \boldsymbol{x}^\top \boldsymbol{V}\boldsymbol{D}^+\boldsymbol{V}^\top \boldsymbol{x}$ and a concave part $g(\boldsymbol{x}) = \boldsymbol{x}^\top \boldsymbol{V}\boldsymbol{D}^-\boldsymbol{V}^\top \boldsymbol{x}$, where $\boldsymbol{D}^+$ and $\boldsymbol{D}^-$ contain the positive and negative eigenvalues of $\boldsymbol{D}$, respectively. In contrast, the standard RLT and RLT/SDP formulations directly operate on the function $\varphi$. Figures 23

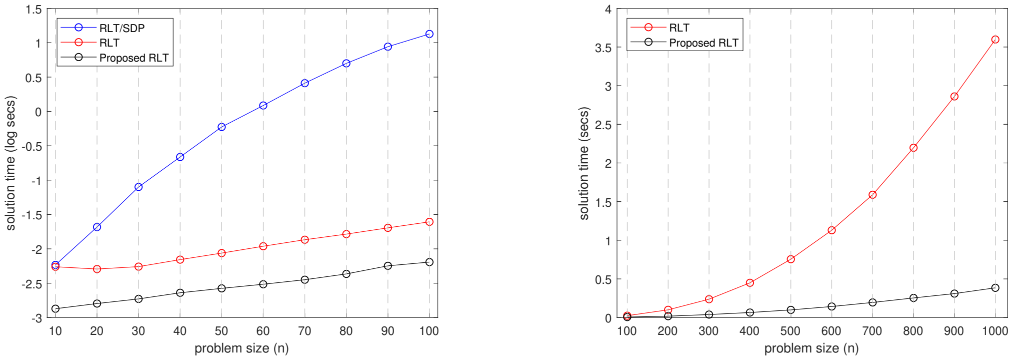


Figure 24: *Median solution times (in* $\log_{10}$ *secs, left, and secs, right) of our RLT formulation ('Proposed RLT') and the standard RLT and RLT/SDP formulations over 25 standard quadratic optimization instances.*

and 24 compare the three approaches in terms of objective values and the required runtimes. The figures show that RLT/SDP tends to provide the tightest relaxations and that our proposed RLT formulation offers tighter relaxations than the standard RLT formulation on type 1 instances, whereas the situation is reversed for type 2 instances. In terms of runtimes, on the other hand, our RLT formulation clearly dominates both alternatives as expected.

# Chapter IV

# Conclusions and Future Work

This thesis explores three interrelated themes. The first two focus on safeguarding data-driven decision making, where safeguards are introduced in the form of ethical constraints and various notions of robustness. My work within these themes frequently involves formulating and solving problems as instances of nonconvex and robust optimization. For example, in the first theme, the task of optimizing a probability distribution for sampling noise that satisfies additive differential privacy gives rise to an infinite-dimensional robust optimization problem. Constructing distributionally and adversarially robust logistic classifiers results in two-stage adjustable robust optimization formulations. Similarly, the Wasserstein classification and regression models lead to optimization problems with constraints that upper bound the supremum of convex (or difference-of-convex) functions, closely mirroring robust optimization techniques for deriving counterparts of constraints that are convex in the uncertain parameters. The third theme of this thesis extends this line of work by contributing new modeling and solution tools for such settings. Overall, I view this thesis as a collection of contributions to the modeling, analysis, and solution of safeguarded data-driven decision making problems.

I am currently working on, or initiating, several projects that extend my research on safeguarded data-driven decision making. These projects broaden both the methodological directions I pursue and the application domains I study. A few are described below.

In the context of decision making subject to ethical constraints, I am starting projects that parallel my earlier work on privacy, now approached from a fairness perspective, where I seek optimal decisions subject to fairness constraints. A notable example is my ongoing collaboration with Bill Tang, Çağıl Koçyiğit, Phebe Vayanos, and Wolfram Wiesemann on optimal and fair housing allocation via weakly coupled dynamic programs, where we impose fairness constraints and subsequently optimize policies for housing allocation. Similarly, in my postdoctoral studies, I will focus on modeling interpretability as a constraint and minimizing the price of interpretability in various tasks.

Within the theme of robust machine learning, I am interested in reducing the conservatism of existing distributionally robust machine learning methods. In particular, I have started a collaboration with Daniel Kuhn, in which we aim to understand how external data sources can be systematically incorporated to achieve this.

From the perspective of nonconvex optimization, I aim to address problem classes that have traditionally been difficult to solve by building on my previous work. To this end, I am currently collaborating with Aharon Ben-Tal and Wolfram Wiesemann on maximizing the sum of maxima of convex functions. We demonstrate that many challenging problems in machine learning and operations management can be formulated in this way, and we develop strong solution methods based on earlier techniques.

# Bibliography

Abadi, M., Chu, A., Goodfellow, I., McMahan, H. B., Mironov, I., Talwar, K., and Zhang, L. (2016). Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, pages 308–318.

Adams, W. P. and Sherali, H. D. (1986). A tight linearization and an algorithm for zero-one quadratic programming problems. *Manag. Sci*, 32(10):1274–1290.

Alley, M., Biggs, M., Hariss, R., Herrmann, C., Li, M. L., and Perakis, G. (2023). Pricing for heterogeneous products: Analytics for ticket reselling. *Manufacturing & Service Operations Management*, 25(2):409–426.

Altschuler, J. and Talwar, K. (2022). Privacy of noisy stochastic gradient descent: More iterations without more privacy loss. *Advances in Neural Information Processing Systems*, 35:3788–3800.

Alvim, M. S., Andrés, M. E., Chatzikokolakis, K., Degano, P., and Palamidessi, C. (2011). Differential privacy: On the trade-off between utility and information leakage. In Barthe, G., Datta, A., and Etalle, S., editors, *Formal Aspects of Security and Trust - 8th International Workshop, FAST 2011, Leuven, Belgium, September 12-14, 2011. Revised Selected Papers*, volume 7140 of *Lecture Notes in Computer Science*, pages 39–54. Springer.

Anderson, E. J. and Nash, P. (1987). *Linear Programming in Infinite-Dimensional Spaces: Theory and Applications*. John Wiley & Sons.

Andrianova, A. A., Korepanova, A. A., and Halilova, I. F. (2016). One algorithm for branch and bound method for solving concave optimization problem. *IOP Conference Series: Materials Science and Engineering*, 158:012005.

Anstreicher, K. M. (2009). Semidefinite programming versus the reformulation-linearization technique for nonconvex quadratically constrained quadratic programming. *J. Global Optim.*, 43(2-3):471–484.

Apple Differential Privacy Team (2017). Learning with privacy at scale. White paper, Apple.

Ardestani-Jaafari, A. and Delage, E. (2016). Robust optimization of sums of piecewise linear functions with application to inventory problems. *Operations Research*, 64(2):474–494.

Arjovsky, M., Chintala, S., and Bottou, L. (2017). Wasserstein generative adversarial networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70.

Asi, H. and Duchi, J. C. (2020a). Instance-optimality in differential privacy via approximate inverse sensitivity mechanisms. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H., editors, *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual.*

Asi, H. and Duchi, J. C. (2020b). Near instance-optimality in differential privacy. *CoRR*, abs/2005.10630.

Audet, C., Hansen, P., and Savard, G. (2005). *Essays and surveys in global optimization.* GERAD. Springer.

Awasthi, P., Jung, C., and Morgenstern, J. (2022). Distributionally robust data join. *arXiv:2202.05797.*

Balle, B. and Wang, Y. (2018). Improving the Gaussian mechanism for differential privacy: Analytical calibration and optimal denoising. In *Proceedings of the 35th International Conference on Machine Learning*, pages 403–412.

Ban, G.-Y. and Rudin, C. (2019). The big data newsvendor: Practical insights from machine learning. *Operations Research*, 67(1):90–108.

Bao, X., Sahinidis, N. V., and Tawarmalani, M. (2011). Semidefinite relaxations for quadratically constrained quadratic programming: A review and comparisons. *Math. Program.*, 129(1):129–157.

Bastani, H. and Bayati, M. (2020). Online decision making with high-dimensional covariates. *Operations Research*, 68(1):276–294.

Bavadekar, S., Boulanger, A., Davis, J., Desfontaines, D., Gabrilovich, E., Gadepalli, K., Ghazi, B., Griffith, T., Gupta, J. P., Kamath, C., Kraft, D., Kumar, R., Kumok, A., Mayer, Y., Manurangsi, P., Patankar, A., Perera, I. M., Scott, C., Shekel, T., Miller, B., Smith, K., Stanton, C., Sun, M., Young, M., and Wellenius, G. (2021). Google COVID-19 vaccination search insights: Anonymization process description. *CoRR*, abs/2107.01179.

Behrendt, A., Savelsbergh, M., and Wang, H. (2023). A prescriptive machine learning method for courier scheduling on crowdsourced delivery platforms. *Transportation Science*, 57(4):889–907.

Belbasi, R., Selvi, A., and Wiesemann, W. (2025). It's all in the mix: Wasserstein machine learning with mixed features. *arXiv:2312.12230.*

Beliakov, G. and Abraham, A. (2002). Global optimisation of neural networks using a deterministic hybrid approach. In *Hybrid Information Systems*, pages 79–92. Springer.

Ben-Tal, A., El Ghaoui, L., and Nemirovski, A. (2009). *Robust Optimization.* Princeton University Press.

Ben-Tal, A., Goryashko, A., Guslitzer, E., and Nemirovski, A. (2004). Adjustable robust solutions of uncertain linear programs. *Mathematical Programming*, 99(2):351–376.

Bennouna, A., Lucas, R., and Van Parys, B. (2023). Certified robust neural networks: Generalization and corruption resistance. In *International Conference on Machine Learning.*

Bennouna, A. and Van Parys, B. (2022). Holistic robust data-driven decisions. *arXiv:2207.09560.*

Benson, H. P. (1995). Concave minimization: theory, applications and algorithms. In *Handbook of global optimization*, pages 43–148. Springer.

Berg, M., Kreveld, M., Overmars, M., and Schwarzkopf, O. C. (2000). *Computational Geometry: Algorithms and Applications*. Springer, 2nd edition.

Bertsekas, D. (2009). *Convex optimization theory*, volume 1. Athena Scientific.

Bertsimas, D., Delarue, A., Jaillet, P., and Martin, S. (2019a). Travel time estimation in the age of big data. *Operations Research*, 67(2):498–515.

Bertsimas, D. and den Hertog, D. (2022). *Robust and Adaptive Optimization*. Dynamic Ideas.

Bertsimas, D., Dunn, J., Pawlowski, C., and Zhuo, Y. D. (2019b). Robust classification. *INFORMS Journal on Optimization*, 1(1):2–34.

Bertsimas, D., Dunning, I., and Lubin, M. (2016a). Reformulation versus cutting-planes for robust optimization: A computational study. *Computational Management Science*, 13(1):195–217.

Bertsimas, D., Goyal, V., and Lu, B. Y. (2015). A tight characterization of the performance of static solutions in two-stage adjustable robust linear optimization. *Mathematical Programming*, 150(2):281–319.

Bertsimas, D. and Kallus, N. (2020). From predictive to prescriptive analytics. *Management Science*, 66(3):1025–1044.

Bertsimas, D., Lauprete, G. J., and Samarov, A. (2004). Shortfall as a risk measure: properties, optimization and applications. *J. Econ. Dyn. Control*, 28(7):1353–1381.

Bertsimas, D., O'Hair, A., Relyea, S., and Silberholz, J. (2016b). An analytics approach to designing combination chemotherapy regimens for cancer. *Management Science*, 62(5):1511–1531.

Bertsimas, D. and Pauphilet, J. (2023). Hospital-wide inpatient flow optimization. *Management Science*, page Available ahead of print.

Bertsimas, D. and Sim, M. (2004). The price of robustness. *Operations Research*, 52(1):35–53.

Bienstock, D. and Özbay, N. (2008). Computing robust basestock levels. *Discrete Optimization*, 5(2):389–414.

Biggio, B., Corona, I., Maiorca, D., Nelson, B., Šrndić, N., Laskov, P., Giacinto, G., and Roli, F. (2013). Evasion attacks against machine learning at test time. In *Machine Learning and Knowledge Discovery in Databases: European Conference*, pages 387–402.

Bishop, C. (2006). *Pattern Recognition and Machine Learning*. Springer.

Blanchet, J. and Kang, Y. (2021). Sample out-of-sample inference based on Wasserstein distance. *Operations Research*, 69(3):985–1013.

Blanchet, J., Kang, Y., and Murthy, K. (2019). Robust Wasserstein profile inference and applications to machine learning. *Journal of Applied Probability*, 56(3):830–857.

Blanchet, J. and Murthy, K. (2019). Quantifying distributional model risk via optimal transport. *Mathematics of Operations Research*, 44(2):565–600.

Bomze, I. M. (1998). On standard quadratic optimization problems. *J. Global Optim.*, 13(4):369–387.

Boyd, S., Kim, S.-J., Vandenberghe, L., and Hassibi, A. (2007). A tutorial on geometric programming. *Optimization and Engineering*, 8(1):67–127.

Boyd, S. and Mattingley, J. (2007). Branch and bound methods. https://stanford.edu/class/ee364b/lectures/bb_notes.pdf.

Boyd, S. and Vandenberghe, L. (2004). *Convex Optimization*. Cambridge university press.

Bromiley, P. (2003). Products and convolutions of Gaussian probability density functions. *Tina-Vision Memo*, 3(4):1.

Bui, T. A., Le, T., Tran, Q., Zhao, H., and Phung, D. (2022). A unified Wasserstein distributional robustness framework for adversarial training. *arXiv:2202.13437*.

Bun, M. and Steinke, T. (2016). Concentrated differential privacy: Simplifications, extensions, and lower bounds. In *Theory of cryptography conference*, pages 635–658. Springer.

Byrd, R. H., Nocedal, J., and Waltz, R. A. (2006). Knitro: An integrated package for nonlinear optimization. In *Large-scale nonlinear optimization*, pages 35–59. Springer.

Cai, N. and Kou, S. (2019). Econometrics with privacy preservation. *Operations Research*, 67(4):905–926.

Canonne, C. L., Kamath, G., and Steinke, T. (2020). The discrete gaussian for differential privacy. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H., editors, *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Carlini, N., Athalye, A., Papernot, N., Brendel, W., Rauber, J., Tsipras, D., Goodfellow, I., Madry, A., and Kurakin, A. (2019). On evaluating adversarial robustness. *arXiv:1902.06705*.

Carmon, Y., Raghunathan, A., Schmidt, L., Duchi, J. C., and Liang, P. S. (2019). Unlabeled data improves adversarial robustness. In *Advances in Neural Information Processing Systems*.

Chan, T. C. Y., Mahmood, R., O'Connor, D. L., Stone, D., Unger, S., Wong, R. K., and Zhu, I. Y. (2023). Got (optimal) milk? Pooling donations in human milk banks with machine learning and optimization. *Manufacturing & Service Operations Management*, page Available ahead of print.

Chang, C.-C. and Lin, C.-J. (2011). LIBSVM: A library for support vector machines. *ACM transactions on Intelligent Systems and Technology*, 2(3):1–27.

Chaudhuri, K. and Monteleoni, C. (2008). Privacy-preserving logistic regression. In Koller, D., Schuurmans, D., Bengio, Y., and Bottou, L., editors, *Advances in Neural Information Processing Systems 21, Proceedings of the Twenty-Second Annual Conference on Neural Information Processing Systems, Vancouver, British Columbia, Canada, December 8-11, 2008*, pages 289–296. Curran Associates, Inc.

Chaudhuri, K., Monteleoni, C., and Sarwate, A. D. (2011). Differentially private empirical risk minimization. *Journal of Machine Learning Research*, 12(3):1069–1109.

Chen, T., Zhang, Z., Liu, S., Chang, S., and Wang, Z. (2020). Robust overfitting may be mitigated by properly learned smoothening. In *International Conference on Learning Representations*.

Chen, X., Miao, S., and Wang, Y. (2023). Differential privacy in personalized pricing with nonparametric demand models. *Operations Research*, 71(2):581–602.

Chen, X., Simchi-Levi, D., and Wang, Y. (2022). Privacy-preserving dynamic personalized pricing with demand learning. *Management Science*, 68(7):4878–4898.

Cohen, G., Afshar, S., Tapson, J., and Van Schaik, A. (2017). EMNIST: Extending MNIST to handwritten letters. In *International Joint Conference on Neural Networks*.

Croce, F., Andriushchenko, M., Sehwag, V., Debenedetti, E., Flammarion, N., Chiang, M., Mittal, P., and Hein, M. (2020). Robustbench: a standardized adversarial robustness benchmark. *arXiv:2010.09670*.

Cuff, P. and Yu, L. (2016). Differential privacy as a mutual information constraint. In *Proceedings of the ACM SIGSAC Conference on Computer and Communications Security*, pages 43–54.

De Klerk, E., Den Hertog, D., and Elabwabi, G. (2008). On the complexity of optimization over the standard simplex. *Eur. J. Oper. Res.*, 191(3):773–785.

Delage, E. and Ye, Y. (2010). Distributionally robust optimization under moment uncertainty with application to data-driven problems. *Operations Research*, 58(3):596–612.

Demelius, L., Kern, R., and Trügler, A. (2025). Recent advances of differential privacy in centralized deep learning: A systematic survey. *ACM Computing Surveys*, 57(6):1–28.

DeMiguel, V. and Nogales, F. J. (2009). Portfolio selection with robust estimation. *Operations Research*, 57(3):560–577.

Deng, Z., Zhang, L., Ghorbani, A., and Zou, J. (2021). Improving adversarial robustness via unlabeled out-of-domain data. In *International Conference on Artificial Intelligence and Statistics*.

Desfontaines, D. (2020). *Lowering the Cost of Anonymization*. PhD thesis, ETH Zurich.

Desfontaines, D. (2021a). A friendly, non-technical introduction to differential privacy. https://desfontain.es/blog/friendly-intro-to-differential-privacy.html. Ted is writing things (personal blog).

Desfontaines, D. (2021b). A list of real-world uses of differential privacy. https://desfontain.es/blog/real-world-differential-privacy.html. Ted is writing things (personal blog).

Desfontaines, D. and Pejó, B. (2020). SoK: Differential privacies. In Chatzikokolakis, K. and Johnson, A., editors, *Proceedings on Privacy Enhancing Technologies*, pages 288–313.

Dinur, I. and Nissim, K. (2003). Revealing information while preserving privacy. In Neven, F., Beeri, C., and Milo, T., editors, *Proceedings of the Twenty-Second ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems, June 9-12, 2003, San Diego, CA, USA*, pages 202–210. ACM.

Dong, J., Roth, A., and Su, W. J. (2022). Gaussian differential privacy. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 84(1):3–37.

Dua, D. and Graff, C. (2017). UCI machine learning repository. http://archive.ics.uci.edu/ml. Accessed: December 17, 2024.

Duchi, J., Hashimoto, T., and Namkoong, H. (2023). Distributionally robust losses for latent covariate mixtures. *Operations Research*, 71(2):649–664.

Duffin, R. J. (1967). *Geometric programming - theory and application*. SIAM.

Dwork, C. (2011). The promise of differential privacy: A tutorial on algorithmic techniques. In Ostrovsky, R., editor, *IEEE 52nd Annual Symposium on Foundations of Computer Science, FOCS 2011, Palm Springs, CA, USA, October 22-25, 2011*, pages 1–2. IEEE Computer Society.

Dwork, C., Kenthapadi, K., McSherry, F., Mironov, I., and Naor, M. (2006a). Our data, ourselves: Privacy via distributed noise generation. In Vaudenay, S., editor, *Advances in Cryptology - EUROCRYPT 2006, 25th Annual International Conference on the Theory and Applications of Cryptographic Techniques, St. Petersburg, Russia, May 28 - June 1, 2006, Proceedings*, volume 4004 of *Lecture Notes in Computer Science*, pages 486–503. Springer.

Dwork, C., McSherry, F., Nissim, K., and Smith, A. D. (2006b). Calibrating noise to sensitivity in private data analysis. In Halevi, S. and Rabin, T., editors, *Theory of Cryptography, Third Theory of Cryptography Conference, TCC 2006, New York, NY, USA, March 4-7, 2006, Proceedings*, volume 3876 of *Lecture Notes in Computer Science*, pages 265–284. Springer.

Dwork, C. and Roth, A. (2014). The algorithmic foundations of differential privacy. *Foundations and Trends® in Theoretical Computer Science*, 9(3–4):211–407.

Dwork, C. and Rothblum, G. N. (2016). Concentrated differential privacy. *CoRR*, abs/1603.01887.

Dwork, C., Smith, A., Steinke, T., and Ullman, J. (2017). Exposed! A survey of attacks on private data. *Annual Review of Statistics and Its Application*, 4(1):61–84.

Enkhbat, R., Barsbold, B., and Kamada, M. (2006). A numerical approach for solving some convex maximization problems. *Journal of Global Optimization*, 35(1):85–101.

Fampa, M., Lee, J., and Melo, W. (2017). On global optimization with indefinite quadratics. *EURO J. Comput. Optim.*, 5(3):309–337.

Feldman, J., Zhang, D. J., Liu, X., and Zhang, N. (2022). Customer choice models vs. machine learning: Finding optimal product displays on Alibaba. *Operations Research*, 70(1):309–328.

Ferreira, K. J., Lee, B. H. A., and Simchi-Levi, D. (2016). Analytics for an online retailer: Demand forecasting and price optimization. *Manufacturing & Service Operations Management*, 18(1):69–88.

Foote, A. D., Machanavajjhala, A., and McKinney, K. (2019). Releasing earnings distributions using differential privacy. *Journal of Privacy and Confidentiality*, 9(2):1–19.

Fournier, N. and Guillin, A. (2015). On the rate of convergence in Wasserstein distance of the empirical measure. *Probability Theory and Related Fields*, 162(3-4):707–738.

Frank, N. S. and Niles-Weed, J. (2024). Existence and minimax theorems for adversarial surrogate risks in binary classification. *Journal of Machine Learning Research*, 25(58):1–41.

Friedman, A. and Schuster, A. (2010). Data mining with differential privacy. In Rao, B., Krishnapuram, B., Tomkins, A., and Yang, Q., editors, *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Washington, DC, USA, July 25-28, 2010*, pages 493–502. ACM.

Friedman, J., Hastie, T., and Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1):1–22.

Gao, R. (2023). Finite-sample guarantees for Wasserstein distributionally robust optimization: Breaking the curse of dimensionality. *Operations Research*, 71(6):2291–2306.

Gao, R., Cai, T., Li, H., Hsieh, C.-J., Wang, L., and Lee, J. D. (2019). Convergence of adversarial training in overparametrized neural networks. In *Advances in Neural Information Processing Systems*.

Gao, R., Chen, X., and Kleywegt, A. J. (2022). Wasserstein distributionally robust optimization and variation regularization. *Operations Research*, page Available ahead of print.

Gao, R. and Kleywegt, A. (2016). Distributionally robust stochastic optimization with Wasserstein distance. `https://arxiv.org/abs/1604.02199`.

Geng, Q., Ding, W., Guo, R., and Kumar, S. (2019). Optimal noise-adding mechanism in additive differential privacy. In Chaudhuri, K. and Sugiyama, M., editors, *The 22nd International Conference on Artificial Intelligence and Statistics, AISTATS 2019, 16-18 April 2019, Naha, Okinawa, Japan*, volume 89 of *Proceedings of Machine Learning Research*, pages 11–20. PMLR.

Geng, Q., Ding, W., Guo, R., and Kumar, S. (2020). Tight analysis of privacy and utility tradeoff in approximate differential privacy. In *The 23rd International Conference on Artificial Intelligence and Statistics*, pages 89–99.

Geng, Q. and Viswanath, P. (2014). The optimal mechanism in differential privacy. In *2014 IEEE International Symposium on Information Theory, Honolulu, HI, USA, June 29 - July 4, 2014*, pages 2371–2375. IEEE.

Geng, Q. and Viswanath, P. (2016). Optimal noise adding mechanisms for approximate differential privacy. *IEEE Trans. Inf. Theory*, 62(2):952–969.

Givens, C. R. and Shortt, R. M. (1984). A class of Wasserstein metrics for probability distributions. *Michigan Mathematical Journal*, 31(2):231–240.

Glaeser, C. K., Fisher, M., and Su, X. (2019). Optimal retail location: Empirical methodology and application to practice. *Manufacturing & Service Operations Management*, 21(1):86–102.

Goemans, M. X. and Williamson, D. P. (1994). .879-approximation algorithms for MAX CUT and MAX 2SAT. In *Proceedings of the Twenty-sixth Annual ACM Symposium on Theory of Computing*, pages 422–431, New York, NY, USA.

Gong, M., Xie, Y., Pan, K., Feng, K., and Qin, A. K. (2020). A survey on differentially private machine learning. *IEEE computational intelligence magazine*, 15(2):49–64.

Goodfellow, I. J., Shlens, J., and Szegedy, C. (2015). Explaining and harnessing adversarial examples. In *International Conference on Learning Representations*.

Gorissen, B. L., Yanıkoğlu, İ., and Den Hertog, D. (2015). A practical guide to robust optimization. *Omega*, 53:124–137.

Gowal, S., Rebuffi, S.-A., Wiles, O., Stimberg, F., Calian, D. A., and Mann, T. A. (2021). Improving robustness using generated data. In *Advances in Neural Information Processing Systems*.

Grötschel, M., Lovász, L., and Schrijver, A. (1988). *Geometric Algorithms and Combinatorial Optimization*. Springer.

Gurobi Optimization, L. (2018). Gurobi optimizer reference manual.

Guslitser, E. (2002). Uncertainty-immunized solutions in linear programming. Master's thesis, Technion – Israeli Institute of Technology.

Hadjiyiannis, M. J., Goulart, P. J., and Kuhn, D. (2011). A scenario approach for estimating the suboptimality of linear decision rules in two-stage robust optimization. In *2011 50th IEEE Conference on Decision and Control and European Control Conference*, pages 7386–7391. IEEE.

Han, S., Tao, M., Topcu, U., Owhadi, H., and Murray, R. M. (2015). Convex optimal uncertainty quantification. *SIAM Journal on Optimization*, 25(3):1368–1387.

Han, S., Topcu, U., and Pappas, G. J. (2016). Differentially private distributed constrained optimization. *IEEE Transactions on Automatic Control*, 62(1):50–64.

Hanasusanto, G. A., Roitch, V., Kuhn, D., and Wiesemann, W. (2015). A distributionally robust perspective on uncertainty quantification and chance constrained programming. *Mathematical Programming*, 151(1):35–62.

Harder, F., Bauer, M., and Park, M. (2020). Interpretable and differentially private predictions. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 4083–4090.

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer.

Heffetz, O. and Ligett, K. (2014). Privacy and data-based research. *Journal of Economic Perspectives*, 28(2):75–98.

Hesthaven, J. S. (1998). From electrostatics to almost optimal nodal sets for polynomial interpolation in a simplex. *SIAM J. Numer. Anal.*, 35:655–676.

Horst, R., Pardalos, P., and Van Thoai, N. (2000). *Introduction to Global Optimization*. Springer.

Horst, R., Phong, T. Q., Thoai, N. V., and de Vries, J. (1991). On solving a DC programming problem by a sequence of linear programs. *Journal of Global Optimization*, 1(2):183–203.

Horst, R. and Thoai, N. V. (1999). DC programming: overview. *Journal of Optimization Theory and Applications*, 103(1):1–43.

Hsu, J., Huang, Z., Roth, A., and Wu, Z. S. (2016). Jointly private convex programming. In Krauthgamer, R., editor, *Proceedings of the Twenty-Seventh Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2016, Arlington, VA, USA, January 10-12, 2016*, pages 580–599. SIAM.

Hsu, J., Roth, A., Roughgarden, T., and Ullman, J. R. (2014). Privately solving linear programs. In Esparza, J., Fraigniaud, P., Husfeldt, T., and Koutsoupias, E., editors, *Automata, Languages, and Programming - 41st International Colloquium, ICALP 2014, Copenhagen, Denmark, July 8-11, 2014, Proceedings, Part I*, volume 8572 of *Lecture Notes in Computer Science*, pages 612–624. Springer.

Hu, H., Salcic, Z., Sun, L., Dobbie, G., Yu, P. S., and Zhang, X. (2022). Membership inference attacks on machine learning: A survey. *ACM Computing Surveys*, 54(11s):1–37.

IBM ILOG CPLEX (2014). V12.6: User's Manual for CPLEX. https://www.ibm.com/support/knowledgecenter/SSSA5P_12.6.2/ilog.odms.studio.help/pdf/usrcplex.pdf.

Ji, Z., Lipton, Z. C., and Elkan, C. (2014). Differential privacy and machine learning: a survey and review. *arXiv:1412.7584*.

Johnson, A. and Shmatikov, V. (2013). Privacy-preserving data exploration in genome-wide association studies. In Dhillon, I. S., Koren, Y., Ghani, R., Senator, T. E., Bradley, P., Parekh, R., He, J., Grossman, R. L., and Uthurusamy, R., editors, *The 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD 2013, Chicago, IL, USA, August 11-14, 2013*, pages 1079–1087. ACM.

Kallus, N. and Udell, M. (2020). Dynamic assortment personalization in high dimensions. *Operations Research*, 68(4):1020–1037.

Kelly, M., Longjohn, R., and Nottingham, K. (2023). The UCI machine learning repository. https://archive.ics.uci.edu.

Khim, J. and Loh, P.-L. (2018). Adversarial risk bounds via function transformation. *arXiv:1810.09519*.

Krizhevsky, A., Hinton, G., et al. (2009). Learning multiple layers of features from tiny images.

Kuhn, D., Mohajerin Esfahani, P., Nguyen, V. A., and Shafieezadeh-Abadeh, S. (2019). Wasserstein distributionally robust optimization: Theory and applications in machine learning. *INFORMS TutORials in Operations Research*, pages 130–169.

Kuhn, D., Shafiee, S., and Wiesemann, W. (2025). Distributionally robust optimization.

Lam, H. (2019). Recovering best statistical guarantees via the empirical divergence-based distributionally robust optimization. *Operations Research*, 6(4):1090–1105.

Le Thi, H. A. and Pham Dinh, T. (2018). DC programming and DCA: thirty years of developments. *Mathematical Programming*, 169(1):5–68.

LeCun, Y., Bottou, L., Bengio, Y., and Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324.

Lee, Y. T., Sidford, A., and Wong, S. C.-W. (2015). A faster cutting plane method and its implications for combinatorial and convex optimization. *2015 IEEE 56th Annual Symposium on Foundations of Computer Science*, pages 1049–1065.

Lei, Y., Miao, S., and Momot, R. (2024). Privacy-preserving personalized revenue management. *Management Science*, 70(7):4875–4892.

Li, B. and Li, Y. (2023). Why clean generalization and robust overfitting both happen in adversarial training. *arXiv:2306.01271*.

Li, L. and Spratling, M. (2023). Understanding and combating robust overfitting via input loss landscape analysis and regularization. *Pattern Recognition*, 136:1–11.

Li, R., Tobey, M., Mayorga, M. E., Caltagirone, S., and Özaltın, O. Y. (2023). Detecting human trafficking: Automated classification of online customer reviews of massage businesses. *Manufacturing & Service Operations Management*, 25(3):1051–1065.

Liberti, L. and Pantelides, C. C. (2006). An exact reformulation algorithm for large nonconvex NLPs involving bilinear terms. *J. Global Optim.*, 36(2):161–189.

Lin, R., Yu, C., Han, B., Su, H., and Liu, T. (2024). Layer-aware analysis of catastrophic overfitting: Revealing the pseudo-robust shortcut dependency. In Salakhutdinov, R., Kolter, Z., Heller, K., Weller, A., Oliver, N., Scarlett, J., and Berkenkamp, F., editors, *Proceedings of the 41st International Conference on Machine Learning*, volume 235, pages 30427–30439.

Lin, R., Yu, C., and Liu, T. (2023). Eliminating catastrophic overfitting via abnormal adversarial examples regularization. *Advances in Neural Information Processing Systems*, 36:67866–67885.

Lipp, T. and Boyd, S. (2016). Variations and extension of the convex–concave procedure. *Optimization and Engineering*, 17(2):263–287.

Löfberg, J. (2004). YALMIP: A toolbox for modeling and optimization in MATLAB. In *2004 IEEE International Conference on Robotics and Automation (IEEE Cat. No. 04CH37508)*, pages 284–289. IEEE.

Löfberg, J. (2016). YALMIP tutorial: Nonlinear operators - integer models. https://yalmip.github.io/tutorial/nonlinearoperatorsmixedinteger/.

Long, D. Z., Sim, M., and Zhou, M. (2023). Robust satisficing. *Operations Research*, 71(1):61–82.

Lopuhaä-Zwakenberg, M., Alishahi, M., Kivits, J., Klarenbeek, J., van der Velde, G., and Zannone, N. (2021). Comparing classifiers' performance under differential privacy. In di Vimercati, S. D. C. and Samarati, P., editors, *Proceedings of the 18th International Conference on Security and Cryptography, SECRYPT 2021, July 6-8, 2021*, pages 50–61. SCITEPRESS.

Lubin, M., Dowson, O., Dias Garcia, J., Huchette, J., Legat, B., and Vielma, J. P. (2023). JuMP 1.0: Recent improvements to a modeling language for mathematical optimization. *Mathematical Programming Computation*, 15(3):581–589.

Machanavajjhala, A., Kifer, D., Gehrke, J., and Venkitasubramaniam, M. (2007). $\ell$-diversity: Privacy beyond $k$-anonymity. *ACM Transactions on Knowledge Discovery from Data*, 1(1):1556–4681.

Madry, A., Makelov, A., Schmidt, L., Tsipras, D., and Vladu, A. (2018). Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations*.

Mangasarian, O. L. (1996). *Machine learning via polyhedral concave minimization*. Physica-Verlag HD, Heidelberg.

Mangasarian, O. L. (2011). Privacy-preserving linear programming. *Optimization Letters*, 5(1):165–172.

Mangasarian, O. L. (2015). Unsupervised classification via convex absolute value inequalities. *Optimization*, 64(1):81–86.

Mangold, P., Bellet, A., Salmon, J., and Tommasi, M. (2022). Differentially private coordinate descent for composite empirical risk minimization. In Chaudhuri, K., Jegelka, S., Song, L., Szepesvári, C., Niu, G., and Sabato, S., editors, *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162 of *Proceedings of Machine Learning Research*, pages 14948–14978. PMLR.

MATLAB (2018). *version 9.5.0 (R2018b)*. The MathWorks Inc., Natick, Massachusetts.

McCormick, G. P. (1976). Computability of global solutions to factorable nonconvex programs: Part I—Convex underestimating problems. *Math. Program.*, 10(1):147–175.

McSherry, F. and Talwar, K. (2007). Mechanism design via differential privacy. In *Proceedings of the 48th Annual IEEE Symposium on Foundations of Computer Science*, pages 94–103.

Meiser, S. (2018). Approximate and probabilistic differential privacy definitions. *Cryptology ePrint Archive, Paper 2018/277*, pages 1–9.

Messing, S., DeGregorio, C., Hillenbrand, B., King, G., Mahanti, S., Mukerjee, Z., Nayak, C., Persily, N., State, B., and Wilkins, A. (2020). Facebook Privacy-Protected Full URLs Data Set. https://doi.org/10.7910/DVN/TDOAPG.

Mezzadri, F. (2006). How to generate random matrices from the classical compact groups. *arXiv preprint arXiv:math-ph/0609050*.

Michaud, R. O. (1989). The Markowitz optimization enigma: Is 'optimized'optimal? *Financial Analysts Journal*, 45(1):31–42.

Misener, R. and Floudas, C. A. (2013). GloMIQO: Global mixed-integer quadratic optimizer. *J. Global Optim.*, 57(1):3–50.

Misener, R. and Floudas, C. A. (2014). ANTIGONE: Algorithms for continuous/integer global optimization of nonlinear equations. *J. Global Optim.*, 59(2-3):503–526.

Mohajerin Esfahani, P. and Kuhn, D. (2018). Data-driven distributionally robust optimization using the Wasserstein metric: performance guarantees and tractable reformulations. *Mathematical Programming*, 171(1–2):1–52.

MOSEK ApS (2019). The MOSEK optimization toolbox for MATLAB manual. Version 9.0. http://docs.mosek.com/9.0/toolbox/index.html.

MOSEK ApS (2023). Modeling cookbook. https://docs.mosek.com/MOSEKModelingCookbook-letter.pdf.

Murphy, K. P. (2022). *Probabilistic Machine Learning: An Introduction*. MIT Press.

Mutapcic, A. and Boyd, S. (2009). Cutting-set methods for robust convex optimization with pessimizing oracles. *Optimization Methods & Software*, 24(3):381–406.

Neelakantan, A., Vilnis, L., Le, Q. V., Sutskever, I., Kaiser, L., Kurach, K., and Martens, J. (2015). Adding gradient noise improves learning for very deep networks. *CoRR*, abs/1511.06807.

Nemhauser, G. L. and Wolsey, L. A. (1988). *Integer and combinatorial optimization*. Wiley-Interscience, New York, NY, USA.

Nergiz, M. E., Atzori, M., and Clifton, C. (2007). Hiding the presence of individuals from shared databases. In Chan, C. Y., Ooi, B. C., and Zhou, A., editors, *Proceedings of the ACM SIGMOD International Conference on Management of Data, Beijing, China, June 12-14, 2007*, pages 665–676. ACM.

Nesterov, Y. (2004). *Introductory Lectures on Convex Optimization: A Basic Course*. Springer.

Nesterov, Y. (2018). *Lectures on Convex Optimization*. Springer, 2nd edition.

Nissim, K., Raskhodnikova, S., and Smith, A. D. (2007). Smooth sensitivity and sampling in private data analysis. In Johnson, D. S. and Feige, U., editors, *Proceedings of the 39th Annual ACM Symposium on Theory of Computing, San Diego, California, USA, June 11-13, 2007*, pages 75–84. ACM.

Owhadi, H., Scovel, C., Sullivan, T. J., McKerns, M., and Ortiz, M. (2013). Optimal uncertainty quantification. *SIAM Review*, 55(2):271–345.

Pang, T., Lin, M., Yang, X., Zhu, J., and Yan, S. (2022). Robustness and accuracy could be reconcilable by (proper) definition. In *International Conference on Machine Learning*.

Pardalos, P. M. and Rosen, J. B. (1986). Methods for global concave minimization: A bibliographic survey. *SIAM Review*, 28(3):367–379.

Pardalos, P. M. and Schnitger, G. (1988). Checking local optimality in constrained quadratic programming is NP-hard. *Operations Research Letters*, 7(1):33–35.

Parikh, N., Boyd, S., et al. (2014). Proximal algorithms. *Foundations and Trends® in Optimization*, 1(3):127–239.

Park, J. (2016). Sparsity-preserving difference of positive semidefinite matrix representation of indefinite matrices. *arXiv preprint arXiv:1609.06762*.

Patki, N., Wedge, R., and Veeramachaneni, K. (2016). The synthetic data vault. In *IEEE International Conference on Data Science and Advanced Analytics*.

Pätzold, J. and Schöbel, A. (2020). Approximate cutting plane approaches for exact solutions to robust optimization problems. *European Journal of Operational Research*, 284(1):20–30.

Pereira, M., Kim, A., Allen, J., White, K., Ferres, J. L., and Dodhia, R. (2021). U.S. broadband coverage data set: A differentially private data release. *CoRR*, abs/2103.14035.

Phan, H., Le, T., Phung, T., Bui, A. T., Ho, N., and Phung, D. (2023). Global-local regularization via distributional robustness. In *International Conference on Artificial Intelligence and Statistics*.

Pydi, M. S. and Jog, V. (2021). The many faces of adversarial risk. In *Advances in Neural Information Processing Systems*.

Qi, M., Cao, Y., and Shen, Z.-J. (2022). Distributionally robust conditional quantile prediction with fixed design. *Management Science*, 68(3):1639–1658.

Rade, R. and Moosavi-Dezfooli, S.-M. (2022). Reducing excessive margin to achieve a better accuracy vs. robustness trade-off. In *International Conference on Learning Representations*.

Raghunathan, A., Xie, S. M., Yang, F., Duchi, J. C., and Liang, P. (2019). Adversarial training can hurt generalization. *arXiv:1906.06032*.

Rahimian, H. and Mehrotra, S. (2022). Frameworks and results in distributionally robust optimization. *Open Journal of Mathematical Optimization*, 3:1–85.

Rebennack, S., Nahapetyan, A., and Pardalos, P. M. (2009). Bilinear modeling solution approach for fixed charge network flow problems. *Optimization Letters*, 3(3):347–355.

Regniez, C., Gidel, G., and Berard, H. (2022). A distributional robustness perspective on adversarial training with the ∞-Wasserstein distance. https://openreview.net/forum?id=z7DAilcTx7.

Rice, L., Wong, E., and Kolter, Z. (2020). Overfitting in adversarially robust deep learning. In *International Conference on Machine Learning*.

Richtárik, P. and Takáč, M. (2014). Iteration complexity of randomized block-coordinate descent methods for minimizing a composite function. *Mathematical Programming*, 144(1):1–38.

Rockafellar, R. T. (1997). *Convex Analysis*. Princeton University Press.

Rogers, R., Cardoso, A. R., Mancuhan, K., Kaura, A., Gahlawat, N., Jain, N., Ko, P., and Ahammad, P. (2020). A members first approach to enabling linkedin's labor market insights at scale. *CoRR*, abs/2010.13981.

Roos, E., Den Hertog, D., Ben-Tal, A., De Ruiter, F., and Zhen, J. (2018). Tractable approximation of hard uncertain optimization problems. *Preprint 6679: Optimization Online*.

Rychener, Y., Esteban-Pérez, A., Morales, J. M., and Kuhn, D. (2024). Wasserstein distributionally robust optimization with heterogeneous data sources. *arXiv:2407.13582*.

Salzberg, S. L. (1997). On comparing classifiers: Pitfalls to avoid and a recommended approach. *Data Mining and Knowledge Discovery*, 1(3):317–328.

Samorani, M., Harris, S. L., Blount, L. G., Lu, H., and Santoro, M. A. (2022). Overbooked and overlooked: Machine learning and racial bias in medical appointment scheduling. *Manufacturing & Service Operations Management*, 24(6):2825–2842.

Sehwag, V., Mahloujifar, S., Handina, T., Dai, S., Xiang, C., Chiang, M., and Mittal, P. (2022). Robust learning meets generative models: Can proxy distributions improve adversarial robustness? In *International Conference on Learning Representations*.

Selvi, A., Belbasi, M. R., Haugh, M., and Wiesemann, W. (2022a). Wasserstein logistic regression with mixed features. *Advances in Neural Information Processing Systems*, 35:16691–16704.

Selvi, A., Ben-Tal, A., Brekelmans, R., and den Hertog, D. (2022b). Convex maximization via adjustable robust optimization. *INFORMS Journal on Computing*, 34(4):2091–2105.

Selvi, A., den Hertog, D., and Wiesemann, W. (2023). A reformulation-linearization technique for optimization over simplices. *Mathematical Programming*, 197(1):427–447.

Selvi, A., Liu, H., and Wiesemann, W. (2025). Differential privacy via distributionally robust optimization. *Operations Research*, 0(0):1–23.

Serrano, S. A. (2015). *Algorithms for unsymmetric cone optimization and an implementation for problems with the exponential cone*. Stanford University.

Shafahi, A., Najibi, M., Ghiasi, M. A., Xu, Z., Dickerson, J., Studer, C., Davis, L. S., Taylor, G., and Goldstein, T. (2019). Adversarial training for free! In *Advances in Neural Information Processing Systems*.

Shafahi, A., Saadatpanah, P., Zhu, C., Ghiasi, A., Studer, C., Jacobs, D. W., and Goldstein, T. (2020). Adversarially robust transfer learning. In *International Conference on Learning Representations*.

Shafieezadeh-Abadeh, S., Aolaritei, L., Dörfler, F., and Kuhn, D. (2023). New perspectives on regularization and computation in optimal transport-based distributionally robust optimization. *arXiv:2303.03900*.

Shafieezadeh-Abadeh, S., Kuhn, D., and Esfahani, P. M. (2019). Regularization via mass transportation. *Journal of Machine Learning Research*, 20(103):1–68.

Shafieezadeh-Abadeh, S., Mohajerin Esfahani, P., and Kuhn, D. (2015). Distributionally robust logistic regression. In *Advances in Neural Information Processing Systems*, volume 28.

Shapiro, A. (2001). On duality theory of conic linear problems. *Nonconvex Optimization and its Applications*, 57:135–155.

Sherali, H. D. and Adams, W. P. (2013). *A reformulation-linearization technique for solving discrete and continuous nonconvex problems*, volume 31. Springer.

Sherali, H. D. and Fraticelli, B. M. (2002). Enhancing RLT relaxations via a new class of semidefinite cuts. *J. Global Optim.*, 22(1-4):233–261.

Sherali, H. D. and Tuncbilek, C. H. (1992). A global optimization algorithm for polynomial programming problems using a reformulation-linearization technique. *J. Global Optim.*, 2(1):101–112.

Sherali, H. D. and Tuncbilek, C. H. (1995). A reformulation-convexification approach for solving nonconvex quadratic programming problems. *J. Global Optim.*, 7(1):1–31.

Shokri, R., Stronati, M., Song, C., and Shmatikov, V. (2017). Membership inference attacks against machine learning models. In *IEEE symposium on security and privacy*, pages 3–18.

Sinha, A., Namkoong, H., and Duchi, J. C. (2018). Certifying some distributional robustness with principled adversarial training. In *International Conference on Learning Representations*.

Smith, J. E. and Winkler, R. L. (2006). The optimizer's curse: Skepticism and postdecision surprise in decision analysis. *Management Science*, 52(3):311–322.

Sommer, D. M. (2021). *Fighting Uphill Battles: Improvements in Personal Data Privacy*. PhD thesis, ETH Zurich.

Song, C., He, K., Wang, L., and Hopcroft, J. E. (2019). Improving the generalization of adversarial training with domain adaptation. In *International Conference on Learning Representations*.

Soria-Comas, J. and Domingo-Ferrer, J. (2013). Optimal data-independent noise for differential privacy. *Information Sciences*, 250(1):200–214.

Staib, M. and Jegelka, S. (2017). Distributionally robust deep learning as a generalization of adversarial training. In *NIPS workshop on Machine Learning and Computer Security*.

Steinke, T. (2022). Composition of differential privacy & privacy amplification by subsampling. *arXiv:2210.00597*.

Subramanyam, A., Gounaris, C. E., and Wiesemann, W. (2020). $K$-adaptability in two-stage mixed-integer robust optimization. *Mathematical Programming Computation*, 12:193–224.

Sweeney, L. (1997). Weaving technology and policy together to maintain confidentiality. *The Journal of Law, Medicine & Ethics*, 25(2-3):98–110.

Sweeney, L. (2001). *Computational Disclosure Control: A Primer on Data Privacy Protection*. PhD thesis, Massachusetts Institute of Technology.

Sweeney, L. (2002). k-anonymity: A model for protecting privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 10(5):557–570.

Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I. J., and Fergus, R. (2014). Intriguing properties of neural networks. In *International Conference on Learning Representations*.

Tanoumand, N., Bodur, M., and Naoum-Sawaya, J. (2023). Data-driven distributionally robust optimization: Intersecting ambiguity sets, performance analysis and tractability. *Optimization Online 22567*.

Taskesen, B., Yue, M.-C., Blanchet, J., Kuhn, D., and Nguyen, V. A. (2021). Sequential domain adaptation by synthesizing distributionally robust experts. In *International Conference on Machine Learning*.

Toland, J. F. (1978). Duality in nonconvex optimization. *Journal of Mathematical Analysis and Applications*, 66(2):399–415.

Tuy, H. (1964). Concave programming under linear constraints. *Soviet Mathematics Doklady*, 5:1437–1440.

Tuy, H. (1986). A general deterministic approach to global optimization via DC programming. In *North-Holland mathematics studies*, volume 129, pages 273–303. Elsevier.

Tuy, H. and Horst, R. (1988). Convergence and restart in branch-and-bound algorithms for global optimization. application to concave minimization and DC optimization problems. *Mathematical Programming*, 41(1-3):161–183.

Uesato, J., O'donoghue, B., Kohli, P., and Oord, A. (2018). Adversarial risk and the dangers of evaluating against weak attacks. In *International Conference on Machine Learning*.

Ullman, J. and Vadhan, S. (2020). PCPs and the hardness of generating synthetic data. *Journal of Cryptology*, 33(4):2078–2112.

Vadhan, S. (2017). The complexity of differential privacy. In *Tutorials on the Foundations of Cryptography*, pages 347–450. Springer.

Vaidya, J., Shafiq, B., Basu, A., and Hong, Y. (2013). Differentially private naïve Bayes classification. In *Proceedings of the 12th IEEE/WIC/ACM International Joint Conferences on Web Intelligence and Intelligent Agent Technologies*, pages 571–576.

Van Parys, B. P. G., Mohajerin Esfahani, P., and Kuhn, D. (2021). From data to decisions: Distributionally robust optimization is optimal. *Management Science*, 67(6):3387–3402.

Vapnik, V. (1999). *The nature of statistical learning theory*. Springer.

Villani, C. (2009). *Optimal transport: Old and new*. Springer.

Wächter, A. and Biegler, L. T. (2006). On the implementation of an interior-point filter line-search algorithm for large-scale nonlinear programming. *Mathematical Programming*, 106(1):25–57.

Wang, T., Chen, N., and Wang, C. (2024). Contextual optimization under covariate shift: A robust approach by intersecting wasserstein balls. *arXiv:2406.02426*.

Wiesemann, W., Kuhn, D., and Sim, M. (2014). Distributionally robust convex optimization. *Operations Research*, 62(6):1358–1376.

Wong, E., Rice, L., and Kolter, J. Z. (2020). Fast is better than free: Revisiting adversarial training. In *International Conference on Learning Representations*.

Wu, D., Xia, S.-T., and Wang, Y. (2020). Adversarial weight perturbation helps robust generalization. In *Advances in Neural Information Processing Systems*.

Xing, Y., Song, Q., and Cheng, G. (2022a). Unlabeled data help: Minimax analysis and adversarial robustness. In *International Conference on Artificial Intelligence and Statistics*.

Xing, Y., Song, Q., and Cheng, G. (2022b). Why do artificially generated data help adversarial robustness. In *Advances in Neural Information Processing Systems*.

Yanıkoğlu, İ., Gorissen, B. L., and den Hertog, D. (2019). A survey of adjustable robust optimization. *European Journal of Operational Research*, 277(3):799–813.

Ye, J., Gao, J., Li, Q., Xu, H., Feng, J., Wu, Z., Yu, T., and Kong, L. (2022). Zerogen: Efficient zero-shot learning via dataset generation. In *Conference on Empirical Methods in Natural Language Processing*.

Yu, C., Han, B., Shen, L., Yu, J., Gong, C., Gong, M., and Liu, T. (2022). Understanding robust overfitting of adversarial training and beyond. In *International Conference on Machine Learning*.

Yue, M., Kuhn, D., and Wiesemann, W. (2021). On linear optimization over Wasserstein balls. *Mathematical Programming*.

Zass, R. and Shashua, A. (2007). Nonnegative sparse PCA. In Schölkopf, B., Platt, J. C., and Hoffman, T., editors, *Advances in Neural Information Processing Systems 19*, pages 1561–1568. MIT Press.

Zhang, H., Yu, Y., Jiao, J., Xing, E., El Ghaoui, L., and Jordan, M. (2019). Theoretically principled trade-off between robustness and accuracy. In *International Conference on Machine Learning*.

Zhao, J., Wang, T., Bai, T., Lam, K., Ren, X., Yang, X., Shi, S., Liu, Y., and Yu, H. (2019). Reviewing and improving the Gaussian mechanism for differential privacy. *CoRR*, abs/1911.12060.

Zhen, J., De Moor, D., and Den Hertog, D. (2021). An extension of the reformulation-linearization technique to nonlinear optimization problems. *Working Paper*.

Zhen, J., De Ruiter, F., and Den Hertog, D. (2017). Robust optimization for models with uncertain SOC and SDP constraints. *Preprint 6371: Optimization Online*.

Zhen, J., Den Hertog, D., and Sim, M. (2018). Adjustable robust optimization via Fourier–Motzkin Elimination. *Operations Research*, 66(4):1086–1100.

Zhong, E., Fan, W., Yang, Q., Verscheure, O., and Ren, J. (2010). Cross validation framework to choose amongst models and datasets for transfer learning. In *Machine Learning and Knowledge Discovery in Databases: European Conference*, volume 6323.

Zwart, P. B. (1974). Global maximization of a convex function with linear inequality constraints. *Operations Research*, 22(3):602–609.