It's All in the Mix:

Wasserstein Classification and Regression with Mixed Features

Reza Belbasi, Aras Selvi, and Wolfram Wiesemann

Imperial College Business School, London, United Kingdom

{r.belbasi21, a.selvi19, ww}@imperial.ac.uk

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.

1

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., 2019; 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., 2016; Bastani and Bayati, 2020; Bertsimas and Pauphilet, 2024).

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, 2023). 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, 2019; Gao et al., 2024), 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 ϕ-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 (Kelly et al., 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 paper 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

4

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 paper 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 papers 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, 2019; Blanchet et al., 2019; Gao et al., 2024). 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., 2022). While that work focuses on logistic regression, the present paper 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 3 in Section 4), an entirely new proof for the absence of regularized problem formulations (*cf.* Theorem 4 in Section 4) that applies to any loss function (as opposed to only the log-loss function in Selvi et al., 2022) and any non-trivial Wasserstein learning instance, as well as a substantially generalized cutting plane scheme (*cf.* Algorithms 1 and 2 as well as Theorem 2 in Section 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 6).

The remainder of this paper is organized as follows. Section 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 presents and analyzes our cutting plane solution approach for this exponential-size convex program. Section 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 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 6, and we offer concluding remarks in Section 7. Exponential-size convex reformulations of the mixed-feature linear Wasserstein classification and regression problems, which are unified by our representation in Section 2, as well as all proofs are relegated to the e-companion. All datasets and source codes accompanying this work are available open source.[1]

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 $\boldsymbol{x}' - \boldsymbol{x} \in \mathcal{C}$

---

[1]Website: `https://anonymous.4open.science/r/Wasserstein-Mixed-Features-088D/`.

and $\boldsymbol{x}' - \boldsymbol{x} \in \mathrm{int}\,\mathcal{C}$, respectively. The dual norm of $\|\cdot\|$ is $\|\boldsymbol{x}\|_* = \sup_{\boldsymbol{x}' \in \mathbb{R}^n}\{\boldsymbol{x}^\top \boldsymbol{x}' \,:\, \|\boldsymbol{x}'\| \leqslant 1\}$, and the cone dual to a cone $\mathcal{C}$ is $\mathcal{C}^* = \{\boldsymbol{x}' \,:\, \boldsymbol{x}^\top \boldsymbol{x}' \geqslant 0 \;\; \forall \boldsymbol{x} \in \mathcal{C}\}$. 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 $\mathrm{lip}(L) = \sup\{|L(\boldsymbol{x}) - L(\boldsymbol{x}')| / \|\boldsymbol{x} - \boldsymbol{x}'\| \,:\, \boldsymbol{x}, \boldsymbol{x}' \in \mathbb{X}, \; \boldsymbol{x} \neq \boldsymbol{x}'\}$. The set $\mathcal{P}_0(\Xi)$ contains all probability distributions supported on Ξ, 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.

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 $k_m \in \mathbb{N}\backslash\{1\}$ denotes the number of values that the m-th discrete feature can attain, $m \in [K]$, and $\mathbb{Z}(s) = \{\boldsymbol{z} \in \mathbb{B}^{s-1} \,:\, \sum_{i \in [s-1]} z_i \leqslant 1\}$ 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

$$
\begin{aligned}
\underset{\boldsymbol{\beta}}{\text{minimize}} \quad & \sup_{\mathbb{Q} \in \mathfrak{B}_\epsilon(\widehat{\mathbb{P}}_N)} \; \mathbb{E}_{\mathbb{Q}}\left[l_{\boldsymbol{\beta}}(\boldsymbol{x}, \boldsymbol{z}, y)\right] \\
\text{subject to} \quad & \boldsymbol{\beta} = (\beta_0, \boldsymbol{\beta}_{\mathrm{x}}, \boldsymbol{\beta}_{\mathrm{z}}) \in \mathbb{R}^{1 + M_{\mathrm{x}} + M_{\mathrm{z}}},
\end{aligned}
\tag{1}
$$

where the ambiguity set $\mathfrak{B}_\epsilon(\widehat{\mathbb{P}}_N) = \{\mathbb{Q} \in \mathcal{P}_0(\Xi) \,:\, \mathrm{W}(\mathbb{Q}, \widehat{\mathbb{P}}_N) \leqslant \epsilon\}$ 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.

7

Definition 1 (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*

$$W(\mathbb{P}, \mathbb{Q}) := \inf_{\Pi \in \mathcal{P}_0(\Xi^2)} \left\{ \int_{\Xi^2} d(\boldsymbol{\xi}, \boldsymbol{\xi}') \, \Pi(\mathrm{d}\boldsymbol{\xi}, \mathrm{d}\boldsymbol{\xi}') \ : \ \Pi(\mathrm{d}\boldsymbol{\xi}, \Xi) = \mathbb{P}(\mathrm{d}\boldsymbol{\xi}), \ \Pi(\Xi, \mathrm{d}\boldsymbol{\xi}') = \mathbb{Q}(\mathrm{d}\boldsymbol{\xi}') \right\},$$

where the ground metric d on Ξ satisfies

$$d(\boldsymbol{\xi}, \boldsymbol{\xi}') = \|\boldsymbol{x} - \boldsymbol{x}'\| + \kappa_z d_z(\boldsymbol{z}, \boldsymbol{z}') + \kappa_y d_y(y, y') \quad \forall \boldsymbol{\xi} = (\boldsymbol{x}, \boldsymbol{z}, y) \in \Xi, \ \boldsymbol{\xi}' = (\boldsymbol{x}', \boldsymbol{z}', y') \in \Xi \quad \text{(2a)}$$

with $\kappa_z, \kappa_y > 0$ as well as, for some $p \geqslant 1$,

$$d_z(\boldsymbol{z}, \boldsymbol{z}') = \left(\sum_{m \in [K]} \mathbb{1}[z_m \neq z'_m] \right)^{1/p} \quad \text{and} \quad d_y(y, y') = \begin{cases} \mathbb{1}[y \neq y'] & \text{if } \mathbb{Y} = \{-1, +1\}, \\ |y - y'| & \text{otherwise.} \end{cases} \quad \text{(2b)}$$

An important special case of Definition 1 arises when $\kappa_y = \infty$. In this setting, any transport between samples with different outputs incurs infinite cost, and hence the ambiguity set only contains distributions that preserve the empirical output values. In other words, $\mathfrak{B}_\epsilon(\widehat{\mathbb{P}}_N)$ then consists of perturbations of the feature distribution conditional on y, while the marginal distribution of y remains fixed.

The loss function $l_{\boldsymbol{\beta}}(\boldsymbol{x}, \boldsymbol{z}, y) : \mathbb{X} \times \mathbb{Z} \times \mathbb{Y} \to \mathbb{R}_+$ in problem (1) satisfies

$$l_{\boldsymbol{\beta}}(\boldsymbol{x}, \boldsymbol{z}, y) = \begin{cases} L(y \cdot [\beta_0 + \boldsymbol{\beta}_x^\top \boldsymbol{x} + \boldsymbol{\beta}_z^\top \boldsymbol{z}]) & \text{if } \mathbb{Y} = \{-1, +1\}, \\ L(\beta_0 + \boldsymbol{\beta}_x^\top \boldsymbol{x} + \boldsymbol{\beta}_z^\top \boldsymbol{z} - y) & \text{otherwise,} \end{cases}$$

where $L : \mathbb{R} \to \mathbb{R}_+$ measures the similarity between the prediction $\beta_0 + \boldsymbol{\beta}_x^\top \boldsymbol{x} + \boldsymbol{\beta}_z^\top \boldsymbol{z}$ 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:

8

(i) L is convex and Lipschitz continuous with Lipschitz modulus $\mathrm{lip}(L)$, $\mathbb{X} = \mathbb{R}^{M_x}$ 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 $\mathbb{X} \subseteq \mathbb{R}^{M_x}$ and $\mathbb{Y} \subseteq \mathbb{R}$ are closed and convex.

In either case, we assume that L is not constant.

The Wasserstein learning problem (1) 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 (1) as a convex optimization problem.

Observation 1. *The Wasserstein learning problem* (1) *admits the equivalent formulation*

$$
\begin{aligned}
\underset{\boldsymbol{\beta}, \lambda, \boldsymbol{s}}{\text{minimize}} \quad & \lambda \epsilon + \frac{1}{N} \sum_{n \in [N]} s_n \\
\text{subject to} \quad & \sup_{(\boldsymbol{x}, y) \in \mathbb{X} \times \mathbb{Y}} \{ l_{\boldsymbol{\beta}}(\boldsymbol{x}, \boldsymbol{z}, y) - \lambda \|\boldsymbol{x} - \boldsymbol{x}^n\| - \lambda \kappa_y d_y(y, y^n) \} - \lambda \kappa_z d_z(\boldsymbol{z}, \boldsymbol{z}^n) \leqslant s_n \\
& \hspace{6cm} \forall n \in [N], \ \forall \boldsymbol{z} \in \mathbb{Z} \\
& \boldsymbol{\beta} = (\beta_0, \boldsymbol{\beta}_x, \boldsymbol{\beta}_z) \in \mathbb{R}^{1 + M_x + M_z}, \ \ \lambda \in \mathbb{R}_+, \ \ \boldsymbol{s} \in \mathbb{R}_+^N.
\end{aligned}
\tag{3}
$$

When $\kappa_y = \infty$, the semi-infinite constraint in Observation 1 simplifies to

$$
\sup_{\boldsymbol{x} \in \mathbb{X}} \{ l_{\boldsymbol{\beta}}(\boldsymbol{x}, \boldsymbol{z}, y^n) - \lambda \|\boldsymbol{x} - \boldsymbol{x}^n\| \} - \lambda \kappa_z d_z(\boldsymbol{z}, \boldsymbol{z}^n) \leqslant s_n \qquad \forall n \in [N], \ \forall \boldsymbol{z} \in \mathbb{Z}.
$$

Problem (3) contains embedded maximization problems, and it comprises exponentially many constraints. We next prove the existence of equivalent reformulations of (3) 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 (3) leverage the unified problem representation

$$
\begin{aligned}
\underset{\boldsymbol{\theta}, \boldsymbol{\sigma}}{\text{minimize}} \quad & f_0(\boldsymbol{\theta}, \boldsymbol{\sigma}) \\
\text{subject to} \quad & f_{ni}(g_{ni}(\boldsymbol{\theta}, \boldsymbol{\xi}_{-\mathbf{z}}^n; \mathbf{z})) - h_{ni}(\boldsymbol{\theta}, \boldsymbol{\xi}_{-\mathbf{z}}^n; d_{\mathbf{z}}(\mathbf{z}, \mathbf{z}^n)) \leqslant \sigma_n \qquad \forall n \in [N],\ \forall i \in \mathcal{I},\ \forall \mathbf{z} \in \mathbb{Z} \\
& \boldsymbol{\theta} \in \Theta,\ \ \boldsymbol{\sigma} \in \mathbb{R}_+^N,
\end{aligned}
\tag{4}
$$

where $\boldsymbol{\xi}_{-\mathbf{z}}^n = (\boldsymbol{x}^n, y^n)$ and $\mathcal{I}$ is a finite index set. To ensure that (4) 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, $g_{ni} : \mathbb{R}^{M_{\boldsymbol{\theta}}} \times (\mathbb{X} \times \mathbb{Y}) \times \mathbb{R}^{M_{\mathbf{z}}} \to \mathbb{R}$ is bi-affine in $\boldsymbol{\theta}$ and $\mathbf{z}$ for every fixed $\boldsymbol{\xi}_{-\mathbf{z}}^n$, $h_{ni} : \mathbb{R}^{M_{\boldsymbol{\theta}}} \times (\mathbb{X} \times \mathbb{Y}) \times \mathbb{R} \to \mathbb{R}$ is concave in $\boldsymbol{\theta}$ for every fixed $\boldsymbol{\xi}_{-\mathbf{z}}^n$ 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 (4) is not unbounded.

Theorem 1. *The following problems admit equivalent reformulations in the form of* (4)*:*

(i) *the mixed-feature Wasserstein classification and regression problems with convex and Lipschitz continuous loss functions L as well as the continuous feature support $\mathbb{X} = \mathbb{R}^{M_{\times}}$.*

(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*

$$
\mathbb{X} = \{\boldsymbol{x} \in \mathbb{R}^{M_{\times}} : \boldsymbol{C}\boldsymbol{x} \preccurlyeq_{\mathcal{C}} \boldsymbol{d}\} \ \text{for some } \boldsymbol{C} \in \mathbb{R}^{r \times M_{\times}}, \boldsymbol{d} \in \mathbb{R}^r
$$
$$
\text{and proper convex cone } \mathcal{C} \subseteq \mathbb{R}^r
$$

for classification problems as well as the support

$$\mathbb{X} \times \mathbb{Y} = \left\{ (\boldsymbol{x}, y) \in \mathbb{R}^{M_{\mathrm{x}}+1} \,:\, \boldsymbol{C}_{\mathrm{x}} \boldsymbol{x} + \boldsymbol{c}_{\mathrm{y}} \cdot y \preccurlyeq_{\mathcal{C}} \boldsymbol{d} \right\} \; \textit{for some } \boldsymbol{C}_{\mathrm{x}} \in \mathbb{R}^{r \times M_{\mathrm{x}}}, \boldsymbol{c}_{\mathrm{y}} \in \mathbb{R}^r, \boldsymbol{d} \in \mathbb{R}^r$$

$$\textit{and proper convex cone } \mathcal{C} \subseteq \mathbb{R}^r$$

for regression problems, assuming that the supports admit Slater points.

The Appendices A and B in the e-companion prove Theorem 1 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 Cutting Plane Solution Scheme

The unified problem representation (4) of the mixed-feature Wasserstein regression and classification problems (*cf.* Theorem 1) 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 1 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 (4), 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 (4). At each iteration, the algorithm identifies and incorporates the most violated constraint from problem (4) with respect to the current 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 (4) in a finite number of iterations. A common challenge in cutting plane techniques is the derivation of

---
Algorithm 1 Cutting Plane Scheme for Problem (4)

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 (4).
   Initialize $(\mathrm{LB}, \mathrm{UB}) = (-\infty, +\infty)$ as lower and upper bounds for problem (4).
   while $\mathrm{LB} < \mathrm{UB}$ **do**
      Let $(\boldsymbol{\theta}^\star, \boldsymbol{\sigma}^\star)$ be optimal in the relaxation of (4) 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

$$\boldsymbol{z}(n, i) \in \arg\max_{\boldsymbol{z} \in \mathbb{Z}} \left\{ f_{ni}(g_{ni}(\boldsymbol{\theta}^\star, \boldsymbol{\xi}_{-\boldsymbol{z}}^n; \boldsymbol{z})) - h_{ni}(\boldsymbol{\theta}^\star, \boldsymbol{\xi}_{-\boldsymbol{z}}^n; d_{\mathrm{z}}(\boldsymbol{z}, \boldsymbol{z}^n)) - \sigma_n^\star \right\}$$

         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 $\vartheta(n, i(n)) > 0$.
      end for
      Define $\boldsymbol{\vartheta}^\star \in \mathbb{R}^N$ via $\vartheta_n^\star = \max\{\vartheta(n, i(n)), 0\}$, $n \in [N]$.
      Update $\mathrm{LB} = f_0(\boldsymbol{\theta}^\star, \boldsymbol{\sigma}^\star)$ and $\mathrm{UB} = \min\{\mathrm{UB}, f_0(\boldsymbol{\theta}^\star, \boldsymbol{\sigma}^\star + \boldsymbol{\vartheta}^\star)\}$.
   end while

---

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 (4) 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 1. *Algorithm 1 terminates in finite time with an optimal solution $(\boldsymbol{\theta}^\star, \boldsymbol{\sigma}^\star)$ to problem (4). Moreover,* LB *and* UB *constitute monotonic sequences of lower and upper bounds on the optimal value of (4) throughout the execution of the algorithm.*

A key step in Algorithm 1 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 2.

Algorithm 2 determines a most violated constraint index, denoted as $\boldsymbol{z}(n, i)$, for a given

---

Algorithm 2 Identification of a Most Violated Constraint in Problem (4)

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 $\mathcal{Z} = \varnothing$.

  Let $(\boldsymbol{w}, w_0) \in \mathbb{R}^{M_z} \times \mathbb{R}$ 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 $\boldsymbol{z}^\star_m \in \arg\max\{\mu \cdot \boldsymbol{w}_m^\top \boldsymbol{z}_m \,:\, \boldsymbol{z}_m \in \mathbb{Z}(k_m) \backslash \{\boldsymbol{z}^n_m\}\}$ for all $m \in [K]$.

     Compute a permutation $\pi : [K] \to [K]$ such that

$$\mu \cdot \boldsymbol{w}_{\pi(m)}^\top (\boldsymbol{z}^\star_{\pi(m)} - \boldsymbol{z}^n_{\pi(m)}) \;\geqslant\; \mu \cdot \boldsymbol{w}_{\pi(m')}^\top (\boldsymbol{z}^\star_{\pi(m')} - \boldsymbol{z}^n_{\pi(m')}) \qquad \forall 1 \leqslant m \leqslant m' \leqslant K.$$

     for $\delta \in [K] \cup \{0\}$ **do**

        Add $\boldsymbol{z}(\delta)$ to $\mathcal{Z}$, where $\boldsymbol{z}_m(\delta) = \boldsymbol{z}^\star_m$ if $\pi(m) \leqslant \delta$; $= \boldsymbol{z}^n_m$ otherwise, $m \in [K]$.

     end for

  end for

  Select $\boldsymbol{z}(n, i) \in \arg\max\{f_{ni}(g_{ni}(\boldsymbol{\theta}^\star, \boldsymbol{\xi}^n_{-\mathbf{z}}; \boldsymbol{z})) - h_{ni}(\boldsymbol{\theta}^\star, \boldsymbol{\xi}^n_{-\mathbf{z}}; d_{\mathrm{z}}(\boldsymbol{z}, \boldsymbol{z}^n)) - \sigma^\star_n \,:\, \boldsymbol{z} \in \mathcal{Z}\}$.

---

incumbent solution $(\boldsymbol{\theta}^\star, \boldsymbol{\sigma}^\star)$ and a given constraint group $(n, i) \in [N] \times \mathcal{I}$ in the optimization problem (4). We obtain such a constraint by maximizing the constraint left-hand side in (4),

$$f_{ni}(g_{ni}(\boldsymbol{\theta}^\star, \boldsymbol{\xi}^n_{-\mathbf{z}}; \boldsymbol{z})) - h_{ni}(\boldsymbol{\theta}^\star, \boldsymbol{\xi}^n_{-\mathbf{z}}; d_{\mathrm{z}}(\boldsymbol{z}, \boldsymbol{z}^n)),$$

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 1). 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 2 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 2 implements.

To maximize $f_{ni}(g_{ni}(\boldsymbol{\theta}^\star, \boldsymbol{\xi}_{-\boldsymbol{z}}^n; \boldsymbol{z}))$ 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 $f_{ni}(g_{ni}(\boldsymbol{\theta}^\star, \boldsymbol{\xi}_{-\boldsymbol{z}}^n; \boldsymbol{z}))$ 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 $g_{ni}(\boldsymbol{\theta}^\star, \boldsymbol{\xi}_{-\boldsymbol{z}}^n; \boldsymbol{z})$ 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 2, which maximizes $\mu \cdot g_{ni}(\boldsymbol{\theta}^\star, \boldsymbol{\xi}_{-\boldsymbol{z}}^n; \boldsymbol{z})$ for $\mu \in \{\pm 1\}$.

We claim that the optimization of $g_{ni}(\boldsymbol{\theta}^\star, \boldsymbol{\xi}_{-\boldsymbol{z}}^n; \boldsymbol{z})$ over $\boldsymbol{z} \in \mathbb{Z}$ with $d_{\mathrm{z}}(\boldsymbol{z}, \boldsymbol{z}^n) = \delta$ 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 $\boldsymbol{z} \mapsto g_{ni}(\boldsymbol{\theta}^\star, \boldsymbol{\xi}_{-\boldsymbol{z}}^n; \boldsymbol{z})$ as $\boldsymbol{w}^\top \boldsymbol{z} + w_0$ for some $\boldsymbol{w} \in \mathbb{R}^{M_{\mathrm{z}}}$ and $w_0 \in \mathbb{R}$. Moreover, the distance metric $d_{\mathrm{z}}(\boldsymbol{z}, \boldsymbol{z}^n)$ only counts the number of subvectors $m \in [K]$ where $\boldsymbol{z}_m$ differs from $\boldsymbol{z}_m^n$, 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 $g_{ni}(\boldsymbol{\theta}^\star, \boldsymbol{\xi}_{-\boldsymbol{z}}^n; \boldsymbol{z})$ over $\boldsymbol{z} \in \mathbb{Z}$ with $d_{\mathrm{z}}(\boldsymbol{z}, \boldsymbol{z}^n) = \delta$ can be broken into two interconnected steps: We need to select the δ 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 $\boldsymbol{z}_m^\star$ that optimizes $\boldsymbol{w}_m^\top (\boldsymbol{z}_m - \boldsymbol{z}_m^n)$ across all $\boldsymbol{z}_m \neq \boldsymbol{z}_m^n$. To this end, Algorithm 2 first computes for every $m \in [K]$ a solution $\boldsymbol{z}_m^\star$ that optimizes $\boldsymbol{w}_m^\top (\boldsymbol{z}_m - \boldsymbol{z}_m^n)$ over $\boldsymbol{z}_m \neq \boldsymbol{z}_m^n$ (first step inside the outer for-loop), it subsequently sorts the contributions $\boldsymbol{w}_m^\top (\boldsymbol{z}_m^\star - \boldsymbol{z}_m^n)$, $m \in [K]$ (second step inside the outer for-loop), and it finally selects the δ largest (smallest) contributions to maximize (minimize) g_{ni} (inner for-loop).

We formalize the above intuition in the next result.

Theorem 2. *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 (4), Algorithm 2 identifies a most violated constraint index $\boldsymbol{z}(n, i)$ in time*

14

$\mathcal{O}(M_{\mathbf{z}} + KT + K \log K)$, *where T is the time required to compute f_{ni}, g_{ni} and h_{ni}.*

4 Complexity Analysis

Despite its exponential size, our unified problem representation (4) admits a polynomial time solution for the classes of loss functions that we consider in this paper. In fact, it follows from Theorem 2 that Algorithm 2 provides a polynomial time separation oracle for problem (4).

Theorem 3 (Complexity of the Unified Problem Representation (4)). *Assume that the subgradients of the functions f_{ni} and $-h_{ni}$ in problem (4) can be computed in polynomial time and that Θ is compact, full-dimensional and admits a polynomial time weak separation oracle. Then problem (4) can be solved to δ-accuracy in polynomial time.*

Recall that an optimization problem is solved to δ-accuracy if a δ-suboptimal solution is identified that satisfies all constraints modulo a violation of at most δ. The consideration of δ-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 Θ can be assumed to be compact in the Wasserstein classification and regression problems that we consider. The regularity assumptions imposed on problem (4) are crucial in this regard. In fact, it follows from Theorem 2 of Selvi et al. (2022) that if the functions f_{ni} in problem (4) 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 3 principally allows us to solve problem (4) 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 $\mathbb{X} = \mathbb{R}^{M_{\mathbf{x}}}$,

the Wasserstein learning problem (1) reduces to a classical learning problem with an additional regularization term in the objective function whenever the output weight κ_y in Definition 1 approaches ∞ (Shafieezadeh-Abadeh et al., 2015, 2019; Blanchet et al., 2019; Gao et al., 2024). It turns out that this reduction no longer holds when discrete features are present.

Theorem 4 (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 1 as well as Appendices A and B) with any number $N \geqslant 1$ training samples, non-zero numbers M_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 $p \geqslant 1$ in the definition of the Wasserstein ground metric. For sufficiently large radii ϵ of the ambiguity set, the objective function*

$$\sup_{\mathbb{Q} \in \mathfrak{B}_\epsilon(\widehat{\mathbb{P}}_N)} \mathbb{E}_{\mathbb{Q}}\left[l_{\boldsymbol{\beta}}(\boldsymbol{x}, \boldsymbol{z}, y)\right],$$

of the Wasserstein learning problem does not admit an equivalent reformulation as a classical regularized learning problem,

$$\mathbb{E}_{\widehat{\mathbb{P}}_N}\left[l_{\boldsymbol{\beta}}(\boldsymbol{x}, \boldsymbol{z}, y)\right] + \mathfrak{R}(\boldsymbol{\beta}) \qquad \text{for any } \mathfrak{R} : \mathbb{R}^{1+M_\mathrm{x}+M_\mathrm{z}} \to \mathbb{R},$$

even when the weight κ_y of the output distance d_y approaches ∞.

We emphasize that Theorem 4 applies to *any* loss function L and *any* reguarlizer $\mathfrak{R}$. We are not aware of any prior results of this form in the literature.

5 Comparison with Continuous-Feature Formulations

We next contrast the mixed-feature Wasserstein learning problem (3) 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 5.1), as well as with a *bounded* continuous-feature formulation that replaces $\mathbb{Z}$ with its convex hull $\text{conv}(\mathbb{Z})$ (Section 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 6).

5.1 Comparison with Unbounded Continuous-Feature Formulation

Our reformulation (3) 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 $\mathbb{R}^{M_z}$, which would allow us to solve problem (3) 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 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 $\text{Prob}(y = z \mid z) = 0.8$. 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 β_0, which in turn will simplify our exposition. We set the label mismatch cost to $\kappa_y = 1$, 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 (1) with an expectation over the empirical distribution $\widehat{\mathbb{P}}_N$.

(ii) *Mixed-feature Wasserstein model.* We solve the Wasserstein learning problem (3) with the discrete support $\mathbb{Z} = \{-1, +1\}$.

(iii) *Unbounded continuous-feature Wasserstein model.* We solve the Wasserstein learning
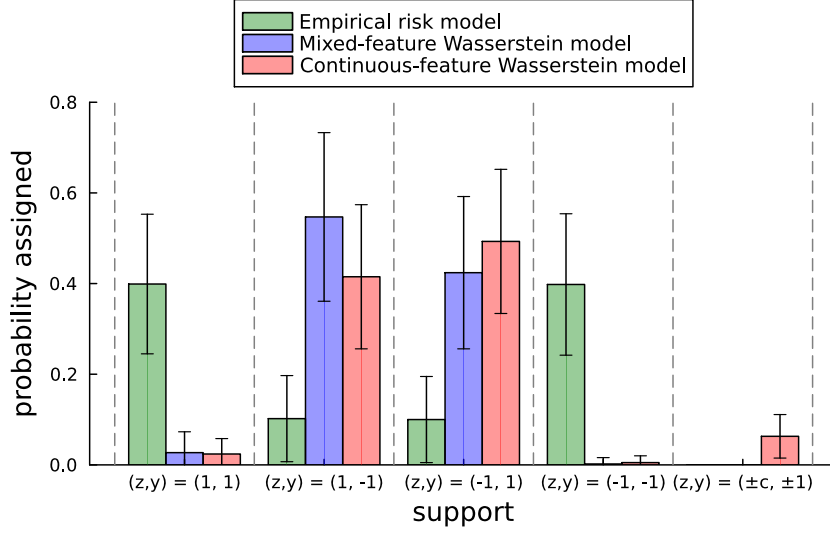
Figure 1: (Worst-case) distributions for our three formulations. The bars represent the mean probabilities, and the whiskers indicate the corresponding standard deviations, computed over 10,000 statistically independent simulations.

problem (3) with the unbounded continuous support $\mathbb{Z} = \mathbb{R}$.

For the fixed hypothesis $\beta_z = 1$ 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 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 ϵ, 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 ϵ increase. In fact, for sufficiently large ϵ the worst-case distributions place a probability mass of 0.1 on a non-sensical atom $(\pm c, \pm 1)$, where c increases with ϵ.

Figure 1 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

18

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 ϵ are chosen via cross-validation.

5.2 Equivalence to Bounded Continuous-Feature Formulation

It is natural to ask whether the mixed-feature Wasserstein learning problem (1) 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 $s \geqslant 2$ and define $\mathbb{U}(s)$ as the convex hull of the one-hot discrete feature encoding $\mathbb{Z}(s)$. One readily observes that

$$\mathbb{U}(s) \;=\; \left\{ \boldsymbol{u} \in [0,1]^{s-1} : \sum_{i \in [s-1]} u_i \leqslant 1 \right\}.$$

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

$$\mathbb{U} \;=\; \text{conv}(\mathbb{Z}) \;=\; \bigtimes_{m \in [K]} \text{conv}(\mathbb{Z}(k_m)) \;=\; \bigtimes_{m \in [K]} \text{conv}(\mathbb{U}(k_m)),$$

where the second identity follows from Exercise 1.19 of Bertsekas (2009). Since the ground

19

metric d_z from Definition 1 is only suitable for pairs of discrete feature vectors, we replace it with the ground metric

$$d_{\mathrm{u}}(\boldsymbol{u}, \boldsymbol{u}') = \left(\sum_{m \in [K]} \frac{1}{2} \|\boldsymbol{u}_m - \boldsymbol{u}'_m\|_1 + \frac{1}{2} |\mathbf{1}^\top (\boldsymbol{u}_m - \boldsymbol{u}'_m)| \right)^{1/p}$$

that extends to pairs of continuous feature vectors. One readily verifies that d_{u} is a metric, that is, it satisfies the identity of indiscernibles, positivity, symmetry and the triangle inequality. The proof of Proposition 2 below will also show that d_z and d_{u} coincide over all pairs of discrete features. We finally define the bounded continuous-feature model as

$$
\begin{aligned}
&\underset{\boldsymbol{\beta}, \lambda, \boldsymbol{s}}{\text{minimize}} \quad \lambda \epsilon + \frac{1}{N} \sum_{n \in [N]} s_n \\
&\text{subject to} \quad \sup_{(\boldsymbol{x}, y) \in \mathbb{X} \times \mathbb{Y}} \{ l_{\boldsymbol{\beta}}(\boldsymbol{x}, \boldsymbol{u}, y) - \lambda \|\boldsymbol{x} - \boldsymbol{x}^n\| - \lambda \kappa_{\mathrm{y}} d_{\mathrm{y}}(y, y^n) \} - \lambda \kappa_{\mathrm{z}} d_{\mathrm{u}}(\boldsymbol{u}, \boldsymbol{z}^n) \leqslant s_n \quad (3') \\
&\hspace{6cm} \forall n \in [N], \ \forall \boldsymbol{u} \in \mathbb{U} \\
&\quad \boldsymbol{\beta} = (\beta_0, \boldsymbol{\beta}_{\mathrm{x}}, \boldsymbol{\beta}_{\mathrm{z}}) \in \mathbb{R}^{1+M_{\mathrm{x}}+M_{\mathrm{z}}}, \ \ \lambda \in \mathbb{R}_+, \ \ \boldsymbol{s} \in \mathbb{R}_+^N,
\end{aligned}
$$

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 2. *The mixed-feature Wasserstein learning problem* (3) *and its corresponding bounded continuous-feature counterpart* (3') *share the same optimal value and the same set of optimal solutions.*

The proof of Proposition 2 crucially relies on the concavity of the family of mappings $\boldsymbol{u} \mapsto d_{\mathrm{u}}(\boldsymbol{u}, \boldsymbol{u}')$, $\boldsymbol{u}' \in \mathbb{Z}$. Indeed, the equivalence stated in the proposition would *not* hold if we were to replace d_{u} with $d'_{\mathrm{u}}(\boldsymbol{u}, \boldsymbol{u}') = \left(\sum_{m \in [K]} \|\boldsymbol{u}_m - \boldsymbol{u}'_m\|_\infty \right)^{1/p}$, despite d'_{u} being a metric on $\mathbb{U}$.

20

Unfortunately, the equivalence of Proposition 2 does not lead to more efficient solution schemes for the mixed-feature Wasserstein learning problem (3). Indeed, we are not aware of any finite-dimensional convex reformulations of the bounded continuous-feature counterpart (3') when the loss function L is convex and Lipschitz continuous. When L is piece-wise affine and convex, on the other hand, problem (3') 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 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.

6   Numerical Results

We compare empirically the performance of our mixed-feature Wasserstein learning problem (1) with classical and regularized learning methods as well as continuous-feature formulations of problem (1). To this end, Section 6.1 compares the out-of-sample losses of our mixed-feature Wasserstein learning problem (1) with those of the unbounded continuous-feature approximation when the Wasserstein radius is selected via cross-validation. Section 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 6.3 compares the out-of-sample performance of our formulation (1) 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.
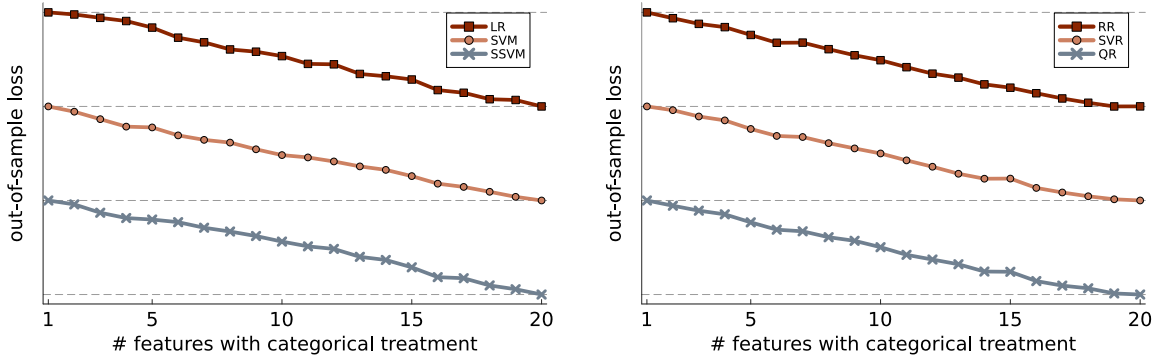
Figure 2: Mean out-of-sample losses for various classification (left) and regression (right) tasks when the number of discrete features that are treated as such varies. All losses are scaled to $[0, 1]$ and shifted so that the curves do not overlap.

6.1 Comparison with Unbounded Continuous-Feature Formulation

Section 5.1 demonstrated that modeling discrete features in the Wasserstein learning problem (1) 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 5.1 have implications on the out-of-sample performance of the mixed-feature and unbounded continuous-feature formulations when the Wasserstein radius ϵ 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), τ-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

22

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 (1) 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 $(\kappa_z, \kappa_y, p) = (1, 1, 1)$ as well as $\|\cdot\| = \|\cdot\|_1$ in our ground metric (2). We refer to the GitHub repository for further details of the instance generation procedure. Figure 2 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 ϵ-grid used for cross-validation; we refer to the GitHub repository for further details.

6.2 Comparison with Bounded Continuous-Feature Formulation

In Section 5.2, we showed that our mixed-feature Wasserstein learning problem (1) 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 (1) is piece-wise 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 1, which compares the monolithic solution of the bounded continuous-feature reformulation with the solution of the mixed-feature model using our cutting plane approach.

23

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 $(\kappa_{\mathrm{z}}, \kappa_{\mathrm{y}}, p) = (1, K, 1)$ 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.

Method	N = 500			N = 1,000			N = 2,500		
	%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 1: 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 τ-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 (1) 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 2 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.

Method	N = 500			N = 1,000			N = 2,500		
	%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 2: 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.

6.3 Performance on Benchmark Instances

In Section 6.1, we observed that the mixed-feature Wasserstein learning problem (1) 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 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 (1) 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 (Kelly et al., 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 3–8 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 (1) using $(\kappa_z, p) = (1, 1)$ and $\|\cdot\| = \|\cdot\|_1$ in our ground metric (2); the results for the corresponding l_2-regularized versions; and the results for the unbounded continuous-feature approximations of problem (1). For the unregularized mixed-feature and continuous-feature Wasserstein learning problems, we jointly cross-validate the hyperparameters (ϵ, κ_y) where $\epsilon \in \{0, 10^{-5}, 10^{-3}, 10^{-1}\}$ and $\kappa_y \in \{1, K, \infty\}$. For the regularized nominal model, we cross-validate the regularization penalty α 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 $(\epsilon, \kappa_y, \alpha)$ from the Cartesian product of the previous three 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 continuous-feature approximation, 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 5.1 and 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.

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 paper, 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

27

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	r-Nom	r-MixF	ConF
balance-scale	625	0	16	4	✗	0.56%	**0.44%** †	0.52%	0.48%	0.44%
					✓	1.65%	**1.47%** †‡	1.92%	1.77%	1.60%
breast-cancer	277	0	42	9	✗	30.10%	**28.46%** †‡	29.46%	**29.36%**	29.45%
					✓	31.87%	**29.88%**	29.58%	**28.31%** †‡	30.48%
credit-approval	690	6	36	9	✗	13.59%	**13.12%** †‡	13.70%	**13.55%**	13.88%
					✓	16.30%	**14.94%**	14.58%	14.94%	15.11%
cylinder-bands	539	19	43	14	✗	22.85%	**22.71%**	22.90%	**22.62%** ‡	23.13%
					✓	27.37%	**26.36%**	27.18%	**26.02%** †	26.41%
lymphography	148	0	42	18	✗	17.24%	**14.66%** †‡	**15.86%**	16.55%	17.41%
					✓	23.98%	**17.73%** ‡	17.96%	**18.07%**	19.14%
primacy	339	0	25	17	✗	**13.51%**	13.96%	14.25%	**13.58%**	14.33%
					✓	15.89%	**14.70%**	14.70%	**14.48%** †	14.68%
spect	267	0	22	22	✗	20.28%	**18.40%**	20.09%	**18.30%**	17.93%
					✓	23.22%	**19.44%**	23.34%	**19.13%** †‡	19.66%
tic-tac-toe	958	0	18	9	✗	1.94%	**1.70%**	**1.70%**	**1.70%**	**1.70%**
					✓	2.75%	**1.65%** †	1.81%	**1.68%**	1.65%

Table 3: Mean classification error for the **logistic regression loss function**. N, M_x, M_z and K refer to the problem parameters from Section 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 (1) 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	r-Nom	r-MixF	ConF
balance-scale	625	0	16	4	✗	**0.44%**	**0.44%** ‡	0.56%	**0.48%**	0.56%
					✓	**1.91%**	1.93%	**2.29%**	2.31%	1.81%
breast-cancer	277	0	42	9	✗	30.27%	**29.27%**	30.46%	**28.36%** †‡	31.27%
					✓	31.66%	**30.24%**	31.05%	**28.98%** †‡	30.57%
credit-approval	690	6	36	9	✗	13.84%	**13.59%**	14.06%	**14.06%**	14.20%
					✓	15.86%	**15.74%**	14.84%	15.33%	15.15%
cylinder-bands	539	19	43	14	✗	23.22%	**22.71%**	23.08%	**22.48%** †‡	23.60%
					✓	28.47%	**27.07%** †	28.55%	27.43%	27.10%
lymphography	148	0	42	18	✗	19.14%	**17.41%** †‡	18.79%	**17.93%**	18.79%
					✓	**20.28%**	21.08%	19.49%	**18.24%** †‡	22.78%
primacy	339	0	25	17	✗	14.03%	**13.81%**	**13.36%**	14.43%	13.96%
					✓	15.96%	**14.88%**	14.53%	**14.53%** ‡	14.88%
spect	267	0	22	22	✗	20.19%	**19.62%**	21.04%	**19.15%** †‡	20.76%
					✓	24.16%	**21.13%**	22.03%	**21.22%**	21.10%
tic-tac-toe	958	0	18	9	✗	**1.70%**	**1.70%**	**1.70%**	**1.70%**	1.70%
					✓	2.58%	**1.65%**	**1.65%**	**1.65%**	1.65%

Table 4: Mean classification error for the **hinge loss function**. We use the same abbreviations as in Table 3.

29

Dataset	N	M_x	M_z	K	Reduced?	Nom	MixF	r-Nom	r-MixF	ConF
balance-scale	625	0	16	4	✗	1.36%	**0.40% †‡**	1.52%	**0.48%**	0.52%
					✓	3.15%	**1.89%**	2.07%	**2.01%**	**1.71%**
breast-cancer	277	0	42	9	✗	29.55%	**29.09% †‡**	30.18%	**30.00%**	30.18%
					✓	31.30%	**30.00%**	28.55%	**28.19% †‡**	31.08%
credit-approval	690	6	36	9	✗	13.44%	**13.37%**	13.44%	**13.33% ‡**	14.42%
					✓	15.59%	**15.31%**	14.65%	**14.43% †‡**	15.09%
cylinder-bands	539	19	43	14	✗	23.27%	**22.06% †‡**	23.04%	**22.66%**	22.62%
					✓	27.43%	**26.49% †**	27.09%	**26.75%**	26.73%
lymphography	148	0	42	18	✗	19.31%	**16.90%**	**16.38%**	**16.38% ‡**	17.76%
					✓	21.31%	**20.11%**	18.47%	**18.07% ‡**	20.17%
primacy	339	0	25	17	✗	13.51%	**13.43%**	**13.13%**	13.66%	14.03%
					✓	15.10%	**14.78%**	14.38%	**14.34%**	14.41%
spect	267	0	22	22	✗	20.28%	**18.30%**	19.25%	**16.79% †‡**	18.11%
					✓	23.16%	**19.91% †**	23.16%	**20.16%**	20.03%
tic-tac-toe	958	0	18	9	✗	1.73%	**1.70%**	**1.70%**	**1.70%**	1.70%
					✓	2.77%	**1.65%**	**1.65%**	**1.65%**	1.65%

Table 5: Mean classification error for the **smooth hinge loss function**. We use the same abbreviations as in Table 3.

Dataset	N	M_x	M_z	K	Reduced?	Nom	MixF	r-Nom	r-MixF	ConF
bike	17,379	4	31	5	✗	**210.23**	**210.23**	**210.23**	**210.23**	210.23
					✓	210.82	**210.81**	**210.84**	**210.84**	213.47
fire	517	10	31	4	✗	**123.33**	**123.33**	123.48	**122.32 †**	122.44
					✓	132.71	**109.84**	114.39	**108.40**	108.12
flare	1,066	0	21	9	✗	**198.46**	**198.46**	198.96	**198.46**	198.44
					✓	197.85	**195.65 †‡**	198.56	**197.15**	198.91
garments	1,197	7	22	4	✗	**363.11**	363.44	**362.55**	362.55	363.36
					✓	890.81	**367.95**	368.15	**367.12 †‡**	368.90
imports	193	14	45	10	✗	442.71	**288.57**	311.08	**269.95 †‡**	288.85
					✓	660.04	**369.06**	352.79	**335.30 †‡**	370.57
student	395	13	26	17	✗	**379.13**	382.07	378.66	**378.12 †‡**	383.78
					✓	**411.98**	414.29	**384.20**	384.46	399.94
vegas	504	5	106	14	✗	232,330.88	**506.61**	**479.91**	**479.91 ‡**	482.06
					✓	287,615.23	**515.22**	**493.60**	496.51	498.68

Table 6: Mean squared errors for the **Huber loss function** (with $\delta = 0.5$). We use the same abbreviations as in Table 3.

Dataset	N	M_x	M_z	K	Reduced?	Nom	MixF	r-Nom	r-MixF	ConF
bike	17,379	4	31	5	✘	217.87	**217.86**	217.87	**217.83**	217.86
					✓	218.17	**218.15**	218.17	**218.12**	218.19
fire	517	10	31	4	✘	124.36	**124.35**	130.45	**124.46**	124.36
					✓	120.54	**109.80**	116.67	**109.85**	109.81
flare	1,066	0	21	9	✘	**218.56**	218.67	218.27	**204.93 †‡**	218.56
					✓	2,230.29	**484.92**	215.22	**203.17 †‡**	216.09
garments	1,197	7	22	4	✘	374.08	**373.93**	373.32	**372.87 †‡**	376.05
					✓	894.07	**376.22 †**	377.78	**376.44**	376.92
imports	193	14	45	10	✘	671.67	**303.77**	351.66	**283.95 †‡**	314.84
					✓	726.46	**386.98**	371.56	**350.52 †‡**	394.28
student	395	13	26	17	✘	399.65	**396.69**	387.07	**386.44 ‡**	404.71
					✓	451.85	**397.77**	403.51	**393.42 †‡**	414.37
vegas	504	5	106	14	✘	232,323.47	**503.93**	493.28	**490.19 †‡**	507.81
					✓	287,607.592	**507.60**	517.85	**498.17 †‡**	510.78

Table 7: Mean squared errors for the **pinball loss function** (with $\tau = 0.5$). We use the same abbreviations as in Table 3.

Dataset	N	M_x	M_z	K	Reduced?	Nom	MixF	r-Nom	r-MixF	ConF
bike	17,379	4	31	5	✘	217.87	**217.61**	217.88	**217.62**	217.61
					✓	218.18	**217.96**	218.15	**217.95**	217.99
fire	517	10	31	4	✘	133.10	**122.58**	136.26	**121.13 †‡**	122.58
					✓	185.04	**108.54**	125.23	**108.28 †**	108.54
flare	1,066	0	21	9	✘	**212.93**	213.55	215.49	**211.32 †‡**	214.00
					✓	1,562.88	**209.84**	232.67	**207.79 †‡**	210.22
garments	1,197	7	22	4	✘	369.10	**368.69**	370.08	**369.04**	368.89
					✓	871.46	**371.12**	**370.72**	371.45	371.79
imports	193	14	45	10	✘	621.38	**293.31 †‡**	331.36	**297.68**	314.97
					✓	761.05	**378.83**	370.97	**338.95 †‡**	382.32
student	395	13	26	17	✘	399.65	**397.63**	383.57	**383.18 ‡**	404.89
					✓	451.93	**397.62 †‡**	410.40	**399.46**	412.98
vegas	504	5	106	14	✘	232,332.44	**502.89**	503.37	**492.94 †‡**	506.67
					✓	287,597.53	**507.90**	507.74	**502.73 †‡**	510.15

Table 8: Mean squared errors for the τ-**insensitive loss function** (with $\tau = 0.1$). We use the same abbreviations as in Table 3.

References

Alley, M., M. Biggs, R. Hariss, C. Herrmann, M. L. Li, and G. Perakis (2023). Pricing for heterogeneous products: Analytics for ticket reselling. *Manufacturing & Service Operations Management 25*(2), 409–426.

Ban, G.-Y. and C. Rudin (2019). The big data newsvendor: Practical insights from machine learning. *Operations Research 67*(1), 90–108.

Bastani, H. and M. Bayati (2020). Online decision making with high-dimensional covariates. *Operations Research 68*(1), 276–294.

Behrendt, A., M. Savelsbergh, and H. Wang (2023). A prescriptive machine learning method for courier scheduling on crowdsourced delivery platforms. *Transportation Science 57*(4), 889–907.

Ben-Tal, A., L. E. Ghaoui, and A. Nemirovski (2009). *Robust Optimization*. Princeton University Press.

Bertsekas, D. (2009). *Convex optimization theory*, Volume 1. Athena Scientific.

Bertsimas, D., A. Delarue, P. Jaillet, and S. Martin (2019). Travel time estimation in the age of big data. *Operations Research 67*(2), 498–515.

Bertsimas, D. and D. den Hertog (2022). *Robust and Adaptive Optimization*. Dynamic Ideas.

Bertsimas, D. and N. Kallus (2020). From predictive to prescriptive analytics. *Management Science 66*(3), 1025–1044.

Bertsimas, D., A. O'Hair, S. Relyea, and J. Silberholz (2016). An analytics approach to designing combination chemotherapy regimens for cancer. *Management Science 62*(5), 1511–1531.

Bertsimas, D. and J. Pauphilet (2024). Hospital-wide inpatient flow optimization. *Manag. Sci. 70*(7), 4893–4911.

Blanchet, J. and Y. Kang (2021). Sample out-of-sample inference based on Wasserstein distance. *Operations Research 69*(3), 985–1013.

Blanchet, J., Y. Kang, and K. Murthy (2019). Robust Wasserstein profile inference and applications to machine learning. *Journal of Applied Probability 56*(3), 830–857.

Blanchet, J. and K. Murthy (2019). Quantifying distributional model risk via optimal transport. *Mathematics of Operations Research 44*(2), 565–600.

Boyd, S. and L. Vandenberghe (2004). *Convex optimization*. Cambridge university press.

Chan, T. C. Y., R. Mahmood, D. L. O'Connor, D. Stone, S. Unger, R. K. Wong, and I. Y. Zhu (2023). Got (optimal) milk? Pooling donations in human milk banks with machine learning and optimization. *Manufacturing & Service Operations Management*, Available ahead of print.

Duchi, J., T. Hashimoto, and H. Namkoong (2023). Distributionally robust losses for latent covariate mixtures. *Operations Research 71*(2), 649–664.

Feldman, J., D. J. Zhang, X. Liu, and N. Zhang (2022). Customer choice models vs. machine learning: Finding optimal product displays on Alibaba. *Operations Research 70*(1), 309–328.

Ferreira, K. J., B. H. A. Lee, and D. Simchi-Levi (2016). Analytics for an online retailer: Demand forecasting and price optimization. *Manufacturing & Service Operations Management 18*(1), 69–88.

Gao, R. (2023). Finite-sample guarantees for wasserstein distributionally robust optimization: Breaking the curse of dimensionality. *Operations Research 71*(6), 2291–2306.

Gao, R., X. Chen, and A. J. Kleywegt (2024). Wasserstein distributionally robust optimization and variation regularization. *Operations Research 72*(3), 1177–1191.

Gao, R. and A. Kleywegt (2023). Distributionally robust stochastic optimization with Wasserstein distance. *Mathematics of Operations Research 48*(2), 603–655.

Glaeser, C. K., M. Fisher, and X. Su (2019). Optimal retail location: Empirical methodology and application to practice. *Manufacturing & Service Operations Management 21*(1), 86–102.

Grötschel, M., L. Lovász, and A. Schrijver (1988). *Geometric Algorithms and Combinatorial Optimization*. Springer.

Hastie, T., R. Tibshirani, and J. Friedman (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer.

Kallus, N. and M. Udell (2020). Dynamic assortment personalization in high dimensions. *Operations Research 68*(4), 1020–1037.

Kelly, M., R. Longjohn, and K. Nottingham (2017). The UCI machine learning repository. `https://archive.ics.uci.edu`.

Kuhn, D., P. Mohajerin Esfahani, V. Nguyen, and S. Shafieezadeh-Abadeh (2019). Wasserstein distributionally robust optimization: Theory and applications in machine learning. *INFORMS TutORials in Operations Research*, 130–169.

Kuhn, D., S. Shafiee, and W. Wiesemann (2025). Distributionally robust optimization. *Acta Numerica 34*, 579–804.

Lam, H. (2019). Recovering best statistical guarantees via the empirical divergence-based distributionally robust optimization. *Operations Research 6*(4), 1090–1105.

Lee, Y. T., A. Sidford, and S. C.-W. Wong (2015). A faster cutting plane method and its implications for combinatorial and convex optimization. In *2015 IEEE 56th Annual Symposium on Foundations of Computer Science*, pp. 1049–1065.

Li, R., M. Tobey, M. E. Mayorga, S. Caltagirone, and O. Y. Özaltın (2023). Detecting human trafficking: Automated classification of online customer reviews of massage businesses. *Manufacturing & Service Operations Management 25*(3), 1051–1065.

Lubin, M., O. Dowson, J. Dias Garcia, J. Huchette, B. Legat, and J. P. Vielma (2023). JuMP 1.0: Recent improvements to a modeling language for mathematical optimization. *Mathematical Programming Computation 15*(3), 581–589.

Michaud, R. O. (1989). The Markowitz optimization enigma: Is 'optimized'optimal? *Financial Analysts Journal 45*(1), 31–42.

Mohajerin Esfahani, P. and D. Kuhn (2018). Data-driven distributionally robust optimization using the Wasserstein metric: Performance guarantees and tractable reformulations. *Mathematical Programming 171*(1-2), 1–52.

Murphy, K. P. (2022). *Probabilistic Machine Learning: An Introduction*. MIT Press.

Qi, M., Y. Cao, and Z.-J. Shen (2022). Distributionally robust conditional quantile prediction with fixed design. *Management Science 68*(3), 1639–1658.

Rahimian, H. and S. Mehrotra (2022). Frameworks and results in distributionally robust optimization. *Open Journal of Mathematical Optimization 3*, 1–85.

Samorani, M., S. L. Harris, L. G. Blount, H. Lu, and M. A. Santoro (2022). Overbooked and overlooked: Machine learning and racial bias in medical appointment scheduling. *Manufacturing & Service Operations Management 24*(6), 2825–2842.

Selvi, A., M. Belbasi, M. Haugh, and W. Wiesemann (2022). Wasserstein logistic regression with mixed features. In *Advances in Neural Information Processing Systems*, Volume 35, pp. 16691–16704.

Shafieezadeh-Abadeh, S., D. Kuhn, and P. Mohajerin Esfahani (2019). Regularization via mass transportation. *Journal of Machine Learning Research 20*(103), 1–68.

Shafieezadeh-Abadeh, S., P. Mohajerin Esfahani, and D. Kuhn (2015). Distributionally robust logistic regression. In *Advances in Neural Information Processing Systems*, Volume 28, pp. 1576–1584.

Smith, J. E. and R. L. Winkler (2006). The optimizer's curse: Skepticism and postdecision surprise in decision analysis. *Management Science 52*(3), 311–322.

Van Parys, B. P. G., P. Mohajerin Esfahani, and D. Kuhn (2021). From data to decisions: Distributionally robust optimization is optimal. *Management Science 67*(6), 3387–3402.

E-Companion to "It's All in the Mix: Wasserstein Machine Learning with Mixed Features"

Appendix A: Wasserstein Classification with Mixed Features

This section derives exponential-size convex reformulations of the mixed-feature Wasserstein learning problem (3) 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 (4), which implies that our results in this appendix prove Theorem 1 for the case of classification problems.

Proposition 3. *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 (3) is equivalent to*

$$
\begin{aligned}
& \underset{\boldsymbol{\beta}, \lambda, \boldsymbol{s}}{\text{minimize}} \quad \lambda\epsilon + \frac{1}{N}\sum_{n\in[N]} s_n \\
& \text{subject to} \quad
\left.
\begin{aligned}
& l_{\boldsymbol{\beta}}(\boldsymbol{x}^n, \boldsymbol{z}, y^n) - \lambda\kappa_{\mathrm{z}}d_{\mathrm{z}}(\boldsymbol{z}, \boldsymbol{z}^n) \leqslant s_n \\
& l_{\boldsymbol{\beta}}(\boldsymbol{x}^n, \boldsymbol{z}, -y^n) - \lambda\kappa_{\mathrm{z}}d_{\mathrm{z}}(\boldsymbol{z}, \boldsymbol{z}^n) - \lambda\kappa_{\mathrm{y}} \leqslant s_n
\end{aligned}
\right\} \; \forall n \in [N], \; \forall \boldsymbol{z} \in \mathbb{Z} \\
& \qquad\qquad\quad \mathrm{lip}(L)\cdot\|\boldsymbol{\beta}_{\mathrm{x}}\|_* \leqslant \lambda \\
& \qquad\qquad\quad \boldsymbol{\beta} = (\beta_0, \boldsymbol{\beta}_{\mathrm{x}}, \boldsymbol{\beta}_{\mathrm{z}}) \in \mathbb{R}^{1+M_{\mathrm{x}}+M_{\mathrm{z}}}, \;\; \lambda \in \mathbb{R}_+, \;\; \boldsymbol{s} \in \mathbb{R}_+^N.
\end{aligned}
\tag{E-1}
$$

When $\kappa_{\mathrm{y}} = \infty$, the second constraint set in Proposition 3 is redundant for any $\lambda > 0$. This is the case whenever λ does not vanish at optimality (*i.e.*, the Wasserstein radius ε actually affects the optimal solution).

Formulation (E-1) emerges as a special case of our unified representation (4) if we set $\mathcal{I} = \{\pm 1\}$, $\boldsymbol{\theta} = (\boldsymbol{\beta}, \lambda)$, $\boldsymbol{\sigma} = \boldsymbol{s}$, $f_0(\boldsymbol{\theta}, \boldsymbol{\sigma}) = \lambda\epsilon + \frac{1}{N}\sum_{n\in[N]} s_n$, $f_{ni}(e) = L(e)$, $g_{ni}(\boldsymbol{\theta}, \boldsymbol{\xi}^n_{-\mathbf{z}}; \boldsymbol{z}) = iy^n \cdot [\beta_0 + \boldsymbol{\beta}_{\mathrm{x}}^\top \boldsymbol{x}^n + \boldsymbol{\beta}_{\mathrm{z}}^\top \boldsymbol{z}]$, $h_{ni}(\boldsymbol{\theta}, \boldsymbol{\xi}^n_{-\mathbf{z}}; d_{\mathrm{z}}(\boldsymbol{z}, \boldsymbol{z}^n)) = \lambda\kappa_{\mathrm{z}}d_{\mathrm{z}}(\boldsymbol{z}, \boldsymbol{z}^n) + \lambda\kappa_{\mathrm{y}} \cdot \mathbb{1}[i = -1]$ and $\Theta = \{\boldsymbol{\theta} : \mathrm{lip}(L)\cdot\|\boldsymbol{\beta}_{\mathrm{x}}\|_* \leqslant \lambda, \lambda \in \mathbb{R}_+\}$. One verifies that this choice of $\mathcal{I}$, $\boldsymbol{\theta}$, $\boldsymbol{\sigma}$, f_0, f_{ni}, g_{ni}, h_{ni} and Θ

satisfies the regularity conditions imposed on problem (4) in the main text.

Proposition 4. *If $\boldsymbol{\beta}$ in problem* (E-1) *is restricted to a bounded hypothesis set $\mathcal{H}$, then λ can be restricted to a bounded set as well.*

Proposition 3 covers Wasserstein support vector machines with a smooth Hinge loss,

$$
L(e) = \begin{cases} 1/2 - e & \text{if } e \leqslant 0, \\[2mm] (1/2) \cdot (1 - e)^2 & \text{if } e \in (0, 1), \\[2mm] 0 & \text{otherwise}, \end{cases}
$$

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 1. *The first set of inequality constraints in* (E-1) *can be reformulated as follows.*

(i) *For the smooth Hinge loss function:*

$$
\left.\begin{array}{l} \dfrac{1}{2} \left(w_{\boldsymbol{z},n}^{+} - y^n \cdot (\beta_0 + \boldsymbol{\beta}_{\mathrm{x}}^{\top}\boldsymbol{x}^n + \boldsymbol{\beta}_{\mathrm{z}}^{\top}\boldsymbol{z})\right)^2 + 1 - w_{\boldsymbol{z},n}^{+} - \lambda\kappa_{\mathrm{z}}d_{\mathrm{z}}(\boldsymbol{z}, \boldsymbol{z}^n) \leqslant s_n \\[4mm] \dfrac{1}{2} \left(w_{\boldsymbol{z},n}^{+} - y^n \cdot (\beta_0 + \boldsymbol{\beta}_{\mathrm{x}}^{\top}\boldsymbol{x}^n + \boldsymbol{\beta}_{\mathrm{z}}^{\top}\boldsymbol{z})\right)^2 - \lambda\kappa_{\mathrm{z}}d_{\mathrm{z}}(\boldsymbol{z}, \boldsymbol{z}^n) \leqslant s_n \end{array}\right\} \forall n \in [N], \ \forall \boldsymbol{z} \in \mathbb{Z}
$$

(ii) *For the log-loss function:*

$$
\log\left(1 + \exp\left[-y^n \cdot (\beta_0 + \boldsymbol{\beta}_{\mathrm{x}}^{\top}\boldsymbol{x}^n + \boldsymbol{\beta}_{\mathrm{z}}^{\top}\boldsymbol{z})\right]\right) - \lambda\kappa_{\mathrm{z}}d_{\mathrm{z}}(\boldsymbol{z}, \boldsymbol{z}^n) \leqslant s_n \quad \forall n \in [N], \ \forall \boldsymbol{z} \in \mathbb{Z}
$$

Here, $w_{\boldsymbol{z},n}^{+} \in \mathbb{R}$ are auxiliary decision variables. The second set of inequality constraints in (E-1) *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 $\lambda\kappa_y$ from the constraint left-hand sides.

We now provide a reformulation of problem (3) without embedded maximizations when the loss function L is piece-wise affine and convex.

Proposition 5. *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 $\mathbb{X} = \{\boldsymbol{x}\in\mathbb{R}^{M_x} : \boldsymbol{Cx} \preccurlyeq_{\mathcal{C}} \boldsymbol{d}\}$ for some $\boldsymbol{C}\in\mathbb{R}^{r\times M_x}$, $\boldsymbol{d}\in\mathbb{R}^r$ and proper convex cone $\mathcal{C}\subseteq\mathbb{R}^r$. If $\mathbb{X}$ admits a Slater point $\boldsymbol{x}^s\in\mathbb{R}^{M_x}$ such that $\boldsymbol{Cx}^s \prec_{\mathcal{C}} \boldsymbol{d}$, then the Wasserstein learning problem (3) is equivalent to*

$$
\begin{aligned}
&\underset{\boldsymbol{\beta},\lambda,\boldsymbol{s},\boldsymbol{q}_{nj}^+,\boldsymbol{q}_{nj}^-}{\text{minimize}} \quad \lambda\epsilon + \frac{1}{N}\sum_{n\in[N]} s_n \\
&\text{subject to} \quad \left.\begin{aligned}
&\boldsymbol{q}_{nj}^{+\top}(\boldsymbol{d} - \boldsymbol{Cx}^n) + a_j y^n \cdot (\beta_0 + \boldsymbol{\beta}_x^\top \boldsymbol{x}^n + \boldsymbol{\beta}_z^\top \boldsymbol{z}) - \lambda\kappa_z d_z(\boldsymbol{z},\boldsymbol{z}^n) + b_j \leqslant s_n \\
&\boldsymbol{q}_{nj}^{-\top}(\boldsymbol{d} - \boldsymbol{Cx}^n) - a_j y^n \cdot (\beta_0 + \boldsymbol{\beta}_x^\top \boldsymbol{x}^n + \boldsymbol{\beta}_z^\top \boldsymbol{z}) - \lambda\kappa_z d_z(\boldsymbol{z},\boldsymbol{z}^n) - \lambda\kappa_y + b_j \leqslant s_n
\end{aligned}\right\} \\
&\hspace{6.5cm} \forall n\in[N],\ \forall j\in[J],\ \forall \boldsymbol{z}\in\mathbb{Z} \\
&\hspace{2cm} \|a_j y^n \cdot \boldsymbol{\beta}_x - \boldsymbol{C}^\top \boldsymbol{q}_{nj}^+\|_* \leqslant \lambda, \quad \|a_j y^n \cdot \boldsymbol{\beta}_x + \boldsymbol{C}^\top \boldsymbol{q}_{nj}^-\|_* \leqslant \lambda \qquad \forall n\in[N],\ \forall j\in[J] \\
&\hspace{2cm} \boldsymbol{\beta} = (\beta_0, \boldsymbol{\beta}_x, \boldsymbol{\beta}_z) \in \mathbb{R}^{1+M_x+M_z}, \quad \lambda\in\mathbb{R}_+, \quad \boldsymbol{s}\in\mathbb{R}_+^N \\
&\hspace{2cm} \boldsymbol{q}_{nj}^+, \boldsymbol{q}_{nj}^- \in \mathcal{C}^*,\ n\in[N]\ \text{and}\ j\in[J].
\end{aligned}
$$

$$(\text{E-2})$$

When $\kappa_y = \infty$, the constraints involving $\boldsymbol{q}_{nj}^-$ are redundant whenever $\lambda > 0$. In this case, the variables $\boldsymbol{q}_{nj}^-$ and all constraints involving $\boldsymbol{q}_{nj}^-$ can be removed from Proposition 5.

Formulation (E-2) emerges as a special case of our unified representation (4) if we set $i = (j,t)$, $\mathcal{I} = [J]\times\{\pm 1\}$, $\boldsymbol{\theta} = (\boldsymbol{\beta},\lambda,\boldsymbol{q}_{ni})$, $\boldsymbol{\sigma} = \boldsymbol{s}$, $f_0(\boldsymbol{\theta},\boldsymbol{\sigma}) = \lambda\epsilon + \frac{1}{N}\sum_{n\in[N]} s_n$, $f_{ni}(e) = a_j e + b_j$, $g_{ni}(\boldsymbol{\theta},\boldsymbol{\xi}_{-z}^n;\boldsymbol{z}) = t y^n \cdot [\beta_0 + \boldsymbol{\beta}_x^\top \boldsymbol{x}^n + \boldsymbol{\beta}_z^\top \boldsymbol{z}]$, $h_{ni}(\boldsymbol{\theta},\boldsymbol{\xi}_{-z}^n;d_z(\boldsymbol{z},\boldsymbol{z}^n)) = -\boldsymbol{q}_{ni}^\top(\boldsymbol{d} - \boldsymbol{Cx}^n) + \lambda\kappa_z d_z(\boldsymbol{z},\boldsymbol{z}^n) + \lambda\kappa_y \cdot \mathbb{1}[t = -1]$ and $\Theta = \{\boldsymbol{\theta} : \|a_j y^n \cdot \boldsymbol{\beta}_x - t\cdot\boldsymbol{C}^\top \boldsymbol{q}_{ni}\|_* \leqslant \lambda\, \forall(n,i)\in[N]\times\mathcal{I},\ \boldsymbol{q}_{ni}\in\mathcal{C}^*\,\forall(n,i)\in[N]\times\mathcal{I},\ \lambda\in\mathbb{R}_+\}$. One verifies that this choice of $\mathcal{I}$, $\boldsymbol{\theta}$, $\boldsymbol{\sigma}$, f_0, f_{ni}, g_{ni}, h_{ni} and Θ satisfies the regularity conditions imposed on problem (4) in the main text.

Proposition 6. *If $\boldsymbol{\beta}$ in problem (E-2) is restricted to a bounded hypothesis set $\mathcal{H}$, then $(\boldsymbol{q}_{ni}^+, \boldsymbol{q}_{ni}^-)_{n,i}$ and λ can be restricted to a bounded set as well.*

Proposition 5 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 2. *For the (non-smooth) Hinge loss function, the first set of inequality constraints in* (E-2) *can be reformulated as*

$$\boldsymbol{q}_n^{+\top}(\boldsymbol{d} - \boldsymbol{C}\boldsymbol{x}^n) - y^n(\beta_0 + \boldsymbol{\beta}_{\mathrm{x}}^\top \boldsymbol{x}^n + \boldsymbol{\beta}_{\mathrm{z}}^\top \boldsymbol{z}) - \lambda \kappa_{\mathrm{z}} d_{\mathrm{z}}(\boldsymbol{z}, \boldsymbol{z}^n) + 1 \leqslant s_n \quad \forall n \in [N],\ \forall \boldsymbol{z} \in \mathbb{Z}.$$

The second set of inequality constraints in (E-2) *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.*

Appendix B: Wasserstein Regression with Mixed Features

This section derives exponential-size convex reformulations of the mixed-feature Wasserstein learning problem (3) 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 (4), which implies that our results in this appendix prove Theorem 1 for the case of regression problems.

Proposition 7. *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 (3) equals*

$$
\begin{aligned}
\underset{\boldsymbol{\beta}, \lambda, \boldsymbol{s}}{\text{minimize}} \quad & \lambda \epsilon + \frac{1}{N} \sum_{n \in [N]} s_n \\
\text{subject to} \quad & l_{\boldsymbol{\beta}}(\boldsymbol{x}^n, \boldsymbol{z}, y^n) - \lambda \kappa_{\mathrm{z}} d_{\mathrm{z}}(\boldsymbol{z}, \boldsymbol{z}^n) \leqslant s_n \qquad \forall n \in [N],\ \forall \boldsymbol{z} \in \mathbb{Z} \\
& \mathrm{lip}(L) \cdot \|\boldsymbol{\beta}_{\mathrm{x}}\|_* \leqslant \lambda \\
& \mathrm{lip}(L) \leqslant \lambda \kappa_{\mathrm{y}} \\
& \boldsymbol{\beta} = (\beta_0, \boldsymbol{\beta}_{\mathrm{x}}, \boldsymbol{\beta}_{\mathrm{z}}) \in \mathbb{R}^{1 + M_{\mathrm{x}} + M_{\mathrm{z}}},\ \ \lambda \in \mathbb{R}_+,\ \ \boldsymbol{s} \in \mathbb{R}_+^N.
\end{aligned}
\tag{E-3}
$$

When $\kappa_{\mathrm{y}} = \infty$ and $\lambda > 0$, the third constraint in Proposition 7 is redundant.

Formulation (E-3) emerges as a special case of our unified representation (4) if we set $\boldsymbol{\theta} = (\boldsymbol{\beta}, \lambda)$, $\boldsymbol{\sigma} = \boldsymbol{s}$, $f_0(\boldsymbol{\theta}, \boldsymbol{\sigma}) = \lambda \epsilon + \frac{1}{N} \sum_{n \in [N]} s_n$, $f_{ni}(e) = L(e)$, $g_{ni}(\boldsymbol{\theta}, \boldsymbol{\xi}_{-\mathbf{z}}^n; \boldsymbol{z}) = \beta_0 + \boldsymbol{\beta}_{\mathrm{x}}^\top \boldsymbol{x}^n + \boldsymbol{\beta}_{\mathrm{z}}^\top \boldsymbol{z} - y^n$, $h_{ni}(\boldsymbol{\theta}, \boldsymbol{\xi}_{-\mathbf{z}}^n; d_{\mathrm{z}}(\boldsymbol{z}, \boldsymbol{z}^n)) = \lambda \kappa_{\mathrm{z}} d_{\mathrm{z}}(\boldsymbol{z}, \boldsymbol{z}^n)$ and $\Theta = \{\boldsymbol{\theta} : \mathrm{lip}(L) \cdot \|\boldsymbol{\beta}_{\mathrm{x}}\|_* \leqslant \lambda, \mathrm{lip}(L) \leqslant \lambda \kappa_{\mathrm{y}}, \lambda \in \mathbb{R}_+\}$. One verifies that this choice of $\boldsymbol{\theta}$, $\boldsymbol{\sigma}$, f_0, f_{ni}, g_{ni}, h_{ni} and Θ satisfies the regularity conditions imposed on problem (4) in the main text.

Corollary 3. *If $\boldsymbol{\beta}$ in problem (E-3) is restricted to a bounded hypothesis set $\mathcal{H}$, then λ can be restricted to a bounded set as well.*

Proposition 7 covers Wasserstein regression with a Huber loss,

$$L(e) = \begin{cases} (1/2) \cdot e^2 & \text{if } |e| \leqslant \delta, \\[2ex] \delta \cdot (|e| - (1/2) \cdot \delta) & \text{otherwise,} \end{cases}$$

where $\delta \in \mathbb{R}_+$ determines the boundary between the quadratic and the absolute loss, implying that $\text{lip}(L) = \delta$. We provide the corresponding reformulation next.

Corollary 4. *For the Huber loss function, the first set of inequality constraints in* (E-3) *can be reformulated as*

$$\left. \begin{aligned} \frac{1}{2}(\beta_0 + \boldsymbol{\beta}_{\mathrm{x}}^\top \boldsymbol{x}^n + \boldsymbol{\beta}_{\mathrm{z}}^\top \boldsymbol{z} - y^n - p_{\boldsymbol{z},n})^2 + \delta p_{\boldsymbol{z},n} - \lambda \kappa_{\mathrm{z}} d_{\mathrm{z}}(\boldsymbol{z}, \boldsymbol{z}^n) \leqslant s_n \\[1ex] \frac{1}{2}(\beta_0 + \boldsymbol{\beta}_{\mathrm{x}}^\top \boldsymbol{x}^n + \boldsymbol{\beta}_{\mathrm{z}}^\top \boldsymbol{z} - y^n - p_{\boldsymbol{z},n})^2 - \delta p_{\boldsymbol{z},n} - \lambda \kappa_{\mathrm{z}} d_{\mathrm{z}}(\boldsymbol{z}, \boldsymbol{z}^n) \leqslant s_n \end{aligned} \right\} \quad \forall n \in [N], \ \forall \boldsymbol{z} \in \mathbb{Z},$$

where $p_{\boldsymbol{z},n} \in \mathbb{R}$ *are auxiliary decision variables.*

We now provide a reformulation of problem (3) without embedded maximizations when the loss function L is piece-wise affine and convex.

Proposition 8. *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* $\mathbb{X} \times \mathbb{Y} = \{(\boldsymbol{x}, y) \in \mathbb{R}^{M_{\mathrm{x}}+1} : \boldsymbol{C}_{\mathrm{x}} \boldsymbol{x} + \boldsymbol{c}_{\mathrm{y}} \cdot y \preccurlyeq_{\mathcal{C}} \boldsymbol{d}\}$ *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 (3) is equivalent to

$$\operatorname*{minimize}_{\boldsymbol{\beta}, \lambda, \boldsymbol{s}, \boldsymbol{q}_{nj}} \quad \lambda\epsilon + \frac{1}{N} \sum_{n\in[N]} s_n$$

subject to $\quad \boldsymbol{q}_{nj}^{\top}(\boldsymbol{d} - \boldsymbol{C}_{\mathrm{x}}\boldsymbol{x}^n - \boldsymbol{c}_{\mathrm{y}} \cdot y^n) + a_j \cdot (\beta_0 + \boldsymbol{\beta}_{\mathrm{x}}^{\top}\boldsymbol{x}^n + \boldsymbol{\beta}_{\mathrm{z}}^{\top}\boldsymbol{z} - y^n) - \lambda\kappa_{\mathrm{z}}d_{\mathrm{z}}(\boldsymbol{z}, \boldsymbol{z}^n) + b_j \leqslant s_n$

$$\forall n \in [N], \ \forall j \in [J], \ \forall \boldsymbol{z} \in \mathbb{Z}$$

$$\|a_j \cdot \boldsymbol{\beta}_{\mathrm{x}} - \boldsymbol{C}_{\mathrm{x}}^{\top}\boldsymbol{q}_{nj}\|_* \leqslant \lambda, \ \ |-a_j - \boldsymbol{c}_{\mathrm{y}}^{\top}\boldsymbol{q}_{nj}| \leqslant \lambda\kappa_{\mathrm{y}} \qquad \forall n \in [N], \ \forall j \in [J]$$

$$\boldsymbol{\beta} = (\beta_0, \boldsymbol{\beta}_{\mathrm{x}}, \boldsymbol{\beta}_{\mathrm{z}}) \in \mathbb{R}^{1+M_{\mathrm{x}}+M_{\mathrm{z}}}, \ \ \lambda \in \mathbb{R}_+, \ \ \boldsymbol{s} \in \mathbb{R}_+^N$$

$$\boldsymbol{q}_{nj} \in \mathcal{C}^{\star}, \ n \in [N] \ \text{and} \ j \in [J].$$

$$\text{(E-4)}$$

When $\kappa_{\mathrm{y}} = \infty$ and $\lambda > 0$, the constraints $|-a_j - \boldsymbol{c}_{\mathrm{y}}^{\top}\boldsymbol{q}_{nj}| \leqslant \lambda\kappa_{\mathrm{y}}$ in Proposition 8 are redundant.

Formulation (E-4) emerges as a special case of our unified representation (4) if we set $\mathcal{I} = [J]$, $\boldsymbol{\theta} = (\boldsymbol{\beta}, \lambda, \boldsymbol{q}_{ni})$, $\boldsymbol{\sigma} = \boldsymbol{s}$, $f_0(\boldsymbol{\theta}, \boldsymbol{\sigma}) = \lambda\epsilon + \frac{1}{N} \sum_{n\in[N]} s_n$, $f_{ni}(e) = a_i e + b_i$, $g_{ni}(\boldsymbol{\theta}, \boldsymbol{\xi}_{-\mathbf{z}}^n; \boldsymbol{z}) = \beta_0 + \boldsymbol{\beta}_{\mathrm{x}}^{\top}\boldsymbol{x}^n + \boldsymbol{\beta}_{\mathrm{z}}^{\top}\boldsymbol{z} - y^n$, $h_{ni}(\boldsymbol{\theta}, \boldsymbol{\xi}_{-\mathbf{z}}^n; d_{\mathrm{z}}(\boldsymbol{z}, \boldsymbol{z}^n)) = -\boldsymbol{q}_{ni}^{\top}(\boldsymbol{d} - \boldsymbol{C}_{\mathrm{x}}\boldsymbol{x}^n - \boldsymbol{c}_{\mathrm{y}} \cdot y^n) + \lambda\kappa_{\mathrm{z}}d_{\mathrm{z}}(\boldsymbol{z}, \boldsymbol{z}^n)$ and $\Theta = \{\boldsymbol{\theta} : \|a_i \cdot \boldsymbol{\beta}_{\mathrm{x}} - \boldsymbol{C}^{\top}\boldsymbol{q}_{ni}\|_* \leqslant \lambda \ \forall (n, i) \in [N] \times \mathcal{I}, \ |-a_i - \boldsymbol{c}_{\mathrm{y}}^{\top}\boldsymbol{q}_{ni}| \leqslant \lambda\kappa_{\mathrm{y}} \ \forall (n, i) \in [N] \times \mathcal{I}, \ \boldsymbol{q}_{ni} \in \mathcal{C}^* \ \forall (n, i) \in [N] \times \mathcal{I}, \ \lambda \in \mathbb{R}_+\}$. One verifies that this choice of $\mathcal{I}$, $\boldsymbol{\theta}$, $\boldsymbol{\sigma}$, f_0, f_{ni}, g_{ni}, h_{ni} and Θ satisfies the regularity conditions imposed on problem (4) in the main text.

Corollary 5. *If $\boldsymbol{\beta}$ in problem* (E-4) *is restricted to a bounded hypothesis set $\mathcal{H}$, then $(\boldsymbol{q}_{ni})_{n,i}$ and λ can be restricted to a bounded set as well.*

Proposition 8 covers Wasserstein support vector regression with an τ-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 $0 \leqslant \tau \leqslant 1$, 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 6. *The first set of inequality constraints in* (E-4) *can be reformulated as follows.*

(i) For the τ-insensitive loss function:

$$\left. \begin{aligned} \boldsymbol{t}_{n1}^\top(\boldsymbol{d}_1 - \boldsymbol{C}_1\boldsymbol{x}^n) + \boldsymbol{v}_{n1}^\top(\boldsymbol{d}_2 - \boldsymbol{C}_2 y^n) + (\beta_0 + \boldsymbol{\beta}_{\mathrm{x}}^\top \boldsymbol{x}^n + \boldsymbol{\beta}_{\mathrm{z}}^\top \boldsymbol{z} - y^n) \\ -\tau - \lambda\kappa_{\mathrm{z}} d_{\mathrm{z}}(\boldsymbol{z}, \boldsymbol{z}^n) \leqslant s_n \\ \boldsymbol{t}_{n2}^\top(\boldsymbol{d}_1 - \boldsymbol{C}_1\boldsymbol{x}^n) + \boldsymbol{v}_{n2}^\top(\boldsymbol{d}_2 - \boldsymbol{C}_2 y^n) - (\beta_0 + \boldsymbol{\beta}_{\mathrm{x}}^\top \boldsymbol{x}^n + \boldsymbol{\beta}_{\mathrm{z}}^\top \boldsymbol{z} - y^n) \\ -\tau - \lambda\kappa_{\mathrm{z}} d_{\mathrm{z}}(\boldsymbol{z}, \boldsymbol{z}^n) \leqslant s_n \end{aligned} \right\} \quad \forall n \in [N],\ \forall \boldsymbol{z} \in \mathbb{Z}$$

(ii) For the pinball loss function:

$$\left. \begin{aligned} \boldsymbol{t}_{n1}^\top(\boldsymbol{d}_1 - \boldsymbol{C}_1\boldsymbol{x}^n) + \boldsymbol{v}_{n1}^\top(\boldsymbol{d}_2 - \boldsymbol{C}_2 y^n) - \tau(\beta_0 + \boldsymbol{\beta}_{\mathrm{x}}^\top \boldsymbol{x}^n + \boldsymbol{\beta}_{\mathrm{z}}^\top \boldsymbol{z} - y^n) \\ -\lambda\kappa_{\mathrm{z}} d_{\mathrm{z}}(\boldsymbol{z}, \boldsymbol{z}^n) \leqslant s_n \\ \boldsymbol{t}_{n2}^\top(\boldsymbol{d}_1 - \boldsymbol{C}_1\boldsymbol{x}^n) + \boldsymbol{v}_{n2}^\top(\boldsymbol{d}_2 - \boldsymbol{C}_2 y^n) + (1 - \tau)(\beta_0 + \boldsymbol{\beta}_{\mathrm{x}}^\top \boldsymbol{x}^n + \boldsymbol{\beta}_{\mathrm{z}}^\top \boldsymbol{z} - y^n) \\ -\lambda\kappa_{\mathrm{z}} d_{\mathrm{z}}(\boldsymbol{z}, \boldsymbol{z}^n) \leqslant s_n \end{aligned} \right\} \quad \forall n \in [N],\ \forall \boldsymbol{z} \in \mathbb{Z}$$

Appendix C: Proofs

Proof of Observation 1. 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 1. We first show that LB and UB constitute lower and upper bounds on the optimal value of (4) throughout the execution of the algorithm, and then we conclude that Algorithm 1 terminates in finite time with an optimal solution $(\boldsymbol{\theta}^\star, \boldsymbol{\sigma}^\star)$ to problem (4).

Algorithm 1 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 (4) 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 (4). 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 (4) throughout the execution of Algorithm 1, we claim that in each iteration, $(\boldsymbol{\theta}^\star, \boldsymbol{\sigma}^\star + \boldsymbol{\vartheta}^\star)$ constitutes a feasible solution to problem (4). 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

$$f_{ni}(g_{ni}(\boldsymbol{\theta}^\star, \boldsymbol{\xi}_{-\mathbf{z}}^n; \boldsymbol{z})) - h_{ni}(\boldsymbol{\theta}^\star, \boldsymbol{\xi}_{-\mathbf{z}}^n; d_{\mathbf{z}}(\boldsymbol{z}, \boldsymbol{z}^n)) \;\leqslant\; \sigma_n^\star + \vartheta(n, i) \;\leqslant\; \sigma_n^\star + \vartheta(n, i(n)) \;\leqslant\; \sigma_n^\star + \vartheta_n^\star,$$

that is, $(\boldsymbol{\theta}^\star, \boldsymbol{\sigma}^\star + \boldsymbol{\vartheta}^\star)$ is indeed feasible in problem (4). Moreover, the sequence of upper bounds UB is monotonic by construction of the updates.

To see that Algorithm 1 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 $\vartheta(n, i(n)) \leqslant 0$ for all $n \in [N]$. In the latter case, however, we have $\boldsymbol{\vartheta}^\star = \boldsymbol{0}$ 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 2. 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

$$\underset{z}{\text{maximize}} \quad f_{ni}(g_{ni}(\boldsymbol{\theta}^\star, \boldsymbol{\xi}^n_{-\mathbf{z}}; \boldsymbol{z})) - h_{ni}(\boldsymbol{\theta}^\star, \boldsymbol{\xi}^n_{-\mathbf{z}}; d_z(\boldsymbol{z}, \boldsymbol{z}^n)) - \sigma_n^\star$$

$$\text{subject to} \quad \boldsymbol{z} \in \mathbb{Z}.$$

We can solve this problem by solving the $K + 1$ problems

$$\left[\begin{array}{ll} \underset{z}{\text{maximize}} & f_{ni}(g_{ni}(\boldsymbol{\theta}^\star, \boldsymbol{\xi}^n_{-\mathbf{z}}; \boldsymbol{z})) - h_{ni}(\boldsymbol{\theta}^\star, \boldsymbol{\xi}^n_{-\mathbf{z}}; \delta) - \sigma_n^\star \\ \text{subject to} & d_z(\boldsymbol{z}, \boldsymbol{z}^n) = \delta \\ & \boldsymbol{z} \in \mathbb{Z} \end{array}\right] \quad \forall \delta \in [K] \cup \{0\},$$

where each problem conditions on a fixed number δ 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 δ-th problem shares its set of optimal solutions with the problem

$$\underset{z}{\text{maximize}} \quad f_{ni}(g_{ni}(\boldsymbol{\theta}^\star, \boldsymbol{\xi}^n_{-\mathbf{z}}; \boldsymbol{z}))$$

$$\text{subject to} \quad d_z(\boldsymbol{z}, \boldsymbol{z}^n) = \delta$$

$$\boldsymbol{z} \in \mathbb{Z}.$$

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 δ-th problem is solved by solving the two problems

$$\left[\begin{array}{ll} \underset{z}{\text{maximize}} & \mu \cdot g_{ni}(\boldsymbol{\theta}^\star, \boldsymbol{\xi}^n_{-\mathbf{z}}; \boldsymbol{z}) \\ \text{subject to} & d_z(\boldsymbol{z}, \boldsymbol{z}^n) = \delta \\ & \boldsymbol{z} \in \mathbb{Z} \end{array}\right] \quad \forall \mu \in \{\pm 1\}$$

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 μ to either value, adopting the

notation for $\boldsymbol{w}$ and w_0 of Algorithm 2 and ignoring constant terms, we can write the problem as

$$\underset{\boldsymbol{z}}{\text{maximize}} \quad \mu \cdot \left[\sum_{m \in [K]} \boldsymbol{w}_m^\top \boldsymbol{z}_m \right]$$

$$\text{subject to} \quad d_{\text{z}}(\boldsymbol{z}, \boldsymbol{z}^n) = \delta$$

$$\boldsymbol{z} \in \mathbb{Z}.$$

The rectangularity of $\mathbb{Z}$ implies that the decisions of this problem admit a decomposition into the selection $\mathcal{M} \subseteq [K]$ of δ 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 $\boldsymbol{z}_m \in \mathbb{Z}(k_m) \backslash \{\boldsymbol{z}_m^n\}$:

$$\underset{\mathcal{M}}{\text{maximize}} \quad \underset{\boldsymbol{z}}{\max} \left\{ \mu \cdot \left[\sum_{m \in [K]} \boldsymbol{w}_m^\top \boldsymbol{z}_m \right] : \begin{bmatrix} \boldsymbol{z}_m \in \mathbb{Z}(k_m) \backslash \{\boldsymbol{z}_m^n\} & \forall m \in \mathcal{M} \\ \boldsymbol{z}_m = \boldsymbol{z}_m^n & \forall m \in [K] \backslash \mathcal{M} \end{bmatrix} \right\}$$

$$\text{subject to} \quad \mathcal{M} \subseteq [K], \ |\mathcal{M}| = \delta.$$

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 2 to obtain the equivalent formulation

$$\underset{\mathcal{M}}{\text{maximize}} \quad \left[\sum_{m \in \mathcal{M}} \mu \cdot \boldsymbol{w}_m^\top \boldsymbol{z}_m^\star \right] + \left[\sum_{m \in [K] \backslash \mathcal{M}} \mu \cdot \boldsymbol{w}_m^\top \boldsymbol{z}_m^n \right]$$

$$\text{subject to} \quad \mathcal{M} \subseteq [K], \ |\mathcal{M}| = \delta.$$

The two summations in the objective function of this problem admit the reformulation

$$\left[\sum_{m \in \mathcal{M}} \mu \cdot \boldsymbol{w}_m^\top \boldsymbol{z}_m^\star \right] + \left[\sum_{m \in [K] \backslash \mathcal{M}} \mu \cdot \boldsymbol{w}_m^\top \boldsymbol{z}_m^n \right] = \left[\sum_{m \in [K]} \mu \cdot \boldsymbol{w}_m^\top \boldsymbol{z}_m^n \right] + \left[\sum_{m \in \mathcal{M}} \mu \cdot \boldsymbol{w}_m^\top (\boldsymbol{z}_m^\star - \boldsymbol{z}_m^n) \right],$$

and ignoring constant terms once more simplifies our optimization problem to

$$\underset{\mathcal{M}}{\text{maximize}} \quad \sum_{m \in \mathcal{M}} \mu \cdot \boldsymbol{w}_m^\top (\boldsymbol{z}_m^\star - \boldsymbol{z}_m^n)$$

$$\text{subject to} \quad \mathcal{M} \subseteq [K], \ |\mathcal{M}| = \delta.$$

This problem is solved by identifying $\mathcal{M}$ with the indices of the δ 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 2 adopts the solution approach just described to determine a maximally violated constraint index $\boldsymbol{z}(n, i)$.

The runtime of Algorithm 2, finally, is dominated by determining the $2K$ maximizers $\boldsymbol{z}_m^\star$, $m \in [K]$ and $\mu \in \{\pm 1\}$, which takes time $\mathcal{O}(M_{\boldsymbol{z}})$ due to the one-hot encoding employed by $\mathbb{Z}$, sorting the $2K$ values $\mu \cdot \boldsymbol{w}_m^\top(\boldsymbol{z}_m^\star - \boldsymbol{z}_m^n)$, 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 $\mathcal{O}(KT)$. $\quad\square$

Proof of Theorem 3. Recall from Grötschel et al. (1988, Corollary 4.2.7) that problem (4) can be solved to δ-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 $\mathcal{S}(\mathcal{K}, \delta) = \{\boldsymbol{r} \in \mathbb{R}^n : \|\boldsymbol{r} - \boldsymbol{r}'\|_2 \leqslant \delta \text{ for some } \boldsymbol{r}' \in \mathcal{K}\}$ is the δ-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 $\boldsymbol{c}^\top \boldsymbol{p} \leqslant \boldsymbol{c}^\top \boldsymbol{q} + \delta$ 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 δ-interior of $\mathcal{K}$ (*i.e.*, $\boldsymbol{c}$ is an *almost* separating hyperplane). We construct a weak separation oracle for problem (4) as follows. Denote the feasible region of problem (4) by $\mathcal{K} \subseteq \Theta \times \mathbb{R}_+^N$, and fix any $\boldsymbol{q}' = (\boldsymbol{\theta}', \boldsymbol{\sigma}') \in \mathbb{R}^{M_\theta} \times \mathbb{R}^N$. If $\sigma_n' < 0$ for any $n \in [N]$, then our algorithm returns the separating hyperplane $\boldsymbol{c}^\top = (\boldsymbol{0}^\top \; -\mathbf{e}_n^\top)$. If $\boldsymbol{\theta}' \notin \Theta$, then we can augment the weakly separating hyperplane $\boldsymbol{c} \in \mathbb{R}^{M_\theta}$ for the set Θ to a weakly separating hyperplane $(\boldsymbol{c}^\top \; \boldsymbol{0}^\top)$ for problem (4). Otherwise, we leverage Algorithm 2 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 2 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 $\vartheta(n,i) \leqslant 0$ for all $(n,i) \in [N] \times \mathcal{I}$, then $\boldsymbol{q}'$ is feasible in problem (4) 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

$$r(\boldsymbol{\theta}, \boldsymbol{\sigma}) \;\; = \;\; f_{n^{\star} i^{\star}}(g_{n^{\star} i^{\star}}(\boldsymbol{\theta}, \boldsymbol{\xi}_{-\mathbf{z}}^{n^{\star}}; \boldsymbol{z}(n^{\star}, i^{\star}))) - h_{n^{\star} i^{\star}}(\boldsymbol{\theta}, \boldsymbol{\xi}_{-\mathbf{z}}^{n^{\star}}; d_{\mathrm{z}}(\boldsymbol{z}(n^{\star}, i^{\star}), \boldsymbol{z}^{n^{\star}})) - \sigma_{n^{\star}}$$

at $(\boldsymbol{\theta}, \boldsymbol{\sigma}) = (\boldsymbol{\theta}', \boldsymbol{\sigma}')$. We have

$$\boldsymbol{c}^{\top}(\boldsymbol{\theta}, \boldsymbol{\sigma}) \;\; \leqslant \;\; \boldsymbol{c}^{\top}(\boldsymbol{\theta}', \boldsymbol{\sigma}') + r(\boldsymbol{\theta}, \boldsymbol{\sigma}) - r(\boldsymbol{\theta}', \boldsymbol{\sigma}') \;\; < \;\; \boldsymbol{c}^{\top}(\boldsymbol{\theta}', \boldsymbol{\sigma}') \qquad \forall (\boldsymbol{\theta}, \boldsymbol{\sigma}) \in \mathcal{K},$$

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 $r(\boldsymbol{\theta}, \boldsymbol{\sigma}) \leqslant 0$ 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 2, 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 (4) 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 Θ 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 Θ is bounded by assumption and $\boldsymbol{\sigma}$ is non-negative by construction. Since the objective function f_0 in (4) is non-decreasing in $\boldsymbol{\sigma}$, we can without loss of generality include in (4) the additional constraints $\sigma_n \leqslant \overline{\sigma}_n$ for

$$\overline{\sigma}_n \;\; = \;\; \max_{\boldsymbol{\theta} \in \Theta} \; \max_{i \in \mathcal{I}} \; \max_{\mathbf{z} \in \mathbb{Z}} \left\{ f_{ni}(g_{ni}(\boldsymbol{\theta}, \boldsymbol{\xi}_{-\mathbf{z}}^{n}; \boldsymbol{z})) - h_{ni}(\boldsymbol{\theta}, \boldsymbol{\xi}_{-\mathbf{z}}^{n}; d_{\mathrm{z}}(\boldsymbol{z}, \boldsymbol{z}^{n})) \right\}, \qquad n \in [N].$$

Note that all $\bar{\sigma}_n$ are finite since Θ 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. $\qquad\square$

Proof of Theorem 4. Fix any Wasserstein classification or regression instance as described in the statement of the theorem, fix any ambiguity radius $\epsilon > \kappa_z K^{1/p}$, any non-zero numbers M_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 (1) 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 $\hat{\boldsymbol{\beta}} = (\hat{\beta}_0, \hat{\boldsymbol{\beta}}_x, \hat{\boldsymbol{\beta}}_z)$ with $\hat{\beta}_0 \in \mathbb{R}$ and $\hat{\boldsymbol{\beta}}_x \neq \boldsymbol{0}$ selected arbitrarily. Fix any discrete feature realization $\boldsymbol{z}^\star \in \mathbb{Z} \backslash \{\boldsymbol{z}^1\}$ as well as the discrete feature coefficients

$$\hat{\beta}_{z,mi} = \begin{cases} -\dfrac{2}{d'} \cdot \kappa_z \mathrm{lip}(L) \cdot \|\hat{\boldsymbol{\beta}}_x\|_* \cdot d_z(\boldsymbol{z}^\star, \boldsymbol{z}^1) & \text{if } z_{mi}^1 = 1, \\[3mm] \dfrac{2}{d'} \cdot \kappa_z \mathrm{lip}(L) \cdot \|\hat{\boldsymbol{\beta}}_x\|_* \cdot d_z(\boldsymbol{z}^\star, \boldsymbol{z}^1) & \text{otherwise} \end{cases} \quad \text{for all } m \in [K] \text{ and } i \in [k_m],$$

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 $d' \cdot \hat{\beta}_{z,mi} \cdot (z_{mi} - z_{mi}^1) \geqslant 0$ 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 $z_{mi}^1 = 0$, then $d' \cdot \hat{\beta}_{z,mi} = 2 \cdot \kappa_z \mathrm{lip}(L) \cdot \|\hat{\boldsymbol{\beta}}_x\|_* \cdot d_z(\boldsymbol{z}^\star, \boldsymbol{z}^1) \geqslant 0$

EC-14

and $z_{mi} - z^1_{mi} \geqslant 0$. If $z^1_{mi} = 1$, on the other hand, then $d' \cdot \hat{\beta}_{z,mi} = -2 \cdot \kappa_z \text{lip}(L) \cdot \|\hat{\beta}_x\|_* \cdot d_z(z^\star, z^1) \leqslant 0$ and $z_{mi} - z^1_{mi} \leqslant 0$. In particular, we have $d' \cdot \hat{\beta}_z^\top(z - z^1) \geqslant 0$ for all $z \in \mathbb{Z}$.

(ii) We have $d' \cdot \hat{\beta}_z^\top(z^\star - z^1) > \kappa_z \text{lip}(L) \cdot \|\hat{\beta}_x\|_* \cdot d_z(z^\star, z^1)$. Indeed, since $z^\star \neq z^1$, 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 $d' \cdot \hat{\beta}_{z,m^\star i^\star} \cdot (z^\star_{m^\star i^\star} - z^1_{m^\star i^\star}) = 2 \cdot \kappa_z \text{lip}(L) \cdot \|\hat{\beta}_x\|_* \cdot d_z(z^\star, z^1) > \kappa_z \text{lip}(L) \cdot \|\hat{\beta}_x\|_* \cdot d_z(z^\star, z^1)$. The claim now follows from the fact that $d' \cdot \hat{\beta}_{z,mi} \cdot (z_{mi} - z^1_{mi}) \geqslant 0$ 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 3 from Appendix A therefore implies that the objective function of the Wasserstein classification problem becomes

$$
\begin{aligned}
\underset{\lambda, s}{\text{minimize}} \quad & \lambda \epsilon + \frac{1}{N} \sum_{n \in [N]} s_n \\
\text{subject to} \quad & \left.\begin{array}{l} l_{\hat{\beta}}(x^n, z, y^n) - \lambda \kappa_z d_z(z, z^n) \leqslant s_n \\ l_{\hat{\beta}}(x^n, z, -y^n) - \lambda \kappa_z d_z(z, z^n) - \lambda \kappa_y \leqslant s_n \end{array}\right\} \forall n \in [N], \ \forall z \in \mathbb{Z} \\
& \text{lip}(L) \cdot \|\hat{\beta}_x\|_* \leqslant \lambda \\
& \lambda \in \mathbb{R}_+, \ \ s \in \mathbb{R}_+^N,
\end{aligned}
$$

and Proposition 7 from Appendix B implies that the objective function of the Wasserstein regression problem becomes

$$
\begin{aligned}
\underset{\lambda, s}{\text{minimize}} \quad & \lambda \epsilon + \frac{1}{N} \sum_{n \in [N]} s_n \\
\text{subject to} \quad & l_{\hat{\beta}}(x^n, z, y^n) - \lambda \kappa_z d_z(z, z^n) \leqslant s_n \qquad \forall n \in [N], \ \forall z \in \mathbb{Z} \\
& \text{lip}(L) \cdot \|\hat{\beta}_x\|_* \leqslant \lambda \\
& \text{lip}(L) \leqslant \lambda \kappa_y \\
& \lambda \in \mathbb{R}_+, \ \ s \in \mathbb{R}_+^N.
\end{aligned}
$$

When $\kappa_y \to \infty$, the second set of constraints in the classification problem and the third constraint

in the regression problem become redundant since $\lambda \geqslant \mathrm{lip}(L) \cdot \|\hat{\boldsymbol{\beta}}_{\mathrm{x}}\|_* > 0$ because $\hat{\boldsymbol{\beta}}_{\mathrm{x}} \neq \boldsymbol{0}$. 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

$$
\begin{aligned}
\underset{\lambda}{\mathrm{minimize}} \quad & \lambda\epsilon + \frac{1}{N} \sum_{n\in[N]} \max_{\boldsymbol{z}\in\mathbb{Z}} \left\{ l_{\hat{\boldsymbol{\beta}}}(\boldsymbol{x}^n, \boldsymbol{z}, y^n) - \lambda\kappa_{\mathrm{z}} d_{\mathrm{z}}(\boldsymbol{z}, \boldsymbol{z}^n) \right\} \\
\mathrm{subject\ to} \quad & \mathrm{lip}(L) \cdot \|\hat{\boldsymbol{\beta}}_{\mathrm{x}}\|_* \leqslant \lambda \\
& \lambda \in \mathbb{R}_+
\end{aligned}
$$

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 $l_{\hat{\boldsymbol{\beta}}}(\boldsymbol{x}^n, \boldsymbol{z}^n, y^n) \geqslant 0$ 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 λ 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

$$
\begin{aligned}
f_1(\hat{\boldsymbol{\beta}}, \{\boldsymbol{\xi}^n\}_{n\in[N]}) &= \lambda^\star\epsilon + \frac{1}{N} \sum_{n\in[N]} \max_{\boldsymbol{z}\in\mathbb{Z}} \left\{ l_{\hat{\boldsymbol{\beta}}}(\boldsymbol{x}^n, \boldsymbol{z}, y^n) - \lambda^\star\kappa_{\mathrm{z}} d_{\mathrm{z}}(\boldsymbol{z}, \boldsymbol{z}^n) \right\} \\
&= \frac{1}{N} \sum_{n\in[N]} l_{\hat{\boldsymbol{\beta}}}(\boldsymbol{\xi}^n) + h_{\hat{\boldsymbol{\beta}}}(\{\boldsymbol{\xi}^n\}_{n\in[N]}),
\end{aligned}
$$

where

$$
h_{\hat{\boldsymbol{\beta}}}(\{\boldsymbol{\xi}^n\}_{n\in[N]}) = \lambda^\star\epsilon + \frac{1}{N} \sum_{n\in[N]} \max_{\boldsymbol{z}\in\mathbb{Z}} \left\{ l_{\hat{\boldsymbol{\beta}}}(\boldsymbol{x}^n, \boldsymbol{z}, y^n) - l_{\hat{\boldsymbol{\beta}}}(\boldsymbol{x}^n, \boldsymbol{z}^n, y^n) - \lambda^\star\kappa_{\mathrm{z}} d_{\mathrm{z}}(\boldsymbol{z}, \boldsymbol{z}^n) \right\}.
$$

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 $n \in [N]\backslash 1$ 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 $\hat{y}^1 \cdot [\hat{\beta}_0 + \hat{\boldsymbol{\beta}}_{\mathrm{x}}^\top \hat{\boldsymbol{x}}^1 + \hat{\boldsymbol{\beta}}_{\mathrm{z}}^\top \hat{\boldsymbol{z}}^1] = x'$ for classification problems and $\hat{\beta}_0 + \hat{\boldsymbol{\beta}}_{\mathrm{x}}^\top \hat{\boldsymbol{x}}^1 + \hat{\boldsymbol{\beta}}_{\mathrm{z}}^\top \hat{\boldsymbol{z}}^1 - \hat{y}^1 = x'$ 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 $\hat{\boldsymbol{\beta}}_{\mathrm{x}} \neq \boldsymbol{0}$. We then have

$$
\begin{aligned}
\max_{\boldsymbol{z}\in\mathbb{Z}} l_{\hat{\boldsymbol{\beta}}}(\hat{\boldsymbol{x}}^1, \boldsymbol{z}, \hat{y}^1) - l_{\hat{\boldsymbol{\beta}}}&(\hat{\boldsymbol{x}}^1, \hat{\boldsymbol{z}}^1, \hat{y}^1) - \lambda^\star \kappa_{\mathrm{z}} d_{\mathrm{z}}(\boldsymbol{z}, \hat{\boldsymbol{z}}^1) \\
&\geqslant\; l_{\hat{\boldsymbol{\beta}}}(\hat{\boldsymbol{x}}^1, \boldsymbol{z}^\star, \hat{y}^1) - l_{\hat{\boldsymbol{\beta}}}(\hat{\boldsymbol{x}}^1, \hat{\boldsymbol{z}}^1, \hat{y}^1) - \lambda^\star \kappa_{\mathrm{z}} d_{\mathrm{z}}(\boldsymbol{z}^\star, \hat{\boldsymbol{z}}^1) \\
&=\; L(x' + \hat{\boldsymbol{\beta}}_{\mathrm{z}}^\top(\boldsymbol{z}^\star - \hat{\boldsymbol{z}}^1)) - L(x') - \lambda^\star \kappa_{\mathrm{z}} d_{\mathrm{z}}(\boldsymbol{z}^\star, \hat{\boldsymbol{z}}^1) \\
&\geqslant\; \hat{\boldsymbol{\beta}}_{\mathrm{z}}^\top(\boldsymbol{z}^\star - \hat{\boldsymbol{z}}^1) \cdot \frac{\mathrm{d}}{\mathrm{d}e} L(e) \big|_{e=x'} - \lambda^\star \kappa_{\mathrm{z}} d_{\mathrm{z}}(\boldsymbol{z}^\star, \hat{\boldsymbol{z}}^1) \\
&=\; d' \cdot \hat{\boldsymbol{\beta}}_{\mathrm{z}}^\top(\boldsymbol{z}^\star - \hat{\boldsymbol{z}}^1) - \kappa_{\mathrm{z}} \mathrm{lip}(L) \cdot \|\hat{\boldsymbol{\beta}}_{\mathrm{x}}\|_* \cdot d_{\mathrm{z}}(\boldsymbol{z}^\star, \hat{\boldsymbol{z}}^1) > 0,
\end{aligned}
$$

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 $\boldsymbol{z} \in \mathbb{Z}\backslash\{\check{\boldsymbol{z}}^1\}$. 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

$$\lim_{x \to -\infty} \frac{\mathrm{d}}{\mathrm{d}e} L(e) \big|_{e=x} \leqslant 0, \tag{E-5}$$

which in turn implies that for all $\boldsymbol{z} \in \mathbb{Z} \setminus \{\check{\boldsymbol{z}}^1\}$, we have

$$\lim_{x \to -\infty} L(x + \hat{\boldsymbol{\beta}}_{\mathrm{z}}^\top (\boldsymbol{z} - \check{\boldsymbol{z}}^1)) - L(x) \ \leqslant \ \lim_{x \to -\infty} \hat{\boldsymbol{\beta}}_{\mathrm{z}}^\top (\boldsymbol{z} - \check{\boldsymbol{z}}^1) \cdot \frac{\mathrm{d}}{\mathrm{d}e} L(e) \big|_{e=x+\hat{\boldsymbol{\beta}}_{\mathrm{z}}^\top (\boldsymbol{z}-\check{\boldsymbol{z}}^1)} \ \leqslant \ 0, \tag{E-6}$$

where the first inequality exploits the fact that the derivative of a convex function is non-decreasing and the second inequality holds because $d' \cdot \hat{\boldsymbol{\beta}}_{\mathrm{z}}^\top (\boldsymbol{z} - \check{\boldsymbol{z}}^1) \geqslant 0$ due to our earlier observation *(i)* and the fact that $\lim_{x \to -\infty} \frac{\mathrm{d}}{\mathrm{d}e} L(e) \big|_{e=x+\hat{\boldsymbol{\beta}}_{\mathrm{z}}^\top (\boldsymbol{z}-\check{\boldsymbol{z}}^1)} = \lim_{x \to -\infty} \frac{\mathrm{d}}{\mathrm{d}e} L(e) \big|_{e=x} \leqslant 0$ by equation (E-5). Moreover, for $n = 1$ and for all $\boldsymbol{z} \in \mathbb{Z} \setminus \{\check{\boldsymbol{z}}^1\}$, we have

$$\lim_{\hat{\boldsymbol{\beta}}_{\mathrm{x}}^\top \check{\boldsymbol{x}}^1 \to -\infty} l_{\hat{\boldsymbol{\beta}}}(\check{\boldsymbol{x}}^1, \boldsymbol{z}, \check{y}^1) - l_{\hat{\boldsymbol{\beta}}}(\check{\boldsymbol{x}}^1, \check{\boldsymbol{z}}^1, \check{y}^1) - \lambda^\star \kappa_{\mathrm{z}} d_{\mathrm{z}}(\boldsymbol{z}, \check{\boldsymbol{z}}^1)$$

$$= \lim_{x'' \to -\infty} L(x'' + \hat{\boldsymbol{\beta}}_{\mathrm{z}}^\top (\boldsymbol{z} - \check{\boldsymbol{z}}^1)) - L(x'') - \lambda^\star \kappa_{\mathrm{z}} d_{\mathrm{z}}(\boldsymbol{z}, \check{\boldsymbol{z}}^1)$$

$$\leqslant -\lambda^\star \kappa_{\mathrm{z}} d_{\mathrm{z}}(\boldsymbol{z}, \check{\boldsymbol{z}}^1) \ < \ 0,$$

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 (E-6), and the last inequality is due to the fact that $\lambda^\star$, κ_{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 $\mathfrak{R}(\hat{\boldsymbol{\beta}})$. $\quad \square$

Proof of Proposition 2. We prove the statement of the proposition in three steps. We first show that our revised ground metric d_{u} coincides with the original ground metric d_{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 (3') reduce to their counterparts in (3). In other words, problem (3') is a restriction of (3) since it contains all of the constraints of (3). 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 (3) is also feasible in the bounded continuous-feature formulation (3').

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

$$
\frac{1}{2}\|\boldsymbol{z}_m - \boldsymbol{z}_m'\|_1 + \frac{1}{2}|\mathbf{1}^\top(\boldsymbol{z}_m - \boldsymbol{z}_m')| \;=\; \frac{1}{2}\left\|\begin{pmatrix} \boldsymbol{z}_m \\ 1-\mathbf{1}^\top\boldsymbol{z}_m \end{pmatrix} - \begin{pmatrix} \boldsymbol{z}_m' \\ 1-\mathbf{1}^\top\boldsymbol{z}_m' \end{pmatrix}\right\|_1 \;=\; \begin{cases} 0 & \text{if } \boldsymbol{z}_m = \boldsymbol{z}_m', \\ 1 & \text{otherwise.} \end{cases}
$$

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

$$
\begin{aligned}
d_{\mathrm{u}}(\boldsymbol{u}, \boldsymbol{z}') \;=\; \Bigg(& \frac{1}{2} \sum_{m \in [K]} \sum_{j \in [k_m-1]} \left[\mathbb{1}[z_{mj}' = 0] \cdot u_{mj} + \mathbb{1}[z_{mj}' = 1] \cdot (1 - u_{mj}) \right] \;+ \\
& \frac{1}{2} \sum_{m \in [K]} \left[\mathbb{1}[\boldsymbol{z}_m' = \mathbf{0}] \cdot \mathbf{1}^\top\boldsymbol{u}_m + \mathbb{1}[\boldsymbol{z}_m' \neq \mathbf{0}] \cdot (1 - \mathbf{1}^\top\boldsymbol{u}_m) \right] \Bigg)^{1/p},
\end{aligned}
$$

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_{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 $g : \mathbb{U} \to \mathbb{R}$ 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 (3). Since

the objective functions of (3) and (3') coincide, we only need to show that $(\boldsymbol{\beta}, \lambda, \boldsymbol{s})$ satisfies the constraints of problem (3'). To this end, we observe that for all $n \in [N]$, we have

$$\sup_{(\boldsymbol{x},y)\in\mathbb{X}\times\mathbb{Y}} \{l_{\boldsymbol{\beta}}(\boldsymbol{x}, \boldsymbol{u}, y) - \lambda\|\boldsymbol{x} - \boldsymbol{x}^n\| - \lambda\kappa_{\mathrm{y}}d_{\mathrm{y}}(y, y^n)\} - \lambda\kappa_{\mathrm{z}}d_{\mathrm{u}}(\boldsymbol{u}, \boldsymbol{z}^n) \leqslant s_n \qquad \forall \boldsymbol{u} \in \mathbb{U}$$

$$\iff \sup_{\boldsymbol{u}\in\mathbb{U}} \{l_{\boldsymbol{\beta}}(\boldsymbol{x}, \boldsymbol{u}, y) - \lambda\|\boldsymbol{x} - \boldsymbol{x}^n\| - \lambda\kappa_{\mathrm{y}}d_{\mathrm{y}}(y, y^n)\} - \lambda\kappa_{\mathrm{z}}d_{\mathrm{u}}(\boldsymbol{u}, \boldsymbol{z}^n) \leqslant s_n \qquad \forall(\boldsymbol{x}, y) \in \mathbb{X} \times \mathbb{Y}$$

$$\iff \sup_{\boldsymbol{u}\in\mathbb{U}} \{l_{\boldsymbol{\beta}}(\boldsymbol{x}, \boldsymbol{u}, y) - \lambda\kappa_{\mathrm{z}}d_{\mathrm{u}}(\boldsymbol{u}, \boldsymbol{z}^n)\} - \lambda\|\boldsymbol{x} - \boldsymbol{x}^n\| - \lambda\kappa_{\mathrm{y}}d_{\mathrm{y}}(y, y^n)\} \leqslant s_n \qquad \forall(\boldsymbol{x}, y) \in \mathbb{X} \times \mathbb{Y}$$

$$\iff \sup_{\boldsymbol{z}\in\mathbb{Z}} \{l_{\boldsymbol{\beta}}(\boldsymbol{x}, \boldsymbol{z}, y) - \lambda\kappa_{\mathrm{z}}d_{\mathrm{u}}(\boldsymbol{z}, \boldsymbol{z}^n)\} - \lambda\|\boldsymbol{x} - \boldsymbol{x}^n\| - \lambda\kappa_{\mathrm{y}}d_{\mathrm{y}}(y, y^n)\} \leqslant s_n \qquad \forall(\boldsymbol{x}, y) \in \mathbb{X} \times \mathbb{Y}$$

$$\iff \sup_{\boldsymbol{z}\in\mathbb{Z}} \{l_{\boldsymbol{\beta}}(\boldsymbol{x}, \boldsymbol{z}, y) - \lambda\kappa_{\mathrm{z}}d_{\mathrm{z}}(\boldsymbol{z}, \boldsymbol{z}^n)\} - \lambda\|\boldsymbol{x} - \boldsymbol{x}^n\| - \lambda\kappa_{\mathrm{y}}d_{\mathrm{y}}(y, y^n)\} \leqslant s_n \qquad \forall(\boldsymbol{x}, y) \in \mathbb{X} \times \mathbb{Y}$$

$$\iff \sup_{(\boldsymbol{x},y)\in\mathbb{X}\times\mathbb{Y}} \{l_{\boldsymbol{\beta}}(\boldsymbol{x}, \boldsymbol{z}, y) - \lambda\kappa_{\mathrm{z}}d_{\mathrm{z}}(\boldsymbol{z}, \boldsymbol{z}^n)\} - \lambda\|\boldsymbol{x} - \boldsymbol{x}^n\| - \lambda\kappa_{\mathrm{y}}d_{\mathrm{y}}(y, y^n)\} \leqslant s_n \qquad \forall \boldsymbol{z} \in \mathbb{Z}.$$

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_{u} and d_{z} agree over all pairs of discrete feature vectors $\boldsymbol{z}, \boldsymbol{z}' \in \mathbb{Z}$. $\qquad\square$

The proof of Proposition 3 utilizes the following lemma, which we state and prove first.

Lemma 1. *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*

$$\sup_{\boldsymbol{x}\in\mathbb{R}^{M_{\mathrm{x}}}} L(\boldsymbol{\alpha}^\top \boldsymbol{x} + \alpha_0) - \lambda\|\boldsymbol{x} - \boldsymbol{x}^n\| = \begin{cases} L(\boldsymbol{\alpha}^\top \boldsymbol{x}^n + \alpha_0) & \textit{if } \mathrm{lip}(L) \cdot \|\boldsymbol{\alpha}\|_* \leqslant \lambda, \\ +\infty & \textit{otherwise.} \end{cases} \qquad \text{(E-7)}$$

Lemma 1 generalizes Lemma 47 of Shafieezadeh-Abadeh et al. (2019) in that it includes a constant α_0 in the argument of the loss function L and that it extends to the case where $\lambda = 0$.

Proof of Lemma 1. 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 (E-7) can then be written as

$$
\sup_{\boldsymbol{x} \in \mathbb{R}^{M_\mathrm{x}}} L(\boldsymbol{\alpha}^\top \boldsymbol{x} + \alpha_0) - \lambda \|\boldsymbol{x} - \boldsymbol{x}^n\| = \sup_{\boldsymbol{w} \in \mathbb{R}^{M_\mathrm{x}}} L(\boldsymbol{\alpha}^\top \boldsymbol{w}) - \lambda \|\boldsymbol{w} - \boldsymbol{d} - (\boldsymbol{w}^n - \boldsymbol{d})\|
$$

$$
= \sup_{\boldsymbol{w} \in \mathbb{R}^{M_\mathrm{x}}} L(\boldsymbol{\alpha}^\top \boldsymbol{w}) - \lambda \|\boldsymbol{w} - \boldsymbol{w}^n\|
$$

If $\lambda > 0$, then Lemma 47 of Shafieezadeh-Abadeh et al. (2019) can be directly applied:

$$
\sup_{\boldsymbol{w} \in \mathbb{R}^{M_\mathrm{x}}} L(\boldsymbol{\alpha}^\top \boldsymbol{w}) - \lambda \|\boldsymbol{w} - \boldsymbol{w}^n\| = \begin{cases} L(\boldsymbol{\alpha}^\top \boldsymbol{w}^n) & \text{if } \mathrm{lip}(L) \cdot \|\boldsymbol{\alpha}\|_* \leqslant \lambda, \\ +\infty & \text{otherwise,} \end{cases}
$$

$$
= \begin{cases} L(\boldsymbol{\alpha}^\top \boldsymbol{x}^n + \alpha_0) & \text{if } \mathrm{lip}(L) \cdot \|\boldsymbol{\alpha}\|_* \leqslant \lambda, \\ +\infty & \text{otherwise.} \end{cases}
$$

Note that Lemma 47 of Shafieezadeh-Abadeh et al. (2019) assumes $\lambda > 0$. For the case where $\lambda = 0$, the left-hand side of (E-7) evaluates to $+\infty$ since L is assumed to be non-constant (*cf.* Section 2) and convex. The right-hand side of (E-7) also evaluates to $+\infty$ since $\mathrm{lip}(L) \cdot \|\boldsymbol{\alpha}\|_* > \lambda$ due to L being non-constant and $\boldsymbol{\alpha} \neq \boldsymbol{0}$. Hence, the equivalence also extends to the case where $\lambda = 0$.

Now consider the case where $\boldsymbol{\alpha} = \boldsymbol{0}$. There is no $\boldsymbol{d}$ such that $\boldsymbol{\alpha}^\top \boldsymbol{d} = \alpha_0$ unless $\alpha_0 = 0$. However, when $\boldsymbol{\alpha} = \boldsymbol{0}$, the left-hand side of (E-7) has the trivial solution $\boldsymbol{x} = \boldsymbol{x}^n$, and the right-hand side of (E-7) simplifies to $L(\alpha_0)$ since $0 \leqslant \lambda$ always holds. Thus, the equivalence still holds, which concludes the proof. $\square$

Proof of Proposition 3. 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 1 to re-express the embedded maximization over $\boldsymbol{x} \in \mathbb{X}$. Details are omitted for the sake of brevity. $\square$

Proof of Proposition 4. For ease of exposition, we adopt the notation introduced after prob-

lem (E-1). By construction of problem (E-1), λ is lower bounded by 0. To see that we can bound λ from above as well, we observe that there are optimal solutions for which λ does not exceed $\overline{\lambda} = \max\{\overline{\lambda}_1, \overline{\lambda}_2\}$, where

$$\overline{\lambda}_1 = \sup_{\beta \in \mathcal{H}} \mathrm{lip}(L) \cdot \|\beta_{\mathrm{x}}\|_*$$

with the bounded set $\mathcal{H} \subseteq \mathbb{R}^{1+M_x+M_z}$ containing all admissible hypotheses β, and

$$\overline{\lambda}_2 = \sup_{\beta \in \mathcal{H}} \max_{n \in [N]} \max_{i \in \mathcal{I}} \max_{z \in \mathbb{Z}} \left\{ \frac{l_\beta(x^n, z, iy^n)}{\kappa_z d_z(z, z^n) + \kappa_y \cdot \mathbb{1}[i = -1]} : (z, i) \neq (z^n, 1) \right\}. \qquad \text{(E-8)}$$

Indeed, selecting $\lambda \geqslant \overline{\lambda}_1$ ensures that all hypotheses $\beta \in \mathcal{H}$ are feasible, and selecting $\lambda \geqslant \overline{\lambda}_2$ implies that for any $\beta \in \mathcal{H}$, all left-hand sides in the first two constraint sets of problem (E-1) are non-positive, and thus all of these constraints are weakly dominated by the non-negativity constraints on s. Note that the numerator in the objective function of (E-8) is bounded since all suprema and maxima in (E-8) operate over bounded sets and l_β is Lipschitz continuous. Moreover, the denominator in the objective function of (E-8) is bounded from below by a strictly positive quantity since $(z, i) \neq (z^n, 1)$ and $\kappa_z, \kappa_y > 0$. $\qquad \square$

Proof of Corollary 1. 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. $\qquad \square$

Proof of Proposition 5. 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. $\qquad \square$

Proof of Proposition 6. For ease of exposition, we adopt the notation introduced after problem (E-2). Our proof proceeds in two steps. We first show that without loss of generality, we can impose bounds on each variable q_{ni}, $n \in [N]$ and $i \in \mathcal{I}$, and we afterwards show that we can impose non-restrictive bounds on λ 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 $\boldsymbol{q}_{ni}^\top(\boldsymbol{d} - \boldsymbol{C}\boldsymbol{x}^n) \geqslant 0$ 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 (E-2). Due to the non-negativity of f_0 in $\boldsymbol{s}$ in problem (E-2), larger values of $\boldsymbol{q}_{ni}$ thus weakly increase the objective function in (E-2). Non-zero values for $\boldsymbol{q}_{ni}$ can therefore only be optimal if they allow to reduce λ via the constraints $\|a_j y^n \cdot \boldsymbol{\beta}_\mathrm{x} - t \cdot \boldsymbol{C}^\top \boldsymbol{q}_{ni}\|_* \leqslant \lambda$. 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 $\boldsymbol{q}_{ni}^\top(\boldsymbol{d} - \boldsymbol{C}\boldsymbol{x}^n) \geqslant 0$ 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 $\boldsymbol{q}_{ni} = \boldsymbol{q}_{ni}^0 + \boldsymbol{q}_{ni}^+$ where $\boldsymbol{q}_{ni}^0 \in \mathrm{Null}(\boldsymbol{C}^\top)$ 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 $\boldsymbol{q}_{ni}' \in \mathrm{Null}(\boldsymbol{C}^\top) \backslash \mathcal{Q}^0$ satisfying $\boldsymbol{q}_{ni}^+ + \boldsymbol{q}_{ni}' \in \mathcal{C}^*$ for some $\boldsymbol{q}_{ni}^+ \in \mathcal{Q}^+$ there is $\boldsymbol{q}_{ni}^0 \in \mathcal{Q}^0$ satisfying $\boldsymbol{q}_{ni}^+ + \boldsymbol{q}_{ni}^0 \in \mathcal{C}^*$ such that $(\boldsymbol{q}_{ni}^+ + \boldsymbol{q}_{ni}^0)^\top(\boldsymbol{d} - \boldsymbol{C}\boldsymbol{x}^n) \leqslant (\boldsymbol{q}_{ni}^+ + \boldsymbol{q}_{ni}')^\top(\boldsymbol{d} - \boldsymbol{C}\boldsymbol{x}^n)$, that is, $\boldsymbol{q}_{ni}^{0\top}\boldsymbol{d} \leqslant \boldsymbol{q}_{ni}'^\top\boldsymbol{d}$, 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 λ can be bounded as well, note that by construction, λ is bounded from below by 0. To see that we can bound λ from above as well, we observe that there are optimal solutions to problem (E-2) for which λ 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 $\mathcal{H} \subseteq \mathbb{R}^{1+M_x+M_z}$ 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 $\lambda \geqslant \overline{\lambda}_1$ ensures that all hypotheses $\boldsymbol{\beta} \in \mathcal{H}$ are feasible, and selecting $\lambda \geqslant \overline{\lambda}_2$ implies that for any $\boldsymbol{\beta} \in \mathcal{H}$, all left-hand sides in the first two constraint sets of problem (E-2) 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 λ in problem (E-2) can be bounded as well. $\qquad\square$

Proof of Corollary 2. The proof is similar to that of Corollary 15 by Shafieezadeh-Abadeh et al. (2019). Details are omitted for the sake of brevity. $\qquad\square$

The proof of Proposition 7 relies on two lemmas that we will state and prove first.

Lemma 2. *The compound norm* $\|[\boldsymbol{\alpha}, \nu]\|_{\mathrm{comp}} = \|\boldsymbol{\alpha}\| + \kappa|\nu|$, $\boldsymbol{\alpha} \in \mathbb{R}^{M_x}$, $\nu \in \mathbb{R}$ *and* $\kappa > 0$, *satisfies*

$$\|[\boldsymbol{\alpha}, \nu]\|_{\mathrm{comp*}} = \max\left\{ \|\boldsymbol{\alpha}\|_*, \frac{|\nu|}{\kappa} \right\}.$$

Proof of Lemma 2. By definition of the dual norm, we have that

$$\|[\boldsymbol{\alpha}, \nu]\|_{\mathrm{comp*}} = \begin{cases} \underset{(\boldsymbol{x},y)\in\mathbb{R}^{M_x}\times\mathbb{R}}{\text{maximize}} & \boldsymbol{\alpha}^\top\boldsymbol{x} + \nu y \\ \\ \text{subject to} & \|\boldsymbol{x}\| + \kappa|y| \leqslant 1. \end{cases}$$

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 $\gamma \geqslant 0$,

$$g(\gamma) = \sup_{(\boldsymbol{x},y)\in\mathbb{R}^{M_{\mathrm{x}}}\times\mathbb{R}} \boldsymbol{\alpha}^\top \boldsymbol{x} + \nu y - \gamma(\|\boldsymbol{x}\| + \kappa|y| - 1).$$

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 $|\nu| \leqslant \gamma\kappa$ 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 1 with L being the identity to obtain the equivalent reformulation

$$g(\gamma) = \begin{cases} \gamma & \text{if } \|\boldsymbol{\alpha}\|_* \leqslant \gamma \ , \ |\nu| \leqslant \gamma\kappa, \\ +\infty & \text{otherwise.} \end{cases}$$

The dual problem therefore is

$$\underset{\gamma}{\text{minimize}} \quad \gamma$$

$$\text{subject to} \quad |\nu| \leqslant \gamma\kappa$$
$$\|\boldsymbol{\alpha}\|_* \leqslant \gamma$$
$$\gamma \in \mathbb{R}_+,$$

which has the optimal solution $\|[\boldsymbol{\alpha},\nu]\|_{\text{comp}*} = \gamma^\star = \max\{\|\boldsymbol{\alpha}\|_*, |\nu|/\kappa\}.$ $\qquad\square$

Lemma 3. *Assume that the loss function L is convex and Lipschitz continuous. For fixed*

$\boldsymbol{\alpha}, \boldsymbol{x}^n \in \mathbb{R}^{M_x}$, $\alpha_0, y^n \in \mathbb{R}$, $\kappa > 0$ *and* $\lambda \in \mathbb{R}_+$, *we have*

$$\sup_{(\boldsymbol{x},y)\in\mathbb{R}^{M_x}\times\mathbb{R}} L(\boldsymbol{\alpha}^\top\boldsymbol{x} - y + \alpha_0) - \lambda\|\boldsymbol{x} - \boldsymbol{x}^n\| - \lambda\kappa|y - y^n|$$

$$= \begin{cases} L(\boldsymbol{\alpha}^\top\boldsymbol{x}^n - y^n + \alpha_0) & \textit{if } \mathrm{lip}(L) \cdot \max\{\kappa\|\boldsymbol{\alpha}\|_*, 1\} \leqslant \lambda\kappa, \\ +\infty & \textit{otherwise.} \end{cases}$$

Proof of Lemma 3. Concatenating the variables $\boldsymbol{x}$ and y to $\boldsymbol{w} = [\boldsymbol{x}^\top, y]^\top \in \mathbb{R}^{M_x+1}$, letting $\boldsymbol{\eta} = [\boldsymbol{\alpha}^\top, -1]^\top$ and $\boldsymbol{w}^n = [(\boldsymbol{x}^n)^\top, y^n]^\top$ and defining the compound norm $\|[\boldsymbol{x}, y]\|_{\mathrm{comp}} = \|\boldsymbol{x}\| + \kappa|y|$, we can write the left-hand side of the equation in the statement of the lemma as

$$\sup_{\boldsymbol{w}\in\mathbb{R}^{M_x+1}} L(\boldsymbol{\eta}^\top\boldsymbol{w} + \alpha_0) - \lambda\|\boldsymbol{w} - \boldsymbol{w}^n\|_{\mathrm{comp}}.$$

Now we can apply Lemma 1 directly to conclude that

$$\sup_{\boldsymbol{w}\in\mathbb{R}^{M_x+1}} L(\boldsymbol{\eta}^\top\boldsymbol{w} + \alpha_0) - \lambda\|\boldsymbol{w} - \boldsymbol{w}^n\|_{\mathrm{comp}} = \begin{cases} L(\boldsymbol{\eta}^\top\boldsymbol{w}^n + \alpha_0) & \text{if } \mathrm{lip}(L) \cdot \|\boldsymbol{\eta}\|_{\mathrm{comp}*} \leqslant \lambda, \\ +\infty & \text{otherwise.} \end{cases}$$

The statement now follows from Lemma 2, which implies that $\mathrm{lip}(L) \cdot \|\boldsymbol{\eta}\|_{\mathrm{comp}*} \leqslant \lambda$ if and only if $\mathrm{lip}(L) \cdot \max\{\|\boldsymbol{\alpha}\|_*, |-1|/\kappa\} \leqslant \lambda$. $\qquad\square$

Proof of Proposition 7. 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 3 to re-express the embedded maximization over $(\boldsymbol{x}, y) \in \mathbb{R}^{M_x} \times \mathbb{R}$. Details are omitted for the sake of brevity. $\quad\square$

Proof of Corollary 3. The proof is similar to that of Proposition 4. Details are omitted for the sake of brevity. $\qquad\square$

Proof of Corollary 4. The proof is similar to that of Corollary 5 by Shafieezadeh-Abadeh et al. (2019). Details are omitted for the sake of brevity. $\qquad\square$

Proof of Proposition 8. 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 5. The proof is similar to that of Proposition 6. Details are omitted for the sake of brevity. □

Proof of Corollary 6. 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. □