
Wasserstein Logistic Regression with Mixed Features

Aras Selvi Mohammad Reza Belbasi Martin B. Haugh Wolfram Wiesemann
Imperial College Business School, Imperial College London, United Kingdom
{a.selvi19, r.belbasi21, m.haugh, ww}@imperial.ac.uk

Abstract

Recent work has leveraged the popular distributionally robust optimization paradigm to combat overfitting in classical logistic regression. While the resulting classification scheme displays a promising performance in numerical experiments, it is inherently limited to numerical features. In this paper, we show that distributionally robust logistic regression with mixed (*i.e.*, numerical and categorical) features, despite amounting to an optimization problem of exponential size, admits a polynomial-time solution scheme. We subsequently develop a practically efficient column-and-constraint approach that solves the problem as a sequence of polynomial-time solvable exponential conic programs. Our model retains many of the desirable theoretical features of previous works, but—in contrast to the literature—it does not admit an equivalent representation as a regularized logistic regression, that is, it represents a genuinely novel variant of logistic regression. We show that our method outperforms both the unregularized and the regularized logistic regression on categorical as well as mixed-feature benchmark instances.

1 Introduction

Consider a data set $(\mathbf{x}^i, y^i)_{i=1}^N$ with feature vectors \mathbf{x}^i and associated binary labels $y^i \in \{-1, 1\}$. Classical logistic regression assumes that the labels depend probabilistically on the features via

$$\text{Prob}(y \mid \mathbf{x}) = [1 + \exp(-y \cdot [\beta_0 + \boldsymbol{\beta}^\top \mathbf{x}])]^{-1},$$

where the parameters $(\beta_0, \boldsymbol{\beta}) \in \mathbb{R}^{1+n}$ are estimated from the empirical risk minimization problem

$$\begin{aligned} & \underset{(\beta_0, \boldsymbol{\beta})}{\text{minimize}} && \frac{1}{N} \sum_{i=1}^N l_{\boldsymbol{\beta}}(\mathbf{x}^i, y^i) \\ & \text{subject to} && (\beta_0, \boldsymbol{\beta}) \in \mathbb{R}^{n+1} \end{aligned}$$

with the *log-loss* function $l_{\boldsymbol{\beta}}(\mathbf{x}, y) := \log(1 + \exp(-y \cdot [\beta_0 + \boldsymbol{\beta}^\top \mathbf{x}]))$. Its compelling performance across many domains, the availability of mature and computationally efficient algorithms as well as its interpretability have all contributed to the widespread adoption of logistic regression [6, 18, 27].

Similar to other machine learning models, logistic regression is prone to overfitting, especially when the number of training samples is small relative to the number of considered features. Moreover, logistic regression can be sensitive to erroneous feature and label values as well as distribution shifts under which the training and test sets stem from different distributions. In recent years, distributionally robust (DR) optimization [3, 4] has been proposed to simultaneously address these challenges. To this end, the DR optimization paradigm models a machine learning task as a zero-sum game between the decision maker, who seeks to obtain the most accurate model (*e.g.*, the parameter vector $\boldsymbol{\beta}$ in a logistic regression), and a fictitious adversary who observes the decision maker’s model and subsequently selects the worst data-generating distribution in the vicinity of the empirical distribution formed from the available training data. In this context, the similarity of distributions is commonly

measured in terms of moment bounds [9, 39], ϕ -divergences such as the Kullback-Leibler divergence [19, 22] or the Wasserstein distance [15, 25]. It has been observed that in many cases, the resulting DR machine learning models are equivalent to regularized versions of the same models, in which case distributionally robustness provides a new perspective on regularizers that are often selected *ad hoc* [21, 41]. Most importantly, DR machine learning models often admit computationally efficient reformulations as finite-dimensional convex optimization problems that can be solved in polynomial time with off-the-shelf solvers.

In this work we study a DR variant of the logistic regression where the distance between the empirical distribution and the unknown true data-generating distribution is measured by the popular Wasserstein (also known as Kantorovich-Rubinstein or earth mover’s) distance [15, 25]. This problem has been first studied by [31], who show that the resulting DR logistic regression problem admits an equivalent reformulation as a polynomial-size convex optimization problem *if all features are numerical*. Applying similar techniques to the mixed-feature logistic regression problem, which appears to be most prevalent in practice, would result in an exponential-size convex optimization problem whose naïve solution with the methods of [31] does not scale to interesting problem sizes (*cf.* Section 4). Our contributions may be summarized as follows.

- (i) On the *theoretical side*, we show that the complexity of DR mixed-feature regression crucially relies on the selected loss function. In particular, for the log-loss function employed in logistic regression, the problem—despite its natural representation as an exponential-size convex optimization problem—admits a polynomial-time solution scheme. We also show that in stark contrast to earlier variants of the problem, our mixed-feature regression does *not* admit an equivalent representation as a regularized problem. This provides compelling evidence that the DR logistic regression problem with mixed features is fundamentally different to the DR problem with only numerical features.
- (ii) On the *computational side*, we propose a column-and-constraint scheme that solves the DR mixed-feature logistic regression as a sequence of polynomial-time solvable exponential conic programs. We show that the key step of our procedure, the identification of the most violated constraint, can be implemented efficiently for a broad range of metrics, despite its natural representation as a combinatorial optimization problem. Indeed, identifying the most violated constraint by brute force is out of the question, while the standard approach [43] of solving the most violated constraint problem would have an unacceptably high runtime.
- (iii) On the *numerical side*, we show that our column-and-constraint scheme drastically reduces computation times over a naïve monolithic implementation of the regression problem. We also show that our model performs favorably on standard categorical and mixed-feature benchmark instances when compared against classical and regularized logistic regression.

The literature on DR machine learning under the Wasserstein distance is vast and rapidly growing. Recent works have considered, among others, the use of Wasserstein DR models in multi-label learning [14], generative adversarial networks [2, 8] and the generation of adversarial examples [40], density estimation [38] and learning Gaussian mixture models [20], graph-based semi-supervised learning [35], supervised dimensionality reduction [13] and reinforcement learning [1]. DR has also been found to help alleviate problems with overfitting [24, 21], label uncertainty [21] and distribution shifts [37, 36]. We refer to [21] for a recent review of the literature. Our work is most closely related to [31], and we compare our findings with the results of that work in Section 2.

We proceed as follows. Section 2 defines and analyzes the mixed-feature DR logistic regression problem, which is solved in Section 3 via column-and-constraint generation. We report numerical results in Section 4. Auxiliary material and all proofs are relegated to the appendix.

Notation. We define $\mathbb{B} = \{0, 1\}$ and $[N] = \{1, \dots, N\}$ for $N \in \mathbb{N}$. The set of all probability distributions supported on a set Ξ is denoted by $\mathcal{P}_0(\Xi)$, while the Dirac distribution placing unit probability mass on $\mathbf{x} \in \mathbb{R}^n$ is denoted by $\delta_{\mathbf{x}} \in \mathcal{P}_0(\mathbb{R}^n)$. The indicator function $\mathbb{1}[\mathcal{E}]$ attains the value 1 (0) whenever the expression \mathcal{E} is (not) satisfied. Finally, we use Roman *d* to denote differentials and to distinguish them from the d used to denote distances.

2 Mixed-Feature DR Logistic Regression

Section 2.1 derives an exponential-size convex programming formulation of the DR logistic regression problem that serves as the basis of our analysis. Section 2.2 compares our formulation with a naïve model that treats categorical features as continuous ones. Section 2.3, finally, shows that our formulation, despite its exponential size, can be solved in polynomial time due to the benign structure of the log-loss function. It also establishes that in stark contrast to the literature, our formulation does not reduce to a regularized non-robust logistic regression, that is, it constitutes a genuinely novel variant of the logistic regression problem.

Our formulation enjoys strong finite-sample and asymptotic performance guarantees from the literature. Moreover, despite its exponential size, our model always accommodates worst-case distributions that exhibit a desirable sparsity pattern. We relegate these results to Appendix A.

2.1 Formulation as an Exponential-Size Convex Program

From now on, we consider a mixed-feature data set $\{\xi^i := (\mathbf{x}^i, \mathbf{z}^i, y^i)\}_{i \in [N]}$ with n numerical features $\mathbf{x}^i = (x_1^i, \dots, x_n^i) \in \mathbb{R}^n$, m categorical features $\mathbf{z}^i = (z_1^i, \dots, z_m^i) \in \mathbb{C}(k_1) \times \dots \times \mathbb{C}(k_m)$ and a binary label $y^i \in \{-1, +1\}$. Here, $\mathbb{C}(s) = \{\mathbf{z} \in \mathbb{B}^{s-1} : \sum_{j \in [s-1]} z_j \leq 1\}$ for $s \in \mathbb{N} \setminus \{1\}$ represents the one-hot encoding of a categorical feature with s possible values; in particular, $\mathbb{C}(2) = \mathbb{B}$ encodes a binary feature. We denote by $\mathbb{C} = \mathbb{C}(k_1) \times \dots \times \mathbb{C}(k_m)$ and $\Xi = \mathbb{R}^n \times \mathbb{C} \times \{-1, +1\}$ the support of the categorical features as well as the data set, respectively, and we let $k = k_1 + \dots + k_m - m$ be the number of slopes used for the categorical features.

If we had access to the true data-generating distribution $\mathbb{P}^0 \in \mathcal{P}_0(\Xi)$, we would solve the (non-robust) logistic regression problem

$$\begin{aligned} & \underset{\beta}{\text{minimize}} && \mathbb{E}_{\mathbb{P}^0} [l_\beta(\mathbf{x}, \mathbf{z}, y)] \\ & \text{subject to} && \beta = (\beta_0, \beta_N, \beta_C) \in \mathbb{R}^{1+n+k}, \end{aligned}$$

where $\mathbb{E}_{\mathbb{P}^0}$ denotes the expectation under \mathbb{P}^0 , and where the log-loss function l_β now takes the form

$$l_\beta(\mathbf{x}, \mathbf{z}, y) := \log(1 + \exp[-y \cdot (\beta_0 + \beta_N^\top \mathbf{x} + \beta_C^\top \mathbf{z})])$$

to account for the presence of categorical features. Since the distribution \mathbb{P}^0 is unknown in practice, the empirical risk minimization problem replaces \mathbb{P}^0 with the empirical distribution $\hat{\mathbb{P}}_N := \frac{1}{N} \sum_{i=1}^N \delta_{\xi^i}$ that places equal probability mass on all observations $\{\xi^i\}_{i \in [N]}$. Standard arguments show that when these observations are i.i.d., the empirical risk minimization problem recovers the logistic regression under \mathbb{P}^0 as $N \rightarrow \infty$. In practice, however, data tends to be scarce, and the empirical risk minimization problem exhibits an ‘optimism bias’ that is also known as overfitting [6, 18, 27], the error maximization effect of optimization [10, 23] or the optimizer’s curse [34].

DR logistic regression combats the aforementioned overfitting phenomenon by solving the semi-infinite optimization problem

$$\begin{aligned} & \underset{\beta}{\text{minimize}} && \sup_{\mathbb{Q} \in \mathfrak{B}_\epsilon(\hat{\mathbb{P}}_N)} \mathbb{E}_{\mathbb{Q}} [l_\beta(\mathbf{x}, \mathbf{z}, y)] \\ & \text{subject to} && \beta = (\beta_0, \beta_N, \beta_C) \in \mathbb{R}^{1+n+k}, \end{aligned} \tag{1}$$

where the *ambiguity set* $\mathfrak{B}_\epsilon(\hat{\mathbb{P}}_N)$ contains all distributions \mathbb{Q} in a (soon to be defined) vicinity of the empirical distribution $\hat{\mathbb{P}}_N$, and where the expectation is taken with respect to \mathbb{Q} . Problem (1) can be interpreted as a zero-sum game between the decision maker, who chooses a logistic regression model parameterized by β , and a fictitious adversary that observes β and subsequently chooses the ‘worst’ distribution (in terms of the incurred log-loss) from $\mathfrak{B}_\epsilon(\hat{\mathbb{P}}_N)$. Contrary to the classical non-robust logistic regression, problem (1) guarantees to *overestimate* the log-loss incurred by β under the unknown true distribution \mathbb{P}^0 as long as \mathbb{P}^0 is contained in $\mathfrak{B}_\epsilon(\hat{\mathbb{P}}_N)$. On the other hand, problem (1) recovers the empirical risk minimization problem when the ambiguity set $\mathfrak{B}_\epsilon(\hat{\mathbb{P}}_N)$ approaches a singleton set that only contains the empirical distribution $\hat{\mathbb{P}}_N$. Note that problem (1) is convex as the convexity of the logistic regression objective $\mathbb{E}_{\mathbb{Q}} [l_\beta(\mathbf{x}, \mathbf{z}, y)]$ is preserved under the supremum operator. That said, the problem typically constitutes a semi-infinite program as it comprises finitely

many decision variables but—if the embedded supremum in the objective is brought to the constraints via an epigraph reformulation—infinately many constraints whenever the ambiguity set harbors infinitely many distributions. As such, it is not obvious how the problem can be solved efficiently.

In this paper, we choose as ambiguity set the Wasserstein ball $\mathfrak{B}_\epsilon(\hat{\mathbb{P}}_N) := \{\mathbb{Q} \in \mathcal{P}_0(\Xi) : W(\mathbb{Q}, \hat{\mathbb{P}}_N) \leq \epsilon\}$ of radius $\epsilon > 0$ that is centered at the empirical distribution $\hat{\mathbb{P}}_N$.

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*

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

where $\xi = (\mathbf{x}, \mathbf{z}, y) \in \Xi$ and $\xi' = (\mathbf{x}', \mathbf{z}', y') \in \Xi$, while $d(\xi, \xi')$ is the ground metric on Ξ .

The Wasserstein distance can be interpreted as the minimum cost of moving \mathbb{Q} to \mathbb{P} when $d(\xi, \xi')$ is the cost of moving a unit mass from ξ to ξ' . The Wasserstein radius ϵ thus imposes a budget on the transportation cost that the adversary can spend on perturbing the empirical distribution $\hat{\mathbb{P}}_N$.

We next define the ground metric d that we use throughout the paper.

Definition 2 (Ground Metric). *We measure the distance between two data-points $\xi = (\mathbf{x}, \mathbf{z}, y) \in \Xi$ and $\xi' = (\mathbf{x}', \mathbf{z}', y') \in \Xi$ with $\mathbf{z} = (z_1, \dots, z_m)$ and $\mathbf{z}' = (z'_1, \dots, z'_m)$ as*

$$d(\xi, \xi') := \|\mathbf{x} - \mathbf{x}'\| + d_C(\mathbf{z}, \mathbf{z}') + \kappa \cdot \mathbb{1}[y \neq y'], \quad (3a)$$

where $\|\cdot\|$ is any rational norm on \mathbb{R}^n , $\kappa > 0$ and the metric d_C on \mathbb{C} satisfies

$$d_C(\mathbf{z}, \mathbf{z}') := \left(\sum_{i \in [m]} \mathbb{1}[z_i \neq z'_i] \right)^{1/p} \quad \text{for some } p > 0. \quad (3b)$$

Intuitively, the ground metric d measures the distance of the numerical features \mathbf{x} and \mathbf{x}' by the norm distance $\|\mathbf{x} - \mathbf{x}'\|$, whereas the distance of the categorical features \mathbf{z} and \mathbf{z}' is measured by (a power of) the number of discrepancies between \mathbf{z} and \mathbf{z}' . Likewise, any discrepancy between the labels y and y' is accounted for by a constant κ , which allows for scaling between the features and the labels.

We first extend the convex optimization model of [31] for the DR continuous-feature logistic regression to an exponential conic model that accommodates for categorical features.

Theorem 1 (Exponential Conic Representation). *The DR logistic regression problem (1) admits the equivalent reformulation*

$$\begin{aligned} & \underset{\beta, \lambda, \mathbf{s}, \mathbf{u}^\pm, \mathbf{v}^\pm}{\text{minimize}} && \lambda\epsilon + \frac{1}{N} \sum_{i \in [N]} s_i \\ & \text{subject to} && \left. \begin{aligned} & u_{i,\mathbf{z}}^+ + v_{i,\mathbf{z}}^+ \leq 1 \\ & (u_{i,\mathbf{z}}^+, 1, -s_i - \lambda d_C(\mathbf{z}, \mathbf{z}^i)) \in \mathcal{K}_{\text{exp}} \\ & (v_{i,\mathbf{z}}^+, 1, -y^i \beta_N^\top \mathbf{x}^i - y^i \beta_C^\top \mathbf{z} - y^i \beta_0 - s_i - \lambda d_C(\mathbf{z}, \mathbf{z}^i)) \in \mathcal{K}_{\text{exp}} \end{aligned} \right\} \forall i \in [N], \forall \mathbf{z} \in \mathbb{C} \\ & && \left. \begin{aligned} & u_{i,\mathbf{z}}^- + v_{i,\mathbf{z}}^- \leq 1 \\ & (u_{i,\mathbf{z}}^-, 1, -s_i - \lambda d_C(\mathbf{z}, \mathbf{z}^i) - \lambda \kappa) \in \mathcal{K}_{\text{exp}} \\ & (v_{i,\mathbf{z}}^-, 1, y^i \beta_N^\top \mathbf{x}^i + y^i \beta_C^\top \mathbf{z} + y^i \beta_0 - s_i - \lambda d_C(\mathbf{z}, \mathbf{z}^i) - \lambda \kappa) \in \mathcal{K}_{\text{exp}} \end{aligned} \right\} \forall i \in [N], \forall \mathbf{z} \in \mathbb{C} \\ & && \|\beta_N\|_* \leq \lambda \\ & && \beta = (\beta_0, \beta_N, \beta_C) \in \mathbb{R}^{1+n+k}, \quad \lambda \geq 0, \quad \mathbf{s} \in \mathbb{R}^N \\ & && (u_{i,\mathbf{z}}^+, v_{i,\mathbf{z}}^+, u_{i,\mathbf{z}}^-, v_{i,\mathbf{z}}^-) \in \mathbb{R}^4, \quad i \in [N] \text{ and } \mathbf{z} \in \mathbb{C} \end{aligned} \quad (4)$$

as a finite-dimensional exponential conic program, where \mathcal{K}_{exp} denotes the exponential cone

$$\mathcal{K}_{\text{exp}} := \text{cl}(\{(a, b, c) : a \geq b \cdot \exp(c/b), a > 0, b > 0\}) \subset \mathbb{R}^3.$$

The s_i 's and λ that appear in (4) are dual variables that arise from dual of the inner (i.e., sup) problem in (1). For each $\mathbf{z} \in \mathbb{C}$ and $i \in [N]$, $(u_{i,\mathbf{z}}^+, v_{i,\mathbf{z}}^+, u_{i,\mathbf{z}}^-, v_{i,\mathbf{z}}^-)$ is a vector of auxiliary variables that we use to model softplus constraints that arise from our intermediate analysis of problem (1) (Appendix E).

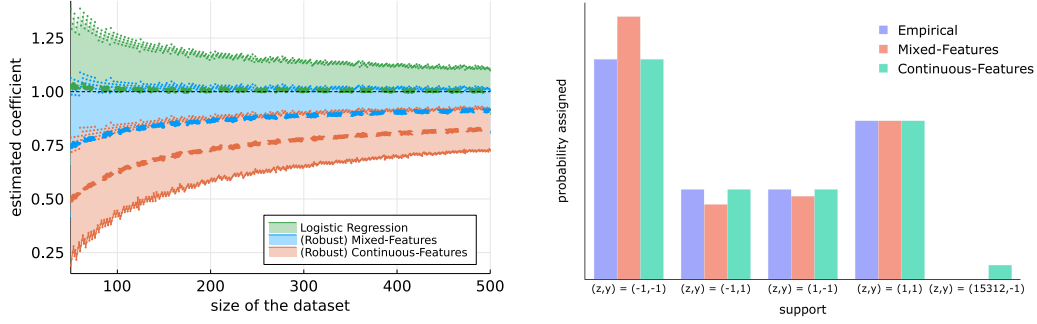


Figure 1: *Left*: Estimates of β for the standard logistic regression, our mixed-feature as well as the continuous-feature model as a function of the size N of the data set. All results are reported as averages over 2,000 statistically independent runs. *Right*: Comparison of the empirical distribution as well as the worst-case distributions for $\beta = 1$ under the two DR models for $N = 250$ samples. The probabilities are plotted on a log-scale to make small values visible.

Problem (4) is finite-dimensional and convex. Moreover, since exponential conic programs admit a self-concordant barrier function, they can be solved in polynomial time relative to their input size [28]. Due to the presence of categorical features in our regression problem, however, problem (4) comprises exponentially many variables and constraints, and a naïve solution of (4) would therefore require an *exponential* amount of time. The exponential size of problem (4) renders our formulation fundamentally different from the model of [31], and it significantly complicates the solution.

2.2 Comparison with Continuous-Feature DR Logistic Regression

Our formulation (4) of the mixed-feature DR logistic regression (1) accounts for categorical features $z \in \mathbb{C}$ at the expense of an exponential number of variables and constraints. It is therefore tempting to treat the categorical features as continuous ones and directly apply the continuous-feature-only reformulation of the DR logistic regression (1) proposed by [31]. In the following, we argue that such a reformulation would hedge against nonsensical worst-case distributions that in turn lead to overly conservative regression models. To see this, consider a stylized setting where the data set $(z^i, y^i)_{i \in [N]}$ comprises a single binary feature $z^i \in \{-1, +1\}$ that impacts the label $y^i \in \{-1, +1\}$ via the logistic model

$$\text{Prob}(y | z) = [1 + \exp(-y \cdot \beta z)]^{-1}$$

with $\beta = 1$. (While we use $z^i \in \{0, 1\}$ in the other sections of the paper, the analysis and results of this subsection and Appendix B assumed $z^i \in \{-1, +1\}$ as this made comparisons with [31] easier.) We attempt to recover this logistic model in two ways:

1. *Mixed-feature model*. We employ our DR logistic regression (1) with a single categorical feature, using $p = 1$ and $\kappa = 1$ in the ground metric (*cf.* Definition 2).
2. *Continuous-feature model*. We employ a variant of problem (1) that treats the categorical feature z_i as a continuous one. We choose as ground metric

$$d(\xi, \xi') := \frac{1}{2}|z - z'| + \mathbb{1}[y \neq y'].$$

We therefore use the feature distances $\mathbb{1}[z \neq z']$ in the mixed-feature case and $\frac{1}{2}|z - z'|$ in the continuous-feature case. Note they both attain the value 0 (1) for equal (different) feature values.

Figure 1 (left) compares the mean values (dashed lines) as well as the 15% and 85% quantiles (shaded regions) of β for the standard logistic regression, our mixed-feature as well as the continuous-feature model as a function of the size N of the data set. For both DR models, we employed the same Wasserstein radius $\epsilon \propto 1/\sqrt{N}$, which is motivated by the finite sample guarantee presented in the next section (*cf.* Theorem 6). The figure shows that the continuous-feature model excessively shrinks its estimates of β , thus leading to overly conservative results. In fact, the true parameter value $\beta = 1$

consistently lies outside the confidence region of the continuous-feature model, independent of the sample size N . This is due to the fact that the continuous-feature model accounts for nonsensical worst-case distributions under which the categorical feature can take values outside its domain $\{-1, +1\}$, see Figure 1 (right). In fact, the continuous-feature model places non-zero probability on an extreme scenario under which the binary feature $z \in \{-1, +1\}$ attains the value 15,312. (In Appendix B we explain why this value of 15,312 arises.) Our mixed-feature model, on the other hand, restricts the worst-case distribution to the domain of the categorical feature and thus hedges against realistic distributions only.

We note that in practice, the radius ϵ of the Wasserstein ball will be chosen via cross-validation (cf. Section 4), in which case our mixed-feature model reliably outperforms the classical logistic regression on standard benchmark instances. While one may argue that the impact of nonsensical worst-case distributions in the continuous-feature model is alleviated by the cross-validated radii, Figure 1 offers strong theoretical and practical reasons against the use of a continuous-feature model for categorical or mixed-feature problems: The continuous-feature model hedges against a clumsily perturbed worst-case distribution that introduces a pronounced bias in the estimation. We also note that another possible approach would be to use the continuous-feature model but to restrict the support of binary features to $[-1, 1]$. Unfortunately, for the log-loss function no tractable reformulation with support constraints appears to be known; see [32]. Moreover, Section 4 will show that the mixed-feature model, despite its exponential size, can be solved quickly and reliably with our novel column-and-constraint approach from Section 3.

2.3 Complexity Analysis

We first show that despite its exponential size, the DR logistic regression problem (1) admits a polynomial-time solution. Key to this perhaps surprising finding is the shape of the loss function: while problem (1) is strongly NP-hard (and thus unlikely to admit a polynomial-time solution scheme) for generic loss functions, it can be solved in polynomial time for the log-loss function l_β employed in logistic regression.

Theorem 2 (Complexity of the DR Logistic Regression (1)).

- (i) *For generic loss functions l_β , problem (1) is strongly NP-hard even if $n = 0$ and $N = 1$.*
- (ii) *For the loss function $l_\beta(\mathbf{x}, \mathbf{z}, y) = \log(1 + \exp[-y \cdot (\beta_0 + \beta_N^\top \mathbf{x} + \beta_C^\top \mathbf{z})])$ and the ground metric of Definition 2, problem (1) 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.

A by now well-known result shows that when $m = 0$ (no categorical features), the DR logistic regression problem (1) reduces to a classical logistic regression with an additional regularization term $\|\beta_x\|_*$ in the objective function when the output label weight κ in Definition 2 approaches ∞ [31, 32]. We next show that this reduction to a classical regularized logistic regression no longer holds in our problem setting when categorical features are present.

Theorem 3 (Absence of a Reformulation as a Regularized Problem). *Even when the output label weight κ approaches ∞ in the DR logistic regression (1), problem (1) does not admit an equivalent reformulation*

$$\begin{aligned} & \underset{\beta}{\text{minimize}} && \mathbb{E}_{\mathbb{P}_N} [l_\beta(\mathbf{x}, \mathbf{z}, y)] + \mathfrak{R}(\beta) \\ & \text{subject to} && \beta = (\beta_0, \beta_N, \beta_C) \in \mathbb{R}^{1+n+k}, \end{aligned}$$

as a classical regularized logistic regression for any regularizer $\mathfrak{R} : \mathbb{R}^{1+n+k} \rightarrow \mathbb{R}$.

Note that Theorem 3 does not only preclude the existence of a specific regularizer, but it excludes the existence of *any* regularizer, no matter how complex its dependence on β might be. We are not aware of any prior results of this form in the literature. Some insight for the result in Theorem 3 may be found via the special case we consider in its proof in Appendix E and in Remark 2 that follows it.

Algorithm 1 Column-and-Constraint Generation Scheme for Problem (4).

Set $\text{LB}_0 = -\infty$ and $\text{UB}_0 = +\infty$
Choose (possibly empty) subsets $\mathcal{W}^+, \mathcal{W}^- \subseteq [N] \times \mathbb{C}$
while $\text{LB}_t \neq \text{UB}_t$ **do**
 Let θ^* be the optimal value and $(\beta, \lambda, \mathbf{s}, \mathbf{u}^\pm, \mathbf{v}^\pm)$ be a minimizer of problem (4) where:
 • the first constraint set is only enforced for all $(i, \mathbf{z}) \in \mathcal{W}^+$;
 • the second constraint set is only enforced for all $(i, \mathbf{z}) \in \mathcal{W}^-$;
 • we only include those $(u_{i,\mathbf{z}}^+, v_{i,\mathbf{z}}^+)$ for which $(i, \mathbf{z}) \in \mathcal{W}^+$;
 • we only include those $(u_{i,\mathbf{z}}^-, v_{i,\mathbf{z}}^-)$ for which $(i, \mathbf{z}) \in \mathcal{W}^-$.
 Identify a most violated index (i^+, \mathbf{z}^+) of the first constraint set in (4) and add it to \mathcal{W}^+
 Identify a most violated index (i^-, \mathbf{z}^-) of the second constraint set in (4) and add it to \mathcal{W}^-
 Let ϑ^+ and ϑ^- be the violations of (i^+, \mathbf{z}^+) and (i^-, \mathbf{z}^-) , respectively
 Update $\text{LB}_t = \theta^*$ and $\text{UB}_t = \min\{\text{UB}_{t-1}, \theta^* + \log(1 + \max\{\vartheta^+, \vartheta^-\})\}$
 Update $t = t + 1$
end while

Algorithm 2 Identification of Most Violated Constraints in the Reduced Problem (4).

for $j \in \{1, \dots, m\}$ **do**
 Find a feature value \mathbf{z}_j^* that minimizes $\mathbf{y}^i \cdot \beta_{\mathbf{C},j}^\top \mathbf{z}_j$ across all $\mathbf{z}_j \in \mathbb{C}(k_j) \setminus \{\mathbf{z}_j^i\}$
end for
Let $\pi : [m] \rightarrow [m]$ be an ordering such that

$$\mathbf{y}^i \cdot \beta_{\mathbf{C},\pi(j)}^\top (\mathbf{z}_{\pi(j)}^* - \mathbf{z}_{\pi(j)}^i) \leq \mathbf{y}^i \cdot \beta_{\mathbf{C},\pi(j')}^\top (\mathbf{z}_{\pi(j')}^* - \mathbf{z}_{\pi(j')}^i) \quad \forall 1 \leq j \leq j' \leq m$$

Set $\mathcal{W} = \emptyset$
for $\delta \in \{0, 1, \dots, m\}$ **do**
 Update $\mathcal{W} = \mathcal{W} \cup \{\mathbf{z}\}$, where $\mathbf{z}_j = \mathbf{z}_j^*$ if $\pi(j) \leq \delta$ and $\mathbf{z}_j = \mathbf{z}_j^i$ otherwise
end for
Determine the most violated constraint from the candidate set \mathcal{W}

3 Column-and-Constraint Solution Scheme

Our DR logistic regression problem (4) comprises exponentially many variables and constraints, which renders its solution as a monolithic exponential conic program challenging. Instead, Algorithm 1 employs a column-and-constraint generation scheme, *e.g.*, [43], that alternates between (i) the solution of relaxations of problem (4) that omit most of its variables and constraints and (ii) adding those variables and constraints that promise to maximally tighten the relaxations.

Theorem 4. *Algorithm 1 solves problem (4) in finitely many iterations. Moreover, LB_t and UB_t constitute monotone sequences of lower and upper bounds on the optimal value of problem (4).*

A key step in Algorithm 1 is the identification of most violated indices $(i, \mathbf{z}) \in [N] \times \mathbb{C}$ of the first and second constraint set in the reduced DR regression problem (4) for a fixed solution $(\beta, \lambda, \mathbf{s}, \mathbf{u}^\pm, \mathbf{v}^\pm)$. For the first constraint set in (4), the identification of such indices requires for each data point $i \in [N]$ the solution of the combinatorial problem

$$\begin{aligned} & \underset{\mathbf{z}}{\text{maximize}} \quad \min_{u^+, v^+} \left\{ u^+ + v^+ : \begin{bmatrix} (u^+, 1, -s_i - \lambda d_{\mathbf{C}}(\mathbf{z}, \mathbf{z}^i)) \in \mathcal{K}_{\text{exp}}, \\ (v^+, 1, -\mathbf{y}^i \beta_{\mathbf{N}}^\top \mathbf{x}^i - \mathbf{y}^i \beta_{\mathbf{C}}^\top \mathbf{z} - \mathbf{y}^i \beta_0 - s_i - \lambda d_{\mathbf{C}}(\mathbf{z}, \mathbf{z}^i)) \in \mathcal{K}_{\text{exp}} \end{bmatrix} \right\} \\ & \text{subject to} \quad \mathbf{z} \in \mathbb{C}, \end{aligned} \tag{5}$$

where an optimal value greater than 1 corresponds to a violated constraint; an analogous problem can be defined for the second constraint set in (4). Despite its combinatorial nature, the above problem can be solved efficiently by means of Algorithm 2.

Theorem 5. *Algorithm 2 identifies a most violated constraint for a given data-point $i \in [N]$ in time $\mathcal{O}(k + n + m^2)$.*

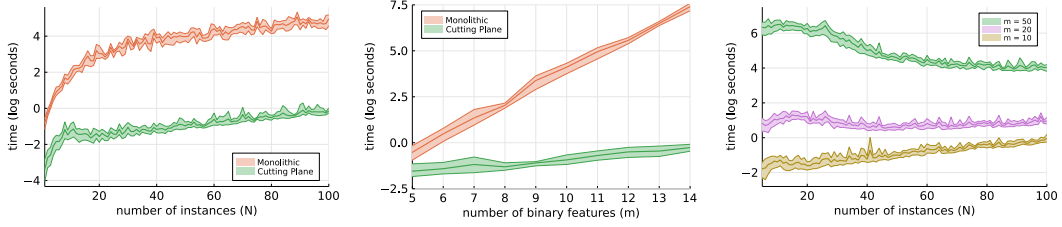


Figure 2: Runtime comparison between our column-and-constraint scheme and a naïve solution of problem (4) as a monolithic exponential conic program. *Left*: Runtimes for $m = 10$ binary features as a function of the number N of data points. *Middle*: Runtimes for $N = 50$ as a function of m . *Right*: Runtimes of our column-and-constraint scheme only for varying combinations of m and N . In all graphs, shaded regions correspond to 10%-90% confidence regions and bold lines report median values over 50 statistically independent runs. Note the log-scale in the plots.

Algorithm 2 first determines for each categorical feature z_j , $j = 1, \dots, m$, the feature value $z_j^* \in \mathbb{C}(k_j) \setminus \{z_j^i\}$ that, *ceteris paribus*, contributes to a maximal constraint violation in time $\mathcal{O}(k)$. The subsequent step sorts the different features $j = 1, \dots, m$ in descending order of their contribution to constraint violations in time $\mathcal{O}(m \log m)$. The last step of Algorithm 2, finally, constructs the most violated constraint z across those in which $\delta = 0, \dots, m$ categorical feature values deviate from z^i and subsequently picks the most violated constraint of these in time $\mathcal{O}(n + m^2)$. We note that the overall complexity of $\mathcal{O}(k + n + m^2)$ can be reduced to $\mathcal{O}(k + n + m \log m)$ by a clever use of data structures in the final step; for ease of exposition, we omit the details. Finally, since Algorithm 2 is applied to each data point $i \in [N]$, the overall complexity increases by a factor of N . We provide additional intuition behind Algorithm 2 immediately before the proof of Theorem 5 in Appendix E.

4 Numerical Results

Section 4.1 first compares the runtimes of our column-and-constraint scheme from Section 3 with those of solving the DR logistic regression problem (4) naïvely as a monolithic exponential conic program. We subsequently compare the classification performance of problem (1) with those of a classical unregularized and regularized logistic regression on standard benchmark instances with categorical features (Section 4.2) and mixed features (Appendix C).

All algorithms were implemented in Julia [5] (MIT license) and executed on Intel Xeon 2.66GHz processors with 8GB memory in single-core mode. We use MOSEK 9.3 [26] (commercial) to solve all exponential conic programs through JuMP [12] (MPL2 License). (We note the open source solvers Ipopt and CVXOpt could be used instead of MOSEK.) All source codes and detailed results are available on GitHub (<https://github.com/selvi-aras/WassersteinLR>).

4.1 Runtime Comparison with Monolithic Formulation

We first compare the computation times of our column-and-constraint scheme from Section 3 with those of solving the DR logistic regression problem (4) naïvely as a monolithic exponential conic program. To this end, we randomly generate synthetic logistic regression instances with varying numbers N of data points and m of binary features. While Figure 2 (left) shows that both approaches scale similarly in the number N of data points, Figure 2 (middle) reveals that the solution of the monolithic formulation scales exponentially in the number m of binary features. In contrast, our column-and-constraint scheme scales gracefully in both the number N of data points and the number m of binary features (*cf.* Figure 2, right). A similar behavior can be observed with more general, non-binary categorical features; we omit the results due to space constraints.

4.2 Performance on Categorical-Feature Instances

We next compare the classification performance (in terms of the out-of-sample classification error) of our unregularized ('DRO') and Lasso-regularized ('r-DRO') DR logistic regression problem (1) with those of a classical unregularized ('LR') as well as Lasso-regularized ('r-LR'), mass transportation-regularized (MT) [31, 32], and robust Wasserstein profile inference-regularized (PI) [7] logistic

Data Set	N	k	m	LR	DRO ($\kappa = 1$)	DRO ($\kappa = m$)	r-LR	r-DRO ($\kappa = 1$)	r-DRO ($\kappa = m$)	MT ($\kappa = 1$)	MT ($\kappa = m$)	PI ($\alpha = 0.05$)
breast-cancer	277	42	9	29.56%	29.15%	28.55% ^{†‡}	29.05%	29.16%	28.82%	29.44%	29.40%	29.33%
spect	267	22	22	18.81%	17.72%	<i>17.51%</i>	17.79%	18.38%	15.83% ^{†‡}	<i>18.49%</i>	18.74%	20.60%
monks-3	554	11	6	2.07%	2.08%	2.10%	2.14%	2.19%	2.13%	2.15%	2.15%	37.87%
tic-tac-toe	958	18	9	1.92%	1.69%	1.67% [†]	1.68%	1.68%	1.69%	1.72%	1.81%	30.02%
kr-vs-kp	3,196	37	36	2.64%	2.66%	2.64%	2.69%	2.66%	2.66%	2.77%	2.74%	12.85%
balance-scale*	625	16	4	0.85%	0.82%	0.72%	0.73%	0.71%	0.64% ^{†‡}	0.72%	0.69%	36.10%
hayes-roth*	160	11	4	17.00%	16.31%	16.09% ^{†‡}	18.66%	18.19%	17.78%	17.50%	16.38%	37.59%
lymphography*	148	42	18	21.17%	17.83%	16.72% ^{†‡}	17.38%	17.52%	17.03%	19.41%	17.83%	20.00%
car*	1,728	15	6	4.73%	4.73%	4.73%	4.99%	4.76%	4.87%	4.73%	4.77%	15.37%
splices*	3,189	229	60	6.80%	5.92% ^{†‡}	6.66%	6.22%	5.99%	6.02%	6.57%	6.87%	11.08%
house-votes-84	435	32	16	6.60%	4.43%	5.37%	4.54%	4.41%	5.16%	4.78%	5.54%	4.26% [†]
hiv	6,590	152	8	5.94%	5.93%	5.92%	5.90% [†]	5.90% [†]	5.90% [†]	6.04%	6.03%	20.45%
primacy-tumor*	339	25	17	13.93%	13.66%	13.76%	14.51%	<i>14.06%</i>	14.19%	13.87%	<i>13.85%</i>	19.18%
audiology*	226	92	69	14.18%	<i>14.03%</i>	14.40%	3.11%	2.80% ^{†‡}	3.91%	12.64%	7.49%	15.49%

Table 1: Classification errors of unregularized as well as Lasso-, mass transportation-, and profile inference-regularized variants of the classical logistic regression and our DR regression on UCI benchmark instances with categorical features only. The smallest error within each model group (unregularized vs. regularized) is highlighted in italics, whereas the smallest error overall (across all groups) is printed in bold. The dagger (†) and double dagger (‡) symbols next to the best model denote statistically significant improvements over LR and the second best model, respectively.

regression on the 14 most popular UCI data sets that only contain categorical features having more than 30 rows [11] (varying licenses). Instances with multiple output labels are indicated with a star; we convert them into instances with binary output labels by distinguishing between the majority class vs. all other classes. The instances vary in the number N of data points, the number m of categorical features as well as, accordingly, the number k of slopes considered in the one-hot encoding of the categorical features (cf. Section 2.1). All results are reported as means over 100 random training set-test set splits (80%:20%). The radius $\epsilon \in \{0, 10^{-5}, \dots, 10^{-4}, \dots, 1\}$ of the Wasserstein ball as well as the Lasso penalty $\gamma \in \{0, \frac{1}{2} \cdot 10^{-5}, \dots, \frac{1}{2} \cdot 10^{-4}, \dots, \frac{1}{2}\}$ are selected via 5-fold cross-validation. We consider two variants of our DR logistic regression that employ a different output label weight ($\kappa = 1$ vs. $\kappa = m$) in the ground metric (cf. Definition 2). The results are reported in Table 1. (Run-times and are provided in Appendix D where we also describe how statistical significance was assessed.) The table shows that for the unregularized model, the classical logistic regression achieves the lowest classification error in 21% of the instances, whereas our DR logistic regression achieve the lowest classification error in 36% and 79% of the instances for $\kappa = 1, m$, respectively. (The double-counting of ties means these percentages do not sum to 100%.) For the regularized model, the results change to 14% (classical logistic regression) vs. 57% (each of our models). Within the numerical-feature Wasserstein DRO benchmarks, PI achieves the lowest classification error in 14% of the instances, whereas MT achieves in 42% and 57% of the instances for $\kappa = 1, m$, respectively. Globally, one of our methods is the winning approach in 12 out of 14 datasets (in the remaining 2 they are the second best approach without statistically significant inferiority). In 9 of the datasets our methods win strictly (without a tie), and in 7 of these datasets the improvements are statistically significant over all other approaches.

5 Conclusions

We proposed a new DR mixed-feature logistic regression model where the proximity between the empirical distribution and the unknown true data-generating distribution is measured by the popular Wasserstein distance. Despite its exponential-size formulation, we prove that the underlying optimization problem can be solved in polynomial time, and we develop a practically efficient column-and-constraint generation scheme for its solution. The promising performance of our model is demonstrated in numerical experiments on standard benchmark instances. We note that our column-and-constraint scheme readily extends to other DR mixed-feature machine learning models such as linear regression, support vector machines and decision trees. Applying our algorithm in those settings is a promising direction for future research. A further interesting direction is the development of tailored solution schemes rather than using off-the-shelf solvers for solving our DR mixed-feature logistic regression problem.

Acknowledgments and Disclosure of Funding

Funding from the EPSRC grant EP/W003317/1 is gratefully acknowledged. The first author acknowledges support from The Alan Turing Institute. The authors are grateful for the comments and suggestions of the anonymous reviewers, which have helped to significantly improve the manuscript. This work has been produced without the involvement of any commercial entities that may cause a conflict of interest or competing interests.

References

- [1] M. A. Abdullah, H. Ren, H. Bou-Ammar, V. Milenkovic, R. Luo, M. Zhang, and J. Wang. Wasserstein robust reinforcement learning. <http://arxiv.org/abs/1907.13196>.
- [2] M. Arjovsky, S. Chintala, and L. Bottou. Wasserstein generative adversarial networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70, 2017.
- [3] A. Ben-Tal, L. El Ghaoui, and A. Nemirovski. *Robust Optimization*. Princeton University Press, 2009.
- [4] D. Bertsimas and D. den Hertog. *Robust and Adaptive Optimization*. Dynamic Ideas, 2022.
- [5] J. Bezanson, A. Edelman, S. Karpinski, and V. B. Shah. Julia: A fresh approach to numerical computing. <https://doi.org/10.1137/141000671>.
- [6] C. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006.
- [7] Jose Blanchet, Yang Kang, and Karthyek Murthy. Robust Wasserstein profile inference and applications to machine learning. *Journal of Applied Probability*, 56(3):830–857, 2019. doi: 10.1017/jpr.2019.49.
- [8] S. Chakrabarti, H. Yiming, T. Li, S. Feizi, and X. Wu. Quantum Wasserstein generative adversarial networks. In *Advances in Neural Information Processing Systems*, volume 32, 2019.
- [9] E. Delage and Y. Ye. Distributionally robust optimization under moment uncertainty with application to data-driven problems. *Operations Research*, 58(3):596–612, 2010.
- [10] V. DeMiguel and F. J. Nogales. Portfolio selection with robust estimation. *Operations Research*, 57(3):560–577, 2009.
- [11] D. Dua and C. Graff. UCI machine learning repository. <http://archive.ics.uci.edu/ml>.
- [12] I. Dunning, J. Huchette, and M. Lubin. JuMP: A modeling language for mathematical optimization. *SIAM Review*, 59(2):295–320, 2017.
- [13] R. Flamary, M. Cuturi, N. Courty, and A. Rakotomamonjy. Wasserstein discriminant analysis. *Machine Learning*, 107(12):1923–1945, 2018.
- [14] C. Frogner, C. Zhang, H. Mobahi, M. Araya, and T. A. Poggio. Learning with a Wasserstein loss. In *Advances in Neural Information Processing Systems*, volume 28, 2015.
- [15] R. Gao and A. Kleywegt. Distributionally robust stochastic optimization with Wasserstein distance. <https://arxiv.org/abs/1604.02199>.
- [16] M. R. Garey and D. S. Johnson. *Computers and Intractability: A Guide to the Theory of NP-Completeness*. W. H. Freeman, 1979.
- [17] M. Grötschel, L. Lovász, and A. Schrijver. *Geometric Algorithms and Combinatorial Optimization*. Springer, 1988.
- [18] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical learning: Data mining, Inference, and Prediction*. Springer, 2009.
- [19] Z. Hu and L. J. Hong. Kullback-Leibler divergence constrained distributionally robust optimization. http://www.optimization-online.org/DB_FILE/2012/11/3677.pdf.

- [20] S. Kolouri, G. K. Rohde, and H. Hoffmann. Sliced Wasserstein distance for learning Gaussian mixture models. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.
- [21] D. Kuhn, P. Mohajerin Esfahani, V.A. Nguyen, and S. Shafieezadeh-Abadeh. Wasserstein distributionally robust optimization: Theory and applications in machine learning. *INFORMS TutORials in Operations Research*, pages 130–169, 2019.
- [22] H. Lam. Recovering best statistical guarantees via the empirical divergence-based distributionally robust optimization. *Operations Research*, 6(4):1090–1105, 2019.
- [23] R. O. Michaud. The Markowitz optimization enigma: Is ‘optimized’ optimal? *Financial Analysts Journal*, 45(1):31–42, 1989.
- [24] P. Mohajerin Esfahani and D. Kuhn. Data-driven distributionally robust optimization using the wasserstein metric: performance guarantees and tractable reformulations. *Math. Program.*, 171: 115—166, 2018.
- [25] P. Mohajerin Esfahani and D. Kuhn. Data-driven distributionally robust optimization using the Wasserstein metric: performance guarantees and tractable reformulations. *Mathematical Programming*, 171(1–2):1–52, 2018.
- [26] MOSEK ApS. The MOSEK optimization toolbox for MATLAB manual. Version 9.0. <http://docs.mosek.com/9.0/toolbox/index.html>, 2019.
- [27] K. P. Murphy. *Probabilistic Machine Learning: An Introduction*. MIT Press, 2022.
- [28] Y. Nesterov. *Lectures on Convex Optimization*. Springer, 2nd edition, 2018.
- [29] Steven L Salzberg. On comparing classifiers: Pitfalls to avoid and a recommended approach. *Data mining and knowledge discovery*, 1(3):317–328, 1997.
- [30] S. A. Serrano. *Algorithms for unsymmetric cone optimization and an implementation for problems with the exponential cone*. Stanford University, 2015.
- [31] S. Shafieezadeh-Abadeh, P. Mohajerin Esfahani, and D. Kuhn. Distributionally robust logistic regression. In *Advances in Neural Information Processing Systems*, volume 28, 2015.
- [32] S. Shafieezadeh-Abadeh, D. Kuhn, and P. Mohajerin Esfahani. Regularization via mass transportation. *Journal of Machine Learning Research*, 20(103):1–68, 2019.
- [33] A. Shapiro. On duality theory of conic linear problems. In *Semi-Infinite Programming*, pages 135–165. Kluwer Academic Publishers, 2001.
- [34] J. E. Smith and R. L. Winkler. The optimizer’s curse: Skepticism and postdecision surprise in decision analysis. *Management Science*, 52(3):311–322, 2006.
- [35] J. Solomon, R. Rustamov, L. Guibas, and A. Butscher. Wasserstein propagation for semi-supervised learning. In *Proceedings of the 31st International Conference on Machine Learning*, 2014.
- [36] Tobias Sutter, Andreas Krause, and Daniel Kuhn. Robust generalization despite distribution shift via minimum discriminating information. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 29754–29767. Curran Associates, Inc., 2021.
- [37] Bahar Taskesen, Man-Chung Yue, Jose Blanchet, Daniel Kuhn, and Viet Anh Nguyen. Sequential domain adaptation by synthesizing distributionally robust experts. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 10162–10172. PMLR, 18–24 Jul 2021.
- [38] J. Weed and Q. Berthet. Estimation of smooth densities in Wasserstein distance. In *Proceedings of the 32nd Conference on Learning Theory*, 2019.

- [39] W. Wiesemann, D. Kuhn, and M. Sim. Distributionally robust convex optimization. *Operations Research*, 62(6):1358–1376, 2014.
- [40] E. Wong, F. Schmidt, and Z. Kolter. Wasserstein adversarial examples via projected Sinkhorn iterations. In *Proceedings of the 36th International Conference on Machine Learning*, 2019.
- [41] H. Xu, C. Caramanis, and S. Mannor. Robustness and regularization of support vector machines. *Journal of Machine Learning Research*, 10(51):1485–1510, 2009.
- [42] M. Yue, D. Kuhn, and W. Wiesemann. On linear optimization over Wasserstein balls. *Mathematical Programming*, 2021.
- [43] Bo Zeng and Long Zhao. Solving two-stage robust optimization problems using a column-and-constraint generation method. *Operations Research Letters*, 41(5):457–461, 2013.
- [44] Jingyi Zhang, Ping Ma, Wenxuan Zhong, and Cheng Meng. Projection-based techniques for high-dimensional optimal transport problems. *WIREs Computational Statistics*, n/a(n/a):e1587.

Appendix A: Performance Guarantees

Wasserstein ambiguity sets benefit from measure concentration results that characterize the rate at which the empirical distribution $\hat{\mathbb{P}}_N$ converges to the unknown true distribution \mathbb{P}^0 . In the following, we review existing results from the literature to characterize the finite sample and asymptotic guarantees of our DR logistic regression (1).

Theorem 6 (Finite Sample Guarantee). *Assume that \mathbb{P}^0 is light-tailed, that is, $\mathbb{E}_{\mathbb{P}^0} [\exp(\|\xi\|^a)] \leq A$ for some $a > 1$ and $A > 0$. Then there are $c_1, c_2 > 0$ only depending on \mathbb{P}^0 through the light-tail parameters a, A and the feature space dimensions (n, m) such that any optimizer β^* to (1) satisfies*

$$[\mathbb{P}^0]^N \left(\mathbb{E}_{\mathbb{P}^0} [l_{\beta^*}(\mathbf{x}, \mathbf{z}, y)] \leq \sup_{\mathbb{Q} \in \mathfrak{B}_\epsilon(\hat{\mathbb{P}}_N)} \mathbb{E}_{\mathbb{Q}} [l_{\beta^*}(\mathbf{x}, \mathbf{z}, y)] \right) \geq 1 - \eta$$

for any confidence level $\eta \in (0, 1)$ and Wasserstein ball radius

$$\epsilon \geq \left(\frac{\log(c_1/\eta)}{c_2 N} \right)^{1/\max\{m+n+1, 2\}} \cdot \mathbb{1} \left[N \geq \frac{\log(c_1/\eta)}{c_2} \right] + \left(\frac{\log(c_1/\eta)}{c_2 N} \right)^{1/\alpha} \cdot \mathbb{1} \left[N < \frac{\log(c_1/\eta)}{c_2} \right].$$

Recall that $[\mathbb{P}^0]^N$ in the statement of Theorem 6 refers to the N -fold product distribution of \mathbb{P}^0 that governs the data set $\{\xi^i\}_{i \in [N]}$ upon which the optimizer(s) β^* of problem (1) depend(s) via $\hat{\mathbb{P}}_N$. Theorem 6 shows that with arbitrarily high probability $1 - \eta$, the optimal value $\sup_{\mathbb{Q} \in \mathfrak{B}_\epsilon(\hat{\mathbb{P}}_N)} \mathbb{E}_{\mathbb{Q}} [l_{\beta^*}(\mathbf{x}, \mathbf{z}, y)]$ of our DR logistic regression (1) overestimates the loss $\mathbb{E}_{\mathbb{P}^0} [l_{\beta^*}(\mathbf{x}, \mathbf{z}, y)]$ incurred by any optimal solution β^* under the unknown true distribution \mathbb{P}^0 as long as the radius ϵ of the Wasserstein ball $\mathfrak{B}_\epsilon(\hat{\mathbb{P}}_N)$ is sufficiently large. Since the categorical features attain finitely many different values, the bound of Theorem 6 can be sharpened by replacing $m + n + 1$ with $n + 1$ if the constants a and A are adapted accordingly. We emphasize that the decay rate of $\mathcal{O}(N^{-1/(n+1)})$ in Theorem 6 is essentially optimal; see [21, §3].

To study the asymptotic consistency of problem (1) as well as the existence of sparse worst-case distributions, we first introduce a technical assumption.

Definition 3 (Growth Condition). *We say that the DR logistic regression (1) satisfies the growth condition if (i) the hypotheses β are restricted to a bounded set $\mathcal{H} \subseteq \mathbb{R}^{1+n+k}$; and (ii) there is $\xi^0 \in \Xi$ and $C > 0$ such that $l_\beta(\xi) \leq C[1 + d(\xi, \xi^0)]$ across all $\beta \in \mathcal{H}$ and $\xi \in \Xi$.*

Lemma 1. *If we restrict the hypotheses β to a bounded set $\mathcal{H} \subseteq \mathbb{R}^{1+n+k}$, then the DR logistic regression (1) satisfies the growth condition of Definition 3.*

We are now in the position to study the asymptotic consistency of problem (1).

Theorem 7 (Asymptotic Consistency). *Under the assumptions of Theorem 6, we have*

$$\sup_{\mathbb{Q} \in \mathfrak{B}_{\epsilon_N}(\hat{\mathbb{P}}_N)} \mathbb{E}_{\mathbb{Q}} [l_{\beta^*}(\mathbf{x}, \mathbf{z}, y)] \xrightarrow{N \rightarrow \infty} \mathbb{E}_{\mathbb{P}^0} [l_{\beta^*}(\mathbf{x}, \mathbf{z}, y)] \quad \mathbb{P}^0\text{-a.s.}$$

whenever (η_N, ϵ_N) is set according to Theorem 6 for all $N \in \mathbb{N}$, $\sum_N \eta_N < \infty$, $\lim_{N \rightarrow \infty} \epsilon_N = 0$, and the growth condition in Definition 3 is satisfied.

Theorem 7 shows that the DR logistic regression (1) achieves asymptotic consistency if the (un-)confidence parameter η and the radius ϵ of the Wasserstein ball are reduced simultaneously. Thus, any optimal solution to (1) converges to the optimal solution of the (non-robust) logistic regression under the unknown true distribution \mathbb{P}^0 when the size of the data set increases.

The proof of Theorem 1 shows that the optimization problem characterizing the worst-case distribution $\mathbb{Q}^* \in \mathfrak{B}_\epsilon(\hat{\mathbb{P}}_N)$ comprises exponentially many decision variables. It is therefore natural to investigate the complexity of worst-case distributions to our DR logistic regression (1). The next result shows that there exist worst-case distributions that exhibit a desirable sparsity pattern: their numbers of atoms scale with the number of data samples.

Theorem 8 (Existence of Sparse Worst-Case Distributions). *Assume that the growth condition in Definition 3 is satisfied. Then there are worst-case distributions $\mathbb{Q}^* \in \mathfrak{B}_\epsilon(\hat{\mathbb{P}}_N)$ satisfying*

$$\mathbb{E}_{\mathbb{Q}^*} [l_{\beta^*}(\mathbf{x}, \mathbf{z}, y)] = \sup_{\mathbb{Q} \in \mathfrak{B}_\epsilon(\hat{\mathbb{P}}_N)} \mathbb{E}_{\mathbb{Q}} [l_{\beta^*}(\mathbf{x}, \mathbf{z}, y)]$$

such that \mathbb{Q}^* is supported on at most $N + 1$ atoms.

Our performance guarantees in this section scale with the dimension of the feature space as we seek for a high confidence of the unknown true distribution being contained in our ambiguity set $\mathfrak{B}_\epsilon(\hat{\mathbb{P}}_N)$. A dimension-independent performance guarantee can be obtained along the lines of [7] if one instead only seeks for a high confidence of the unknown true model β^* being contained in the union of optimal classifiers corresponding to the individual distributions \mathbb{P} contained in $\mathfrak{B}_\epsilon(\hat{\mathbb{P}}_N)$. We refer to [44] for a detailed review of the achievable performance guarantees of Wasserstein ambiguity sets with respect to the dimension of the feature space.

Appendix B: Finding the Worst Distribution in Example of Section 2.2

We provide further details here on the derivation of the worst distribution associated with the benchmark in Section 2.2 where we treat the (only) binary feature as a numerical feature and use the DR logistic regression algorithm of [31]. We have a single binary feature $z \in \{-1, 1\}$ and there is no intercept term so the log-loss function is $l_\beta(z, y) = \log(1 + \exp(-\beta \cdot y \cdot z))$. The true unknown value of β is $\beta = 1$, we have a data set $(z^i, y^i)_{i \in [N]}$, and we take $p = 1$, $\kappa = 1$ (distance metric parameters), and $\epsilon = 1/(2\sqrt{N})$. We would like to solve the worst distribution problem

$$\sup_{\mathbb{Q} \in \mathfrak{B}_\epsilon(\hat{\mathbb{P}}_N)} \mathbb{E}_{\mathbb{Q}} [l_\beta(z, y)]. \quad (6)$$

We can do this by considering the following problem taken from [32, Thm 20] and adopted for our specific setting:

$$\begin{aligned} & \underset{\theta, \{\alpha_i\}_{i \in [N]}}{\text{maximize}} && \theta + \frac{1}{N} \sum_{i=1}^N [(1 - \alpha_i)l_\beta(z^i, y^i) + \alpha_i l_\beta(z^i, -y^i)] \\ & \text{subject to} && \theta + \frac{1}{N} \sum_{i=1}^N \alpha_i = \epsilon - \gamma \\ & && 0 \leq \alpha_i \leq 1, \quad i \in [N] \\ & && \theta \geq 0. \end{aligned} \quad (7)$$

It is parametrized by some $\gamma \in [0, \min\{\epsilon, 1\}]$ and [32, Thm 20] show that its optimal value for the case $\gamma = 0$ coincides with the optimal value of problem (6). Furthermore, if we denote by $(\theta^*(\gamma), \{\alpha_i^*(\gamma)\}_{i \in [N]})$ the optimal solutions to (7), then the sequence of probability distributions

$$\begin{aligned} \mathbb{Q}_\gamma = & \frac{1}{N} \sum_{i=2}^N [(1 - \alpha_i^*(\gamma))\delta_{(z^i, y^i)} + \alpha_i^*(\gamma)\delta_{(z^i, -y^i)}] + \frac{\eta(\gamma)}{N} \delta_{(z^1 + \frac{\theta^*(\gamma)N}{\eta(\gamma)}, y^1)} \\ & + \frac{1 - \eta(\gamma)}{N} [(1 - \alpha_1^*(\gamma))\delta_{(z^1, y^1)} + \alpha_1^*(\gamma)\delta_{(z^1, -y^1)}] \end{aligned} \quad (8)$$

where $\eta(\gamma) := \gamma/(\theta^*(\gamma) + 2 - \epsilon + \gamma)$, constructs an asymptotically optimal solution to problem (6) as $\gamma \downarrow 0$. In order to explicitly characterize this sequence, we derive a closed-form solution to problem (7). Using the equality constraint to substitute for θ in the objective in (8) yields

$$\begin{aligned} & \underset{\{\alpha_i\}_{i \in [N]}}{\text{maximize}} && \epsilon - \gamma - \frac{1}{N} \sum_{i=1}^N \alpha_i + \frac{1}{N} \sum_{i=1}^N [(1 - \alpha_i)l_\beta(z^i, y^i) + \alpha_i l_\beta(z^i, -y^i)] \\ & \text{subject to} && 0 \leq \alpha_i \leq 1, \quad i \in [N] \\ & && \epsilon - \gamma \geq \frac{1}{N} \sum_{i=1}^N \alpha_i. \end{aligned}$$

Ignoring the constant terms in the objective and re-arranging terms yields

$$\begin{aligned}
& \underset{\{\alpha_i\}_{i \in [N]}}{\text{maximize}} && \sum_{i=1}^N \alpha_i [-1 + l_\beta(z^i, -y^i) - l_\beta(z^i, y^i)] \\
& \text{subject to} && 0 \leq \alpha_i \leq 1, \quad i \in [N] \\
& && \epsilon - \gamma \geq \frac{1}{N} \sum_{i=1}^N \alpha_i.
\end{aligned}$$

Since $z^i \in \{-1, +1\}$, $i \in [N]$ holds, we have $l_\beta(z^i, -y^i) - l_\beta(z^i, y^i) \in \{-1, +1\}$, which implies that the coefficients of α_i in the objective function are non-positive. Hence, $\alpha_i^* = 0$, $i \in [N]$ is an optimal solution to the problem. This implies $\alpha_i^*(\gamma) = 0$, $i \in [N]$ and $\theta^*(\gamma) = \epsilon - \gamma$ are optimal in problem (7).

We can then observe from (8) that the worst distribution places $1/N$ mass on each data point $i = 2, \dots, N$, and the remaining $1/N$ mass is distributed as: $(1 - \eta(\gamma))/N$ mass on the data point $i = 1$ and $\eta(\gamma)/N$ mass on the point $(z^1 + \frac{(\epsilon - \gamma)N}{\gamma/2}, y^1)$ which is **not** in the data-set and in fact is an infeasible point. In our experiments, we take $\gamma = 10^{-3}$, because the optimal value of problem (7) numerically converges after this value, and one can then verify that in the setting we work with ($N = 250$) a mass of $\frac{10^{-3}}{500}$ is placed on the point with feature $z^1 + \frac{(\epsilon - \gamma)N}{\gamma/2} = 1 + \frac{(\frac{1}{1\sqrt{N}} - 10^{-3})N}{10^{-3/2}} \approx 15,312$ and label y^1 . This summarizes the specific approach we took to obtain the worst distribution in Figure 1 and demonstrates the key problem that arises with treating categorical variables as numerical.

Appendix C: Numerical Results on UCI Data-Sets with Mixed Features

We repeat the experiment of the Section 4.2 for the five most popular mixed-feature instances of the UCI data set [11]. The results are reported in Table 2. All results are reported as medians over 20 random training set-test set splits (80%:20%). Cross-validation is applied for the same set of parameters as Section 4.2 over identical grids. The conclusions are qualitatively similar to those of Section 4.2. We highlight, however, that now only two of the lowest classification errors are achieved by one of the non-robust models, whereas the lowest classification error in each instance is obtained by at least one of the robust models we propose.

Data Set	N	n	k	m	LR	DRO ($\kappa = 1$)	DRO ($\kappa = m$)	r-LR	r-DRO ($\kappa = 1$)	r-DRO ($\kappa = m$)	MT ($\kappa = 1$)	MT ($\kappa = m$)	PI ($\alpha = 0.05$)
credit-approval	690	6	36	9	14.13%	13.77%	13.04% ^{†‡}	14.13%	14.13%	13.04% ^{†‡}	14.13%	14.13%	13.41%
annealing*	798	6	46	32	2.52%	2.83%	2.52%	2.20%	1.89% ^{†‡}	1.89% ^{†‡}	2.52%	2.20%	12.58%
contraceptive*	1,473	2	15	7	33.16%	33.16%	33.16%	32.82% [†]	32.99%	32.82% [†]	33.16%	33.16%	39.63%
hepatite	155	6	23	13	16.13%	16.13%	19.35%	19.35%	17.74%	16.13%	16.13%	16.13%	17.74%
cylinder-bands	539	19	43	15	23.36%	21.50% ^{†‡}	21.50% ^{†‡}	23.36%	21.50% ^{†‡}	22.43%	22.43%	22.43%	30.84%

Table 2: Classification errors of unregularized and Lasso-regularized variants of the classical logistic regression and our DR regression on UCI benchmark instances with mixed features. We use the same notation and highlighting conventions as in Table 1.

Data Set	N	n	k	m	LR	DRO ($\kappa = 1$)	DRO ($\kappa = m$)	r-LR	r-DRO ($\kappa = 1$)	r-DRO ($\kappa = m$)	MT ($\kappa = 1$)	MT ($\kappa = m$)	PI ($\alpha = 0.05$)
credit-approval	690	6	36	9	0.11	53.46	55.89	0.09	36.77	43.58	0.36	0.56	0.07 + 0.08
annealing*	798	6	46	32	0.18	62.96	71.09	43.58	22.91	0.24	0.36	0.58	0.10 + 0.12
contraceptive*	1,473	2	15	7	0.14	86.06	81.70	0.15	91.10	84.04	0.52	0.74	0.08 + 0.03
hepatite	155	6	23	13	0.05	9.01	10.59	0.02	2.49	4.86	0.06	0.10	0.01 + 0.05
cylinder-bands	539	19	43	15	0.13	96.10	106.48	0.09	67.09	41.63	0.32	0.59	0.06 + 0.12

Table 3: Median runtimes (in seconds) associated with Table 2.

Appendix D: Further Details on Numerical Experiments

Throughout the numerical experiments, we fixed $p = 1$ in the ground metric (cf. Definition 2).

Synthetic data sets: In Section 4.1 we generate synthetic data sets with N data points and m binary features. The data generation process is summarized in Algorithm 3.

Algorithm 3 Construction of synthetic data sets in Section 4.1.

Sample the components of β_0 and β_C i.i.d. from a standard normal distribution
 Normalize β_0 and β_C by dividing them with $\|\beta\|_2$, where $\beta = (\beta_0, \beta_C)$
for $i \in \{1, \dots, N\}$ **do**
 Construct $\mathbf{z}^i \in \mathbb{C}(2) \times \dots \times \mathbb{C}(2)$ by sampling $\mathbf{z}^i \sim \{0, 1\}^m$ uniformly at random
 Find p^i , the probability of the i th data point having label $+1$, by using the ‘true’ β_0 and β_C :

$$p^i = [1 + \exp(-[\beta_0 + \beta_C^\top \mathbf{z}^i])]^{-1}$$

 Sample $y^i \in \{-1, 1\}$ from a Bernoulli distribution with parameter p^i
end for
 The synthetic data set is the collection of data points $\{(\mathbf{z}^i, y^i)\}_{i \in [N]}$ constructed above

UCI data sets: The datasets of Sections 4.2 and Appendix C were taken from the UCI repository [11]. Missing values (NaNs) were encoded as a new category of the corresponding feature; the dataset breast-cancer is an exception, where rows with missing values were dropped. Some data sets include features that are derivatives of the labels, primary keys of the instances (*e.g.*, the audiology data set), or they have features with only one possible category (*e.g.*, the annealing and cylinder-bands data sets). We removed such features manually. We also removed rows that were readily identified as erroneous (*e.g.*, two rows in cylinder-bands had less columns than required). Data sets with multi-class (*i.e.*, non-binary) labels were converted to binary labels by distinguishing between the majority label class and all other classes. If a data set includes separate training and test sets, we merged them and subsequently applied our training set-test set split as described in the main paper. As per standard practice, we randomly permuted the rows of each data set before conducting the splits into training and test sets. For the detailed data processing steps, we refer to the GitHub repository accompanying this paper, which also contains a sample Python 3 script.¹

Implementation of the column-and-constraint generation scheme: In our implementation, we switch between solving the primal and dual exponential conic reformulation of the DR logistic regression problem, as we found that sometimes the dual problem can be solved more easily than the primal. We have further implemented an ‘easing’ step in the column-and-constraint generation scheme that periodically deletes constraints from \mathcal{W}^+ and \mathcal{W}^- whose slacks exceed a pre-specified threshold. To this end, note that the constraints $u_{i,\mathbf{z}}^+ + v_{i,\mathbf{z}}^+ \leq 1$, $(i, \mathbf{z}) \in \mathcal{W}^+$, in the relaxations of the exponential conic reformulation (4) have a slack ranging between 0 and 1 by construction. We implemented variants of our column-and-constraint generation scheme that conduct easing steps either every 200 iterations or in iteration $t = 100 \cdot 1.5^k$, $k \in \mathbb{N}$. We have also implemented variants that keep the slack threshold constant at 0.05 and where this threshold starts at 0.02 and is subsequently increased by 0.02 in each easing step. In all of our variants, a constraint is deleted at most once, that is, it is no longer considered for deletion if it has been reintroduced after a prior deletion. This ensures that Algorithm 1 terminates in finite time without cycling. Analogous steps have been implemented for the constraints $u_{i,\mathbf{z}}^- + v_{i,\mathbf{z}}^- \leq 1$, $(i, \mathbf{z}) \in \mathcal{W}^-$.

Determining statistical Significance: In the numerical experiments of Section 4.2 we compare the means of 100 out-of-sample errors attained by several logistic regression methods and also identify which methods appeared to be statistically significant. In Table 1, for example, a dagger (†) symbol next to the winning approach denotes statistically significant error improvement over the standard logistic regression (‘LR’) and a double dagger (‡) symbol denotes such improvement over the second best approach. Note that if the winning approach is a variant of the DRO methods we propose, then the second best approach is taken over the methods excluding our methods for a fair comparison.

We add a dagger (†) to the winning approach according to the following approach. Firstly, we subtract (element-wise) the vector of errors attained by standard logistic regression from the vector of errors attained by the winning approach. Each element of the new vector is the ‘additional error’ the winning approach has compared to the standard logistic regression. We then try to reject the hypothesis (at 5%-significance level) that this additional error is non-negative (*i.e.*, we try to reject

¹<https://github.com/selvi-aras/WassersteinLR>

Data Set	N	k	m	LR	DRO ($\kappa = 1$)	DRO ($\kappa = m$)	r-LR	r-DRO ($\kappa = 1$)	r-DRO ($\kappa = m$)	MT ($\kappa = 1$)	MT ($\kappa = m$)	PI ($\alpha = 0.05$)
<u>breast-cancer</u>	277	42	9	0.06	218.61	214.83	0.07	15.19	20.52	0.08	0.09	0.02 + 0.07
<u>spect</u>	267	22	22	0.08	32.05	38.95	0.07	20.81	35.50	0.07	0.09	0.02 + 0.03
<u>monks-3</u>	554	11	6	0.12	46.83	42.59	0.08	31.04	42.53	0.12	0.14	0.03 + 0.02
<u>tic-tac-toe</u>	958	18	9	0.20	71.26	69.80	0.13	61.88	64.33	0.23	0.27	0.06 + 0.04
<u>kr-vs-kp</u>	3,196	37	36	1.65	217.98	40.64	0.71	21.79	11.29	2.11	2.87	0.34 + 0.08
<u>balance-scale</u> *	625	16	4	0.15	13.60	17.41	0.19	19.74	24.04	0.21	0.19	0.03 + 0.03
<u>hayes-roth</u> *	160	11	4	0.06	3.67	3.92	0.02	1.12	0.85	0.06	0.06	0.01 + 0.02
<u>lymphography</u> *	148	42	18	0.09	145.46	154.69	0.03	4.83	11.91	0.08	0.09	0.01 + 0.07
<u>car</u> *	1,728	15	6	0.46	168.09	112.33	0.28	210.35	181.63	0.57	0.52	0.11 + 0.05
<u>splices</u> *	3,189	229	60	4.25	4,246.36	4,393.97	1.48	1,169.80	453.73	6.13	10.21	1.55 + 1.77
<u>house-votes-84</u>	435	32	16	0.18	88.56	105.66	0.06	21.6	20.71	0.12	0.16	0.03 + 0.08
<u>hiv</u>	6,590	152	8	8.44	2,874.62	1,598.83	2.20	2,119.26	768.01	10.94	12.93	2.18 + 1.14
<u>primacy-tumor</u> *	339	25	17	0.22	61.75	57.81	0.06	11.93	13.21	0.08	0.12	0.02 + 0.04
<u>audiology</u> *	226	92	69	0.05	18.52	16.10	0.04	9.19	5.28	0.11	0.13	0.04 + 0.16

Table 4: Mean runtimes (in seconds) associated with Table 1.

the hypothesis that the standard logistic regression is at least as good). To this end, we compute the t-statistic for the mean of the additional errors vector with a hypothesis mean of 0 and sample size of 100. We then compute the cumulative probability of this value via a one-sided t-test (with $100 - 1$ degrees of freedom) to obtain a p-value. If this value is less than 0.05, then we reject the hypothesis, concluding that the improvement is significant. Note that, with this approach we implicitly assume the out-of-sample errors are independent, which is typically not the case [29]. Hence, we acknowledge that these tests of significance are only approximate. The presence of a double dagger is determined analogously.

Computing environment: We implemented all algorithms in Julia using MOSEK’s exponential cone solver as well as JuMP to interact with the solver. We used the high performance computing cluster of Imperial College London, which runs a Linux operating system as well as Portable Batch System (PBS) for scheduling the jobs. We ran our experiments as batch jobs on Intel Xeon 2.66GHz processors with 8GB memory in single-core and single-thread mode. The job descriptions as well as all PBS commands are included in the GitHub repository.

Runtimes corresponding to Table 1: Table 4 presents the mean runtimes of each method on every dataset for the experiments corresponding to Table 1. Here, the times reported for ‘LR’ and ‘MT’ are the times the solver took to solve the corresponding optimization problems. The times corresponding to ‘DRO’ methods we propose are the total solver times summed for each sub-problem solved during the column-and-constraint generation scheme, including the identification of the most violated constraints. The times corresponding to ‘PI’ display the solution time to solve the regularized logistic regression problems plus the time it takes to identify the regularization parameter as proposed in [7]. The columns of Table 4 are identical to those in Table 1.

Error bars of the experiments: Figures 3 and 4 report the error distributions corresponding to the experiments of Tables 1 and 2, respectively. In these figures, each of the six sub-plots reports the errors of a specific model (*e.g.*, regularized DRO with $\kappa = 1$). In each sub-plot, the horizontal axis lists the considered data sets, and the vertical axis visualizes the 100 test errors in box-and-whisker representation (where the boxes enclose the 25% and 75% quartiles).

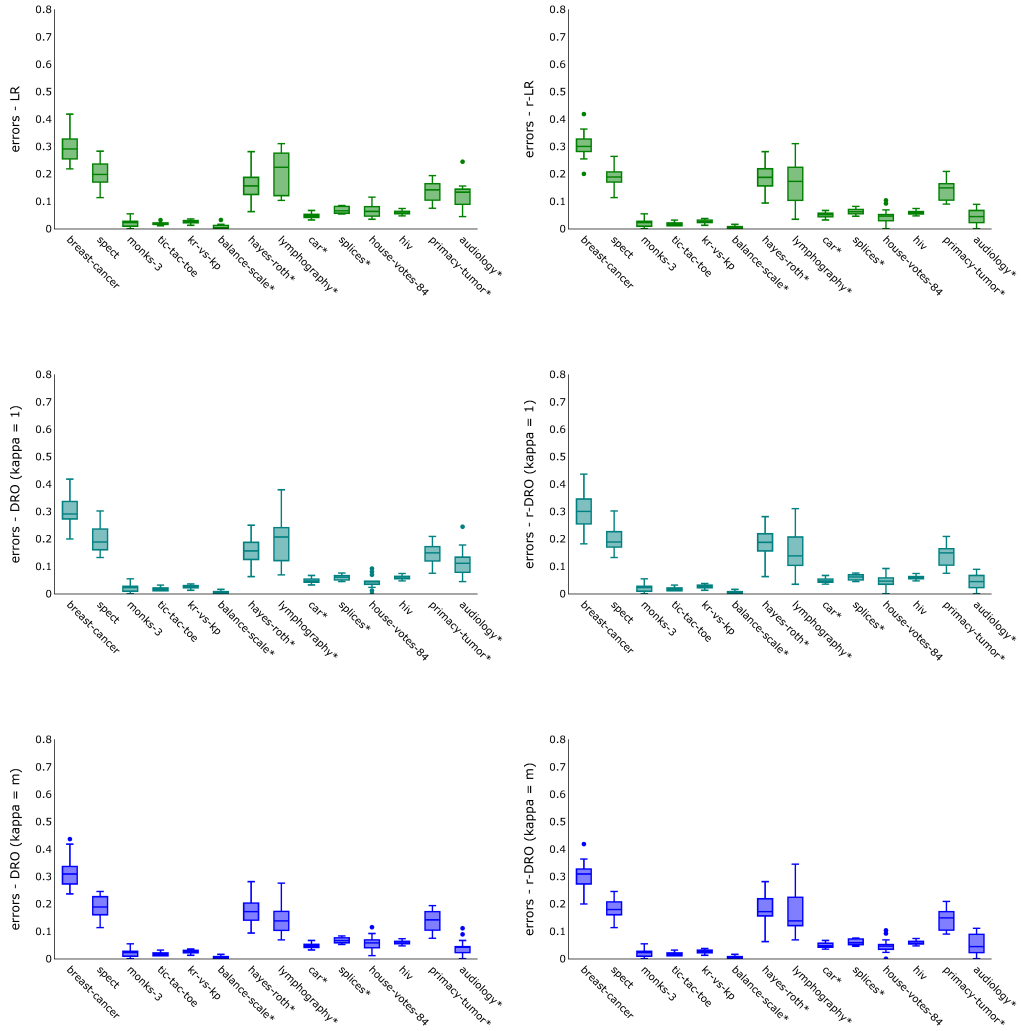


Figure 3: Error bars for the unregularized (left column) and regularized (right column) methods on the considered UCI datasets with categorical features.

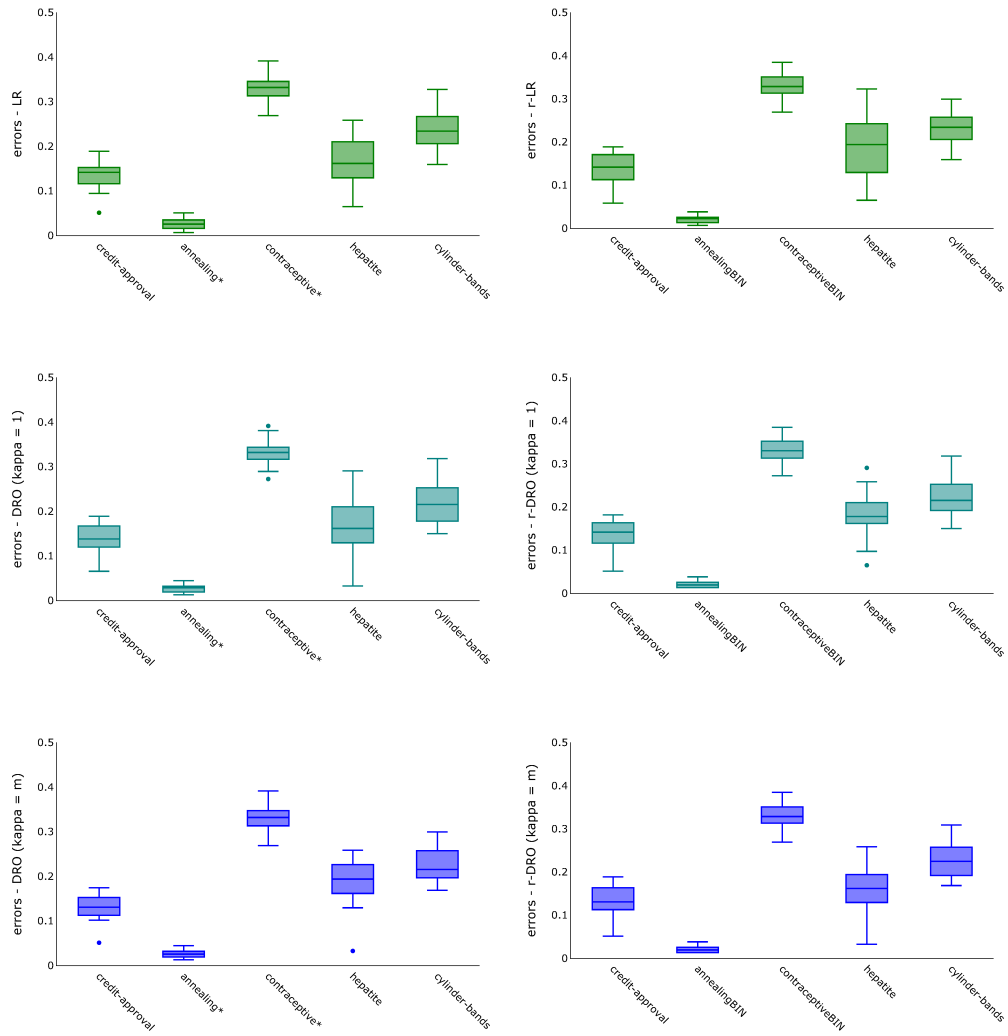


Figure 4: Error bars for the unregularized (left column) and regularized (right column) methods on the considered UCI datasets with mixed features.

Appendix E: Proofs

Our proof of Theorem 1 relies on the following result from the literature, which we state first.

Lemma 2. Consider the convex function $h_\gamma(\mathbf{w}) := \log(1 + \exp(-\gamma^\top \mathbf{w} + v))$ for $\gamma, \mathbf{w} \in \mathbb{R}^n$ and $v \in \mathbb{R}$. Then

$$\sup_{\mathbf{w} \in \mathbb{R}^n} h_\gamma(\mathbf{w}) - \lambda \|\hat{\mathbf{w}} - \mathbf{w}\| = \begin{cases} h_\gamma(\hat{\mathbf{w}}) & \text{if } \|\gamma\|_* \leq \lambda, \\ +\infty & \text{otherwise} \end{cases}$$

for every $\lambda > 0$, where $\|\cdot\|_*$ is the dual norm of $\|\cdot\|$, that is, $\|\gamma\|_* := \sup_{\|\mathbf{w}\| \leq 1} \gamma^\top \mathbf{w}$.

Proof. The statement immediately follows from Lemma 47 of [32]. \square

Our proof of Theorem 1 follows [31] with two main changes: (i) we handle the categorical features in \mathbf{z} , and (ii) we reformulate the optimization problem as an exponential conic program. Despite the similarities to the existing result, we include the entire proof to keep the paper self-contained.

Proof of Theorem 1: Recall that $\boldsymbol{\xi} = (\mathbf{x}, \mathbf{z}, y) \in \Xi = \mathbb{R}^n \times \mathbb{C} \times \{-1, +1\}$. To simplify our notation, we will use $\boldsymbol{\xi}$ and $(\mathbf{x}, \mathbf{z}, y)$ interchangeably. The inner problem in (1) can be written as

$$\left[\begin{array}{ll} \underset{\mathbb{Q}}{\text{maximize}} & \mathbb{E}_{\mathbb{Q}}[l_\beta(\mathbf{x}, \mathbf{z}, y)] \\ \text{subject to} & \mathbb{Q} \in \mathfrak{B}_\epsilon(\hat{\mathbb{P}}_N) \end{array} \right] = \left[\begin{array}{ll} \underset{\mathbb{Q}}{\text{maximize}} & \int_{\boldsymbol{\xi} \in \Xi} l_\beta(\boldsymbol{\xi}) \mathbb{Q}(\mathrm{d}\boldsymbol{\xi}) \\ \text{subject to} & \mathbb{Q} \in \mathfrak{B}_\epsilon(\hat{\mathbb{P}}_N) \end{array} \right],$$

and replacing $\mathfrak{B}_\epsilon(\hat{\mathbb{P}}_N)$ with its definition (cf. Definition 1) yields

$$\begin{aligned} & \underset{\mathbb{Q}, \Pi}{\text{maximize}} && \int_{\boldsymbol{\xi} \in \Xi} l_\beta(\boldsymbol{\xi}) \mathbb{Q}(\mathrm{d}\boldsymbol{\xi}) \\ & \text{subject to} && \int_{(\boldsymbol{\xi}, \boldsymbol{\xi}') \in \Xi^2} d(\boldsymbol{\xi}, \boldsymbol{\xi}') \Pi(\mathrm{d}\boldsymbol{\xi}, \mathrm{d}\boldsymbol{\xi}') \leq \epsilon \\ & && \int_{\boldsymbol{\xi} \in \Xi} \Pi(\mathrm{d}\boldsymbol{\xi}, \mathrm{d}\boldsymbol{\xi}') = \hat{\mathbb{P}}_N(\mathrm{d}\boldsymbol{\xi}') \quad \forall \boldsymbol{\xi}' \in \Xi \\ & && \int_{\boldsymbol{\xi}' \in \Xi} \Pi(\mathrm{d}\boldsymbol{\xi}, \mathrm{d}\boldsymbol{\xi}') = \mathbb{Q}(\mathrm{d}\boldsymbol{\xi}) \quad \forall \boldsymbol{\xi} \in \Xi \\ & && \mathbb{Q} \in \mathcal{P}_0(\Xi), \quad \Pi \in \mathcal{P}_0(\Xi^2). \end{aligned}$$

We can substitute \mathbb{Q} using the second equality constraint to obtain

$$\begin{aligned} & \underset{\Pi}{\text{maximize}} && \int_{\boldsymbol{\xi} \in \Xi} l_\beta(\boldsymbol{\xi}) \int_{\boldsymbol{\xi}' \in \Xi} \Pi(\mathrm{d}\boldsymbol{\xi}, \mathrm{d}\boldsymbol{\xi}') \\ & \text{subject to} && \int_{(\boldsymbol{\xi}, \boldsymbol{\xi}') \in \Xi^2} d(\boldsymbol{\xi}, \boldsymbol{\xi}') \Pi(\mathrm{d}\boldsymbol{\xi}, \mathrm{d}\boldsymbol{\xi}') \leq \epsilon \\ & && \int_{\boldsymbol{\xi} \in \Xi} \Pi(\mathrm{d}\boldsymbol{\xi}, \mathrm{d}\boldsymbol{\xi}') = \hat{\mathbb{P}}_N(\mathrm{d}\boldsymbol{\xi}') \quad \forall \boldsymbol{\xi}' \in \Xi \\ & && \Pi \in \mathcal{P}_0(\Xi^2). \end{aligned}$$

Denoting by $\mathbb{Q}^i(\mathrm{d}\boldsymbol{\xi}) := \Pi(\mathrm{d}\boldsymbol{\xi} | \boldsymbol{\xi}^i)$ the conditional distribution of Π upon the realization of $\boldsymbol{\xi}' = \boldsymbol{\xi}^i$ and exploiting the fact that $\hat{\mathbb{P}}_N$ is a discrete distribution supported on the N atoms $\{\boldsymbol{\xi}^i\}_{i \in [N]}$, we can use the marginalized representation $\Pi(\mathrm{d}\boldsymbol{\xi}, \mathrm{d}\boldsymbol{\xi}') = \frac{1}{N} \sum_{i=1}^N \delta_{\boldsymbol{\xi}^i}(\mathrm{d}\boldsymbol{\xi}') \mathbb{Q}^i(\mathrm{d}\boldsymbol{\xi})$ to obtain the equivalent reformulation

$$\begin{aligned} & \underset{\mathbb{Q}^i}{\text{maximize}} && \frac{1}{N} \sum_{i=1}^N \int_{\boldsymbol{\xi} \in \Xi} l_\beta(\boldsymbol{\xi}) \mathbb{Q}^i(\mathrm{d}\boldsymbol{\xi}) \\ & \text{subject to} && \frac{1}{N} \sum_{i=1}^N \int_{\boldsymbol{\xi} \in \Xi} d(\boldsymbol{\xi}, \boldsymbol{\xi}^i) \mathbb{Q}^i(\mathrm{d}\boldsymbol{\xi}) \leq \epsilon \\ & && \mathbb{Q}^i \in \mathcal{P}_0(\Xi), \quad i \in [N]. \end{aligned}$$

We can now decompose the N variables (distributions) \mathbb{Q}^i into $2N \cdot \prod_j k_j$ measures $\mathbb{Q}_{\mathbf{z},y}^i$:

$$\begin{aligned} \underset{\mathbb{Q}_{\mathbf{z},y}^i}{\text{maximize}} \quad & \frac{1}{N} \sum_{i=1}^N \sum_{(\mathbf{z},y) \in \mathbb{C} \times \{-1,+1\}} \int_{\mathbf{x} \in \mathbb{R}^n} l_{\beta}(\xi) \mathbb{Q}_{\mathbf{z},y}^i(d\mathbf{x}) \\ \text{subject to} \quad & \frac{1}{N} \sum_{i=1}^N \sum_{(\mathbf{z},y) \in \mathbb{C} \times \{-1,+1\}} \int_{\mathbf{x} \in \mathbb{R}^n} d(\xi, \xi^i) \mathbb{Q}_{\mathbf{z},y}^i(d\mathbf{x}) \leq \epsilon \\ & \sum_{(\mathbf{z},y) \in \mathbb{C} \times \{-1,+1\}} \int_{\mathbf{x} \in \mathbb{R}^n} \mathbb{Q}_{\mathbf{z},y}^i(d\mathbf{x}) = 1 \quad \forall i \in [N] \\ & \mathbb{Q}_{\mathbf{z},y}^i \in \mathcal{M}_+(\mathbb{R}^n), i \in [N] \text{ and } (\mathbf{z}, y) \in \mathbb{C} \times \{-1, +1\}, \end{aligned}$$

where $\mathcal{M}_+(\mathbb{R}^n)$ denotes the space of non-negative measures supported on \mathbb{R}^n . This infinite-dimensional linear program admits the dual

$$\begin{aligned} \underset{\lambda, \mathbf{u}}{\text{minimize}} \quad & \lambda\epsilon + \sum_{i=1}^N u_i \\ \text{subject to} \quad & \frac{1}{N} \cdot \sup_{\mathbf{x} \in \mathbb{R}^n} \{l_{\beta}(\mathbf{x}, \mathbf{z}, y) - \lambda d((\mathbf{x}, \mathbf{z}, y), \xi^i)\} \leq u_i \quad \forall i \in [N] \\ & \forall (\mathbf{z}, y) \in \mathbb{C} \times \{-1, +1\} \\ & \lambda \geq 0, \quad \mathbf{u} \in \mathbb{R}^N. \end{aligned}$$

Strong duality holds for any $\epsilon > 0$ due to Proposition 3.4 of [33]. We rewrite the dual problem further by substituting $s_i := Nu_i$, using Definition 2 to write out d explicitly, as well as breaking down the dual constraints into separate cases for $y = -1$ and $y = +1$:

$$\begin{aligned} \underset{\lambda, \mathbf{s}}{\text{minimize}} \quad & \lambda\epsilon + \frac{1}{N} \sum_{i=1}^N s_i \\ \text{subject to} \quad & \sup_{\mathbf{x} \in \mathbb{R}^n} \{l_{\beta}(\mathbf{x}, \mathbf{z}, +1) - \lambda \|\mathbf{x} - \mathbf{x}^i\|\} - \lambda\kappa \cdot \mathbb{1}[y^i \neq 1] - \lambda d_{\mathbb{C}}(\mathbf{z}, \mathbf{z}^i) \leq s_i \quad \forall i \in [N], \mathbf{z} \in \mathbb{C} \\ & \sup_{\mathbf{x} \in \mathbb{R}^n} \{l_{\beta}(\mathbf{x}, \mathbf{z}, -1) - \lambda \|\mathbf{x} - \mathbf{x}^i\|\} - \lambda\kappa \cdot \mathbb{1}[y^i \neq -1] - \lambda d_{\mathbb{C}}(\mathbf{z}, \mathbf{z}^i) \leq s_i \quad \forall i \in [N], \mathbf{z} \in \mathbb{C} \\ & \lambda \geq 0, \quad \mathbf{s} \in \mathbb{R}^N \end{aligned}$$

Applying Lemma 2 to the suprema of the above problem results in the reformulation

$$\begin{aligned} \underset{\lambda, \mathbf{s}}{\text{minimize}} \quad & \lambda\epsilon + \frac{1}{N} \sum_{i=1}^N s_i \\ \text{subject to} \quad & l_{\beta}(\mathbf{x}^i, \mathbf{z}, +1) - \lambda\kappa \cdot \mathbb{1}[y^i \neq 1] - \lambda d_{\mathbb{C}}(\mathbf{z}, \mathbf{z}^i) \leq s_i \quad \forall i \in [N], \forall \mathbf{z} \in \mathbb{C} \\ & l_{\beta}(\mathbf{x}^i, \mathbf{z}, -1) - \lambda\kappa \cdot \mathbb{1}[y^i \neq -1] - \lambda d_{\mathbb{C}}(\mathbf{z}, \mathbf{z}^i) \leq s_i \quad \forall i \in [N], \forall \mathbf{z} \in \mathbb{C} \\ & \|\beta_{\mathbf{N}}\|_* \leq \lambda \\ & \lambda \geq 0, \quad \mathbf{s} \in \mathbb{R}^N. \end{aligned}$$

For $y^i = +1$, the identities

$$\begin{aligned} l_{\beta}(\mathbf{x}^i, \mathbf{z}, +1) - \lambda\kappa \cdot \mathbb{1}[y^i \neq 1] &= l_{\beta}(\mathbf{x}^i, \mathbf{z}, y^i) \\ l_{\beta}(\mathbf{x}^i, \mathbf{z}, -1) - \lambda\kappa \cdot \mathbb{1}[y^i \neq -1] &= l_{\beta}(\mathbf{x}^i, \mathbf{z}, -y^i) - \lambda\kappa \end{aligned}$$

hold; similar identities hold for $y^i = -1$. The above optimization problem thus simplifies to

$$\begin{aligned} \underset{\lambda, \mathbf{s}}{\text{minimize}} \quad & \lambda\epsilon + \frac{1}{N} \sum_{i=1}^N s_i \\ \text{subject to} \quad & l_{\beta}(\mathbf{x}^i, \mathbf{z}, y^i) - \lambda d_{\mathbb{C}}(\mathbf{z}, \mathbf{z}^i) \leq s_i \quad \forall i \in [N], \forall \mathbf{z} \in \mathbb{C} \\ & l_{\beta}(\mathbf{x}^i, \mathbf{z}, -y^i) - \lambda\kappa - \lambda d_{\mathbb{C}}(\mathbf{z}, \mathbf{z}^i) \leq s_i \quad \forall i \in [N], \forall \mathbf{z} \in \mathbb{C} \\ & \|\beta_{\mathbf{N}}\|_* \leq \lambda \\ & \lambda \geq 0, \quad \mathbf{s} \in \mathbb{R}^N. \end{aligned} \tag{9}$$

Plugging this problem into the overall optimization problem over $\beta \in \mathbb{R}^{1+n+k}$ therefore yields

$$\begin{aligned}
& \underset{\beta, \lambda, s}{\text{minimize}} && \lambda\epsilon + \frac{1}{N} \sum_{i=1}^N s_i \\
& \text{subject to} && l_\beta(\mathbf{x}^i, \mathbf{z}, y^i) - \lambda d_C(\mathbf{z}, \mathbf{z}^i) \leq s_i && \forall i \in [N], \forall \mathbf{z} \in \mathbb{C} \\
& && l_\beta(\mathbf{x}^i, \mathbf{z}, -y^i) - \lambda\kappa - \lambda d_C(\mathbf{z}, \mathbf{z}^i) \leq s_i && \forall i \in [N], \forall \mathbf{z} \in \mathbb{C} \\
& && \|\beta_N\|_* \leq \lambda \\
& && \beta = (\beta_0, \beta_N, \beta_C) \in \mathbb{R}^{1+n+k}, \quad \lambda \geq 0, \quad s \in \mathbb{R}^N.
\end{aligned} \tag{10}$$

Finally, we discuss how to reformulate the *softplus constraints* (i.e., convex constraints with a log-loss function on the left-hand side and a linear function on the right-hand side) as exponential cone constraints. To this end, recall that the *exponential cone* is defined as

$$\mathcal{K}_{\text{exp}} := \text{cl}(\{(a, b, c) : a \geq b \exp(c/b), a > 0, b > 0\}) \subset \mathbb{R}^3,$$

where $\text{cl}(\cdot)$ denotes the closure. Observe further that

$$\log(1 + \exp(c)) \leq a \iff 1 + \exp(c) \leq \exp(a) \iff \exp(-a) + \exp(c - a) \leq 1.$$

Using the auxiliary variables u, v , we can reformulate this constraint as

$$\begin{bmatrix} u + v \leq 1 \\ \exp(-a) \leq u \\ \exp(c - a) \leq v \end{bmatrix} \iff \begin{bmatrix} u + v \leq 1 \\ (u, 1, -a) \in \mathcal{K}_{\text{exp}} \\ (v, 1, c - a) \in \mathcal{K}_{\text{exp}} \end{bmatrix},$$

where the second system of equations uses the definition of \mathcal{K}_{exp} . Applying this reformulation to both softplus constraints in our minimization problem results in problem (4) and thus concludes the proof of the theorem. \square

Remark 1 (Exponential Cone Reformulation). *The use of exponential conic constraints has a significant impact on the theoretical complexity of problem (4). We refer to [30] for an overview of exponential conic programs and to [26] for modeling techniques, respectively. We also note that the difference in theoretical complexity carries over to a significant difference in practical solvability: In our experiments, we observe that using the exponential cone solver of MOSEK drastically speeds up the solution times of our DR logistic regression problem.*

Proof of Theorem 2: In view of the first statement, we recall that the strongly NP-hard integer programming problem is defined as follows [16]:

<p>0/1 INTEGER PROGRAMMING.</p> <p>Instance. Given are $\mathbf{F} \in \mathbb{Z}^{\mu \times \nu}$, $\mathbf{g} \in \mathbb{Z}^\mu$, $\mathbf{c} \in \mathbb{Z}^\nu$, $\zeta \in \mathbb{Z}$.</p> <p>Question. Is there a vector $\chi \in \{0, 1\}^\nu$ such that $\mathbf{F}\chi \leq \mathbf{g}$ and $\mathbf{c}^\top \chi \geq \zeta$?</p>
--

Here, \mathbb{Z} denotes the set of integers. We claim that the answer to the integer programming problem is affirmative if and only if the optimal objective value for the DR logistic regression (1) with $n = 0$, $m = \nu$, $k_1 = \dots = k_m = 2$, $N = 1$ and the loss function

$$l_\beta(\mathbf{z}, y) = \begin{cases} \mathbf{c}^\top \mathbf{z} & \text{if } \mathbf{F}\mathbf{z} \leq \mathbf{g}, \\ -M & \text{otherwise} \end{cases}$$

and M and ϵ sufficiently large is greater than or equal to ζ . This then immediately implies that the DR logistic regression (1) is strongly NP-hard for generic loss functions (which include the above loss function as a special case).

To see this, note that our reformulation from the proof of Theorem 1, under the assumption that $n = 0$ and $N = 1$, shows that problem (1) is equivalent to

$$\begin{aligned}
& \underset{\beta, \lambda, s}{\text{minimize}} && \lambda\epsilon + s \\
& \text{subject to} && l_\beta(\mathbf{z}, y^1) - \lambda d_C(\mathbf{z}, \mathbf{z}^1) \leq s && \forall \mathbf{z} \in \mathbb{C} \\
& && l_\beta(\mathbf{z}, -y^1) - \lambda\kappa - \lambda d_C(\mathbf{z}, \mathbf{z}^1) \leq s && \forall \mathbf{z} \in \mathbb{C} \\
& && \beta = (\beta_0, \beta_C) \in \mathbb{R}^{k+1}, \quad \lambda \geq 0, \quad s \in \mathbb{R}.
\end{aligned}$$

Assuming that ϵ is sufficiently large, the optimal λ vanishes, and the problem simplifies further to

$$\begin{aligned} & \underset{\beta, s}{\text{minimize}} && s \\ & \text{subject to} && \mathbf{c}^\top \mathbf{z} \leq s \quad \forall \mathbf{z} \in \mathbb{C} : \mathbf{F}\mathbf{z} \leq \mathbf{g} \\ & && -M \leq s \quad \text{if } \exists \mathbf{z} \in \mathbb{C} : \mathbf{F}\mathbf{z} \not\leq \mathbf{g} \\ & && \beta = (\beta_0, \beta_C) \in \mathbb{R}^{k+1}, \quad s \in \mathbb{R} \end{aligned}$$

where we have also plugged in the aforementioned loss function. One readily observes that the optimal value of this problem equals $\max\{\mathbf{c}^\top \boldsymbol{\chi} : \mathbf{F}\boldsymbol{\chi} \leq \mathbf{g}, \boldsymbol{\chi} \in \{0, 1\}^\nu\}$ whenever $\{\mathbf{F}\boldsymbol{\chi} \leq \mathbf{g}, \boldsymbol{\chi} \in \{0, 1\}^\nu\}$ is non-empty and $-M$ otherwise. In particular, the optimal value of the problem is greater than or equal to ζ if and only if the answer to the integer programming problem is affirmative.

In view of the second statement, we note that the feasible region of problem (1) can be circumscribed by a convex body as per Definition 2.1.16 of [17], and that Algorithm 2 in Section 3 constitutes a weak separation oracle for problem (1). The statement thus follows from Corollary 4.2.7 of [17]. \square

Proof of Theorem 3: Consider a class of instances of problem (1) with a Wasserstein radius $\epsilon < 1$, $n = 1$ numerical feature, and $m = 1$ categorical feature with $k_1 = 2$ so that the categorical feature is binary. We have $N = 1$ sample with $z_1^1 = 0$ and $y_1^1 = -1$. This instance class is therefore parameterized by the value of the numerical feature x_1^1 in the data sample. We now derive the optimal objective value of problem (1) analytically as a function of x_1^1 , and show that this functional dependence on x_1^1 cannot be represented by the objective function of *any* regularized logistic regression.

For our instance class, the reformulation of our DR logistic regression (1) simplifies to

$$\begin{aligned} & \underset{\lambda, s_1, \beta}{\text{minimize}} && \lambda\epsilon + s_1 \\ & \text{subject to} && l_\beta(x_1^1, z, -1) - \lambda z \leq s_1 \quad \forall z \in \mathbb{B} \\ & && l_\beta(x_1^1, z, 1) - \lambda\kappa - \lambda z \leq s_1 \quad \forall z \in \mathbb{B} \\ & && \lambda \geq |\beta_N|, \quad s_1 \in \mathbb{R}. \end{aligned} \tag{11}$$

Assuming further that κ approaches ∞ , the constraint $\lambda \geq 0$ (implied by the constraint $\lambda \geq |\beta_N|$) implies that the second constraint is redundant. Removing this second constraint and substituting s_1 into the objective function results in

$$\begin{aligned} & \underset{\lambda, \beta}{\text{minimize}} && \lambda\epsilon + \max \{l_\beta(x_1^1, 0, -1), l_\beta(x_1^1, 1, -1) - \lambda\} \\ & \text{subject to} && \lambda \geq |\beta_N|. \end{aligned} \tag{12}$$

We now break the minimization over (λ, β) in (12) into a minimization over λ first followed by the minimization over β . That leads to re-writing (12) as

$$\underset{\beta}{\text{minimize}} \quad f(\beta, x_1^1) \tag{13}$$

where

$$f(\beta, x_1^1) = \begin{cases} \underset{\lambda}{\text{minimize}} & \lambda\epsilon + \max \{l_\beta(x_1^1, 0, -1), l_\beta(x_1^1, 1, -1) - \lambda\} \\ \text{subject to} & \lambda \geq |\beta_N| \end{cases}. \tag{14}$$

Since the Wasserstein radius satisfies $\epsilon < 1$, it is possible to achieve a reduction in the objective function choosing values larger than $|\beta_N|$ for λ . Focusing on the minimization over λ in (14), we recognize that the optimal solution is $\lambda^* = g(\beta, x_1^1) := l_\beta(x_1^1, 1, -1) - l_\beta(x_1^1, 0, -1)$ if $g(\beta, x_1^1) \geq |\beta_N|$, and $\lambda^* = |\beta_N|$ otherwise. Substituting this optimal value for λ into (14), we obtain

$$\begin{aligned} f(\beta, x_1^1) &= \begin{cases} \epsilon \cdot g(\beta, x_1^1) + l_\beta(x_1^1, 0, -1) & \text{if } |\beta_N| \leq g(\beta, x_1^1) \\ \epsilon \cdot |\beta_N| + l_\beta(x_1^1, 0, -1) & \text{otherwise} \end{cases} \\ &= l_\beta(x_1^1, 0, -1) + h(\beta, x_1^1), \end{aligned}$$

where

$$h(\beta, x_1^1) := \epsilon \cdot \begin{cases} g(\beta, x_1^1) & \text{if } |\beta_N| \leq g(\beta, x_1^1) \\ |\beta_N| & \text{otherwise.} \end{cases} = \epsilon \cdot \max \{g(\beta, x_1^1), |\beta_N|\}.$$

The DR logistic regression problem (13) now becomes

$$\underset{\beta}{\text{minimize}} \quad l_{\beta}(x_1^1, 0, -1) + h(\beta, x_1^1).$$

Using the definition of the empirical distribution, the regularized logistic regression for the instance problem has the form

$$\underset{\beta}{\text{minimize}} \quad l_{\beta}(x_1^1, 0, -1) + \mathfrak{R}(\beta). \quad (15)$$

The proof concludes by noting that $\mathfrak{R}(\beta)$ remains constant across all the instances parameterized by x_1^1 as a regularizer is data-independent. Hence, the objective function of the regularized logistic regression cannot capture the dependency on x_1^1 that we observe in $h(\beta, x_1^1)$ (it is easy to confirm that $h(\beta, x_1^1)$ is not constant in x_1^1). We further notice that the objective function of DR logistic regression is a non-smooth function of x_1^1 while the objective function of regularized logistic regression is a smooth function of x_1^1 by construction. \square

Remark 2. *It is informative to see how the counter-example in the proof of Theorem 3 breaks down if we assume there are no categorical features. In this case we use the ground metric $d(\xi, \xi') := \|\mathbf{x} - \mathbf{x}'\| + \kappa \cdot \mathbb{1}[y \neq y']$. The constraints in the problem formulation (10) will no longer be indexed over \mathbf{z} and all terms involving \mathbf{z} and $d_C(\mathbf{z}, \mathbf{z}^i)$ will disappear. The reformulation with only numerical features then coincides with that of Corollary 17 in [32]. Our instance class problem from the proof of Theorem 3 above will then be*

$$\begin{aligned} &\underset{\lambda, s_1, \beta}{\text{minimize}} \quad \lambda\epsilon + s_1 \\ &\text{subject to} \quad l_{\beta}(x_1^1, -1) \leq s_1 \\ &\quad \quad \quad l_{\beta}(x_1^1, 1) - \lambda\kappa \leq s_1 \\ &\quad \quad \quad \lambda \geq |\beta_N|, \quad s_1 \in \mathbb{R}. \end{aligned}$$

Taking κ to infinity, we see the second constraint becomes redundant. As λ only appears in one term in the objective function, the optimal solution is $\lambda^* = |\beta_N|$. Substituting for s_1 in the resulting objective becomes

$$\underset{\beta}{\text{minimize}} \quad l_{\beta}(x_1^1, -1) + \epsilon|\beta_N|$$

which is in the form of regularized logistic regression (15). This of course is consistent with [31] who showed that DR logistic regression without categorical features could be formulated as a regularized logistic regression problem.

Proof of Theorem 4: It follows from the proof of Theorem 1 that the optimization problems solved inside the while-loop of Algorithm 1 are equivalent to the relaxations

$$\begin{aligned} &\underset{\beta, \lambda, \mathbf{s}}{\text{minimize}} \quad \lambda\epsilon + \frac{1}{N} \sum_{i=1}^N s_i \\ &\text{subject to} \quad l_{\beta}(\mathbf{x}^i, \mathbf{z}, y^i) - \lambda d_C(\mathbf{z}, \mathbf{z}^i) \leq s_i \quad \forall (i, \mathbf{z}) \in \mathcal{W}^+ \\ &\quad \quad \quad l_{\beta}(\mathbf{x}^i, \mathbf{z}, -y^i) - \lambda\kappa - \lambda d_C(\mathbf{z}, \mathbf{z}^i) \leq s_i \quad \forall (i, \mathbf{z}) \in \mathcal{W}^- \\ &\quad \quad \quad \|\beta_N\|_* \leq \lambda \\ &\quad \quad \quad \beta = (\beta_0, \beta_N, \beta_C) \in \mathbb{R}^{1+n+k}, \quad \lambda \geq 0, \quad \mathbf{s} \in \mathbb{R}^N \end{aligned}$$

of problem (10) that is itself equivalent to our DR logistic regression problem (4); the only difference between our relaxations above and problem (10) is that the index sets $(i, \mathbf{z}) \in [N] \times \mathbb{C}$ in the first two constraint sets of (10) are replaced with the subsets \mathcal{W}^+ and \mathcal{W}^- above. This shows that each value LB_t indeed constitutes a lower bound on the optimal value of problem (4), and that the sequence $\{\text{LB}_t\}_t$ of lower bounds is monotonically non-decreasing.

To see that the sequence $\{\text{UB}_t\}_t$ bounds the optimal value of our DR logistic regression problem (4) from above, note that a maximum constraint violation of ϑ^+ in the first constraint set of (4) implies that

$$\left[\begin{array}{l} u_{i,\mathbf{z}}^+ + v_{i,\mathbf{z}}^+ \leq 1 + \vartheta^+ \\ (u_{i,\mathbf{z}}^+, 1, -s_i - \lambda d_C(\mathbf{z}, \mathbf{z}^i)) \in \mathcal{K}_{\text{exp}} \\ (v_{i,\mathbf{z}}^+, 1, -y^i \beta_N^\top \mathbf{x}^i - y^i \beta_C^\top \mathbf{z} - y^i \beta_0 - s_i - \lambda d_C(\mathbf{z}, \mathbf{z}^i)) \in \mathcal{K}_{\text{exp}}, \end{array} \right] \quad \forall (i, \mathbf{z}) \in [N] \times \mathbb{C},$$

and, by the definition of the exponential cone, this means that for all $(i, z) \in [N] \times \mathbb{C}$, we have

$$\exp[-s_i - \lambda d_C(z, z^i)] + \exp[-y^i \beta_N^\top x^i - y^i \beta_C^\top z - y^i \beta_0 - s_i - \lambda d_C(z, z^i)] \leq 1 + \vartheta^+.$$

Dividing both sides by $1 + \vartheta^+ = \exp[\log(1 + \vartheta^+)]$, the inequality becomes

$$\exp[-s'_i - \lambda d_C(z, z^i)] + \exp[-y^i \beta_N^\top x^i - y^i \beta_C^\top z - y^i \beta_0 - s'_i - \lambda d_C(z, z^i)] \leq 1,$$

where $s'_i := s_i + \log(1 + \vartheta^+)$. The (potentially) infeasible solution to the relaxed problem of Algorithm 1 thus allows us to construct a solution to problem (4) that satisfies all members of the first constraint set by replacing s_i with $s'_i := s_i + \log(1 + \vartheta^+)$, $i \in [N]$. Compared to the solution of the relaxed problem, the objective value of the newly created solution increased by $\log(1 + \vartheta^+)$. A similar argument can be made for the members of the second constraint set, which shows that a solution to problem (4) satisfying both constraint sets can be constructed by increasing the objective value by no more than $\log(1 + \max\{\vartheta^+, \vartheta^-\})$. This shows the validity of the upper bound $\theta^* + \log(1 + \max\{\vartheta^+, \vartheta^-\})$. The fact that the sequence $\{\text{UB}_t\}_t$ of upper bounds is monotonically non-increasing, on the other hand, holds by construction since we take the minimum of the derived upper bound $(\theta^* + \log(1 + \max\{\vartheta^+, \vartheta^-\}))$ and the upper bound UB_{t-1} of the previous iteration.

To see that Algorithm 1 terminates in finite time, note that every iteration t either identifies a violated constraint $(i, z) \in ([N] \times \mathbb{C}) \setminus \mathcal{W}^+ \cup ([N] \times \mathbb{C}) \setminus \mathcal{W}^-$ that is subsequently added to \mathcal{W}^+ and/or \mathcal{W}^- (and will thus never again be identified as violated for the same constraint set), or the update $\text{LB}_t = \text{UB}_t = \theta^*$ is conducted, in which case the algorithm terminates in the next iteration. Finite termination thus holds since the set $[N] \times \mathbb{C}$ of potentially violated constraints is finite. \square

Additional Intuition Behind Algorithm 2

In the combinatorial optimization problem (5), the decision variables $z \in \mathbb{C}$ only appear in (monotone transformations of) the terms $-y^i \beta_C^\top z$ and $-\lambda d_C(z, z^i)$. Moreover, the expression $[d_C(z, z^i)]^p$ can only attain one of the values $\delta \in \{0, \dots, m\}$ (cf. Definition 2). Conditioning on each possible value of δ , which records the number of categorical features along which z and z^i disagree, we can therefore maximize the constraint violation by maximizing $-y^i \beta_C^\top z$ along all $z \in \mathbb{C}$ that differ from z^i in exactly δ categorical features. Finally, since $-y^i \beta_C^\top z$ is linearly separable, that is, $-y^i \beta_C^\top z = -y^i \beta_{C,1}^\top z_1 - \dots - y^i \beta_{C,m}^\top z_m$, we can choose the categorical features along which z and z^i disagree iteratively by considering each expression $-y^i \beta_C^\top z$ separately and choosing the disagreeing feature values greedily based on their individual contribution to the overall constraint violation.

Proof of Theorem 5 Recall that Algorithm 2 aims to solve the optimization problem

$$\underset{z}{\text{maximize}} \quad \min_{u^+, v^+} \left\{ u^+ + v^+ : \begin{bmatrix} (u^+, 1, -s_i - \lambda d_C(z, z^i)) \in \mathcal{K}_{\text{exp}}, \\ (v^+, 1, -y^i \beta_N^\top x^i - y^i \beta_C^\top z - y^i \beta_0 - s_i - \lambda d_C(z, z^i)) \in \mathcal{K}_{\text{exp}} \end{bmatrix} \right\}$$

subject to $z \in \mathbb{C}$.

It immediately follows from the definition of \mathcal{K}_{exp} that the above problem can be represented as

$$\underset{z}{\text{maximize}} \quad \exp[-s_i - \lambda d_C(z, z^i)] + \exp[-y^i \beta_N^\top x^i - y^i \beta_C^\top z - y^i \beta_0 - s_i - \lambda d_C(z, z^i)]$$

subject to $z \in \mathbb{C}$.

In this problem, the optimization variable z appears inside the exponential terms as $-y^i \beta_C^\top z$ and as $-\lambda d_C(z, z^i)$. Since increasing the value of one term may decrease the value of the other term, the above problem does not admit a trivial solution. We next discuss how the above maximization problem can be decomposed into $m + 1$ sub-problems, each of which can be solved efficiently.

Although z may take exponentially many values, notice that the expression $d_C(z, z^i) = (\sum_{j \in [m]} \mathbb{1}[z_j \neq z_j^i])^{1/p}$ counts the number of disagreements between the features of z and z^i , which takes a value from the set $\{0, 1, \dots, m\}$. Hence, Algorithm 2 decomposes the above combinatorial optimization problem into $m + 1$ sub-problems in which the number $\delta \in \{0, 1, \dots, m\}$ of disagreements between z and z^i is fixed and we only need to maximize $-y^i \beta_C^\top z$ for a given number δ of disagreements. If we can solve each of these sub-problem efficiently, then we can compare the constraint violations of the optimal z to each sub-problem and pick the largest one as

the most-violated constraint. We next investigate how each of the $m + 1$ sub-problems can be solved efficiently.

The sub-problem corresponding to a fixed $\delta \in \{0, 1, \dots, m\}$ can be formulated as

$$\begin{aligned} & \underset{\mathbf{z}}{\text{maximize}} && -y^i \beta_{\mathbf{C}}^\top \mathbf{z} \\ & \text{subject to} && d_{\mathbf{C}}(\mathbf{z}, \mathbf{z}^i) = \delta \\ & && \mathbf{z} \in \mathbb{C}. \end{aligned}$$

Denote by $\mathcal{J} \subseteq [m] : |\mathcal{J}| = \delta$ the set of features where \mathbf{z} disagrees with \mathbf{z}^i at optimality. Denote further by \mathbf{z}_j^* the maximizer of $-y^i \cdot \beta_{\mathbf{C},j}^\top \mathbf{z}_j$ over $\mathbf{z}_j \in \mathbb{C}(k_j) \setminus \{\mathbf{z}_j^i\}$ (cf. the first step of Algorithm 2). The above sub-problem can then be written as

$$\begin{aligned} & \underset{\mathcal{J}}{\text{maximize}} && \left(\sum_{j=1}^m -y^i \cdot \beta_{\mathbf{C},j}^\top \mathbf{z}_j^i \right) + \left(\sum_{j \in \mathcal{J}} (-y^i \cdot \beta_{\mathbf{C},j}^\top \mathbf{z}_j^*) - (-y^i \cdot \beta_{\mathbf{C},j}^\top \mathbf{z}_j^i) \right) \\ & \text{subject to} && \mathcal{J} \subseteq [m], |\mathcal{J}| = \delta. \end{aligned}$$

The first summation above is constant, and the solution of this problem can therefore be uniquely identified from the solution of the following problem:

$$\begin{aligned} & \underset{\mathcal{J}}{\text{maximize}} && \sum_{j \in \mathcal{J}} (-y^i \cdot \beta_{\mathbf{C},j}^\top (\mathbf{z}_j^* - \mathbf{z}_j^i)) \\ & \text{subject to} && \mathcal{J} \subseteq [m], |\mathcal{J}| = \delta. \end{aligned}$$

This problem can be solved greedily by selecting the δ largest values of $(-y^i \cdot \beta_{\mathbf{C},j}^\top (\mathbf{z}_j^* - \mathbf{z}_j^i))$ across all $j \in [m]$, and those j indexes will be included in \mathcal{J} (cf. the ordering $\pi : [m] \mapsto [m]$ and the selection according to $\pi(j) \leq \delta$ in Algorithm 2). Instead of sorting $(-y^i \cdot \beta_{\mathbf{C},j}^\top (\mathbf{z}_j^* - \mathbf{z}_j^i))$, $j \in [m]$, for each sub-problem $\delta \in \{0, 1, \dots, m\}$, Algorithm 2 sorts those values once in the beginning.

The discussion so far establishes the correctness of Algorithm 2. In view of its runtime, we note that the aforementioned sorting takes time $\mathcal{O}(m \log m)$ as each value $(-y^i \cdot \beta_{\mathbf{C},j}^\top (\mathbf{z}_j^* - \mathbf{z}_j^i))$, $j \in [m]$ can be computed in constant time, determining \mathbf{z}_j^* for all $j \in [m]$ takes time $\mathcal{O}(k)$ since the variables corresponding to each feature vanish in all but one (known) location, and determining the set \mathcal{J} for each $\delta \in \{0, 1, \dots, m\}$ takes time $\mathcal{O}(n + m^2)$ since the expression $-y^i \beta_{\mathbf{N}}^\top \mathbf{x}^i$ only needs to be computed once. \square

Proof of Theorem 6: The statement of the theorem follows immediately from Theorems 18 and 19 of [21]. \square

Proof of Lemma 1: The assumption in the statement of the lemma allows us to choose $M \in \mathbb{R}$ such that $\mathcal{H} \subseteq [-M, +M]^{1+n+k}$. We then show that there is $\xi^0 \in \Xi$ and $C > 0$ such that

$$l_{\beta}(\xi) \leq C[1 + d(\xi, \xi^0)] \quad \forall \beta \in [-M, M]^{1+n+k}, \quad \forall \xi \in \Xi,$$

that is, for all $\beta \in [-M, M]^{1+n+k}$ and all $\xi \in \Xi$ we have

$$\begin{aligned} & \log(1 + \exp[-y \cdot (\beta_0 + \beta_{\mathbf{N}}^\top \mathbf{x} + \beta_{\mathbf{C}}^\top \mathbf{z})]) \\ & \leq C \left[1 + \|\mathbf{x} - \mathbf{x}^0\| + \left(\sum_{i \in [m]} \mathbb{1}[z_i \neq z_i^0] \right)^{1/p} + \kappa \cdot \mathbb{1}[y \neq y^0] \right]. \end{aligned}$$

To see this, fix any $\xi^0 = (\mathbf{0}, \mathbf{z}^0, y^0) \in \Xi$. Since $\kappa \cdot \mathbb{1}[y \neq y^0] \geq 0$ and $\sum_{i \in [m]} \mathbb{1}[z_i \neq z_i^0]^{1/p} \geq 0$ on the right-hand side of the above inequality, it suffices to show that

$$\log(1 + \exp[-y \cdot (\beta_0 + \beta_{\mathbf{N}}^\top \mathbf{x} + \beta_{\mathbf{C}}^\top \mathbf{z})]) \leq C[1 + \|\mathbf{x}\|] \quad \forall \beta \in [-M, M]^{1+n+k}, \quad \forall \xi \in \Xi.$$

Note that the left-hand side of this inequality can be bounded from above by

$$\begin{aligned} 1 + \exp[-y \cdot (\beta_0 + \beta_{\mathbf{N}}^\top \mathbf{x} + \beta_{\mathbf{C}}^\top \mathbf{z})] & \leq 1 + \exp[|\beta_0 + \beta_{\mathbf{N}}^\top \mathbf{x} + \beta_{\mathbf{C}}^\top \mathbf{z}|] \\ & \leq 2 \exp[|\beta_0 + \beta_{\mathbf{N}}^\top \mathbf{x} + \beta_{\mathbf{C}}^\top \mathbf{z}|]. \end{aligned}$$

It is therefore sufficient to show that

$$\begin{aligned}
& \log(2 \exp [|\beta_0 + \beta_N^\top \mathbf{x} + \beta_C^\top \mathbf{z}|]) \leq C[1 + \|\mathbf{x}\|] & \forall \beta \in [-M, M]^{1+n+k}, \\
& \iff \log(2) + |\beta_0 + \beta_N^\top \mathbf{x} + \beta_C^\top \mathbf{z}| \leq C[1 + \|\mathbf{x}\|] & \forall \xi \in \Xi \\
& \iff \log(2) + \max_{\beta \in [-M, M]^{1+n+k}} \{|\beta_0 + \beta_N^\top \mathbf{x} + \beta_C^\top \mathbf{z}|\} \leq C[1 + \|\mathbf{x}\|] & \forall \beta \in [-M, M]^{1+n+k}, \\
& \iff \log(2) + (M + M\|\mathbf{x}\| + Mk) \leq C[1 + \|\mathbf{x}\|] & \forall \xi \in \Xi \\
& \iff \log(2) + M(1 + k) + M\|\mathbf{x}\| \leq C + C\|\mathbf{x}\| & \forall \xi \in \Xi.
\end{aligned}$$

The last condition, however, is readily seen to be satisfied by any $C \geq \log(2) + M(1 + k)$. \square

Proof of Theorem 7: The statement of the theorem follows directly from Theorem 20 of [21]. \square

Proof of Theorem 8: The existence of an optimal solution follows from Theorem 3 of [42], and the sparsity of the optimal solution is due to Theorem 4 of [42]. Note that the assumptions of those theorems are satisfied since the Wasserstein ball $\mathfrak{B}_\epsilon(\hat{\mathbb{P}}_N)$ is centered at the empirical distribution $\hat{\mathbb{P}}_N$, which has finite moments by construction. \square