DATA-DRIVEN OPTIMAL TRANSPORT COST SELECTION FOR DISTRIBUTIONALLY ROBUST OPTIMIZATION

BLANCHET, J., KANG, Y., ZHANG, F., AND MURTHY, K.

Abstract. Recently, Blanchet et al. (2016) showed that several machine learning algorithms, such as square-root Lasso, Support Vector Machines, and regularized logistic regression, among many others, can be represented exactly as distributionally robust optimization (DRO) problems. The distributional uncertainty is defined as a neighborhood centered at the empirical distribution. We propose a methodology which learns such neighborhood in a natural data-driven way. We show rigorously that our framework encompasses adaptive regularization as a particular case. Moreover, we demonstrate empirically that our proposed methodology is able to improve upon a wide range of popular machine learning estimators.

1. Introduction

A Distributionally Robust Optimization (DRO) problem takes the general form

$$\min_{\beta} \max_{P \in \mathcal{U}_\delta} \mathbb{E}_P \left[l\left(X, Y, \beta\right) \right], \tag{1}$$

where β is a decision variable, (X, Y) is a random element, and $l(x, y, \beta)$ measures a suitable loss or cost incurred when $(X, Y) = (x, y)$ and the decision β is taken. The expectation $\mathbb{E}_P[\cdot]$ is taken under the probability model P. The set $\mathcal{U}_\delta$ is called the distributional uncertainty set and it is indexed by the parameter $\delta > 0$, which measures the size of the distributional uncertainty.

The DRO problem is said to be *data-driven* if the uncertainty set $\mathcal{U}_\delta$ is informed by empirical observations. One natural way to supply this information is by letting the "center" of the uncertainty region be placed at the empirical measure, P_n, induced by a data set $\{X_i, Y_i\}_{i=1}^n$, which represents an empirical sample of realizations of W. In order to emphasize the data-driven nature of a DRO formulation such as (1), when the uncertainty region is informed by an empirical sample, we write $\mathcal{U}_\delta = \mathcal{U}_\delta(P_n)$. To the best of our knowledge, the available data is utilized in the DRO literature only by defining the center of the uncertainty region $\mathcal{U}_\delta(P_n)$ as the empirical measure P_n.

Our goal in this paper is to discuss a data-driven framework to inform the *shape* of $\mathcal{U}_\delta(P_n)$. Throughout this paper, we assume that the class of functions to fit, indexed by β, is given and that a sensible loss function $l\left(x, y, \beta\right)$ has been selected for the problem at hand. Our contribution concerns the construction of the uncertainty region in a fully data-driven way and the implications of this design in machine learning applications. Before providing our construction, let us discuss the significance of data-driven DRO in the context of machine learning.

Recently, Blanchet et al. (2016) showed that many prevailing machine learning estimators can be represented exactly as a data-driven DRO formulation in (1). For example, suppose that $X \in \mathbb{R}^d$ and $Y \in \{-1, 1\}$. Further, let $l(x, y, \beta) = \log(1 + \exp(-y\beta^T x))$ be the log-exponential loss associated to a logistic regression model where $Y \sim Ber(1/(1 + \exp(-\beta_*^T x))$, and β_* is the underlying parameter to learn. Then, given a set of empirical samples $\mathcal{D}_n = \{(X_i, Y_i)\}_{i=1}^n$, and a judicious

Date: May 23, 2017.

choice of the distributional uncertainty set $\mathcal{U}_\delta(P_n)$, Blanchet et al. (2016) shows that

$$\min_{\beta} \max_{P \in \mathcal{U}_\delta(P_n)} \mathbb{E}_P[l(X,Y,\beta)] = \min_{\beta} \left(\mathbb{E}_{P_n}[l(X,Y,\beta)] + \delta \|\beta\|_p \right), \tag{2}$$

where $\|\cdot\|_p$ is the ℓ_p-norm in $\mathbb{R}^d$ for $p \in [1,\infty)$ and $\mathbb{E}_{P_n}[l(X,Y,\beta)] = n^{-1}\sum_{i=1}^n l(X_i,Y_i,\beta)$.

The definition of $\mathcal{U}_\delta(P_n)$ turns out to be informed by the dual norm $\|\cdot\|_q$ with $1/p + 1/q = 1$. If $p = 1$ we see that (2) recovers L_1 regularized logistic regression (see Friedman et al. (2001)). Other estimators such as Support Vector Machines and sqrt-Lasso are shown in Blanchet et al. (2016) to admit DRO representations analogous to (2) – provided that the loss function and the uncertainty region are judiciously chosen. Note that the parameter δ in $\mathcal{U}_\delta(P_n)$ is precisely the regularization parameter in the right hand side of (2). So, the data-driven DRO representation (2) provides a direct interpretation of the regularization parameter as the size of the probabilistic uncertainty around the empirical evidence.

An important element to all of the DRO representations obtained in Blanchet et al. (2016) is that the design of the uncertainty region $\mathcal{U}_\delta(P_n)$ is based on optimal transport theory. In particular, we have that

$$\mathcal{U}_\delta(P_n) = \{P : D_c(P,P_n) \leq \delta\}, \tag{3}$$

and $D_c(P,P_n)$ is the minimal cost of rearranging (i.e. transporting the mass of) the distribution P_n into the distribution P. The rearrangement mechanism has a transportation cost $c(u,w) \geq 0$ for moving a unit of mass from location u in the support of P_n to location w in the support of P. For instance, in the setting of (2) we have that

$$c\big((x,y),(x',y')\big) = \|x - x'\|_q^2 \, I\left(y = y'\right) + \infty \cdot I\left(y \neq y'\right). \tag{4}$$

In the end, as we discuss in Section 3, $D_c(P,P_n)$ can be easily computed as the solution of a linear programming (LP) problem which is known as Kantorovich's problem (see Villani (2008)).

Other discrepancy notions between probability models have been considered, typically using the Kullback-Leibler divergence and other divergence based notions Hu and Hong (2013). Using divergence (or likelihood ratio) based discrepancies to characterize the uncertainty region $\mathcal{U}_\delta(P_n)$ forces the models $P \in \mathcal{U}_\delta(P_n)$ to share the same support with P_n, which may restrict generalization properties of a DRO-based estimator, and such restriction may induce overfitting problem (see the discussions in Esfahani and Kuhn (2015) and Blanchet et al. (2016)).

In summary, data-driven DRO via optimal transport has been shown to encompass a wide range of prevailing machine learning estimators. However, so far the cost function $c(\cdot)$ has been taken as a given, and not chosen in a data-driven way.

Our main contribution in this paper is to propose a comprehensive approach for designing the uncertainty region $\mathcal{U}_\delta(P_n)$ in a fully data-driven way, using the convenient role of $c(\cdot)$ in the definition of the optimal transport discrepancy $D_c(P,P_n)$. Our modeling approach further underscores, beyond the existence of representations such as (2), the convenience of working with an optimal transport discrepancy for the design of data-driven DRO machine learning estimators. In other words, because one can select $c(\cdot)$ in a data driven way, it is sensible to use our data-driven DRO formulation even if one is not able to simplify the inner optimization in order to achieve a representation such as (2).

Our idea is to apply metric-learning procedures to estimate $c(\cdot)$ from the training data. Then, use such data-driven $c(\cdot)$ in the definition of $D_c(P,P_n)$ and the construction $\mathcal{U}_\delta(P_n)$ in (3). Finally, solve the DRO problem (1), using cross-validation to choose δ.

The intuition behind our proposal is the following. By using a metric learning procedure we are

able to calibrate a cost function $c(\cdot)$ which attaches relatively high transportation costs to (u, w) if transporting mass between these locations substantially impacts performance (e.g. in the response variable, so increasing the expected risk). In turn, the adversary, with a given budget δ, will carefully choose the data which is to be transported. The mechanism will then induce enhanced out-of-sample performance focusing precisely on regions of relevance, while improving generalization error.

One of the challenges for the implementation of our idea is to efficiently solve (1). We address this challenge by proposing a stochastic gradient descent algorithm which takes advantage of a duality representation and fully exploits the nature of the LP structure embedded in the definition of $D_c(P, P_n)$, together with a smoothing technique.

Another challenge consists in selecting the type of cost $c(\cdot)$ to be used in practice, and the methodology to fit such cost. To cope with this challenge, we rely on metric-learning procedures. We do not contribute any novel metric learning methodology; rather, we discuss various parametric cost functions and methods developed in the metric-learning literature. And we discuss their use in the context of fully data-drive DRO formulations for machine learning problems – which is what we propose in this paper. The choice of $c(\cdot)$ ultimately will be influenced by the nature of the data and the application at hand. For example, in the setting of image recognition, it might be natural to use a cost function related to similarity notions.

In addition to discussing intuitively the benefits of our approach in Section 2, we also show that our methodology provides a way to naturally estimate various parameters in the setting of adaptive regularization. For example, Theorem 1 below, shows that choosing $c(\cdot)$ using a suitable weighted norm, allows us to recover an adaptive regularized ridge regression estimator Ishwaran and Rao (2014). In turn, using standard techniques from metric learning we can estimate $c(\cdot)$. Hence, our technique connects metric learning tools to estimate the parameters of adaptive regularized estimators.

More broadly, we compare the performance of our procedure with a number of alternatives in the setting of various data sets and show that our approach exhibits consistently superior performance.

2. Data-Driven DRO: Intuition and Interpretations

One of the main benefits of DRO formulations such as (1) and (2) is their interpretability. For example, we can readily see from the left hand side of (2) that the regularization parameter corresponds precisely to the size of the *data-driven* distributional uncertainty.

The data-driven aspect is important because we can employ statistical thinking to optimally characterize the size of the uncertainty, δ. This readily implies an optimal choice of the regularization parameter, as explained in Blanchet et al. (2016), in settings such as (2). Elaborating, we can interpret $\mathcal{U}_\delta(P_n)$ as the set of plausible variations of the empirical data, P_n. Consequently, for instance, in the linear regression setting leading to (2), the estimate $\beta_P = \arg\min_\beta \mathbb{E}_P(l(X, Y, \beta))$ is a plausible estimate of the regression parameter β_* as long as $P \in \mathcal{U}_\delta(P_n)$. Hence, the set

$$\Lambda_\delta(P_n) = \{\beta_P : P \in \mathcal{U}_\delta(P_n)\}$$

is a natural confidence region for β_* which is non-decreasing in δ. Thus, a statistically minded approach for choosing δ is to fix a confidence level, say $(1-\alpha)$, and choose an optimal δ $(\delta_*(n))$ via

$$\delta_*(n) := \inf\{\delta : P(\beta_* \in \Lambda_\delta(P_n)) \geq 1-\alpha\}. \tag{5}$$

Note that the random element in $P(\beta_* \in \Lambda_\delta(P_n))$ is given by P_n. In Blanchet et al. (2016) this optimization problem is solved asymptotically as $n \to \infty$ under standard assumptions on the data generating process. If the underlying model is correct, one would typically obtain, as in Blanchet

et al. (2016), that $\delta_*(n) \to 0$ at a suitable rate. For instance, in the linear regression setting corresponding to (2), if the data is i.i.d. with finite variance and the linear regression model holds then $\delta_*(n) = \chi_{1-\alpha}\left(1 + o\left(1\right)\right)/n$ as $n \to \infty$ (where χ_α is the α quantile of an explicitly characterized distribution).

In practice, one can also choose δ by cross-validation. The work of Blanchet et al. (2016) compares the asymptotically optimal choice $\delta_*(n)$ against cross-validation, concluding that the performance is comparable in the experiments performed. In this paper, we use cross validation to choose δ, but the insights behind the limiting behavior of (5) are useful, as we shall see, to inform the design of our algorithms.

More generally, the DRO formulation (1) is appealing because the distributional uncertainty endows the estimation of β directly with a mechanism to enhance generalization properties. To wit, we can interpret (1) as a game in which we (the outer player) choose a decision β, while the adversary (the inner player) selects a model which is a perturbation, P, of the data (encoded by P_n). The amount of the perturbation is dictated by the size of δ which, as discussed earlier, is data driven. But the type of perturbation and how the perturbation is measured is dictated by $D_c(P, P_n)$. It makes sense to inform the design of $D_c(\cdot)$ using a data-driven mechanism, which is our goal in this paper. The intuition is to allow the types of perturbations which focus the effort and budget of the adversary mostly on out-of-sample exploration over regions of relevance.

The type of benefit that is obtained by informing $D_c\left(P, P_n\right)$ with data is illustrated in Figure 1(a) below. Figure 1(a) illustrates a classification task. The data roughly lies on a lower dimensional

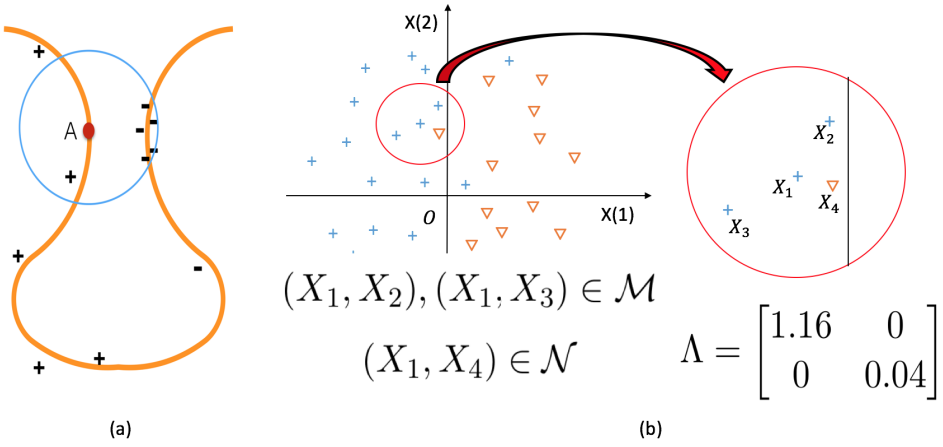

FIGURE 1. Stylized examples illustrating the need for data-driven cost function.

non-linear manifold. Some data which is classified with a negative label is seen to be "close" to data which is classified with a positive label when seeing the whole space (i.e. $\mathbb{R}^2$) as the natural ambient domain of the data. However, if we use a distance similar to the geodesic distance intrinsic to the manifold we would see that the negative instances are actually far from the positive instances. So, the generalization properties of the algorithm would be enhanced relative to using a standard metric in the ambient space, because with a given budget δ the adversarial player would be allowed perturbations mostly along the intrinsic manifold where the data lies and instances which are surrounded (in the intrinsic metric) by similarly classified examples will naturally carry significant impact in testing performance. A quantitative example to illustrate this point will be discussed in the sequel.

3. Background on Optimal Transport and Metric Learning Procedures

In this section we quickly review basic notions on optimal transport and metric learning methods so that we can define $D_c(P, P_n)$ and explain how to calibrate the function $c(\cdot)$.

3.1. **Defining Optimal Transport Distances and Discrepancies.** Assume that the cost function $c : \mathbb{R}^{d+1} \times \mathbb{R}^{d+1} \to [0, \infty]$ is lower semicontinuous. We also assume that $c(u, v) = 0$ if and only if $u = v$. Given two distributions P and Q, with supports $\mathcal{S}_P$ and $\mathcal{S}_Q$, respectively, we define the optimal transport discrepancy, D_c, via

$$D_c\left(P, Q\right) = \inf\left\{\mathbb{E}_\pi\left[c(U, V)\right] : \pi \in \mathcal{P}\left(\mathcal{S}_{_P} \times \mathcal{S}_{_Q}\right),\ \pi_{_U} = P,\ \pi_{_V} = Q\right\}, \tag{6}$$

where $\mathcal{P}(\mathcal{S}_{_P} \times \mathcal{S}_{_Q})$ is the set of probability distributions π supported on $\mathcal{S}_P \times \mathcal{S}_{_Q}$, and $\pi_{_U}$ and $\pi_{_V}$ denote the marginals of U and V under π, respectively. Because $c(\cdot)$ is non-negative we have that $D_c(P, Q) \geq 0$. Moreover, requiring that $c(u, v) = 0$ if and only if $u = v$ guarantees that $D_c(P, Q) = 0$ if and only $P = Q$. If, in addition, $c(\cdot)$ is symmetric (i.e. $c(u, v) = c(v, u)$), and there exists $\varrho \geq 1$ such that $c^{1/\varrho}(u, w) \leq c^{1/\varrho}(u, v) + c^{1/\varrho}(v, w)$ (i.e. $c^{1/\varrho}(\cdot)$ satisfies the triangle inequality) then it can be easily verified (see Villani (2008)) that $D_c^{1/\varrho}\left(P, Q\right)$ is a metric. For example, if $c(u, v) = \|u - v\|_q^\varrho$ for $q \geq 1$ (where $\|u - v\|_q$ denotes the l_q norm in $\mathbb{R}^{d+1}$) then $D_c(\cdot)$ is known as the Wasserstein distance of order ϱ. Observe that (6) is a linear program in the variable π.

3.2. **On Metric Learning Procedures.** In order to keep the discussion focused, we use a few metric learning procedures, but we emphasize that our approach can be used in combination with virtually any method in the metric learning literature, see the survey paper Bellet et al. (2013) that contains additional discussion on metric learning procedures. The procedures that we consider, as we shall see, can be seen to already improve significantly upon natural benchmarks. Moreover, as we shall see, these metric families can be related to adaptive regularization. This connection will be useful to further enhance the intuition of our procedure.

3.2.1. *The Mahalanobis Distance.* The Mahalanobis metric is defined as

$$d_\Lambda\left(x, x'\right) = \left(\left(x - x'\right)^T \Lambda \left(x - x'\right)\right)^{1/2},$$

where Λ is symmetric and positive semi-definite and we write $\Lambda \in PSD$. Note that $d_\Lambda(x, x')$ is the metric induced by the norm $\|x\|_\Lambda = \sqrt{x^T \Lambda x}$.

For a discussion, assume that our data is of the form $\mathcal{D}_n = \{(X_i, Y_i)\}_{i=1}^n$ and $Y_i \in \{-1, +1\}$. The prediction variables are assumed to be standardized. Motivated by applications such as social networks, in which there is a natural graph which can be used to connect instances in the data, we assume that one is given sets $\mathcal{M}$ and $\mathcal{N}$, where $\mathcal{M}$ is the set of the pairs that should be close (so that we can connect them) to each other, and $\mathcal{N}$, on contrary, is characterizing the relations that the pairs should be far away (not connected), we define them as

$$\begin{aligned}\mathcal{M} &:= \{(X_i, X_j) \mid X_i \text{ and } X_j \text{ must connect}\},\\ \mathcal{N} &:= \{(X_i, X_j) \mid X_i \text{ and } X_j \text{ should not connect}\}.\end{aligned}$$

While it is typically assumed that $\mathcal{M}$ and $\mathcal{N}$ are given, one may always resort to k-Nearest-Neighbor (k-NN) method for the generation of these sets. This is the approach that we follow in our numerical experiments. But we emphasize that choosing any criterion for the definition of $\mathcal{M}$ and $\mathcal{N}$ should be influenced by the learning task in order to retain both interpretability and performance.

In our experiments we let (X_i, X_j) belong to $\mathcal{M}$ if, in addition to being sufficiently close (i.e. in

the k-NN criterion), $Y_i = Y_j$. If $Y_i \neq Y_j$, then we have that $(X_i, X_j) \in \mathcal{N}$.
The work of Xing et al. (2002), one of the earlier reference on the subject, suggests considering

$$\min_{\Lambda \in PSD} \sum_{(X_i,X_j)\in\mathcal{M}} d_\Lambda^2 (X_i, X_j) \tag{7}$$

$$s.t. \sum_{(X_i,X_j)\in\mathcal{N}} d_\Lambda^2 (X_i, X_j) \geq \bar{\lambda}. \tag{8}$$

In words, the previous optimization problem minimizes the total distance between pairs that should be connect, while keeping the pairs that should not connect well separated. The constant $\bar{\lambda} > 0$ is somewhat arbitrary (given that Λ can be normalized by $\bar{\lambda}$, we can choose $\bar{\lambda} = 1$).

The optimization problem (8) is an LP problem on the convex cone PSD and it has been widely studied. Since $\Lambda \in PSD$, we can always write $\Lambda = LL^T$, and therefore $d_\Lambda(X_i, X_j) = \|X_i - X_j\|_\Lambda := \|LX_i - LX_j\|_2$. There are various techniques which can be used to exploit the PSD structure to efficiently solve (8); see, for example, Xing et al. (2002) for a projection-based algorithm; or Schultz and Joachims (2004), which uses a factorization-based procedure; or the survey paper Bellet et al. (2013) for the discussion of a wide range of techniques.

We have chosen formulation (8) to estimate Λ because it is intuitive and easy to state, but the topic of learning Mahalanobis distances is an active area of research and there are different algorithms which can be implemented (see Li et al. (2016)).

3.2.2. *Using Mahalanobis Distance in Data-Driven DRO* . Let us assume that the underlying data takes the form $\mathcal{D}_n = \{(X_i, Y_i)\}_{i=1}^n$, where $X_i \in R^d$ and $Y_i \in R$ and the loss function, depending on a decision variable $\beta \in R^m$, is given by $l(x, y, \beta)$. Note that we are not imposing any linear structure on the underlying model or in the loss function. Then, motivated by the cost function (4), we may consider

$$c_\Lambda\left((x, y), (x', y')\right) = d_\Lambda^2 \left(x, x'\right) I\left(y = y'\right) + \infty I\left(y \neq y'\right), \tag{9}$$

for $\Lambda \in PSD$. The infinite contribution in the definition of c_Λ (i.e. $\infty \cdot I\left(y \neq y'\right)$) indicates that the adversarial player in the DRO formulation is not allowed to perturb the response variable.
Even in this case, since the definitions of $\mathcal{M}$ and $\mathcal{N}$ depend on $W_i = (X_i, Y_i)$ (in particular, on the response variable), cost function $c_\Lambda(\cdot)$ (once Λ is calibrated using, for example, the method discussed in the previous subsection), will be informed by the Y_is. Then, we estimate β via

$$\min_\beta \sup_{P: D_{c_\Lambda}(P, P_n) \leq \delta} \mathbb{E}[l(X, Y, \beta)]. \tag{10}$$

It is important to note that Λ has been applied only to the definition of the cost function.

The intuition behind the formulation can be gained in the context of a logistic regression setting, see the example in Figure 1(b): Suppose that $d = 2$, and that Y depends only on $X(1)$ (i.e. the first coordinate of X). Then, the metric learning procedure in (8) will induce a relatively low transportation cost across the $X(2)$ direction and a relatively high transportation cost in the $X(1)$ direction, whose contribution, being highly informative, is reasonably captured by the metric learning mechanism. Since the $X(1)$ direction is most impactful, from the standpoint of expected loss estimation, the adversarial player will reach a compromise, between transporting (which is relatively expensive) and increasing the expected loss (which is the adversary's objective). Out of this compromise the DRO procedure localizes the out-of-sample enhancement, and yet improves generalization.

3.2.3. *Mahalanobis Metrics on a Non-Linear Feature Space.* In this section, we consider the case in which the cost function is defined after applying a non-linear transformation, $\Phi : R^d \to R^l$, to the data. Assume that the data takes the form $\mathcal{D}_n = \{(X_i, Y_i)\}_{i=1}^n$, where $X_i \in \mathbb{R}^d$ and $Y_i \in \mathbb{R}$ and the loss function, depending on decision variable $\beta \in R^m$, is given by $l(x, y, \beta)$. Once again, motivated by the cost function (4), we may define

$$c_\Lambda^\Phi\left((x,y),(x',y')\right) = d_\Lambda^2\left(\Phi(x), \Phi\left(x'\right)\right) I\left(y = y'\right) + \infty I\left(y \neq y'\right), \tag{11}$$

for $\Lambda \in PSD$. To preserve the properties of a cost function (i.e. non-negativity, lower semicontinuity and $c_\Lambda^\Phi(u, w) = 0$ implies $u = w$), we assume that $\Phi(\cdot)$ is continuous and that $\Phi(w) = \Phi(u)$ implies that $w = u$. Then we can apply a metric learning procedure, such as the one described in (8), to calibrate Λ. The intuition is the same as the one provided in the linear case in Section 3.2.2.

4. Data Driven Cost Selection and Adaptive Regularization

In this section we establish a direct connection between our fully data-driven DRO procedure and adaptive regularization. Moreover, our main result here, together with our discussion from the previous section, provides a direct connection between the metric learning literature and adaptive regularized estimators. As a consequence, the methods from the metric learning literature can be used to estimate the parameter of adaptively regularized estimators.

Throughout this section we consider again a data set of the form $\mathcal{D}_n = \{(X_i, Y_i)\}_{i=1}^n$ with $X_i \in \mathbb{R}^d$ and $Y_i \in \mathbb{R}$. Motivated by the cost function (4) we define the cost function $c_\Lambda(\cdot)$ as in (9). Using (9) we obtain the following result, which is proved in the appendix.

Theorem 1 (DRO Representation for Generalized Adaptive Regularization)**.** *Assume that $\Lambda \in R^{d\times d}$ in (9) is positive definite. Given the data set $\mathcal{D}_n$, we obtain the following representation*

$$\min_\beta \max_{P: D_{c_\Lambda}(P, P_n) \leq \delta} \mathbb{E}_P^{1/2}\left[\left(Y - X^T\beta\right)^2\right] = \min_\beta \sqrt{\frac{1}{n}\sum_{i=1}^n \left(Y_i - X_i^T\beta\right)^{1/2}} + \sqrt{\delta}\,\|\beta\|_{\Lambda^{-1}}\,. \tag{12}$$

Moreover, if $Y \in \{-1, +1\}$ in the context of adaptive regularized logistic regression, we obtain the following representation

$$\min_\beta \max_{P: D_{c_\Lambda}(P, P_n) \leq \delta} \mathbb{E}\left[\log\left(1 + e^{-Y(X^T\beta)}\right)\right] = \min_\beta \frac{1}{n}\sum_{i=1}^n \log\left(1 + e^{-Y_i(X_i^T\beta)}\right) + \delta\,\|\beta\|_{\Lambda^{-1}}\,. \tag{13}$$

In order to recover a more familiar setting in adaptive regularization, assume that Λ is a diagonal positive definite matrix. In which case we obtain, in the setting of (12),

$$\min_\beta \max_{P: D_{c_\Lambda}(P, P_n) \leq \delta} \mathbb{E}_P^{1/2}\left[\left(Y - X^T\beta\right)^2\right] = \min_\beta \sqrt{\frac{1}{n}\sum_{i=1}^n \left(Y_i - X_i^T\beta\right)^2} + \sqrt{\delta}\sqrt{\sum_{i=1}^d \beta_i^2/\Lambda_{ii}}. \tag{14}$$

The adaptive regularization method was first derived as a generalization for ridge regression in Hoerl and Kennard (1970b) and Hoerl and Kennard (1970a). Recent work shows that adaptive regularization can improve the predictive power of its non-adaptive counterpart, specially in high-dimensional settings (see in Zou (2006) and Ishwaran and Rao (2014)).

In view of (14), our discussion in Section 3.2.1 uncovers tools which can be used to estimate the coefficients $\{1/\Lambda_{ii} : 1 < i \leq d\}$ using the connection to metric learning procedures. To complement the intuition given in Figure 1(b), note that in the adaptive regularization literature one often choose $\Lambda_{ii} \approx 0$ to induce $\beta_i \approx 0$ (i.e., there is a high penalty to variables with low explanatory power). This, in our setting, would correspond to transport costs which are low in such low explanatory directions.

5. Solving Data Driven DRO Based on Optimal Transport Discrepancies

In order to fully take advantage of the combination synergies between metric learning methodology and our DRO formulation, it is crucial to have a methodology which allows us to efficiently estimate β in DRO problems such as (1). In the presence of a simplified representation such as (2) or (14), we can apply standard stochastic optimization results (see Lei and Jordan (2016)).

Our objective in this section is to study algorithms which can be applied for more general loss and cost functions, for which a simplified representation might not be accessible.

Throughout this section, once again we assume that the data is given in the form $\mathcal{D}_n = \{(X_i, Y_i)\}_{i=1}^n \subset \mathbb{R}^{d+1}$. The loss function is written as $\{l(x, y, \beta) : (x, y) \in \mathbb{R}^{d+1}, \beta \in \mathbb{R}^m\}$. We assume that for each (x, y), the function $l(x, y, \cdot)$ is convex and continuously differentiable. Further, we shall consider cost functions of the form

$$\bar{c}\left((x, y), (x', y')\right) = c(x, x') I(y = y') + \infty I(y \neq y'),$$

as this will simplify the form of the dual representation in the inner optimization of our DRO formulation. To ensure boundedness of our DRO formulation, we impose the following assumption.

Assumption 1. There exists $\Gamma(\beta, y) \in (0, \infty)$ such that $l(u, y, \beta) \leq \Gamma(\beta, y) \cdot (1 + c(u, x))$, for all $(x, y) \in \mathcal{D}_n$, Under Assumption 1, we can guarantee that

$$\max_{P: D_c(P, P_n) \leq \delta} \mathbb{E}_P [l(X, Y, \beta)] \leq (1 + \delta) \max_{i=1,\ldots,n} \Gamma(\beta, Y_i) < \infty.$$

Using the strong duality theorem for semi-infinity linear programming problem in Appendix B of Blanchet et al. (2016),

$$\max_{P: D_c(P, P_n) \leq \delta} \mathbb{E}_P [l(X, Y, \beta)] = \min_{\lambda \geq 0} \frac{1}{n} \sum_{i=1}^n \phi(X_i, Y_i, \beta, \lambda), \tag{15}$$

where $\psi(u, X, Y, \beta, \lambda) := l(u, Y, \beta) - \lambda(c(u, X) - \delta)$, $\phi(X, Y, \beta, \lambda) := \max_{u \in \mathbb{R}^d} \psi(u, X, Y, \beta, \lambda)$. Therefore,

$$\min_{\beta} \max_{P: D_{c_\Lambda}(P, P_n) \leq \delta} \mathbb{E}_P [l(X, Y, \beta)] = \min_{\lambda \geq 0, \beta} \{\mathbb{E}_{P_n} [\phi(X, Y, \beta, \lambda)]\}. \tag{16}$$

The optimization in (16) is minimize over β and λ, which we can consider stochastic approximation algorithm if the gradient of $\phi(\cdot)$ with respect to β and λ exist. However, $\phi(\cdot)$ is given in the form of the value function of a maximization problem, of which the gradient is not easy accessible. We will discuss the detailed algorithm and the validity of the smoothing approximation below.

We consider a smoothing approximation technique to remove the maximization problem $\phi(\cdot)$ using soft-max counterpart, $\phi_{\epsilon,f}(\cdot)$. The smoothing soft-max approximation has been explored and applied to approximately solve the DRO problem for the discrete case, where we restrict the distributionally uncertainty set only contains probability measures support on finite set (i.e., labeled training data and unlabeled training data with pseudo labels), we refer Blanchet and Kang (2017) for further details.

However, due to the continuous-infinite support constraint, the soft-max approximation is a non-trivial generalization of the finite-discrete analogue. The smoothing approximation for $\phi(\cdot)$ is defined as,

$$\phi_{\epsilon,f}(X, Y, \beta, \lambda) = \epsilon \log \left(\int_{\mathbb{R}^d} \exp\left([\psi(u, X, Y, \beta, \lambda)] / \epsilon \right) f(u) \, du \right),$$

where $f(\cdot)$ is a probability density in $\mathbb{R}^d$; for example, we can consider a multivariate normal distribution and ϵ is a small positive number regarded as smoothing parameter.

Theorem 2 below allows to quantify the error due to smoothing approximation.

Theorem 2. *Under mild technical assumptions (see Assumption 1-4 in Appendix B), there exists* $\epsilon_0 > 0$ *such that for every* $\epsilon < \epsilon_0$*, we have*

$$\phi(X, Y, \beta, \lambda) \geq \phi_{\epsilon,f}(X, Y, \beta, \lambda) \geq \phi(X, Y, \beta, \lambda) - d\epsilon \log(1/\epsilon)$$

The proof of Theorem 2 is given in Appendix B.

After applying smooth approximation, the optimization problem turns into a standard stochastic optimization problem and we can apply mini-batch based stochastic approximation (SA) algorithm to solve it. As we can notice, as a function and β and λ, the gradient of $\phi_{\epsilon,f}(\cdot)$ satisfies

$$\nabla_\beta \phi_{\epsilon,f}(X, Y, \beta, \lambda) = \frac{\mathbb{E}_{U\sim f}\left[\exp\left(\psi\left(U, X, Y, \beta, \lambda\right)/\epsilon\right) \nabla_\beta l\left(f_\beta\left(U\right), Y\right)\right]}{\mathbb{E}_{U\sim f}\left[\exp\left(\psi\left(U, X, Y, \beta, \lambda\right)/\epsilon\right)\right]},$$

$$\nabla_\lambda \phi_{\epsilon,f}(X, Y, \beta, \lambda) = \frac{\mathbb{E}_{U\sim f}\left[\exp\left(\psi\left(u, X, Y, \beta, \lambda\right)/\epsilon\right) \left(\delta - c_{\mathcal{D}_n}\left(u, X\right)\right)\right]}{\mathbb{E}_{U\sim f}\left[\exp\left(\psi\left(U, X, Y, \beta, \lambda\right)/\epsilon\right)\right]}.$$

However, since the gradients are still given in the form of expectation, we can apply a simple Monte Carlo sampling algorithm to approximate the gradient, i.e., we sample U_i's from $f(\cdot)$ and evaluate the numerators and denominators of the gradient using Monte Carlo separately. For more details of the SA algorithm, please see in Algorithm 1.

Algorithm 1 Stochastic Gradient Descent with Continuous State

1: **Initialize** $\lambda = 0$, and β to be empirical risk minimizer, $\epsilon = 0.5$, tracking error $Error = 100$.
2: **while** $Error > 10^{-3}$ **do**
3: **Sample** a mini-batch uniformly from observations $\left\{X_{(j)}, Y_{(j)}\right\}_{j=1}^{M}$, with $M \leq n$.
4: For each $j = 1, \ldots, M$, sample i.i.d. $\{U_k^{(j)}\}_{k=1}^{L}$ from $\mathcal{N}\left(0, \sigma^2 I_{d\times d}\right)$.
5: We denote f_L^j as empirical distribution for $U_k^{(j)}$'s, and estimate the batched as

$$\nabla_\beta \phi_{\epsilon,f} = \frac{1}{M}\sum_{j=1}^{M} \nabla_\beta \phi_{\epsilon,f_L^j}\left(X_{(j)}, Y_{(j)}, \beta, \lambda\right), \nabla_\lambda \phi_{\epsilon,f} = \frac{1}{M}\sum_{j=1}^{M} \nabla_\lambda \phi_{\epsilon,f_L^j}\left(X_{(j)}, Y_{(j)}, \beta, \lambda\right).$$

6: Update β and λ using $\beta = \beta + \alpha_\beta \nabla_\beta \phi_{\epsilon,f}$ and $\lambda = \lambda + \alpha_\lambda \nabla_\lambda \phi_{\epsilon,f}$.
7: Update tracking error $Error$ as the norm of difference between latest parameter and average of last 50 iterations.
8: **Output** β.

6. Numerical Experiments

We validate our data-driven cost function based DRO using 5 real data examples from the UCI machine learning database Lichman (2013). We focus on a DRO formulation based on the log-exponential loss for a linear model. We use the linear metric learning framework explained in equation (8), which then we feed into the cost function, c_Λ, as in (9), denoting by DRO-L. In addition, we also fit a cost function c_Λ^Φ, as explained in (11) using linear and quadratic transformations of the data; the outcome is denote as (DRO-NL). We compare our DRO-L and DRO-NL with logistic regression (LR), and regularized logistic regression (LRL1). For each iteration and each data set, the data is split randomly into training and test sets. We fit the models on the training and evaluate the performance on test set. The regularization parameter is chosen via $5-$fold cross-validation for LRL1, DRO-L and DRO-NL. We report the mean and standard deviation for training and testing log-exponential error and testing accuracy for 200 independent experiments for each data set. The details of the numerical results and basic information of the data is summarized in Table 1.

		BC	BN	QSAR	Magic	MB	SB
LR	Train	0 ± 0	$.008 \pm .003$	$.026 \pm .008$	$.213 \pm .153$	0 ± 0	0 ± 0
	Test	8.75 ± 4.75	2.80 ± 1.44	35.5 ± 12.8	17.8 ± 6.77	18.2 ± 10.0	14.5 ± 9.04
	Accur	$.762 \pm .061$	$.926 \pm .048$	$.701 \pm .040$	$.668 \pm .042$	$.678 \pm .059$	$.789 \pm .035$
LRL1	Train	$.185 \pm .123$	$.080 \pm .030$	$.614 \pm .038$	$.548 \pm .087$	$.401 \pm .167$	$.470 \pm .040$
	Test	$.428 \pm .338$	$.340 \pm .228$	$.755 \pm .019$	$.610 \pm .050$	$.910 \pm .131$	$.588 \pm .140$
	Accur	$.929 \pm .023$	$.930 \pm .042$	$.646 \pm .036$	$.665 \pm .045$	$.717 \pm .041$	$.811 \pm .034$
DRO-L	Train	$.022 \pm .019$	$.197 \pm .112$	$.402 \pm .039$	$.469 \pm .064$	$.294 \pm .046$	$.166 \pm .031$
	Test	$.126 \pm .034$	$.275 \pm .093$	$.557 \pm .023$	$.571 \pm .043$	$.613 \pm .053$	$.333 \pm .018$
	Accur	$.954 \pm .015$	$.919 \pm .050$	$.733 \pm .0.026$	$.727 \pm .039$	$.714 \pm .032$	$.887 \pm .011$
DRO-NL	Train	$.032 \pm .015$	$.113 \pm .035$	$.339 \pm .044$	$.381 \pm .084$	$.287 \pm .049$	$.195 \pm .034$
	Test	$.119 \pm .044$	$.194 \pm .067$	$.554 \pm .032$	$.576 \pm ..049$	$.607 \pm .060$	$.332 \pm .015$
	Accur	$.955 \pm .016$	$.931 \pm .036$	$.736 \pm .027$	$.730 \pm .043$	$.716 \pm .054$	$.889 \pm .009$
Num Predictors		30	4	30	10	20	56
Train Size		40	20	80	30	30	150
Test Size		329	752	475	9990	125034	2951

TABLE 1. Numerical results for real data sets.

7. Conclusion and Discussion

Our fully data-driven DRO procedure combines a semiparametric approach (i.e. the metric learning procedure) with a parametric procedure (expected loss minimization) to enhance the generalization performance of the underlying parametric model. We emphasize that our approach is applicable to any DRO formulation and is not restricted to classification tasks. An interesting research avenue that might be considered include the development of a semisupervised framework as in Blanchet and Kang (2017), in which unlabeled data is used to inform the support of the elements in $\mathcal{U}_\delta(P_n)$.

References

[1] Bellet, A., Habrard, A., and Sebban, M. (2013). A survey on metric learning for feature vectors and structured data. *arXiv preprint arXiv:1306.6709*.

[2] Blanchet, J. and Kang, Y. (2017). Distributionally robust semi-supervised learning. *arXiv preprint arXiv:1702.08848*.

[3] Blanchet, J., Kang, Y., and Murthy, K. (2016). Robust wasserstein profile inference and applications to machine learning. *arXiv preprint arXiv:1610.05627*.

[4] Esfahani, P. M. and Kuhn, D. (2015). Data-driven distributionally robust optimization using the wasserstein metric: Performance guarantees and tractable reformulations. *arXiv preprint arXiv:1505.05116*.

[5] Friedman, J., Hastie, T., and Tibshirani, R. (2001). *The elements of statistical learning*, volume 1. Springer series in statistics Springer, Berlin.

[6] Hoerl, A. E. and Kennard, R. W. (1970a). Ridge regression: applications to nonorthogonal problems. *Technometrics*, 12(1):69–82.

[7] Hoerl, A. E. and Kennard, R. W. (1970b). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67.

[8] Hu, Z. and Hong, L. J. (2013). Kullback-leibler divergence constrained distributionally robust optimization. *Available at Optimization Online*.

[9] Ishwaran, H. and Rao, J. S. (2014). Geometry and properties of generalized ridge regression in high dimensions. *Contemp. Math*, 622:81–93.

[10] Lei, L. and Jordan, M. I. (2016). Less than a single pass: Stochastically controlled stochastic gradient method. *arXiv preprint arXiv:1609.03261*.
[11] Li, L., Sun, C., Lin, L., Li, J., and Jiang, S. (2016). A mahalanobis metric learning-based polynomial kernel for classification of hyperspectral images. *Neural Computing and Applications*, pages 1–11.
[12] Lichman, M. (2013). UCI machine learning repository.
[13] Schultz, M. and Joachims, T. (2004). Learning a distance metric from relative comparisons. In *Advances in neural information processing systems*, pages 41–48.
[14] Shafieezadeh-Abadeh, S., Esfahani, P. M., and Kuhn, D. (2015). Distributionally robust logistic regression. In *Advances in Neural Information Processing Systems*, pages 1576–1584.
[15] Villani, C. (2008). *Optimal transport: old and new*, volume 338. Springer Science & Business Media.
[16] Xing, E. P., Ng, A. Y., Jordan, M. I., and Russell, S. (2002). Distance metric learning with application to clustering with side-information. In *NIPS*, volume 15, page 12.
[17] Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American statistical association*, 101(476):1418–1429.

Appendix A. Proof of Theorem 1

We first state and prove Lemma 1 which will be useful in proving Theorem 1.

Lemma 1. *If Λ is a is positive definite matrix and we define $\|x\|_\Lambda = \left(x^T \Lambda x\right)^{1/2}$, then $\|\cdot\|_{\Lambda^{-1}}$ is the dual norm of $\|\cdot\|_\Lambda$. Furthermore, we have*

$$u^T w \leq \|u\|_\Lambda \|w\|_{\Lambda^{-1}},$$

where the equality holds if and only if, there exists non-negative constant τ, s.t $\tau \Lambda u = \Lambda^{-1} w$ or $\tau \Lambda^{-1} w = \Lambda u$.

Proof for Lemma 1. This result is a direct generalization of l_2 norm in Euclidean space. Note that

$$u^T w = (\Lambda u)^T \left(\Lambda^{-1} w\right) \leq \|\Lambda u\|_2 \left\|\Lambda^{-1} w\right\|_2 = \|u\|_\Lambda \|w\|_{\Lambda^{-1}}. \tag{17}$$

The inequality in the above is Cauchy-Schwartz inequality for $\mathbb{R}^d$ applies to Λu and $\Lambda^{-1} w$, and the equality holds if and only if there exists nonnegative τ, s.t. $\tau \Lambda u = \Lambda^{-1} w$ or $\tau \Lambda^{-1} w = \Lambda u$. Now, by the definition of the dual norm,

$$\|w\|_\Lambda^* = \sup_{u:\|u\|_\Lambda \leq 1} u^T w = \sup_{u:\|u\|_\Lambda \leq 1} \|u\|_\Lambda \|w\|_{\Lambda^{-1}} = \|w\|_{\Lambda^{-1}}.$$

While the first equality follows from the definition of dual norm, the second equality is due to Cauchy-Schwartz inequality (17), and the equality condition therein, and the last equality are immediate after maximizing. □

Proof for Theorem 1. The technique is a generalization of the method used in proving Theorem 1 in Blanchet et al. (2016). We can apply the strong duality result, see Proposition 6 in Appendix of Blanchet et al. (2016), for worst-case expected loss function, which is a semi-infinite linear programming problem, to obtain

$$\sup_{P: D_{c_\Lambda}(P, P_n) \leq \delta} \mathbb{E}_P \left[\left(Y - X^T \beta\right)^2\right] = \min_{\gamma \geq 0} \left\{ \gamma \delta - \frac{1}{n} \sum_{i=1}^n \sup_u \left\{ \left(y_i - u^T \beta\right)^2 - \gamma \|x_i - u\|_\Lambda^2 \right\} \right\}.$$

For the inner suprema , let us denote $\Delta = u - X_i$ and $e_i = Y_i - X_i^T\beta$ for notation simplicity. The inner optimization problem associated with (X_i, Y_i) becomes,

$$\begin{aligned}
&\sup_u \left\{ \left(y_i - u^T\beta\right)^2 - \gamma \left\|x_i - u\right\|_\Lambda^2 \right\} \\
&\quad = e_i^2 + \sup_\Delta \left\{ \left(\Delta^T\beta\right)^2 - 2e_i\Delta^T\beta - \gamma \left\|\Delta\right\|_\Lambda^2 \right\}, \\
&\quad = e_i^2 + \sup_\Delta \left\{ \left(\sum_j \left|\Delta_j\right| \left|\beta_j\right| \right)^2 + 2\left|e_i\right| \sum_j \left|\Delta_j\right| \left|\beta_j\right| - \gamma \left\|\Delta\right\|_\Lambda^2 \right\}, \\
&\quad = e_i^2 + \sup_{\left\|\Delta\right\|_\Lambda} \left\{ \left\|\Delta\right\|_\Lambda^2 \left\|\beta\right\|_{\Lambda^{-1}}^2 + 2\left|e_i\right| \left\|\Delta\right\|_\Lambda \left\|\beta\right\|_{\Lambda^{-1}} - \gamma \left\|\Delta\right\|_\Lambda^2 \right\}, \\
&\quad = \begin{cases} e_i^2 \frac{\gamma}{\gamma - \left\|\beta\right\|_{\Lambda^{-1}}^2} & \text{if } \gamma > \left\|\beta\right\|_{\Lambda^{-1}}^2, \\ +\infty & \text{if } \gamma \le \left\|\beta\right\|_{\Lambda^{-1}}^2. \end{cases}
\end{aligned}$$

While the first equality is due to the change of variable, the second equality is because we are working on a maximization problem, and the last term only depends on the magnitude rather than sign of Δ, thus the optimization problem will always pick Δ that satisfying the equality. Considering the third equality, for the optimization problem, we can first apply the Cauchy-Schwartz inequality in Lemma 1 and we know that the maximization problem is to take Δ satisfying the equality constraint. For the last equality, if $\gamma \le \left\|\beta\right\|_{\Lambda^{-1}}^2$, the optimization problem is unbounded, while $\gamma > \left\|\beta\right\|_{\Lambda^{-1}}^2$, we can solve the quadratic optimization problem and it leads to the final equality.
For the outer minimization problem over γ, as the inner suprema equal infinity if $\gamma \le \left\|\beta\right\|_{\Lambda^{-1}}^2$, the worst-case expected loss becomes,

$$\begin{aligned}
&\sup_{P: D_{c_{\mathcal{D}_n}}(P, P_n) \le \delta} \mathbb{E}_P \left[\left(Y - X^T\beta\right)^2 \right] \\
&\quad = \min_{\gamma > \left\|\beta\right\|_{\alpha\text{-}(p,s)}^2} \left\{ \gamma\delta - \frac{1}{n} \sum_{i=1}^n \left(Y_i - X_i^T\beta\right) \frac{\gamma}{\gamma - \left\|\beta\right\|_{\Lambda^{-1}}^2} \right\}, \\
&\quad = \left(\sqrt{\frac{1}{n} \sum_{i=1}^n \left(Y_i - X_i^T\beta\right)} + \sqrt{\delta} \left\|\beta\right\|_{\Lambda^{-1}} \right)^2.
\end{aligned} \tag{18}$$

The first equality follows the discussion above for restricting $\gamma > \left\|\beta\right\|_{\Lambda^{-1}}^2$. We can observe that the objective function in the right hand side of (18) is convex and differentiable and as $\gamma \to \infty$ and $\gamma \to \left\|\beta\right\|_\Lambda^2$, the value function will be infinity. We know the optimizer could be uniquely characterized via first order optimality condition. Solving for γ in this way (through first order optimality), it is straightforward to obtain the last equality in (18). If we take square root on both sides, we prove the claim for linear regression.
For the log-exponential loss function, the proof follows a similar strategy. By applying strong duality results for semi-infinity linear programming problem in Blanchet et al. (2016), we can write the worst case expected loss function as,

$$\begin{aligned}
&\sup_{P: D_{c_{\mathcal{D}_n}}(P, P_n) \le \delta} \mathbb{E}_P \left[\log\left(1 + \exp\left(-Y\beta^T X\right)\right) \right] \\
&= \min_{\gamma \ge 0} \left\{ \gamma\delta - \frac{1}{n} \sum_{i=1}^n \sup_u \left\{ \log\left(1 + \exp\left(-Y_i\beta^T u\right)\right) - \gamma \left\|X_i - u\right\|_\Lambda \right\} \right\}.
\end{aligned}$$

For each i, we can apply Lemma 1 in Shafieezadeh-Abadeh et al. (2015) and dual-norm result in Lemma 1 to deal with the inner optimization problem. It gives us,

$$\sup_u \left\{\log\left(1+\exp\left(-Y_i\beta^T u\right)\right) - \gamma\left\|X_i-u\right\|_\Lambda\right\} = \begin{cases} \log\left(1+\exp\left(-Y_i\beta^T X_i\right)\right) & \text{if} \quad \|\beta\|_{\Lambda^{-1}} \le \gamma, \\ \infty & \text{if} \quad \|\beta\|_{\Lambda^{-1}} > \gamma. \end{cases}$$

Moreover, since the outer optimization is trying to minimize, following the same discussion for the proof for linear regression case, we can plug-in the result above and it leads the first equality below,

$$\begin{aligned} &\min_{\gamma\ge 0}\left\{\gamma\delta - \frac{1}{n}\sum_{i=1}^n \sup_u \left\{\log\left(1+\exp\left(-Y_i\beta^T u\right)\right) - \gamma\left\|X_i-u\right\|_\Lambda\right\}\right\} \\ &\quad = \min_{\gamma\ge\|\beta\|_{\Lambda^{-1}}}\left\{\delta\gamma + \frac{1}{n}\sum_{i=1}^n \log\left(1+\exp\left(-Y_i\beta^T X_i\right)\right)\right\} \\ &\quad = \frac{1}{n}\sum_{i=1}^n \log\left(1+\exp\left(-Y_i\beta^T X_i\right)\right) + \delta\left\|\beta\right\|_{\Lambda^{-1}}. \end{aligned}$$

We know that the target function is continuous and monotone increasing in γ, thus we can notice it is optimized by taking $\gamma = \|\beta\|_{\Lambda^{-1}}$, which leads to second equality above. This proves the claim for logistic regression in the statement of the theorem. □

Appendix B. Proof of Theorem 2

Let us begin by listing the assumptions required to prove Theorem 2. First, we begin by recalling Assumption 1 from Section 5.

Assumption 1. There exists $\Gamma(\beta, y) \in (0,\infty)$ such that $l(u, y, \beta) \le \Gamma(\beta, y)\cdot(1 + c(u, x))$, for all $(x, y) \in \mathcal{D}_n$,

We now introduce Assumptions 2-4 below.

Assumption 2. $\psi(\cdot, X, Y, \beta, \lambda)$ is twice continuously differentiable and the Hessian of $\psi(\cdot, X, Y, \beta, \lambda)$ evaluated at u^*, $D_u^2\psi(u^*, X, Y, \beta, \lambda)$, is positive definite. In particular, we can find $\theta > 0$ and $\eta > 0$, such that

$$\psi(u, X, Y, \beta, \lambda) \ge \psi(u^*, X, Y, \beta, \lambda) - \frac{\theta}{2}\|u-u^*\|_2^2, \quad \forall u \text{ with } \|u-u^*\|_2 \le \eta.$$

Assumption 3. For a constant $\lambda_0 > 0$ such that $\phi(X, Y, \beta, \lambda_0) < \infty$, let $K = K(X, Y, \beta, \lambda_0)$ be any upper bound for $\phi(X, Y, \beta, \lambda_0)$.

Assumption 4. In addition to the lower semicontinuity of $c(\cdot) \ge 0$, we assume that $c(\cdot, X)$ is coercive in the sense that $c(u, X) \to \infty$ as $\|u\|_2 \to \infty$.

For any set S, the r-neighborhood of S is defined as the set of all points in $\mathbb{R}^d$ that are at distance less than r from S, i.e. $S_r = \cup_{u\in S}\{\bar{u} : \|\bar{u}-u\|_2 \le r\}$.

Proof of Theorem 2. The first part of the inequality is easy to derive. For the second part, we proceed as follows: Under Assumptions 3 and 4, we can define the compact set

$$\mathcal{C} = \mathcal{C}(X, Y, \beta, \lambda) = \{u : c(u, X) \le l(X, Y, \beta) - K + \lambda_0/(\lambda - \lambda_0)\}.$$

It is easy to check that $\arg\max\{\psi(u, X, Y, \lambda)\} \subset \mathcal{C}$. Owing to optimality of u^* and from Assumption 2 that $K \ge \phi(X, Y, \beta, \lambda_0)$, we can see that

$$\begin{aligned} l(X, Y) &\le l(u^*, Y)) - \lambda c(u, X) \\ &= l(u^*, Y) - \lambda_0 c(u^*, X) - (\lambda - \lambda_0)c(u^*, X) \\ &\le K - \lambda_0 - (\lambda - \lambda_0)c(u^*, X). \end{aligned}$$

Thus by definition of $\mathcal{C} = \mathcal{C}(X, Y, \beta, \lambda)$, it follows easily that $u^* \in \mathcal{C}$, which further implies $\{u : \|u - u^*\|_2 \leq \eta\} \subset \mathcal{C}_\eta$. Then we combine the strongly convexity assumption in Assumption 2 and the definition of $\phi_{\epsilon,f}(u, X, Y, \beta, \lambda)$, which yields

$$\begin{aligned}
\phi_{\epsilon,f}(X, Y, \beta, \lambda) &\geq \epsilon \log \left(\int_{\|u-u^*\|_2 \leq \eta} \exp\left(\left[\phi(X, Y, \beta, \lambda) - \frac{\theta}{2}\|u - u^*\|_2^2 \right] / \epsilon \right) f(u) du \right) \\
&= \epsilon \log \left(\exp\left(\phi(X, Y, \beta, \lambda) / \epsilon \right) \right) \int_{\|u-u^*\|_2 \leq \eta} \exp\left(-\frac{\theta}{2}\|u - u^*\|_2^2 / \epsilon \right) f(u) du \\
&= \phi(X, Y, \beta, \lambda) + \epsilon \log \int_{\|u-u^*\|_2 \leq \eta} \exp\left(-\frac{\theta \|u - u^*\|_2^2}{2\epsilon} \right) f(u) du.
\end{aligned}$$

As $\{u : \|u - u^*\|_2 \leq \eta\} \subset \mathcal{C}_\eta$, we can use the lower bound of $f(\cdot)$ to deduce that

$$\begin{aligned}
\int_{\|u-u^*\|_2 \leq \eta} \exp\left(-\frac{\theta \|u - u^*\|_2^2}{2\epsilon} \right) f(u) du &\geq \inf_{u \in \mathcal{C}_\eta} f(u) \times \int_{\|u-u^*\|_2 \leq \eta} \exp\left(-\frac{\theta \|u - u^*\|_2^2}{2\epsilon} \right) du \\
&= \inf_{u \in \mathcal{C}_\eta} f(u) \times (2\pi\epsilon/\theta)^{d/2} P(Z_d \leq \eta^2 \theta / \epsilon),
\end{aligned}$$

where Z_d is a chi-squared random variable of d degrees of freedom. To conclude, recall that $\epsilon \in (0, \eta^2 \theta \chi_\alpha)$, the lower bound of $\phi_{\epsilon,f}(\cdot)$ can be written as

$$\phi_{\epsilon,f}(X, Y, \beta, \lambda) \geq \phi(X, Y, \beta, \lambda) - \frac{d}{2}\epsilon \log(1/\epsilon) + \frac{d}{2}\epsilon \log\left((2\pi\alpha/\theta) \inf_{u \in \mathcal{C}_\eta} f(u) \right).$$

This completes the proof of Theorem 2. □

Columbia University, Department of Statistics and Department of Industrial Engineering & Operations Research, New York, NY 10027, United States.
E-mail address: jose.blanchet@columbia.edu

Columbia University, Department of Statistics. New York, NY 10027, United States.
E-mail address: yang.kang@columbia.edu

Columbia University, Department of Industrial Engineering & Operations Research. New York, NY 10027, United States.
E-mail address: fz2222@columbia.edu

Columbia University, Department of Industrial Engineering & Operations Research,Mudd Building, 500 W. 120 Street, New York, NY 10027, United States.
E-mail address: karthyek.murthy@columbia.edu