Proceedings of Machine Learning Research 77:97–112, 2017 ACML 2017

Distributionally Robust Groupwise Regularization Estimator

Jose Blanchet JOSE.BLANCHET@COLUMBIA.EDU

Yang Kang YANG.KANG@COLUMBIA.EDU

1255 Amsterdam Ave. Department of Statistics, Columbia University, New York, NY, USA. 10027.

Editors: Yung-Kyun Noh and Min-Ling Zhang

Abstract

Regularized estimators in the context of group variables have been applied successfully in model and feature selection in order to preserve interpretability. We formulate a Distributionally Robust Optimization (DRO) problem which recovers popular estimators, such as Group Square Root Lasso (GSRL). Our DRO formulation allows us to interpret GSRL as a game, in which we learn a regression parameter while an adversary chooses a perturbation of the data. We wish to pick the parameter to minimize the expected loss under any plausible model chosen by the adversary - who, on the other hand, wishes to increase the expected loss. The regularization parameter turns out to be precisely determined by the amount of perturbation on the training data allowed by the adversary. In this paper, we introduce a data-driven (statistical) criterion for the optimal choice of regularization, which we evaluate asymptotically, in closed form, as the size of the training set increases. Our easy-to-evaluate regularization formula is compared against cross-validation, showing comparable performance.

Keywords: Distributionally Robust Optimization, Group Lasso, Optimal Transport.

1. Introduction

Group Lasso (GR-Lasso) estimator is a generalization of the Lasso estimator (see Tibshirani (1996)). The method focuses on variable selection in settings where some predictive variables, if selected, must be chosen as a group. For example, in the context of the use of dummy variables to encode a categorical predictor, the application of the standard Lasso procedure might result in the algorithm including only a few of the variables but not all of them, which could make the resulting model difficult to interpret. Another example, where the GR-Lasso estimator is particularly useful, arises in the context of feature selection. Once again, a particular feature might be represented by several variables, which often should be considered as a group in the variable selection process.

The GR-Lasso estimator was initially developed for the linear regression case (see Yuan and Lin (2006)), but a similar group-wise regularization was also applied to logistic regression in Meier et al. (2008). A brief summary of GR-Lasso technique type of methods can be found in Friedman et al. (2010).

Recently, Bunea et al. (2014) developed a variation of the GR-Lasso estimator, called the Group-Square-Root-Lasso (GSRL) estimator, which is very similar to the GR-Lasso estimator. The GSRL is to the GR-Lasso estimator what sqrt-Lasso, introduced in Belloni et al. (2011), is to the standard Lasso estimator. In particular, GSRL has a superior advantage over GR-Lasso, namely, that the regularization parameter can be chosen independently from the standard deviation of the regression error in order to guarantee the statistical consistency of the regression estimator (see Belloni et al. (2011), and Bunea et al. (2014)). Our contribution in this paper is to provide a DRO representation for the GSRL estimator, which is rich in interpretability and which provides insights to optimally select (using a natural criterion) the regularization parameter without the need of time-consuming cross-validation. We compute the optimal regularization choice (based on a simple formula we derive in this paper) and evaluate its performance empirically. We will show that our method for the regularization parameter is comparable, and sometimes superior, to cross-validation.

In order to describe our contributions more precisely, let us briefly describe the GSRL estimator. We choose the context of linear regression to simplify the exposition, but an entirely analogous discussion applies to the context of logistic regression.

Consider a given set of training data $\{(X_1, Y_1), \ldots, (X_n, Y_n)\}$. The input $X_i \in \mathbb{R}^d$ is a vector of d predicting variables, and $Y_i \in \mathbb{R}$ is the response variable. (Throughout the paper any vector is understood to be a column vector and the transpose of x is denoted by x^T.) We use (X, Y) to denote a generic sample from the training data set. It is postulated that

$$Y_i = X_i^T \beta^* + e_i,$$

for some $\beta^* \in \mathbb{R}^d$ and errors $\{e_1, ..., e_n\}$. Under suitable statistical assumptions (such as independence of the samples in the training data), one may be interested in estimating β^*. Underlying, we consider the square loss function, i.e. $l(x, y; \beta) = \left(y - \beta^T x\right)^2$, for the purpose of this discussion but this choice, as we shall see, is not necessary.

Throughout the paper we will assume the following group structure for the space of predictors. There are $\bar{d} \leq d$ mutually exclusive groups, which form a partition. More precisely, suppose that $G_1, \ldots, G_{\bar{d}}$ satisfies that $G_i \cap G_j = \varnothing$ for $i \neq j$, that $G_1 \cup ... \cup G_{\bar{d}} = \{1, ..., d\}$, and the G_i's are non-empty. We will use g_i to denote the cardinality of G_i and shall write G for a generic set in the partition and let g denote the cardinality of G.

We shall denote by $x(G) \in \mathbb{R}^g$ the sub-vector $x \in \mathbb{R}^d$ corresponding to G. So, if $G = \{i_1, ..., i_g\}$, then $x(G) = \left(X_{i_1}, \ldots, X_{i_g}\right)^T$.

Next, given $p, s \geq 1$, and $\alpha \in \mathbb{R}^{\bar{d}}_{++}$ (i.e. $\alpha_i > 0$ for $1 \leq i \leq \bar{d}$) we define for each $x \in \mathbb{R}^d$,

$$\|x\|_{\alpha\text{-}(p,s)} = \Big(\sum_{i=1}^{\bar{d}} \alpha_i^s \left\|x(G_i)\right\|_p^s\Big)^{1/s}, \tag{1}$$

where $\|x(G_i)\|_p$ denotes the p-norm of $x(G_i)$ in $\mathbb{R}^{g_i}$. (We will study fundamental properties of $\|x\|_{\alpha\text{-}(p,s)}$ as a norm in Proposition 5.)

Let P_n be the empirical distribution function, namely,

$$P_n(dx, dy) := \frac{1}{n}\sum_{i=1}^{n} \delta_{\{(X_i, Y_i)\}}(dx, dy).$$

Throughout the paper we use the notation $\mathbb{E}_P[\cdot]$ to denote expectation with respect to a probability distribution P.

The GSRL estimator takes the form

$$\min_{\beta}\Big[\frac{1}{n}\sum_{i=1}^{n} l\left(X_i,Y_i;\beta\right)\Big]^{1/2}+\lambda\left\|\beta\right\|_{\sqrt{\tilde{g}}-(2,1)}=\min_{\beta}\left(\mathbb{E}_{P_n}^{1/2}\left[l\left(X,Y;\beta\right)\right]+\lambda\left\|\beta\right\|_{\sqrt{\tilde{g}}-(2,1)}\right),$$

where λ is the so-called regularization parameter and $\sqrt{\tilde{g}}=(\sqrt{g_1},\ldots,\sqrt{g_{\bar{d}}})$ with $\sqrt{g_i}=card(G_i)$. The previous optimization problem can be easily solved using standard convex optimization techniques as explained in Belloni et al. (2011) and Bunea et al. (2014).
Our contributions in this paper can now be explicitly stated. We introduce a notion of discrepancy, $\mathcal{D}_c\left(P,P_n\right)$, discussed in Section 2, between P_n and any other probability measure P, such that

$$\min_{\beta}\max_{P:D_c(P,P_n)\leq\delta}\mathbb{E}_P^{1/2}\left[l\left(X,Y;\beta\right)\right]=\min_{\beta}\left(\mathbb{E}_{P_n}^{1/2}\left[l\left(X,Y;\beta\right)\right]+\delta^{1/2}\left\|\beta\right\|_{\alpha-(p,s)}\right). \tag{2}$$

Using this representation, which we formulate, together with its logistic regression analogue, in Section 2.2.1 and Section 2.2.2, we are able to draw the following insights:

I) GSRL can be interpreted as a game in which we choose a parameter (i.e. β) and an adversary chooses a "plausible" perturbation of the data (i.e. P); the parameter δ controls the degree in which P_n is allowed to be perturbed to produce P. The value of the game is dictated by the expected loss, under E_P, of the decision variable β.

II) The set $\mathcal{U}_\delta\left(P_n\right)=\{P:\mathcal{D}_c\left(P,P_n\right)\leq\delta\}$ denotes the set of distributional uncertainty. It represents the set of plausible variations of the underlying probabilistic model which are reasonably consistent with the data.

III) The DRO representation (2) exposes the role of the regularization parameter. In particular, because $\lambda=\delta^{1/2}$, we conclude that λ directly controls the size of the distributionally uncertainty and should be interpreted as the parameter which dictates the degree to which perturbations or variations of the available data should be considered.

IV) As a consequence of I) to III), the DRO representation (2) endows the GSRL estimator with desirable generalization properties. The GSRL aims at choosing a parameter, β, which should perform well for *all* possible probabilistic descriptions which are plausible given the data.

Naturally, it is important to understand what types of variations or perturbations are measured by the discrepancy $\mathcal{D}_c\left(P,P_n\right)$. For example, a popular notion of the discrepancy is the Kullback-Leibler (KL) divergence. However, KL divergence has the limitation that only considers probability distributions which are supported precisely on the available training data, and therefore potentially ignores plausible variations of the data which could have an adverse impact on generalization risk.
In the rest of the paper we answer the following questions. First, in Section 2 we explain the nature of the discrepancy $\mathcal{D}_c\left(P,P_n\right)$, which we choose as an Optimal Transport discrepancy. We will see that $\mathcal{D}_c\left(P,P_n\right)$ can be computed using a linear program.
Intuitively, $\mathcal{D}_c\left(P,P_n\right)$ represents the minimal transportation cost for moving the mass encoded by P_n into a sinkhole which is represented by P. The cost of moving mass from location $u=(x,y)$ to $w=(x',y')$ is encoded by a cost function $c\left(u,w\right)$ which we shall discuss and this will depend on the α-(p,s) norm that we defined in (1). The subindex c in $\mathcal{D}_c\left(P,P_n\right)$ represents the dependence on the chosen cost function.

The next item of interest is the choice of δ, again the discussion of items I) to III) of the DRO formulation (2) provides a natural way to optimally choose δ. The idea is that every model $P \in \mathcal{U}_\delta(P_n)$ should intuitively represent a plausible variation of P_n and therefore $\beta^P = \arg\min\{\mathbb{E}_P[l(X,Y;\beta)] : \beta\}$ is a plausible estimate of β^*. The set $\{\beta^P : P \in \mathcal{U}_\delta(P_n)\}$ therefore yields a confidence region for β^* which is increasing in size as δ increases. Hence, it is natural to minimize δ to guarantee a target confidence level (say 95%). In Section 3 we explain how this optimal choice can be asymptotically computed as $n \to \infty$.

Finally, it is of interest to investigate if the optimal choice of δ (and thus of λ) actually performs well in practice. We compare performance of our (asymptotically) optimal choice of λ against cross-validation empirically in Section 4. We conclude that our choice is quite comparable to cross validation.

Before we continue with the program that we have outlined, we wish to conclude this Introduction with a brief discussion of work related to the methods discussed in this paper. Connections between regularized estimators and robust optimization formulations have been studied in the literature. For example, the work of Xu et al. (2009) investigates determinist perturbations on the predictor variables to quantify uncertainty. In contrast, our DRO approach quantifies perturbations from the empirical measure. This distinction allows us to statistical theory which is key to optimize the size of the uncertainty, δ (and thus the regularization parameter) in a data-driven way. The work of Shafieezadeh-Abadeh et al. (2015) provides connections to regularization in the setting of logistic regression in an approximate form. More importantly, Shafieezadeh-Abadeh et al. (2015) propose a data-driven way to choose the size of uncertainty, δ, which is based on the concentration of measure results. The concentration of measure method depends on restrictive assumptions and leads to sub-optimal choices, which deteriorate poorly in high dimensions.

The present work is a continuation of the line of research development in Blanchet et al. (2016), which concentrates only on classical regularized estimators without the group structure. Our current contributions require the development of duality results behind the α-(p,t) norm which closely parallels that of the standard duality between l_p and l_q spaces (with $1/p + 1/q = 1$) and a number of adaptations and interpretations that are special to the group setting only.

2. Optimal Transport and DRO

2.1. Defining the optimal transport discrepancy

Let $c : \mathbb{R}^{d+1} \times \mathbb{R}^{d+1} \to [0,\infty]$ be lower semicontinuous and we assume that $c(u,w) = 0$ if and only if $u = w$. For reasons that will become apparent in the sequel, we will refer to $c(\cdot)$ as a cost function.

Given two distributions P and Q, with supports $\mathcal{S}_P \subseteq \mathbb{R}^{d+1}$ and $\mathcal{S}_Q \subseteq \mathbb{R}^{d+1}$, respectively, we define the optimal transport discrepancy, $\mathcal{D}_c$, via

$$\mathcal{D}_c(P,Q) = \inf_\pi \{E_\pi[c(U,W)] : \pi \in \mathcal{P}(\mathcal{S}_P \times \mathcal{S}_Q),\ \pi_U = P,\ \pi_W = Q\}, \tag{3}$$

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 π_U and π_W denote the marginals of U and W under π, respectively.

If, in addition, $c(\cdot)$ is symmetric (i.e. $c(u,w) = c(w,u)$), 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), it can be easily verified (see Villani (2008)) that $\mathcal{D}_c^{1/\varrho}(P,Q)$ is a metric on the probability measures. For example, if $c(u,w) = \|u-w\|_q^{\varrho}$ for $q \geq 1$ (where $\|u-w\|_q$ denotes the l_q norm in $\mathbb{R}^{d+1}$) then $\mathcal{D}_c(\cdot)$ is known as the Wasserstein distance of order ϱ.
Observe that (3) is obtained by solving a linear programming problem. For example, suppose that $Q = P_n$, so $\mathcal{S}_{P_n} = \{(X_i, Y_i)\}_{i=1}^n$, and let P be supported in some finite set $\mathcal{S}_P$ then, using $U = (X,Y)$, we have that $D_c(P,P_n)$ is obtained by computing

$$\begin{aligned} \{ \min_{\pi} & \sum_{u\in\mathcal{S}_P}\sum_{w\in\mathcal{S}_{P_n}} c(u,w)\,\pi(u,w) : \text{s.t.} \sum_{u\in\mathcal{S}_P} \pi(u,w) = \frac{1}{n}\ \forall\ w \in \mathcal{S}_{P_n} \\ & \sum_{w\in\mathcal{S}_{P_n}} \pi(u,w) = P(\{u\})\ \ \forall\ u\in\mathcal{S}_P, \pi(u,w) \geq 0\ \forall\ \ (u,w)\in\mathcal{S}_P\times\mathcal{S}_{P_n} \} \end{aligned} \tag{4}$$

A completely analogous linear program (LP), albeit an infinite dimensional one, can be defined if $\mathcal{S}_P$ has infinitely many elements. This LP has been extensively studied in great generality in the context of Optimal Transport under the name of Kantorovich's problem (see Villani (2008))).
Note that Kantorovich's problem is always feasible (take π with independent marginals, for example). Moreover, under our assumptions on c, if the optimal value is finite, then there is an optimal solution π^* (see Chapter 1 of Villani (2008))).
It is clear from the formulation of the previous LP that $\mathcal{D}_c(P,P_n)$ can be interpreted as the minimal cost of transporting mass from P_n to P, assuming that the marginal cost of transporting the mass from $u \in \mathcal{S}_P$ to $w \in \mathcal{S}_{P_n}$ is $c(u,w)$. It is also not difficult to realize from the assumption that $c(u,w) = 0$ if and only if $u = w$ that $\mathcal{D}_c(P,P_n) = 0$ if and only if $P = P_n$. We shall discuss, for instance, how to choose $c(\cdot)$ to recover (2) and the corresponding logistic regression formulation of GR-Lasso.

2.2. DRO Representation of GSRL Estimators

In this section, we will construct a cost function $c(\cdot)$ to obtain the GSRL (or GR-Lasso) estimators. We will follow an approach introduced in Blanchet et al. (2016) for the context of square-root Lasso (SR-Lasso) and regularized logistic regression estimators.

2.2.1. GSRL Estimators for Linear Regression

We start by assuming precisely the linear regression setup described in the Introduction and leading to (2). Given $\alpha = (\alpha_1, ..., \alpha_{\bar{d}})^T \in R_{++}^{\bar{d}}$ define $\alpha^{-1} = \left(\alpha_1^{-1}, ..., \alpha_{\bar{d}}^{-1}\right)^T$. Now, underlying there is a partition $G_1, ..., G_{\bar{d}}$ of $\{1, ..., d\}$ and given $q, t \in [1,\infty]$ we introduce the cost function

$$c\left((x,y),(x',y')\right) = \|x-x'\|_{\alpha^{-1}\text{-}(q,t)}^{\varrho} I_{y=y'} + \infty I_{y\neq y'}, \tag{5}$$

where, following (1), we have that

$$\|x-x'\|_{\alpha^{-1}\text{-}(q,t)}^{\varrho} = (\sum_{i=1}^{\bar{d}} \alpha_i^{-t} \|x(G_i) - x'(G_i)\|_q^t)^{\varrho/t}.$$

Then, we obtain the following result.

Theorem 1 (DRO Representation for Linear Regression GSRL) *Suppose that* $q, t \in [1, \infty]$ *and* $\alpha \in R_{++}^{\bar{d}}$ *are given and* $c(\cdot)$ *is defined as in (5) for* $\varrho = 2$. *Then, if* $l(x, y; \beta) = \left(y - x^T\beta\right)^2$ *we obtain*

$$\min_{\beta \in \mathbb{R}^d} \sup_{P: D_c(P, P_n) \le \delta} \left(\mathbb{E}_P\left[l\left(X, Y; \beta\right)\right]\right)^{1/2} = \min_{\beta \in \mathbb{R}^d} \left(E_{P_n}\left[l\left(X, Y; \beta\right)\right]\right)^{1/2} + \sqrt{\delta}\, \|\beta\|_{\alpha\text{-}(p,s)},$$

where $1/p + 1/q = 1$, *and* $1/s + 1/t = 1$.

We remark that choosing $p = q = 2$, $t = \infty$, $s = 1$, and $\alpha_i = \sqrt{g_i}$ for $i \in \{1, ..., \bar{d}\}$ we end up obtaining the GSRL estimator formulated in Bunea et al. (2014)).
We note that the cost function $c(\cdot)$ only allows mass transportation on the predictors (i.e X), but no mass transportation is allowed on the response variable Y. This implies that the GSRL estimator implicitly assumes that distributional uncertainty is only present on prediction variables (i.e. variations on the data only occurs through the predictors).

2.2.2. GR-Lasso Estimators for Logistic Regression

We now discuss GR-Lasso for classification problems. We consider a training data set of the form $\{(X_1, Y_1), \ldots, (X_n, Y_n)\}$. Once again, the input $X_i \in \mathbb{R}^d$ is a vector of d predictor variables, but now the response variable $Y_i \in \{-1, 1\}$ is a categorical variable. In this section we shall consider as our loss function the log-exponential function, namely,

$$l\left(x, y; \beta\right) = \log\left(1 + \exp\left(-y\beta^T x\right)\right). \tag{6}$$

This loss function is motivated by a logistic regression model which we shall review in the sequel. But for the DRO representation formulation it is not necessary to impose any statistical assumption. We then obtain the following theorem.

Theorem 2 (DRO Representation for Logistic Regression GR-Lasso) *Suppose that* $q, t \in [1, \infty]$ *and* $\alpha \in R_{++}^{\bar{d}}$ *are given and* $c(\cdot)$ *is defined as in (5) for* $\varrho = 1$. *Then, if* $l(x, y; \beta)$ *is defined as in (6) we obtain*

$$\min_{\beta \in \mathbb{R}^d} \sup_{P: D_c(P, P_n) \le \delta} \mathbb{E}_P\left[l\left(X, Y; \beta\right)\right] = \min_{\beta \in \mathbb{R}^d} E_{P_n}\left(l\left(X, Y; \beta\right)\right) + \delta\, \|\beta\|_{\alpha\text{-}(p,s)},$$

where $1 \le q, t \le \infty$, $1/p + 1/q = 1$ *and* $1/s + 1/t = 1$.

We note that by taking $p = q = 2$, $t = \infty$, $s = 1$, $\alpha_i = \sqrt{g_i}$ for $i \in \{1, ..., \bar{d}\}$, and $\lambda = \delta$ we recover the GR-Lasso logistic regression estimator from Meier et al. (2008).
As discussed in the previous subsection, the choice of $c(\cdot)$ implies that the GR-Lasso estimator implicitly assumes that distributionally uncertainty is only present on prediction variables.

3. Optimal Choice of Regularization Parameter

Let us now discuss the mathematical formulation of the optimal criterion that we discussed for choosing δ (and therefore the regularization parameter λ). We define

$$\Lambda_{\delta}\left(P_n\right) = \{\beta^P : P \in \mathcal{U}_{\delta}\left(P_n\right)\},$$

as discussed in the Introduction, $\Lambda_{\delta}\left(P_n\right)$ is a natural confidence region for β^* because each element P in the distributional uncertainty set $\mathcal{U}_{\delta}\left(P_n\right)$ can be interpreted as a plausible variation of the empirical data P_n. Then, given a confidence level $1-\chi$ (say $1-\chi = .95$) we wish to choose

$$\delta_n^* = \inf\{\delta : P\left(\beta^* \in \Lambda_{\delta}\left(P_n\right)\right) > 1-\chi\}.$$

Note that in the evaluation of $P\left(\beta^* \in \Lambda_{\delta}\left(P_n\right)\right)$ the random element is P_n. So, we shall impose natural probabilistic assumptions on the data generating process in order to *asymptotically evaluate* δ_n^* as $n \to \infty$.

3.1. The Robust Wasserstein Profile Function

In order to asymptotically evaluate δ_n^* we must recall basic properties of the so-called Robust Wassertein Profile function (RWP function) introduced in Blanchet et al. (2016).
Suppose for each (x, y), the loss function $l\left(x, y; \cdot\right)$ is convex and differentiable, then under natural moment assumptions which guarantee that expectations are well defined, we have that for

$$P \in \mathcal{U}_{\delta}\left(P_n\right) = \{P : D_c\left(P, P_n\right) \leq \delta\},$$

the parameter β^P must satisfy

$$\mathbb{E}_P\left[\nabla_{\beta} l\left(X, Y; \beta^P\right)\right] = 0. \tag{7}$$

Now, for any given β, let us define

$$\mathcal{M}\left(\beta\right) = \left\{P : \mathbb{E}_P\left[\nabla_{\beta} l\left(X, Y; \beta\right)\right] = 0\right\},$$

which is the set of probability measures P, under which β is the optimal risk minimization parameter. We would like to choose δ as small as possible so that

$$\mathcal{U}_{\delta}\left(P_n\right) \cap \mathcal{M}\left(\beta^*\right) \neq \varnothing \tag{8}$$

with probability at least $1-\chi$. But note that (8) holds if and only if there exists P such that $D_c\left(P, P_n\right) \leq \delta$ and $\mathbb{E}_P\left[\nabla_{\beta} l\left(X, Y; \beta^*\right)\right] = 0$.
The RWP function is defined

$$R_n\left(\beta\right) = \min\{D_c\left(P, P_n\right) : \mathbb{E}_P\left[\nabla_{\beta} l\left(X, Y; \beta\right)\right] = 0\}. \tag{9}$$

In view of our discussion following (8), it is immediate that $\beta^* \in \Lambda_{\delta}\left(P_n\right)$ if and only if $R_n\left(\beta^*\right) \leq \delta$, which then implies that

$$\delta_n^* = \inf\{\delta : P\left(R_n\left(\beta^*\right) \leq \delta\right) > 1-\chi\}.$$

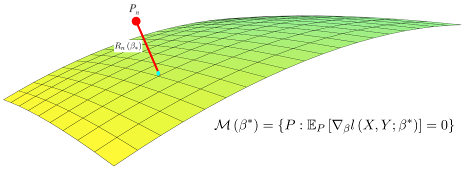
P_n
$R_n\left(\beta_*\right)$
$\mathcal{M}\left(\beta^*\right)=\left\{P:\mathbb{E}_P\left[\nabla_\beta l\left(X,Y;\beta^*\right)\right]=0\right\}$

Figure 1: Intuitive Plot for the RWP function $R_n\left(\beta\right)$ and the set $\mathcal{M}\left(\beta\right)$.

Consequently, we conclude that δ_n^* can be evaluated asymptotically in terms of the $1-\chi$ quantile of $R_n\left(\beta^*\right)$ and therefore we must identify the asymptotic distribution of $R_n\left(\beta^*\right)$ as $n\to\infty$. We illustrate intuitively the role of the RWP function and $\mathcal{M}\left(\beta\right)$ in Figure 1, where RWP function $R_n\left(\beta^*\right)$ could be interpreted as the discrepancy distance between empirical measure P_n and the manifold $\mathcal{M}\left(\beta^*\right)$ associated with β^*.

Typically, under assumptions supporting the underlying model (as in the generalized linear setting we considered), we will have that β^* is characterized by the estimating equation (7). Therefore, under natural statistical assumptions one should expect that $R_n\left(\beta^*\right)\to 0$ as $n\to\infty$ at a certain rate and therefore $\delta_n^*\to 0$ at a certain (optimal) rate. This then yields an optimal rate of convergence to zero for the underlying regularization parameter. The next subsections will investigate the precise rate of convergence analysis of δ_n^*.

3.2. Optimal Regularization for GSRL Linear Regression

We assume, for simplicity, that the training data set $\{(X_1,Y_1),\ldots,(X_n,Y_n)\}$ is i.i.d. and that the linear relationship $Y_i={\beta^*}^TX_i+e_i$, holds with the errors $\{e_1,...,e_n\}$ being i.i.d. and independent of $\{X_1,\ldots,X_n\}$. Moreover, we assume that both the entries of X_i and the errors have finite second moment and the errors have zero mean.

Since in our current setting $l\left(x,y;\beta\right)=\left(y-x^T\beta\right)^2$, then the RWP function (9) for linear regression model is given as,

$$R_n\left(\beta\right)=\min_P\left\{D_c\left(P,P_n\right):\mathbb{E}_P\left[X\left(Y-X^T\beta\right)\right]=0\right\}. \tag{10}$$

Theorem 3 (RWP Function Asymptotic Results: Linear Regression) *Under the assumptions imposed in this subsection and the cost function as given in (5), with $\varrho=2$,*

$$nR_n\left(\beta^*\right)\Rightarrow L_1:=\max_{\zeta\in\mathbb{R}^d}\left\{2\sigma\zeta^TZ-\mathbb{E}\left[\left\|e\zeta-\left(\zeta^TX\right)\beta^*\right\|_{\alpha\text{-}(p,s)}^2\right]\right\},$$

as $n\to\infty$, where $\Rightarrow$ means convergence in distribution and $Z\sim\mathcal{N}\left(0,\Sigma\right)$ with $\Sigma=Var(X)$. Moreover, we can observe the more tractable stochastic upper bound,

$$L_1\overset{D}{\leq}L_2:=\frac{\mathbb{E}\left[e^2\right]}{\mathbb{E}\left[e^2\right]-\left(\mathbb{E}\left[|e|\right]\right)^2}\left\|Z\right\|_{\alpha^{-1}\text{-}(q,t)}^2.$$

We now explain how to use Theorem 3 to set the regularization parameter in GSRL linear regression:

1. Estimate the $1-\chi$ quantile of $\|Z\|_{\alpha^{-1}\text{-}(q,t)}^2$. We use use $\hat{\eta}_{1-\chi}$ to denote the estimator for this quantile. This step involves estimating Σ from the training data.

2. The regularization parameter λ in the GSRL linear regression takes the form

$$\lambda = \sqrt{\delta} = \hat{\eta}_{1-\chi}^{1/2} \left(n(1 - (\mathbb{E}\,|e|)^2 / \mathbb{E}e^2) \right)^{-1/2}.$$

   Note that the denominator in the previous expression must be estimated from the training data.

Remark that the regularization parameter for GSRL for linear regression chosen via our RWPI asymptotic result does not depends on the magnitude of error e (see also the discussion in Bunea et al. (2014)). It is also possible to formulate the optimal regularization results for high-dimension setting, where the number of predictors growth with sample size. Due to the space constraints, we include the results in the supplimentary material, namely Section B.3.

3.3. Optimal Regularization for GR-Lasso Logistic Regression

We assume that the training data set $\{(X_1, Y_1), \ldots, (X_n, Y_n)\}$ is i.i.d.. In addition, we assume that the X_i's have a finite second moment and also that they possess a density with respect to the Lebesgue measure. Moreover, we assume a logistic regression model; namely,

$$P\left(Y_i = 1 | X_i\right) = 1/\left(1 + \exp\left(-X_i^T \beta^*\right)\right), \tag{11}$$

and $P\left(Y_i = -1 | X_i\right) = 1 - P\left(Y_i = 1 | X_i\right)$.
In the logistic regression setting, we consider the log-exponential loss defined in (6). Therefore, the RWP function, (9), for logistic regression is

$$R_n\left(\beta\right) = \min\left\{ D_c\left(P, P_n\right) : \mathbb{E}_P\left[\frac{YX}{1 + \exp\left(YX^T\beta\right)}\right] = 0 \right\}. \tag{12}$$

Theorem 4 (RWP Function Asymptotic Results: Logistic Regression) *Under the assumptions imposed in this subsection and the cost function as given in (5), with $\varrho = 1$,*

$$\sqrt{n} R_n\left(\beta^*\right) \Rightarrow L_3 := \sup_{\zeta \in A} \quad \zeta^T Z,$$

as $n \to \infty$, where

$$Z \sim \mathcal{N}(0, \mathbb{E}[\frac{XX^T}{\left(1 + \exp\left(YX^T\beta^*\right)\right)^2}])$$

and

$$A = \left\{ \zeta \in \mathbb{R}^d : \operatorname*{ess\,sup}_{X,Y} \left\| \zeta^T \frac{y\left(1 + \exp\left(YX^T\beta^*\right)\right) I_{d\times d} - XX^T}{\left(1 + \exp\left(YX^T\beta^*\right)\right)^2} \right\|_{\alpha\text{-}(p,s)} \leq 1 \right\}.$$

Further, the limit law L_3 follows the simpler stochastic bound,

$$L_3 \overset{D}{\leq} L_4 := \left\| \tilde{Z} \right\|_{\alpha^{-1}\text{-}(q,t)},$$

where $\tilde{Z} \sim \mathcal{N}\left(0, \Sigma\right)$.

We now explain how to use Theorem 4 to set the regularization parameter in GR-Lasso logistic regression.

1. Estimate the $1-\chi$ quantile of L_4. We use use $\hat{\eta}_{1-\chi}$ to denote the estimator for this quantile. This step involves estimating Σ from the training data.
2. We set the regularization parameter λ in the GR-Lasso to be $\lambda = \delta = \hat{\eta}_{1-\chi}/\sqrt{n}$.

4. Numerical Experiments

We proceed to numerical experiments on both simulated and real data to verify the performance of our method for choosing the regularization parameter. We apply "grpreg" in R, from Breheny and Breheny (2016), to solve GR-Lasso for logistic regression. For GSRL for linear regression, we consider apply the "grpreg" solver for the GR-Lasso problem combined with the iterative procedure discussed in Section 2 of Sun and Zhang (2011) (see also Section 5 of Li et al. (2015) for the SR-Lasso counterpart of such numerical procedure).
Data preparation for simulated experiments: We borrow the setting from example III in Yuan and Lin (2006), where the group structure is determined by the third order polynomial expansion. More specifically, we assume that we have 17 random variables $Z_1, \ldots, Z_{16}$ and W, they are i.i.d. and follow the normal distribution. The covariates $X_1, \ldots, X_{16}$ are given as $X_i = \left(Z_i + W\right)/\sqrt{2}$. For the predictors, we consider each covariate and its second and third order polynomial, i.e. X_i, X_i^2 and X_i^3. In total, we have 48 predictors.
For linear regression: The response Y is given by

$$Y = \beta_{3,1}X_3 + \beta_{3,2}X_3^2 + \beta_{3,3}X_3^3 + \beta_{5,1}X_5 + \beta_{5,2}X_5^2 + \beta_{5,3}X_5^3 + e,$$

where $\beta_{(\cdot,\cdot)}$ coefficients draw randomly and e represents an independent random error.
For classification: We consider Y simulated by a Bernoulli distribution, i.e.

$$Y \sim Ber\left(1/\left[1 + \exp\left(-\left(\beta_{3,1}X_3 + \beta_{3,2}X_3^2 + \beta_{3,3}X_3^3 + \beta_{5,1}X_5 + \beta_{5,2}X_5^2 + \beta_{5,3}X_5^3\right)\right)\right]\right).$$

We compare the following methods for linear regression and logistic regression: 1) groupwise regularization with asymptotic results (in Theorem 3, 4) selected tuning parameter (RWPI GRSL and RWPI GR-Lasso), 2) groupwise regularization with cross-validation (CV GRSL and CV GR-Lasso), and 3) ordinary least square and logistic regression (OLS and LR).
We report the error as the square loss for linear regression and log-exponential loss for logistic regression. The training error is calculated via the training data. The size of the training data is taken to be $n = 50, 100, 500$ and 1000. The testing error is evaluated using a simulated data set of size 1000 using the same data generating process described earlier. The mean and standard deviation of the error are reported via 200 independent runs of the whole experiment, for each sample size n.
The detailed results are summarized in Table 1 for linear regression and Table 2 for logistic regression. We can see that our procedure is very comparable to cross validation, but it is significantly less time consuming and all of the data can be directly used to estimate the model parameter, by-passing significant data usage in the estimation of the regularization parameter via cross validation. We also validated our method using the Breast Cancer classification problem with data from the UCI machine learning database discussed in Lichman (2013). The data set contains 569 samples with one binary response and 30 predictors.

	RWPI GSRL		CV GSRL		OLS	
Sample Size	Training	Testing	Training	Testing	Training	Testing
$n = 50$	5.64 ± 1.16	9.15 ± 3.58	3.18 ± 1.07	7.66 ± 2.69	0.07 ± 0.09	80.98 ± 30.53
$n = 100$	4.67 ± 0.70	5.83 ± 1.38	3.61 ± 0.74	5.22 ± 1.05	2.09 ± 0.44	73.35 ± 16.51
$n = 500$	4.09 ± 0.29	4.16 ± 0.27	3.93 ± 0.3	4.12 ± 0.27	3.63 ± 0.27	73.08 ± 10.40
$n = 1000$	4.02 ± 0.19	4.11 ± 0.26	3.95 ± 0.19	4.11 ± 0.26	3.82 ± 0.19	72.28 ± 8.05

Table 1: Linear Regression Simulation Results.

	RWPI GR-Lasso		CV GR-Lasso		Logistic Regression	
Sample Size	Training	Testing	Training	Testing	Training	Testing
$n = 50$	$.683 \pm .016$	$.702 \pm .014$	$.459 \pm .118$	$.628 \pm .099$	$.002 \pm .001$	5.288 ± 1.741
$n = 100$	$.593 \pm .038$	$.618 \pm .029$	$.450 \pm .061$	$.551 \pm .037$	$.042 \pm .041$	4.571 ± 1.546
$n = 500$	$.513 \pm .021$	$.518 \pm .019$	$.461 \pm .025$	$.493 \pm .018$	$.083 \pm .057$	$1.553 \pm .355$
$n = 1000$	$.492 \pm .016$	$.488 \pm .017$	$.491 \pm .017$	$.488 \pm .019$	$.442 \pm .018$	$.510 \pm .028$

Table 2: Logistic Regression Simulation Results.

We consider all the predictors and their first, second, and third order polynomial expansion. Thus, we end up having 90 predictors divided into 30 groups. For each iteration, we randomly split the data into a training set with 112 samples and the rest in the testing set. We repeat the experiment 500 times to observe the log-exponential loss function for the training and testing error. We compare our asymptotic results based GR-Lasso logistic regression (RWPI GR-Lasso), cross-validation based GR-Lasso logistic regression (CV GR-Lasso), vanilla logistic regression (LR), and regularized logistic regression (LRL1). We can observe, even when the sample size is small as in the example, our method still provides very comparable results (see in Table 3).

LR		LRL1		RWPI GR-Lasso		CV GR-Lasso	
Training	Testing	Training	Testing	Training	Testing	Training	Testing
0.0 ± 0.0	15.267 ± 5.367	$.510 \pm .215$	$.414 \pm .173$	$.186 \pm .032$	$.240 \pm .098$	$.198 \pm .041$	$.213 \pm .041$

Table 3: Numerical results for breast cancer data set.

It is intuitively clear that our regularization prescription (based on RWPI) can be implemented much more quickly than k-fold cross-validation. Calibrating the quantile of the RWPI asymptotic distribution requires a computational effort that is not worst than solving the regularized risk minimization problem. Once this quantile has been calibrated, then our approach involves solving the regularized risk minimization problem only once. So, the time to implement our approach with the RWPI-based optimal regularization selection is not worst than solving the regularized risk minimization problem twice. In contrast, k-fold cross validation requires solving such problem at least k times (k is often taken to be 10).

5. Conclusion and Extensions

Our discussion of GSRL as a DRO problem has exposed rich interpretations which we have used to understand GSRL's generalization properties by means of a game theoretic formulation. Moreover, our DRO representation also elucidates the crucial role of the regularization parameter in measuring the distributional uncertainty present in the data.

Finally, we obtained asymptotically valid formulas for optimal regularization parameters under a criterion which is naturally motivated, once again, thanks to our DRO formulation. Our easy-to-implement formulas are shown to perform comparable to cross validation.
We strongly believe that our discussion in this paper can be easily extended to a wide range of machine learning estimators. We envision formulating the DRO problem considering different types of models and cost functions. Currently, we consider the linear model with square loss function and the log-exponential loss function. However, the DRO formulation could be applied to more general modeling frameworks, for example, the neural network based cross-entropy loss function. We plan to investigate algorithms which solve the DRO problem directly (even if no direct regularization representation exists).

References

Alexandre Belloni, Victor Chernozhukov, and Lie Wang. Square-root Lasso: pivotal recovery of sparse signals via conic programming. *Biometrika*, 98(4):791–806, 2011.

Jose Blanchet, Yang Kang, and Karthyek Murthy. Robust Wasserstein profile inference and applications to machine learning. *arXiv preprint arXiv:1610.05627*, 2016.

Patrick Breheny and Maintainer Patrick Breheny. Package grpreg. 2016.

Florentina Bunea, Johannes Lederer, and Yiyuan She. The group square-root Lasso: Theoretical properties and fast algorithms. *IEEE Transactions on Information Theory*, 60 (2):1313–1325, 2014.

Jerome Friedman, Trevor Hastie, and Robert Tibshirani. A note on the group Lasso and a sparse group Lasso. *arXiv preprint arXiv:1001.0736*, 2010.

Xingguo Li, Tuo Zhao, Xiaoming Yuan, and Han Liu. The Flare package for high dimensional linear regression and precision matrix estimation in R. *JMLR*, 16:553–557, 2015.

M. Lichman. UCI machine learning repository, 2013. URL archive.ics.uci.edu/ml.

Lukas Meier, Sara Van De Geer, and Peter Bühlmann. The group Lasso for logistic regression. *Journal of the Royal Statistical Society: Series B*, 70(1):53–71, 2008.

Soroosh Shafieezadeh-Abadeh, Peyman Mohajerin Esfahani, and Daniel Kuhn. Distributionally robust logistic regression. In *Advances in NIPS 28*, pages 1576–1584. 2015.

Tingni Sun and Cun-Hui Zhang. Scaled sparse linear regression. *arXiv preprint arXiv:1104.4595*, 2011.

Robert Tibshirani. Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society. Series B*, 58(1):267–288, 1996. ISSN 00359246.

Cdric Villani. *Optimal Transport: old and new*. Springer Science & Business Media, 2008.

Huan Xu, Constantine Caramanis, and Shie Mannor. Robust regression and Lasso. In *Advances in NIPS 21*, pages 1801–1808. 2009.

Ming Yuan and Yi Lin. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B*, 68(1):49–67, 2006.

Appendix A. Appendix: Technical Proofs

We will first derive some properties for α-(p,s) norm we defined in (1), then we move to the proof for DRO problem in Section A.2 and the optimal selection of regularization parameter in Section A.3. Due the space constraint, we will focus on the proof for linear regression and leave the part for logistic regression, which follows the similar techniques, in the supplementary material.

A.1. Basic Properties of the α-(p,s) Norm

The following Proposition, which describes basic properties of the α-(p,s) norm, will be very useful in our proofs.

Proposition 5 *For $\alpha-(p,s)$ norm defined for $\mathbb{R}^d$ as in (1) and the notations therein, we have the following properties:*
I) *The dual norm of $\alpha-(p,s)$ norm is α^{-1}-(q,t) norm, where $\alpha^{-1}=(1/\alpha_1,\ldots,1/\alpha_{\bar{d}})^T$, $1/p+1/q=1$, and $1/s+1/t=1$ (i.e. p,q are conjugate and s,t are conjugate).*
II) *The Hölder inequality holds for the α-(p,s) norm, i.e. for $a,b\in\mathbb{R}^d$, we have,*

$$a^Tb\le\|a\|_{\alpha\text{-}(p,s)}\|b\|_{\alpha^{-1}\text{-}(q,t)},$$

where the equality holds if and only if $sign(a(G_j)_i)=sign(b(G_j)_i)$ and

$$|\alpha_j a(G_j)_i|\|\frac{1}{\alpha_j}b(G_j)\|_q^{q/p-t/s}\|b\|_{\alpha^{-1}\text{-}(q,t)}^{t/s}=|\frac{1}{\alpha_j}b(G_j)_i|^{q/p},$$

is true for all $j=1,\ldots,\bar{d}$ and $i=1,\ldots,g_j$.
The triangle inequality holds, i.e. for $a,b\in\mathbb{R}^d$ and $a\neq 0$, we have

$$\|a\|_{\alpha\text{-}(p,s)}+\|b\|_{\alpha\text{-}(p,s)}\ge\|a+b\|_{\alpha\text{-}(p,s)},$$

where the equality holds if and only if, there exists nonnegative τ, such that $\tau a=b$.

Proof [Proof of Proposition 5]We first proceed to prove II). Let us consider any $a,b\in\mathbb{R}^d$. We can assume $a,b\neq 0$, otherwise the claims are immediate. The inner product (or dot product) of a and b an be written as:

$$a^Tb=\sum_{j=1}^{\bar{d}}[\sum_{i=1}^{g_j}a(G_j)_ib(G_j)_i]\le\sum_{j=1}^{\bar{d}}[\sum_{i=1}^{g_j}|a(G_j)_i|\cdot|b(G_j)_i|].$$

The equality holds for the above inequality if and only if $a(G_j)_i$ and $b(G_j)_i$ shares the same sign. For each fixed $j=1,\ldots,\bar{d}$, we consider the term in the bracket,

$$\sum_{i=1}^{g_j}|a(G_j)_i|\cdot|b(G_j)_i|=\sum_{i=1}^{g_j}\alpha_j\,|a(G_j)_i|\cdot|b(G_j)_i|\,/\alpha_j\le\|\alpha_j a(G_j)\|_p\cdot\left\|\frac{1}{\alpha_j\cdot}b(G_j)\right\|_q.$$

The above inequality is due to Hölder's inequality for $p-$norm and the equality holds if and only if

$$\left\|\frac{1}{\alpha_{j\cdot}}b(G_j)\right\|_q^q |\alpha_j a(G_j)_i|^p = \|\alpha_j a(G_j)\|_p^p \left|\frac{1}{\alpha_j}b(G_j)_i\right|^q,$$

is true for all $i=\overline{1,g_j}$. Combining the above result for each $j=1,\dots,\bar{d}$, we have,

$$a^Tb \le \sum_{j=1}^{\bar{d}}\left\|\alpha_j a(G_j)\right\|_p\cdot\left\|\frac{1}{\alpha_j}b(G_j)\right\|_q \le \|a\|_{\alpha\text{-}(p,s)}\cdot\|b\|_{\alpha^{-1}\text{-}(q,t)},$$

where the final inequality is due to Hölder inequality applied to the vectors

$$\tilde{a} = \Big(\alpha_1\left\|a(G_1)\right\|_p,\dots,\alpha_{\bar{d}}\left\|a(G_{\bar{d}})\right\|_p\Big)^T, \text{ and } \tilde{b} = \left(\frac{1}{\alpha_1}\left\|b_{G_1}\right\|_q,\dots,\frac{1}{\alpha_{\bar{d}}}\left\|b(G_{\bar{d}})\right\|_q\right)^T. \quad (13)$$

This proves the Hölder type inequality stated in the theorem. We can further observe that the final inequality becomes equality if and only if

$$\|b\|^t_{\alpha^{-1}-(q,t)}\left\|\alpha_j a(G_j)\right\|_p^s = \|a\|^s_{\alpha-(p,s)}\,\|\frac{1}{\alpha_j}b(G_j)\|_q^t,$$

holds for all $j=1,\dots,\bar{d}$. Combining the conditions for equalities hold for each inequality, we conclude condition II) in the statement of the proposition. Next we proceed to prove I). Recall the definition of a dual norm, i.e. $\|b\|^*_{\alpha\text{-}(p,s)} = \sup_{a:\|a\|_{\alpha\text{-}(p,s)}=1} a^Tb$. Now, choose $b\in\mathbb{R}^d$, $b\neq 0$, and let us take a satisfying, $\|a\|_{\alpha-(p,s)}=1$ and

$$a(G_j)_i = sign(b(G_j)_i)\alpha_j^{-1}|\frac{1}{\alpha_j}b(G_j)_i|^{q/p}\|\frac{1}{\alpha_j}b(G_j)\|_q^{t/s-q/p}\|b\|^{-t/s}_{\alpha^{-1}\text{-}(q,t)}.$$

By part II), we have that

$$\|b\|^*_{\alpha-(p,s)} = \sup_{a:\|a\|_{\alpha-(p,s)}=1} a^Tb = \|a\|_{\alpha-(p,s)}\|b\|_{\alpha^{-1}-(q,t)} = \|b\|_{\alpha^{-1}-(q,t)}.$$

Thus we proved part I). Finally, let us discuss the triangle inequality. For any $a,b\in\mathbb{R}^d$ and $a,b\neq 0$ we have

$$\|a\|_{\alpha\text{-}(p,s)} + \|b\|_{\alpha\text{-}(p,s)} = \big[\sum_{j=1}^{\bar{d}}\alpha_j\left\|a(G_j)\right\|_p^s\big]^{1/s} + \big[\sum_{j=1}^{\bar{d}}\alpha_j\left\|b(G_j)\right\|_p^s\big]^{1/s}$$

$$\ge \big[\sum_{j=1}^{\bar{d}}\alpha_j\Big(\left\|a(G_j)\right\|_p^s + \left\|b(G_j)\right\|_p^s\Big)\big]^{1/s} \ge \big[\sum_{j=1}^{\bar{d}}\alpha_j\left\|a(G_j)+b(G_j)\right\|_p^s\big]^{1/s} = \|a+b\|_{\alpha\text{-}(p,s)}.$$

For the above derivation, the first equality is due to definition in (1), Second equality is applying the triangle inequality of s-norm for $\tilde{a}$ and $\tilde{b}$ defined in (13), where the equality holds if and only if, there exist positive number $\tilde{\tau}$, such that $\tilde{\tau}\tilde{a}=\tilde{b}$. Third inequality is due to triangle equality of p-norm to $a(G_j)$ and $b(G_j)$ for each $j=1,\dots,\bar{d}$, where the equality holds if and only if, there exists nonnegative numbers τ_j, such that $\tau_j a(G_j) = b(G_j)$.

Combining the equality condition for second and third estimate above, we can conclude the equality condition for the triangle inequality for α-(p,s) norm is if and only if there exists a non-negative number τ, such that $\tau a = b$. ■

A.2. Proof of DRO for Linear Regression

Proof [Proof of Theorem 1]Let us apply the strong duality results, as in the Appendix of Blanchet et al. (2016), for worst-case expected loss function, which is a semi-infinity linear programming problem, and write the worst-case loss as,

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

For each i, let us consider the inner optimization problem over u. We can denote $\Delta = u - x_i$ and $e_i = y_i - x_i^T\beta$ for notation simplicity, then the $i-$th inner optimization problem becomes,

$$\begin{aligned}
&e_i^2 + \sup_{\Delta}\left\{\left(\Delta^T\beta\right)^2 - 2e_i\Delta^T\beta - \gamma\left\|\Delta\right\|^2_{\alpha^{-1}-(q,t)}\right\}\\
=&e_i^2 + \Delta\{(\sum_j|\Delta_j|\,|\beta_j|)^2 + 2\,|e_i|\sum_j|\Delta_j|\,|\beta_j| - \gamma\left\|\Delta\right\|^2_{\alpha^{-1}-(q,t)}\}\\
=&e_i^2 \sup_{\left\|\Delta\right\|_{\alpha^{-1}\text{-}(q,t)}}\left\{\left\|\beta\right\|^2_{\alpha\text{-}(p,s)}\left\|\Delta\right\|^2_{\alpha^{-1}\text{-}(q,t)} + 2\left\|\beta\right\|_{\alpha\text{-}(p,s)}|e_i|\left\|\Delta\right\|_{\alpha^{-1}\text{-}(q,t)} - \gamma\left\|\Delta\right\|^2_{\alpha^{-1}\text{-}(q,t)}\right\}\\
=&e_i^2\frac{\gamma}{\gamma - \left\|\beta\right\|^2_{\alpha\text{-}(p,s)}}I_{\gamma>\left\|\beta\right\|^2_{\alpha\text{-}(p,s)}} + \infty I_{\gamma\leq\left\|\beta\right\|^2_{\alpha\text{-}(p,s)}},
\end{aligned}\tag{14}$$

where the development uses the duality results developed in Proposition 5. The last equality is optimize over Δ for two different cases of λ.
Since optimization over γ is a minimization, the outer player will always select γ that avoids an infinite value of the game. Then we can write the worst-case expected loss function as,

$$\begin{aligned}
\sup_{P:D_c(P,P_n)\leq\delta}\mathbb{E}_P\left[\left(Y - X^T\beta\right)^2\right] &= \min_{\gamma>\left\|\beta\right\|^2_{\alpha\text{-}(p,s)}}\left\{\gamma\delta - \gamma\frac{E_{P_n}l\left(X,Y;\beta\right)}{\gamma - \left\|\beta\right\|^2_{\alpha\text{-}(p,s)}}\right\}\\
&= \left(\sqrt{E_{P_n}l\left(X,Y;\beta\right)} + \sqrt{\delta}\left\|\beta\right\|_{\alpha\text{-}(p,s)}\right)^2.
\end{aligned}\tag{15}$$

The first equality in (15) is a plug-in from the result in (14). For the second equality, we can observe the target function is convex and differentiable and as $\gamma \to \infty$ and $\gamma \to \left\|\beta\right\|^2_{\alpha\text{-}(p,s)}$, the value function will be infinity. We can solve this convex optimization problem which leads to the result above. We further take square root and take minimization over β on both sides, we proved the claim of the theorem. ■

A.3. Proof for Optimal Selection of Regularization for Linear Regression

Proof [Proof for Theorem 3]For linear regression with the square loss function, if we apply the strong duality result for semi-infinity linear programming problem as in Section B of

Blanchet et al. (2016), we can write the scaled RWP function for linear regression as

$$nR_n\left(\beta_*\right) = \sup_{\zeta}\left\{-\zeta^T Z_n - \mathbb{E}_{P_n}\phi\left(X_i, Y_i, \beta^*, \zeta\right)\right\}, \tag{16}$$

where $Z_n = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} e_i X_i$ and

$$\phi\left(X_i, Y_i, \beta^*, \zeta\right) = \sup_{\Delta}\left\{e_i\zeta^T\Delta - \left(\beta_*^T\Delta\right)\left(\zeta^T X_i\right) - \left(\|\Delta\|^2_{\alpha^{-1}\text{-}(q,t)} + n^{-1/2}\left(\beta_*^T\Delta\right)\left(\zeta^T\Delta\right)\right)\right\}.$$

Follow the similar discussion in the proof of Theorem 4 in Blanchet et al. (2016). Applying Lemma 2 in Blanchet et al. (2016), we argue that the optimizer ζ can be restrict on a compact set asymptotically with high probability. We apply the uniform law of large number estimate as in Lemma 3 of Blanchet et al. (2016) to the second term in (16) and obtain

$$nR_n\left(\beta_*\right) = \sup_{\zeta}\{-\zeta^T Z_n - \mathbb{E}\phi\left(X, Y, \beta, \zeta\right)]\} + o_P(1). \tag{17}$$

For any fixed X, Y, as $n \to \infty$, we can simplify the contribution of $\phi(\cdot)$ inside sup in (17). This is done by applying the duality result (Hölder-type inequality) in Proposition 5 and noting that $\phi(\cdot)$ becomes quadratic in $\|\Delta\|_{\alpha^{-1}\text{-}(q,t)}$. This results in the simplified expression

$$nR_n\left(\beta_*\right) = \sup_{\zeta}\left\{-\zeta^T Z_n - \mathbb{E}\left[\left\|e\zeta - (\zeta^T X)\beta^*\right\|^2_{\alpha\text{-}(p,s)}\right]\right\} + o_P(1).$$

Since we observe $Z_n \Rightarrow \sigma Z$ as $n \to \infty$, we proved the first argument. We need to show that the feasible region can be compactified with high probability. This compactification argument is done with a technique similar to Lemma 2 in Blanchet et al. (2016).

By the definition of L_1, we can apply Hölder inequality to the first term, and split the optimization into optimizing over direction $\|\zeta'\|_{\alpha\text{-}(p,s)} = 1$ and magnitude $a \geq 0$. Thus, we have

$$L_1 \leq \max_{\zeta' : \|\zeta'\|_{\alpha\text{-}(p,s)}=1}\max_{a\geq 0}\left\{2a\sigma\|Z\|_{\alpha^{-1}\text{-}(q,t)} - a^2\mathbb{E}\left[\left\|e\zeta' - (\zeta'^T X)\beta^*\right\|^2_{\alpha\text{-}(p,s)}\right]\right\}.$$

It is easy to solve the quadratic programming problem in a and we conclude that

$$L_1 \leq \frac{\sigma^2\|Z\|^2_{\alpha^{-1}\text{-}(q,t)}}{\min_{\zeta' : \|\zeta'\|_{\alpha\text{-}(p,s)}=1}\mathbb{E}\left[\left\|e\zeta' - (\zeta'^T X)\beta^*\right\|^2_{\alpha\text{-}(p,s)}\right]}.$$

For the denominator, we have estimates as follows:

$$\min_{\zeta' : \|\zeta'\|_{\alpha\text{-}(p,s)}=1}\mathbb{E}\left[\left\|e\zeta' - (\zeta'^T X)\beta^*\right\|^2_{\alpha\text{-}(p,s)}\right] \geq \min_{\zeta' : \|\zeta'\|_{\alpha\text{-}(p,s)}=1}\mathbb{E}\left[|e| - \left|\zeta^T X\right|\|\beta^*\|_{\alpha\text{-}(p,s)}\right]^2$$

$$\geq Var(|e|) + \min_{\zeta' : \|\zeta'\|_{\alpha\text{-}(p,s)}=1}\left(\|\beta^*\|_{\alpha\text{-}(p,s)}\mathbb{E}\left|\zeta'^T X\right| - \mathbb{E}|e|\right)^2 \geq Var(|e|).$$

The first estimate is due to the triangle inequality in Proposition 5, the second estimate follows using Jensen's inequality, the last inequality is immediate. Combining these inequalities we conclude

$$L_1 \leq \sigma^2\|Z\|^2_{\alpha^{-1}\text{-}(q,t)}/Var(|e|).$$

■