# SEMI-SUPERVISED LEARNING BASED ON DISTRIBUTIONALLY ROBUST OPTIMIZATION

JOSE BLANCHET AND YANG KANG

Abstract. We propose a novel method for semi-supervised learning (SSL) based on data-driven distributionally robust optimization (DRO) using optimal transport metrics. Our proposed method enhances generalization error by using the unlabeled data to restrict the support of the worst case distribution in our DRO formulation. We enable the implementation of our DRO formulation by proposing a stochastic gradient descent algorithm which allows to easily implement the training procedure. We demonstrate that our Semi-supervised DRO method is able to improve the generalization error over natural supervised procedures and state-of-the-art SSL estimators. Finally, we include a discussion on the large sample behavior of the optimal uncertainty region in the DRO formulation. Our discussion exposes important aspects such as the role of dimension reduction in SSL.

## 1. Introduction

We propose a novel method for semi-supervised learning (SSL) based on data-driven distributionally robust optimization (DRO) using an optimal transport metric – also known as earth-moving distance (see [19]).

Our approach enhances generalization error by using the unlabeled data to restrict the support of the models which lie in the region of distributional uncertainty. The intuition is that our mechanism for fitting the underlying model is automatically tuned to generalize beyond the training set, but only over potential instances which are relevant. The expectation is that predictive variables often lie in lower dimensional manifolds embedded in the underlying ambient space; thus, the shape of this manifold is informed by the unlabeled data set (see Figure 1 for an illustration of this intuition).

To enable the implementation of the DRO formulation we propose a stochastic gradient descent (SGD) algorithm which allows to implement the training procedure at ease. Our SGD construction includes a procedure of independent interest which, we believe, can be used in more general stochastic optimization problems.

We focus our discussion on semi-supervised classification but the modeling and computational approach that we propose can be applied more broadly as we shall illustrate in Section 4.

We now explain briefly the formulation of our learning procedure. Suppose that the training set is given by $\mathcal{D}_n = \{(Y_i, X_i)\}_{i=1}^n$, where $Y_i \in \{-1, 1\}$ is the label of the $i$-th observation and we assume that the predictive variable, $X_i$, takes values in $\mathbb{R}^d$. We use $n$ to denote the number of labeled data points.

In addition, we consider a set of unlabeled observations, $\{X_i\}_{i=n+1}^N$. We build the set $\mathcal{E}_{N-n} = \{(1, X_i)\}_{i=n+1}^N \cup \{(-1, X_i)\}_{i=n+1}^N$. That is, we replicate each unlabeled data point twice, recognizing

*Date*: August 7, 2017.

*Key words and phrases.* Distributionally Robust Optimization, Semisupervised Learning, Stochastic Gradient Descent.

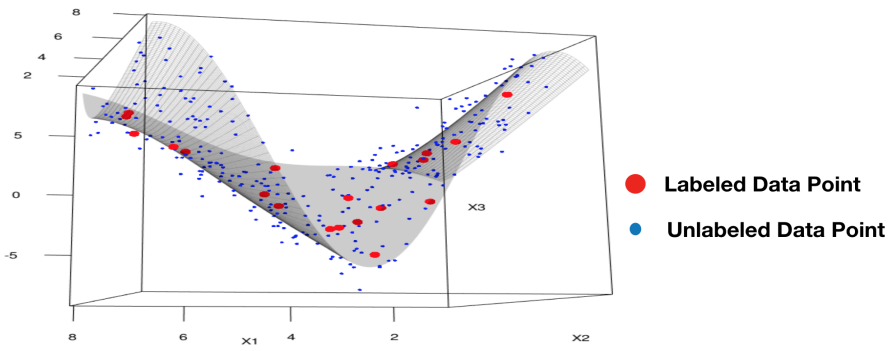


FIGURE 1. Idealization of the way in which the unlabeled predictive variables provide a proxy for an underlying lower dimensional manifold. Large red dots represent labeled instances and small blue dots represent unlabeled instances.

that the missing label could be any of the two available alternatives. We assume that the data must be labeled either -1 or 1.

We then construct the set $\mathcal{X}_N = \mathcal{D}_n \cup \mathcal{E}_{N-n}$ which, in simple words, is obtained by just combining both the labeled data and the unlabeled data with all the possible labels that can be assigned. The cardinality of $\mathcal{X}_N$, denoted as $|\mathcal{X}_N|$, is equal to $2(N-n)+n$ (for simplicity we assume that all of the data points and the unlabeled observations are distinct).

Let us define $\mathcal{P}(\mathcal{X}_N)$ to be the space of probability measures whose support is contained in $\mathcal{X}_N$. We use $P_n$ to denote the empirical measure supported on the set $\mathcal{D}_n$, so $P_n \in \mathcal{P}(\mathcal{X}_N)$. In addition, we write $E_P(\cdot)$ to denote the expectation associated with a given probability measure $P$.

Let us assume that we are interested in fitting a classification model by minimizing a given expected loss function $l(X,Y,\beta)$, where $\beta$ is a parameter which uniquely characterizes the underlying model. We shall assume that $l(X,Y,\cdot)$ is a convex function for each fixed $(X,Y)$. The empirical risk associated to the parameter $\beta$ is

$$E_{P_n}(l(X,Y,\beta)) = \frac{1}{n}\sum_{i=1}^{n} l(X_i,Y_i,\beta).$$

In this paper, we propose to estimate $\beta$ by solving the DRO problem

$$\min_{\beta} \max_{P\in\mathcal{P}(\mathcal{X}_N):D_c(P,P_n)\leq\delta^*} E_P[l(X,Y,\beta)], \tag{1}$$

where $D_c(\cdot)$ is a suitably defined discrepancy between $P_n$ and any probability measure $P \in \mathcal{P}(\mathcal{X}_N)$ which is within a certain tolerance measured by $\delta^*$.

So, intuitively, (1) represents the value of a game in which the outer player (we) will choose $\beta$ and the adversary player (nature) will rearrange the support and the mass of $P_n$ within a budget measured by $\delta^*$. We then wish to minimize the expected risk regardless of the way in which the adversary might corrupt (within the prescribed budget) the existing evidence. In formulation (1), the adversary is crucial to ensure that we endow our mechanism for selecting $\beta$ with the ability to cope with the risk impact of out-of-sample (i.e. out of the training set) scenarios. We denote the formulation in (1) as semi-supervised distributionally robust optimization (SSL-DRO) or semi-supervised learning based on distributionally robust optimization.

The criterion that we use to define $D_c(\cdot)$ is based on the theory of optimal transport and it is closely related to the concept of Wasserstein distance, see Section 3. The choice of $D_c(\cdot)$ is motivated by recent results which show that popular estimators such as regularized logistic regression, Support

Vector Machines (SVMs) and square-root Lasso (SR-Lasso) admit a DRO representation *exactly equal to* (1) in which the support $\mathcal{X}_N$ is replaced by $\mathbb{R}^{d+1}$ (see [6] and also equation (9) in this paper.)

In view of these representation results for supervised learning algorithms, the inclusion of $\mathcal{X}_N$ in our DRO formulation (1) provides a natural SSL approach in the context of classification and regression. The goal of this paper is to enable the use of the distributionally robust training framework (1) as a SSL technique. We will show that estimating $\beta$ via (1) may result in a significant improvement in generalization relative to natural supervised learning counterparts (such as regularized logistic regression and SR-Lasso). The potential improvement is illustrated in Section 4. Moreover, we show via numerical experiments in Section 5, that our method is able to improve upon state-of-the-art SSL algorithms.

As a contribution of independent interest, we construct a stochastic gradient descent algorithm to approximate the optimal selection, $\beta_N^*$, minimizing (1).

An important parameter when applying (1) is the size of the uncertainty region, which is parameterized by $\delta^*$. We apply cross-validation to calibrate $\delta^*$, but we also discuss the non-parametric behavior of an optimal selection of $\delta^*$ (according to a suitably defined optimality criterion explained in Section 6) as $n, N \to \infty$.

In Section 2, we provide a broad overview of alternative procedures in the SSL literature, including recent approaches which are related to robust optimization. A key role in our formulation is played by $\delta^*$, which can be seen as a regularization parameter. This identification is highlighted in the form of (1) and the DRO representation of regularized logistic regression which we recall in (9). The optimal choice of $\delta^*$ ensures statistical consistency as $n, N \to \infty$.

Similar robust optimization formulations to (1) for machine learning have been investigated in the literature recently. For example, connections between robust optimization and machine learning procedures such as Lasso and SVMs have been studied in the literature, see [23]. In contrast to this literature, the use of distributionally robust uncertainty allows to discuss the optimal size of the uncertainty region as the sample size increases (as we shall explain in Section 6). The work of [20] is among the first to study DRO representations based on optimal transport, they do not study the implications of these types of DRO formulations in SSL as we do here.

We close this Introduction with a few important notes. First, our SSL-DRO is not a robustifying procedure for a given SSL algorithm. Instead, our contribution is in showing how to use unlabeled information on top of DRO to enhance traditional supervised learning methods. In addition, our SSL-DRO formulation, as stated in (1) , is not restricted to logistic regression, instead DRO counterpart could be formulated for general supervised learning methods with various choice of loss function.

The rest of the paper is structured as follows. We will quickly review the alternative related state-of-the-art SSL algorithms. In Section 3 we discuss the elements of our DRO formulation, including the definition of optimal transport metric and the implementation of a stochastic gradient descent algorithm for the solution of (1). In Section 4 we explore the improvement in out-of-sample performance of our method relative to regularized logistic regression. In Section 5, we compare our procedure against alternative SSL estimators, both in the context of some binary classification real data sets. In Section 6, we explore the behavior of the optimal uncertainty size $\delta^*$ as the sample size increases, especially we discuss certain asymptotic results on how to pick up the distributional uncertainty size optimally with asymptotic consistency. Section 7 contains final considerations and further discussions. In Appendix A and the supplementary material, we provide more technical details for the asymptotic results stated in Section 6.

## 2. Alternative Semisupervised Learning Procedures

We shall briefly discuss alternative procedures which are known in the SSL literature, which is quite substantial. We refer the reader to the excellent survey of [24] for a general overview of the area. Our goal here is to expose the similarities and connections between our approach and some of the methods that have been adopted in the community.

For example, broadly speaking graph-based methods [7] and [9] attempt to construct a graph which represents a sketch of a lower dimensional manifold in which the predictive variables lie. Once the graph is constructed, a regularization procedure is performed, which seeks to enhance generalization error along the manifold while ensuring continuity in the prediction regarding an intrinsic metric. Our approach bypasses the construction of the graph, which we see as a significant advantage of our procedure. However, we believe that the construction of the graph can be used to inform the choice of cost function $c(\cdot)$ which should reflect high transportation costs for moving mass away from the manifold sketched by the graph.

Some recent SSL estimators are based on robust optimization, such as the work of [1]. The difference between data driven DRO and robust optimization is that the inner maximization in (1) for robust optimization is not over probability models which are variations of the empirical distribution. Instead, in robust optimization, one attempts to minimize the risk of the worst case performance of potential outcomes inside a given uncertainty set.

In [1], the robust uncertainty set is defined in terms of constraints obtained from the testing set. The problem with the approach in [1] is that there is no clear mechanism which informs an optimal size of the uncertainty set (which in our case is parameterized by $\delta^*$). In fact, in the last paragraph of Section 2.3, [1] point out that the size of the uncertainty could have a significant detrimental impact in practical performance.

We conclude with a short discussion on the work of [14], which is related to our approach. In the context of linear discriminant analysis, [14] also proposes a distributionally robust optimization estimator, although completely different from the one we propose here. More importantly, we provide a way (both in theory and practice) to study the optimal size of the distributional uncertainty (i.e. $\delta^*$), which allows us to achieve asymptotic consistency of our estimator.

## 3. Semi-supervised Learning based on DRO

This section is divided into two parts. First, we provide the elements of our DRO formulation. Then we will explain how to solve the SSL-DRO problem, i.e. find optimal $\beta$ in (1).

### 3.1. Defining the optimal transport discrepancy:

Assume that the cost function $c : \mathbb{R}^{d+1} \times \mathbb{R}^{d+1} \to [0, \infty]$ is lower semicontinuous. As mentioned in the Introduction, we also assume that $c(u, v) = 0$ if and only if $u = v$.

Now, given two distributions $P$ and $Q$, with supports $\mathcal{S}_P \subseteq \mathcal{X}_N$ and $\mathcal{S}_Q \subseteq \mathcal{X}_N$, respectively, we define the optimal transport discrepancy, $D_c$, via

$$D_c(P, Q) = \inf\{E_\pi[c(U, V)] : \pi \in \mathcal{P}(\mathcal{S}_P \times \mathcal{S}_Q),\ \pi_U = P,\ \pi_V = Q\}, \tag{2}$$

where $\mathcal{P}(\mathcal{S}_P \times \mathcal{S}_Q)$ is the set of probability distributions $\pi$ supported on $\mathcal{S}_P \times \mathcal{S}_Q$, and $\pi_U$ and $\pi_V$ denote the marginals of $U$ and $V$ under $\pi$, respectively.

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), it can be easily verified (see [22]) that $D_c^{1/\varrho}(P, Q)$ 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 $\varrho$.

Observe that (2) is obtained by solving a linear programming problem. For example, suppose that $Q = P_n$, and let $P \in \mathcal{P}(\mathcal{X}_N)$ then, using $U = (X, Y)$, we have that $D_c(P, P_n)$ is obtained by computing

$$\min_{\pi} \Big\{ \sum_{u \in \mathcal{X}_N} \sum_{v \in \mathcal{D}_n} c(u,v)\pi(u,v) : \text{ s.t. } \sum_{u \in \mathcal{X}_N} \pi(u,v) = \frac{1}{n} \ \forall\ v \in \mathcal{D}_n, \sum_{v \in \mathcal{D}_N} \pi(u,v) = P(\{u\}) \ \forall\ u \in \mathcal{X}_N, \pi(u,v) \geq 0 \ \forall\ (u,v) \in \mathcal{X}_N \times \mathcal{D}_n \Big\} \tag{3}$$

We shall discuss, for instance, how the choice of $c(\cdot)$ in formulations such as (1) can be used to recover popular machine learning algorithms.

3.2. **Solving the SSL-DRO formulation:** A direct approach to solve (1) would involve alternating between minimization over $\beta$, which can be performed by, for example, stochastic gradient descent and maximization which is performed by solving a linear program similar to (3). Unfortunately, the large scale of the linear programming problem, which has $O(N)$ variables and $O(n)$ constraints, makes this direct approach rather difficult to apply in practice.

So, our goal here is to develop a direct stochastic gradient descent approach which can be used to approximate the solution to (1).

First, it is useful to apply linear programming duality to simplify (1). Note that, given $\beta$, the inner maximization in (1) is simply

$$\max_{\pi} \Big\{ \sum_{u \in \mathcal{X}_N} \sum_{v \in \mathcal{D}_N} l(u,\beta)\pi(u,v) : \text{ s.t. } \sum_{u \in \mathcal{X}_N} \pi(u,v) = \frac{1}{n} \ \forall\ v \in \mathcal{D}_n \sum_{u \in \mathcal{X}_N} \sum_{v \in \mathcal{D}_n} c(u,v)\,\pi(u,v) \leq \delta, \pi(u,v) \geq 0 \ \forall\ (u,v) \in \mathcal{X}_N \times \mathcal{D}_n \Big\}. \tag{4}$$

Of course, the feasible region in this linear program is always non-empty because the probability distribution $\pi(u,v) = I(u = v)\, I(v \in \mathcal{D}_n)/n$ is a feasible choice. Also, the feasible region is clearly compact, so the dual problem is always feasible and by strong duality its optimal value coincides with that of the primal problem, see [2], [3] and [6].

The dual problem associated to (4) is given by

$$\min \Big\{ \sum_{v \in \mathcal{D}_N} \gamma(v)/n + \lambda\delta \text{ s.t. } \gamma(v) \in \mathbb{R} \ \forall\ v \in \mathcal{D}_n,\ \lambda \geq 0, \gamma(v) \geq l(u,\beta) - \lambda c(u,v) \ \forall\ (u,v) \in \mathcal{X}_N \times \mathcal{D}_n. \Big\} \tag{5}$$

Maximizing over $u \in \mathcal{X}_N$ in the inequality constraint, for each $v$, and using the fact that we are minimizing the objective function, we obtain that (5) can be simplified to

$$E_{P_n}[\max_{u \in \mathcal{X}_N} \{l(u,\beta) - \lambda c(u,(X,Y)) + \lambda\delta^*\}] .$$

Consequently, defining $\phi(X, Y, \beta, \lambda) = \max_{u \in \mathcal{X}_N} \{l(u,\beta) - \lambda c(u,(X,Y)) + \lambda\delta^*\}$, we have that (1) is equivalent to

$$\min_{\lambda \geq 0, \beta} E_{P_n}[\phi(X, Y, \beta, \lambda)] . \tag{6}$$

Moreover, if we assume that $l(u,\cdot)$ is a convex function, then we have that the mapping $(\beta, \lambda) \hookrightarrow l(u,\beta) - \lambda c(u,(X,Y)) + \lambda\delta^*$is convex for each $u$ and therefore, $(\beta, \lambda) \hookrightarrow \phi(X, Y, \beta, \lambda)$, being the maximum of convex mappings is also convex.

A natural approach consists in directly applying stochastic sub-gradient descent (see [8] and [17]). Unfortunately, this would involve performing the maximization over all $u \in \mathcal{X}_N$ in each iteration. This approach could be prohibitively expensive in typical machine learning applications where $N$ is large.

So, instead, we perform a standard smoothing technique, namely, we introduce $\epsilon > 0$ and define

$$\phi_\epsilon(X,Y,\beta,\lambda) = \lambda\delta^* + \epsilon \log\Big(\sum_{u\in\mathcal{X}_N} \exp\left(\{l(u,\beta) - \lambda c(u,(X,Y))\}/\epsilon\right)\Big).$$

It is easy to verify (using Hölder inequality) that $\phi_\epsilon(X,Y,\cdot)$ is convex and it also follows that

$$\phi(X,Y,\beta,\lambda) \le \phi_\epsilon(X,Y,\beta,\lambda) \le \phi(X,Y,\beta,\lambda) + \log(|\mathcal{X}_N|)\epsilon.$$

Hence, we can choose $\epsilon = O(1/\log N)$ in order to control the bias incurred by replacing $\phi$ by $\phi_\epsilon$. Then, defining

$$\tau_\epsilon(X,Y,\beta,\lambda,u) = \exp\left(\{l(u,\beta) - \lambda c(u,(X,Y))\}/\epsilon\right),$$

we have (assuming differentiability of $l(u,\beta)$) that

$$\nabla_\beta \phi_\epsilon(X,Y,\beta,\lambda) = \frac{\sum_{u\in\mathcal{X}_N} \tau_\epsilon(X,Y,\beta,\lambda,u)\nabla_\beta l(u,\beta)}{\sum_{v\in\mathcal{X}_N}\tau_\epsilon(X,Y,\beta,\lambda,v)}, \tag{7}$$

$$\frac{\partial\phi_\epsilon(X,Y,\beta,\lambda)}{\partial\lambda} = \delta^* - \frac{\sum_{u\in\mathcal{X}_N} \tau_\epsilon(X,Y,\beta,\lambda,u)\, c(u,(X,Y))}{\sum_{v\in\mathcal{X}_N}\tau_\epsilon(X,Y,\beta,\lambda,v)}.$$

In order to make use of the gradient representations (7) for the construction of a stochastic gradient descent algorithm, we must construct unbiased estimators for $\nabla_\beta\phi_\epsilon(X,Y,\beta,\lambda)$ and $\partial\phi_\epsilon(X,Y,\beta,\lambda)/\partial\lambda$, given $(X,Y)$. This can be easily done if we assume that one can simulate directly $u \in \mathcal{X}_N$ with probability proportional to $\tau(X,Y,\beta,\lambda,u)$. Because of the potential size of $\mathcal{X}_N$ and specially because such distribution depends on $(X,Y)$ sampling with probability proportional to $\tau_\epsilon(X,Y,\beta,\lambda,u)$ can be very time consuming.

So, instead, we apply a strategy discussed in [4] and explained in Section 2.2.1. The proposed method produces random variables $\Lambda(X,Y,\beta,\lambda)$ and $\Gamma(X,Y,\beta,\lambda)$, which can be simulated easily by drawing i.i.d. samples from the uniform distribution over $\mathcal{X}_N$, and such that

$$E(\Lambda(X,Y,\beta,\lambda)|X,Y) = \partial_\lambda\phi_\epsilon(X,Y,\beta,\lambda), \quad E(\Gamma(X,Y,\beta,\lambda)|X,Y) = \nabla_\beta\phi_\epsilon(X,Y,\beta,\lambda).$$

Using this pair of random variables, then we apply the stochastic gradient descent recursion

$$\beta_{k+1} = \beta_k - \alpha_{k+1}\Gamma(X_{k+1},Y_{k+1},\beta_k,\lambda_k), \lambda_{k+1} = (\lambda_k - \alpha_{k+1}\Lambda(X_{k+1},Y_{k+1},\beta_k,\lambda_k))^+, \tag{8}$$

where learning sequence, $\alpha_k > 0$ satisfies the standard conditions, namely, $\sum_{k=1}^\infty \alpha_k = \infty$ and $\sum_{k=1}^\infty \alpha_k^2 < \infty$, see [21].

We apply a technique from [4] to construct the random variables $\Lambda$ and $\Gamma$, which originates from Multilevel Monte Carlo introduced in [10], and associated randomization methods [16],[18].

First, define $\bar{P}_N$ to be the uniform measure on $\mathcal{X}_N$ and let $W$ be a random variable with distribution $\bar{P}_N$. Note that, given $(X,Y)$,

$$\nabla_\beta\phi_\epsilon(X,Y,\beta,\lambda) = \frac{E_{\bar{P}_N}(\tau_\epsilon(X,Y,\beta,\lambda,W)\nabla_\beta l(W,\beta)\mid X,Y)}{E_{\bar{P}_N}(\tau_\epsilon(X,Y,\beta,\lambda,W)\mid X,Y)},$$

$$\partial_\lambda\phi_\epsilon(X,Y,\beta,\lambda) = \delta^* - \frac{E_{\bar{P}_N}(\tau_\epsilon(X,Y,\beta,\lambda,W)\, c(W,(X,Y))\mid X,Y)}{E_{\bar{P}_N}(\tau_\epsilon(X,Y,\beta,\lambda,W)\mid X,Y)}.$$

Note that both gradients can be written in terms of the ratios of two expectations. The following results from [4] can be used to construct unbiased estimators of functions of expectations. The function of interest in our case is the ratio of expectations.

Let us define:

$$h_0(W) = \tau_\epsilon(X, Y, \beta, \lambda, W),$$
$$h_1(W) = h_0(W)\, c(W, (X, Y)),$$
$$h_2(W) = h_0(W)\, \nabla_\beta l(W, \beta).$$

Then, we can write the gradient estimator as

$$\partial_\lambda \phi_\epsilon(X, Y, \beta, \lambda) = \frac{E_{\bar{P}_N}(h_1(W) \mid X, Y)}{E_{\bar{P}_N}(h_0(W) \mid X, Y)}, \text{ and } \nabla_\beta \phi_\epsilon(X, Y, \beta, \lambda) = \frac{E_{\bar{P}_N}(h_2(W) \mid X, Y)}{E_{\bar{P}_N}(h_0(W) \mid X, Y)}.$$

The procedure developed in [4] proceeds as follows. First, define for a given $h(W)$, and $n \geq 0$, the average over odd and even labels to be

$$\bar{S}^E_{2^n}(h) = \frac{1}{2^n}\sum_{i=1}^{2^n} h(W_{2i}), \quad \bar{S}^O_{2^n}(h) = \frac{1}{2^n}\sum_{i=1}^{2^n} h(W_{2i-1}),$$

and the total average to be $\bar{S}_{2^{n+1}}(h) = \frac{1}{2}\left(\bar{S}^E_{2^n}(h) + \bar{S}^O_{2^n}(h)\right)$. We then state the following algorithm for sampling unbiased estimators of $\partial_\lambda \phi_\epsilon(X, Y, \beta, \lambda)$ and $\nabla_\beta \phi_\epsilon(X, Y, \beta, \lambda)$ in Algorithm 1.

**Algorithm 1** Unbiased Gradient

1: Given $(X, Y, \beta)$ the function outputs $(\Lambda, \Gamma)$ such that $E(\Lambda) = \partial_\lambda \phi_\epsilon(X, Y, \beta, \lambda)$ and $E(\Gamma) = \nabla_\beta \phi_\epsilon(X, Y, \beta, \lambda)$.
2: **Step1:** Sample $G$ distributed geometric with success parameter $p_G = 1 - 2^{-3/2}$.
3: **Step2:** Sample $W_0, W_1, ..., W_{2^{G+1}}$ i.i.d. copies of $W$ independent of $G$.
4: **Step3:** Compute

$$\Delta^\lambda = \frac{\bar{S}_{2^{G+1}}(h_1)}{\bar{S}_{2^{G+1}}(h_0)} - \frac{1}{2}\left(\frac{\bar{S}^O_{2^{G+1}}(h_1)}{\bar{S}^O_{2^{G+1}}(h_0)} + \frac{\bar{S}^E_{2^G}(h_1)}{\bar{S}^E_{2^G}(h_0)}\right),$$
$$\Delta^\beta = \frac{\bar{S}_{2^{G+1}}(h_2)}{\bar{S}_{2^{G+1}}(h_0)} - \frac{1}{2}\left(\frac{\bar{S}^O_{2^{G+1}}(h_2)}{\bar{S}^O_{2^{G+1}}(h_0)} + \frac{\bar{S}^E_{2^G}(h_2)}{\bar{S}^E_{2^G}(h_0)}\right).$$

5: **Output:**

$$\Lambda = \delta^* - \frac{\Delta^\lambda}{p_G(1-p_G)^G} - \frac{h_1(W_0)}{h_0(W_0)}, \qquad \Gamma = \frac{\Delta^\beta}{p_G(1-p_G)^G} + \frac{h_2(W_0)}{h_0(W_0)}.$$

## 4. Error Improvement of Our SSL-DRO Formulation

Our goal in this section is to intuitively discuss why, owing to the inclusion of the constraint $P \in \mathcal{P}(\mathcal{X}_N)$, we expect desirable generalization properties of the SSL-DRO formulation (1). Moreover, our intuition suggests strongly why our SSL-DRO formulation should possess better generalization performance than natural supervised counterparts. We restrict the discussion for logistic regression due to the simple form of regularization connection we will make in (9), however, the error improvement discussion should also apply to general supervised learning setting.

As discussed in the Introduction using the game-theoretic interpretation of (1), by introducing $\mathcal{P}(\mathcal{X}_N)$, the SSL-DRO formulation provides a mechanism for choosing $\beta$ which focuses on potential out-of-sample scenarios which are more relevant based on available evidence.

Suppose that the constraint $P \in \mathcal{P}(\mathcal{X}_N)$ was not present in the formulation. So, the inner maximization in (1) is performed over all probability measures $\mathcal{P}\left(\mathbb{R}^{d+1}\right)$ (supported on some subset

of $\mathbb{R}^{d+1}$). As indicated earlier, we assume that $l(X,Y;\cdot)$ is strictly convex and differentiable, so the first order optimality condition $E_P(\nabla_\beta l(X,Y;\beta)) = 0$ characterizes the optimal choice of $\beta$ assuming the validity of the probabilistic model $P$. It is natural to assume that there exists an actual model underlying the generation of the training data, which we denote as $P_\infty$. Moreover, we may also assume that there exists a unique $\beta^*$ such that $E_{P_\infty}(\nabla_\beta l(X,Y;\beta^*)) = 0$.
The set

$$\mathcal{M}(\beta_*) = \{P \in \mathcal{P}\left(\mathbb{R}^{d+1}\right) : E_P(\nabla_\beta l(X,Y;\beta^*)) = 0\}$$

corresponds to the family of all probability models which correctly estimate $\beta^*$. Clearly, $P_\infty \in \mathcal{M}(\beta_*)$, whereas, typically, $P_n \notin \mathcal{M}(\beta_*)$. Moreover, if we write $\mathcal{X}_\infty = supp(P_\infty)$ we have that

$$P_\infty \in m(N,\beta^*) := \{P \in \mathcal{P}(\mathcal{X}_\infty) : \quad E_P(\nabla_\beta l(X,Y;\beta^*)) = 0\} \subset \mathcal{M}(\beta_*).$$

Since $\mathcal{X}_N$ provides a sketch of $\mathcal{X}_\infty$, then we expect to have that the extremal (i.e. worst case) measure, denoted by $P_N^*$, will be in some sense a better description of $P_\infty$.

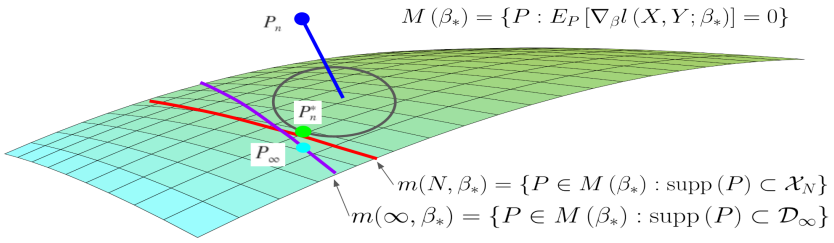


FIGURE 2. Pictorial representation of the role that the support constraint plays in the SSL-DRO approach and how its presence enhances the out-of-sample performance.

Figure 2 provides a pictorial representation of the previous discussion. In the absence of the constraint $P \in \mathcal{P}(\mathcal{X}_N)$, the extremal measure chosen by nature can be interpreted as a projection of $P_n$ onto $\mathcal{M}(\beta_*)$. In the presence of the constraint $P \in \mathcal{P}(\mathcal{X}_N)$, we can see that $P_N^*$ may bring the learning procedure closer to $P_\infty$. Of course, if $N$ is not large enough, the schematic may not be valid because one may actually have $m(N,\beta^*) = \varnothing$.
The previous discussion is useful to argue that our SSL-DRO formulation should be superior to the DRO formulation which is not informed by the unlabeled data. But this comparison may not directly apply to alternative supervised procedures that are mainstream in machine learning, which should be considered as the natural benchmark to compare with. Fortunately, replacing the constraint that $P \in \mathcal{P}(\mathcal{X}_N)$ by $P \in \mathcal{P}\left(\mathbb{R}^{d+1}\right)$ in the DRO formulation recovers exactly supervised learning algorithms such as regularized logistic regression.
Recall from [6] that if $l(x,y,\beta) = \log(1+\exp(-y\cdot\beta^T x))$ and if we define

$$c\left((x,y),(x',y')\right) = \|x-x'\|_q I(y=y') + \infty I(y\neq y'),$$

for $q \geq 1$ then, according to Theorem 3 in [6], we have that

$$\min_\beta \max_{D_c(P,P_n)\leq\bar{\delta}} E_P[l(X,Y,\beta)] = \min_{\beta\in\mathbb{R}^d}\left\{E_{P_n}[l(X,Y,\beta)] + \bar{\delta}\,\|\beta\|_p\right\}, \tag{9}$$

where $q$ satisfies $1/p + 1/q = 1$. Formulation (1) is, therefore, the natural SSL extension of the standard regularized logistic regression estimator.
We conclude that, for logistic regression, SSL-DRO as formulated in (1), is a natural SSL extension of the standard regularized logistic regression estimator, which would typically induce superior

generalization abilities over its supervised counterparts, and similar discussion should apply to most supervised learning methods.

## 5. Numerical Experiments

We proceed to numerical experiments to verify the performance of our SSL-DRO method empirically using six binary classification real data sets from UCI machine learning data base [13].
We consider our SSL-DRO formulation based on logistic regression and compare with other state-of-the-art logistic regression based SSL algorithms, entropy regularized logistic regression with $L_1$ regulation (ERLRL1) [11] and regularized logistic regression based self-training (STLRL1) [12]. In addition, we also compare with its supervised counterpart, which is regularized logistic regression (LRL1). For each iteration of a data set, we randomly split the data into labeled training, unlabeled training and testing set, we train the models on training sets and evaluate the testing error and accuracy with testing set. We report the mean and standard deviation for training and testing error using log-exponential loss and the average testing accuracy, which are calculated via 200 independent experiments for each data set. We summarize the detailed results, the basic information of the data sets, and our data split setting in Table 1.
We can observe that our SSL-DRO method has the potential to improve upon these state-of-the-art SSL algorithms.

Table 1. Numerical experiments for real data sets.

| | | Breast Cancer | qsar | Magic | Minibone | Spambase |
|---|---|---|---|---|---|---|
| LRL1 | Train | $.185 \pm .123$ | $.614 \pm .038$ | $.548 \pm .087$ | $.401 \pm .167$ | $.470 \pm .040$ |
| | Test | $.428 \pm .338$ | $.755 \pm .019$ | $.610 \pm .050$ | $.910 \pm .131$ | $.588 \pm .141$ |
| | Accur | $.929 \pm .023$ | $.646 \pm .036$ | $.665 \pm .045$ | $.717 \pm .041$ | $.811 \pm .034$ |
| ERLRL1 | Train | $.019 \pm .010$ | $.249 \pm .050$ | $2.37 \pm .987$ | $.726 \pm .353$ | $.008 \pm .028$ |
| | Test | $.265 \pm .146$ | $.720 \pm .029$ | $4.28 \pm 1.51$ | $1.98 \pm .678$ | $.505 \pm .108$ |
| | Accur | $.944 \pm .018$ | $.731 \pm .026$ | $.721 \pm .056$ | $.708 \pm .071$ | $.883 \pm .018$ |
| STLRL1 | Train | $.089 \pm .019$ | $.498 \pm .120$ | $3.05 \pm .987$ | $1.50 \pm .706$ | $.370 \pm .082$ |
| | Test | $.672 \pm .034$ | $2.37 \pm .860$ | $8.03 \pm 1.51$ | $4.81 \pm .732$ | $1.47 \pm .316$ |
| | Accur | $.955 \pm .023$ | $.694 \pm .038$ | $.692 \pm .056$ | $.704 \pm .033$ | $.843 \pm .023$ |
| DROSSL | Train | $.045 \pm .023$ | $.402 \pm .039$ | $.420 \pm .075$ | $.287 \pm .047$ | $.221 \pm .028$ |
| | Test | $.120 \pm .029$ | $.555 \pm .025$ | $.561 \pm .039$ | $.609 \pm .054$ | $.333 \pm .012$ |
| | Accur | $.956 \pm .016$ | $.734 \pm .025$ | $.733 \pm .034$ | $.710 \pm .032$ | $.892 \pm .009$ |
| Num Predictors | | 30 | 30 | 10 | 20 | 56 |
| Labeled Size | | 40 | 80 | 30 | 30 | 150 |
| Unlabeled Size | | 200 | 500 | 9000 | 5000 | 1500 |
| Testing Size | | 329 | 475 | 9990 | 125034 | 2951 |

## 6. Discussion on the Size of the Uncertainty Set

One of the advantages of DRO formulations such as (1) and (9) is that they lead to a natural criterion for the optimal choice of the parameter $\delta^*$ or, in the case of (9), the choice of $\bar{\delta}$ (which incidentally corresponds to the regularization parameter). The optimality criterion that we use to select the size of $\delta^*$ is motivated by Figure 2.
First, interpret the uncertainty set

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

as the set of plausible models which are consistent with the empirical evidence encoded in $P_n$ and $\mathcal{X}_N$. Then, for every plausible model $P$, we can compute

$$\beta(P) = \arg\min_{\beta} E_P[l(X, Y, \beta)]$$

and therefore the set

$$\Lambda_\delta(P_n, \mathcal{X}_N) = \{\beta(P) = \arg\min E_P[l(X, Y, \beta)] : P \in \mathcal{U}_\delta(P_n, \mathcal{X}_N)\}$$

can be interpreted as a confidence region. It is then natural to select a confidence level $\alpha \in (0, 1)$ and compute $\delta^* := \delta^*_{N,n}$ by solving

$$\min\{\delta : P(\beta^* \in \Lambda_\delta(P_n, \mathcal{X}_N)) \geq 1 - \alpha\}. \tag{10}$$

Similarly, for the supervised version, we can select $\bar{\delta} = \bar{\delta}_n$ by solving the problem

$$\min\{\delta : P\left(\beta^* \in \Lambda_\delta\left(P_n, \mathbb{R}^{d+1}\right)\right) \geq 1 - \alpha\}. \tag{11}$$

It is easy to see that $\bar{\delta}_n \leq \delta^*_{N,n}$. Now, we let $N = \gamma n$ for some $\gamma > 0$ and consider $\delta^*_{N,n}$, $\bar{\delta}_n$ as $n \to \infty$. This analysis is relevant because we are attempting to sketch $supp(P_\infty)$ using the set $\mathcal{X}_N$, while considering large enough plausible variations to be able to cover $\beta^*$ with $1 - \alpha$ confidence. More precisely, following the discussion in [6] for the supervised case in finding $\bar{\delta}_n$ in (10) using Robust Wasserstein Profile (RWP) function, solving (11) for $\delta^*_{N,n}$ is equivalent to finding the $1 - \alpha$ quantile of the asymptotic distribution of the RWP function, defined as

$$R_n(\beta) = \min_{\pi}\Big\{ \sum_{u\in\mathcal{X}_n}\sum_{v\in\mathcal{D}_n} c(u, v)\pi(u, v), \sum_{u\in\mathcal{X}_n} \pi(u, v) = \frac{1}{n}, \forall v\mathcal{D}_n, \\ \pi \subset \mathcal{P}(\mathcal{X}_n \times \mathcal{D}_n), \sum_{u\in\mathcal{X}_n}\sum_{v\in\mathcal{D}_n} \nabla_\beta l(u; \beta)\pi(u, v) = 0.\Big\}. \tag{12}$$

The RWP function is the distance, measured by the optimal transport cost function, between the empirical distribution and the manifold of probability measures for which $\beta_*$ is the optimal parameter. A pictorial representation is given in Figure 2. Additional discussion on the RWP function and its interpretations can be found in [6, 5].

In the setting of the DRO formulation for (9) it is shown in [6], that $\bar{\delta}_n = O\left(n^{-1}\right)$ for (9) as $n \to \infty$. Intuitively, we expect that if the predictive variables possess a positive density supported in a lower dimensional manifold of dimension $\bar{d} < d$, then sketching $supp(P_\infty)$ with $O(n)$ data points will leave relatively large portions of the manifold unsampled (since, on average, $O\left(n^{\bar{d}}\right)$ sampled points are needed to be within distance $O(1/n)$ of a given point in box of unit size in $\bar{d}$ dimensions). The optimality criterion will recognize this type of discrepancy between $\mathcal{X}_N$ and $supp(P_\infty)$. Therefore, we expect that $\delta^*_{\gamma n,n}$ will converge to zero at a rate which might deteriorate slightly as $\bar{d}$ increases. This intuition is given rigorous support in Theorem 1 for the case of linear regression with square loss function and $L_2$ cost function for DRO. In turn, Theorem 1 follows as a corollary to the results in [5]. Detailed assumptions are given in the appendix. To make our discussion self-contained we have added a sketch of its proof in the appendix and supplementary material.

*Theorem* 1. Assume the linear regression model $Y = \beta^* X + e$ with square loss function, i.e. $l(X, X; \beta) = \left(Y - \beta^T X\right)^2$, and transport cost

$$c\left((x, y), (x', y')\right) = \left\|x - x'\right\|_2^2 I_{y=y'} + \infty I_{y\neq y'}.$$

Assume $N = \gamma n$ and under mild assumptions on $(X, Y)$, if we denote $\tilde{Z} \sim \mathcal{N}(0, E[V_1])$, we have:

- When $d = 1$,
$$nR_n(\beta_*) \Rightarrow \kappa_1 \chi_1^2.$$
- When $d = 2$,
$$nR_n(\beta_*) \Rightarrow F_2\left(\tilde{Z}\right),$$
where $F_2(\cdot)$ is a continuous function and $F_2(z) = O(\|z\|_2^2)$ as $\|z\|_2 \to \infty$.
- When $d \geq 3$,
$$n^{1/2+\frac{3}{2d+2}} R_n(\beta_*) \Rightarrow F_d\left(\tilde{Z}\right),$$
where $F_d(\cdot)$ is a continuous function (depending on $d$) and $F_d(z) = O\big(\|z\|_2^{d/2+1}\big)$.

It is shown in Theorem 1 for SSL linear regression that when $q = 2$, $\delta^*_{\gamma n,n} = O\big(n^{-1/2-3/(2\bar{d}+2)}\big)$ for $\bar{d} \geq 3$, and $\delta^*_{\gamma n,n} = O\left(n^{-1}\right)$ for $\bar{d} = 1, 2$. A similar argument can be made for logistic regression as well. We believe that this type of analysis and its interpretation is of significant interest and we expect to report a more complete picture in the future, including the case $q \geq 1$ (which we believe should obey the same scaling).

## 7. Conclusions

We have shown that our SSL-DRO, as a SSL method, is able to enhance the generalization predicting power versus its supervised counterpart. Our numerical experiments show superior performance of our SSL-DRO method when compared to state-of-the-art SSL algorithms such as ERLRL1 and STLRL1. We would like to emphasize that our SSL-DRO method is not restricted to linear and logistic regressions. As we can observe from the DRO formulation and the algorithm. If a learning algorithm has an accessible loss function and the loss gradient can be computed, we are able to formulate the SSL-DRO problem and benefit from unlabeled information. Finally, we discussed a stochastic gradient descent technique for solving DRO problems such as (1), which we believe can be applied to other settings in which the gradient is a non-linear function of easy-to-sample expectations.

## References


[1] Akshay Balsubramani and Yoav Freund. Scalable semi-supervised aggregation of classifiers. In *Advances in Neural Information Processing Systems*, pages 1351–1359, 2015.

[2] Dimitris Bertsimas, David Brown, and Constantine Caramanis. Theory and applications of robust optimization. *SIAM review*, 53(3):464–501, 2011.

[3] Dimitris Bertsimas, Vishal Gupta, and Nathan Kallus. Data-driven robust optimization. *arXiv preprint arXiv:1401.0212*, 2013.

[4] Jose Blanchet and Peter Glynn. Unbiased Monte Carlo for optimization and functions of expectations via multi-level randomization. In *Proceedings of the 2015 Winter Simulation Conference*, pages 3656–3667. IEEE Press, 2015.

[5] Jose Blanchet and Yang Kang. Sample out-of-sample inference based on wasserstein distance. *arXiv preprint arXiv:1605.01340*, 2016.

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

[7] Avrim Blum and Shuchi Chawla. Learning from labeled and unlabeled data using graph mincuts. 2001.

[8] Stephen Boyd and Lieven Vandenberghe. *Convex optimization*. Cambridge university press, 2004.

[9] Olivier Chapelle, Bernhard Scholkopf, and Alexander Zien. Semi-supervised learning. *IEEE Transactions on Neural Networks*, 20(3):542–542, 2009.
[10] Michael Giles. Multilevel Monte Carlo path simulation. *Operations Research*, 56(3), 2008.
[11] Yves Grandvalet and Yoshua Bengio. Semi-supervised learning by entropy minimization. In *Advances in NIPS*, pages 529–536, 2005.
[12] Yuanqing Li, Cuntai Guan, Huiqi Li, and Zhengyang Chin. A self-training semi-supervised svm algorithm and its application in an eeg-based brain computer interface speller system. *Pattern Recognition Letters*, 29(9):1285–1294, 2008.
[13] Moshe. Lichman. UCI machine learning repository, 2013.
[14] Marco Loog. Contrastive pessimistic likelihood estimation for semi-supervised classification. *IEEE transactions on pattern analysis and machine intelligence*, 38(3):462–475, 2016.
[15] David G Luenberger. *Introduction to linear and nonlinear programming*, volume 28. Addison-Wesley Reading, MA, 1973.
[16] Don McLeish. A general method for debiasing a Monte Carlo estimator. *Monte Carlo Meth. and Appl.*, 17(4):301–315, 2011.
[17] Sundhar Ram, Angelia Nedić, and Venugopal Veeravalli. Distributed stochastic subgradient projection algorithms for convex optimization. *Journal of optimization theory and applications*, 147(3):516–545, 2010.
[18] Chang-han Rhee and Peter Glynn. Unbiased estimation with square root convergence for SDE models. *Operations Research*, 63(5):1026–1043, 2015.
[19] Yossi Rubner, Carlo Tomasi, and Leonidas Guibas. The earth mover's distance as a metric for image retrieval. *International journal of computer vision*, 40(2):99–121, 2000.
[20] Soroosh Shafieezadeh-Abadeh, Peyman Mohajerin Esfahani, and Daniel Kuhn. Distributionally robust logistic regression. In *Advances in Neural Information Processing Systems*, pages 1576–1584, 2015.
[21] Alexander Shapiro, Darinka Dentcheva, et al. *Lectures on stochastic programming: modeling and theory*, volume 16. Siam, 2014.
[22] Cédric Villani. *Optimal transport: old and new*, volume 338. Springer Science & Business Media, 2008.
[23] Huan Xu, Constantine Caramanis, and Shie Mannor. Robust regression and lasso. In *Advances in Neural Information Processing Systems*, pages 1801–1808, 2009.
[24] Xiaojin Zhu, John Lafferty, and Ronald Rosenfeld. *Semi-supervised learning with graphs*. Carnegie Mellon University, 2005.

## Appendix A. Technical Details for Theorem 1

In this appendix, we first state the general assumptions to guarantee the validity of the asymptotically optimal selection for the distributional uncertainty size in Section A.1. In Section A.2 we provide a roadmap for the proof of Theorem 1. In the supplementary material, we are more complete details.

A.1. **Assumptions of Theorem 1.** For linear regression model, let us assume we have a collection of labeled data $\mathcal{D}_n = \{(X_i, Y_i)\}_{i=1}^n$ and a collection of unlabeled data $\{X_i\}_{i=n+1}^N$. We consider the set $\mathcal{X}_N = \{X_i\}_{i=1}^N \times \{Y_i\}_{i=1}^n$, to be the cross product of all the predictors from labeled and unlabeled data and the labeled responses. In order to have proper asymptotic results holds for the RWP function, we require some mild assumptions on the density and moments of $(X, Y)$ and estimating equation $\nabla_\beta l\left(X, Y; \beta\right) = \left(Y - \beta_*^T\right) X$. We state them explicitly as follows:

**A)** We assume the predictors $X_i$'s for the labeled and unlabeled data are i.i.d. from the same distribution with positive differentiable density $f_X(\cdot)$ with bounded bounded gradients.
**B)** We assume the $\beta_* \in \mathbb{R}^d$ is the true parameter and under null hypothesis of the linear regression model satisfying $Y = \beta_*^T X + e$, where $e$ is a random error independent of $X$.
**C)** We assume $E\left[X^T X\right]$ exists and is positive definite and $E\left[e^2\right] < \infty$.
**D)** For the true model of labeled data, we have $E_{P_*}\left[X\left(Y - \beta_*^T X\right)\right] = 0$ (where $P_*$ denotes the actual population distribution which is unknown).
The first two assumptions, namely Assumption A and B, are elementary assumptions for linear regression model with an additive independent random error. The requirements for the differentiable positive density for the predictor $X$, is because when $d \geq 3$, the density function appears in the asymptotic distribution. Assumption C is a mild requirement on the moments exist for predictors and error, and Assumption D is to guarantee true parameter $\beta_*$ could be characterized via first order optimality condition, i.e. the gradient of the square loss function. Due to the simple structure of the linear model, with the above four assumptions, we can prove Theorem 1 and we show a sketch in the following subsection.

A.2. **Sketch of the Proof of Theorem 1.** Theorem 1 is a corollary of Theorem 3 in [5], although its proof requires some adaptations. The proof of Theorem 1 follows the 6-step procedure explained in Section 3 of [5]. We highlight the main differences in deriving the duality of the RWP function in this section. To make the paper more self-contained, we include more technical details borrowed from [5] in the supplementary material.

*Sketch of the Proof of Theorem 1.* **Deriving Strong Duality From for RWP Function.** For $u \in \mathcal{D}_n$ and $v \in \mathcal{X}_N$, let us denote $u_x, u_y$ and $v_x, v_y$ to be its subvectors for the predictor and response. By the definition of RWP function as in (12), we can write it as a linear program (LP), given as

$$R_n(\beta_*) = \min_{\pi}\Big\{ \sum_{u\in\mathcal{D}_n}\sum_{v\in\mathcal{X}_N} \pi(u,v)\left(\|u_x - v_x\|_2^2 I_{v_y=u_y} + \infty I_{v_y\neq u_y}\right) \ \textbf{s.t.} \quad \pi \in \mathcal{P}(\mathcal{X}_N \times \mathcal{D}_n),$$
$$\sum_{u\in\mathcal{D}_n}\sum_{v\in\mathcal{X}_N} \pi(u,v)\, v_x\left(v_y - \beta_*^T v_x\right) = 0, \sum_{v\in\mathcal{X}_N}\pi(u,v) = 1/n, \forall u \in \mathcal{D}_n.\Big\}$$

For as $n$ large enough the LP is finite and feasible (because $P_n$ approaches $P_*$, and $P_*$ is feasible). Thus, for $n$ large enough we can write

$$R_n(\beta_*) = \min_{\pi}\Big\{ \sum_{u\in\mathcal{D}_n}\sum_{v_x\in\{X_i\}_{i=1}^N} \pi(u,v_x)\|u_x - v_x\|_2^2 \quad \textbf{s.t.} \quad \pi \in \mathcal{P}(\mathcal{X}_N \times \mathcal{D}_n)$$
$$\sum_{u\in\mathcal{D}_n}\sum_{v\in\mathcal{X}_N} \pi(u,v)\, v_x\left(u_y - \beta_*^T v_x\right) = 0, \sum_{v\in\mathcal{X}_N}\pi(u,v) = 1/n, \forall u \in \mathcal{D}_n.\Big\}$$

We can apply strong duality theorem for LP, see [15], and write the RWP function in dual form:

$$R_n(\beta_*) = \max_{\lambda}\left\{\frac{1}{n}\sum_{i=1}^n \min_{j=\overline{1,N}}\left\{-\lambda^T X_j\left(Y_i - \beta_*^T X_j\right) + \|X_i - X_j\|_2^2\right\}\right\},$$
$$= \max_{\lambda}\Big\{\frac{1}{n}\sum_{i=1}^n -\lambda^T X_i\left(Y_i - \beta_*^T X_i\right)$$
$$+ \min_{j=\overline{1,N}}\left\{\lambda^T X_i\left(Y_i - \beta_*^T X_j\right) - \lambda^T X_j\left(Y_i - \beta_*^T X_j\right) + \|X_i - X_j\|_2^2\right\}\Big\},.$$

This finishes Step 1 as in the 6-step proving technique introduced in Section 3 of [5].

In Step 2 and Step 3, after rescaling the RWP function by $n$ for $d = 1$ and 2 and rescaling by $n^{\frac{1}{2}+\frac{3}{2d+2}}$ for $d \geq 3$, we can quantify the difference between the inner minimization problem for each $i$,

$$\min_{j=\overline{1,N}} \left\{ \lambda^T X_i \left(Y_i - \beta_*^T X_i\right) - \lambda^T X_j \left(Y_i - \beta_*^T X_j\right) + \|X_i - X_j\|_2^2 \right\}$$

and its lower bound,

$$\min_{a} \left\{ \lambda^T X_i \left(Y_i - \beta_*^T X_i\right) - \lambda^T a \left(Y_i - \beta_*^T a\right) + \|X_i - a\|_2^2 \right\},$$

by defining a family of auxiliary, weakly dependent, Poisson point processes (indexed by $i$).

Applying the results in Step 3, we can prove the asymptotic distribution for $d = 1$ in Step 4, $d = 2$ in Step 5, and $d \geq 3$ in Step 6 using the Central Limit Theorem (CLT) and the Continuous Mapping Theorem. More details are shown in the supplementary material Section B.2.

□

## Appendix B. Supplementary Material

In this supplementary material, we will restate Theorem 1 more explicitly to show how the asymptotic distribution varies for different dimension $d$ in Section B.1. In Section B.2, we will feed more technical details in proving Theorem 1.

B.1. **Revisit Theorem 1.** In this section, we revisit the asymptotic result for optimally choosing uncertainty size for semi-supervised learning for the linear regression model. We assume that, under the null hypothesis, $Y = \beta_*^T X + e$, where $X \in \mathbb{R}^d$ is the predictors, $e$ is independent of $X$ as random error, and $\beta_* \in \mathbb{R}^d$ is the true parameter. We consider the square loss function and assume that $\beta_*$ is the minimizer to the square loss function, i.e.

$$\beta_* = \arg\min_{\beta} E\left[\left(Y - \beta^T X\right)^2\right].$$

If we can assume the second-moment exists for $X$ and $e$, then we can switch the order of expectation and derivative w.r.t. $\beta$, then optimal $\beta$ could be uniquely characterized via the first order optimality condition,

$$E\left[X\left(Y - \beta_*^T X\right)\right] = 0.$$

As we discussed in Section 6, the optimal distributional uncertainty size $\delta_{n,N}^*$ at confidence level $1-\alpha$, is simply the $1-\alpha$ quantile of the RWP function defined in (12). In turn, the asymptotic limit of the RWP function is characterized in Theorem 1, which we restate more explicitly here.

*Theorem* 1 (Restate of Theorem 1 in Section 6). For linear regression model we defined above and square loss function, if we take cost function for DRO formulation to be

$$c\left((x,y),\left(x',y'\right)\right) = \left\|x - x'\right\|_2^2 I_{y=y'} + \infty I_{y\neq y'}.$$

If we assume Assumptions A,B, and D stated in Section A.1 to be true and number of unlabeled data satisfing $N = \gamma n$. Furthermore, let us denote: $V_i = \left(e_i I - X_i \beta_*^T\right)\left(e_i I - \beta_* X_i^T\right)$, where $e_i = Y_i - \beta_*^T X_i$ being the residual under the null hypothesis. Then, we have:

- When $d = 1$,

$$nR_n(\beta_*) \Rightarrow \frac{E\left[X_1^2 e_1^2\right]}{E\left[\left(e_1 - \beta_*^T X_1\right)^2\right]}\chi_1^2.$$

- When $d = 2$,

$$nR_n(\beta_*) \Rightarrow 2\tilde{\zeta}(\tilde{Z})^T\tilde{Z} - \tilde{\zeta}\left(\tilde{Z}\right)^T \tilde{G}_2\left(\tilde{\zeta}\left(\tilde{Z}\right)\right)\tilde{\zeta}\left(\tilde{Z}\right),$$

  where $\tilde{Z} \sim \mathcal{N}\left(0, E[V_1]\right)$, $\tilde{G}_2 : \mathbb{R}^2 \to \mathbb{R}^2 \times \mathbb{R}^2$ is a continuous mapping defined as

$$\tilde{G}_2(\zeta) = E\left[V_1 \max\left(1 - \tau/(\zeta^T V_1 \zeta), 0\right)\right],$$

  and $\tilde{\zeta} : \mathbb{R}^2 \to \mathbb{R}^2$ is a continuous mapping, such that $\tilde{\zeta}(\tilde{Z})$ is the unique solution to

$$\tilde{Z} = -E\left[V_1 I_{(\tau \leq \zeta^T V_1 \zeta)}\right]\zeta.$$

- When $d \geq 3$,

$$n^{1/2+\frac{3}{2d+2}} R_n(\beta_*) \Rightarrow -2\tilde{\zeta}(\tilde{Z})^T\tilde{Z} - \frac{2}{d+2}\tilde{G}_3\left(\tilde{\zeta}(\tilde{Z})\right),$$

where $\tilde{Z} \sim \mathcal{N}(0, E[V_1])$, $\tilde{G}_2 : \mathbb{R}^d \to \mathbb{R}$ is a deterministic continuous function defined as

$$\tilde{G}_2(\zeta) = E\left[\frac{\pi^{d/2}\gamma f_X(X_1)}{\Gamma(d/2+1)}\left(\zeta^T V_1 \zeta\right)^{d/2+1}\right],$$

and $\tilde{\zeta} : \mathbb{R}^d \to \mathbb{R}^d$ js a continuous mapping, such that $\tilde{\zeta}(\tilde{Z})$ is the unique solution to

$$\tilde{Z} = -E\left[V_1 \frac{\pi^{d/2}\gamma f_X(X_1)}{\Gamma(d/2+1)}\left(\zeta^T V_1 \zeta\right)^{d}\right]\zeta.$$

B.2. **Proof of Theorem 1.** In this section, we complete the proof for Theorem 1 in addition to the scratch in Section A.2. As we discussed before, Theorem 1 could be treated as a non-trivial corollary of Theorem 3 in [5] and the proving techniques follow the 6-step proof for Sample-out-of-Sample (SoS) Theorem, namely Theorem 1 and Theorem 3 in [5].

*Proof of Theorem 1.* We derived the duality formulation for RWP function in Section A.2 as the Step 1 of the proof.

**Step 2 and Step 3**, When $d = 1$ and 2, we consider scaling the RWP function by $n$ and let define $\zeta = \sqrt{n}\lambda/2$ and denote $W_n = n^{-1/2}\sum_{i=1}^n X_i e_i$, we have the scaled RWP function becomes,

$$\begin{aligned} nR_n(\beta_*) = &\max_{\zeta}\Big\{ -\zeta^T W_n \\ &+ \sum_{i=1}^n \min_{j=\overline{1,N}}\{-2\frac{\zeta^T}{\sqrt{n}}X_j\left(Y_i - \beta_*^T X_j\right) + 2\frac{\zeta^T}{\sqrt{n}}X_i\left(Y_i - \beta_*^T X_i\right) + \|X_i - X_j\|_2^2\}\Big\}. \end{aligned}$$

For each fixed $i$, let us consider the inner minimization problem,

$$\min_{j=\overline{1,N}}\{-2\frac{\zeta^T}{\sqrt{n}}X_j\left(Y_i - \beta_*^T X_j\right) + 2\frac{\zeta^T}{\sqrt{n}}X_i\left(Y_i - \beta_*^T X_i\right) + \|X_i - X_j\|_2^2\}$$

Similar to Section 3 in [5], we would like to solve the minimization problem by first replacing $X_j$ by $a$, which is a free variable without support constraint in $\mathbb{R}^d$, then quantify the gap. We then obtain a lower bound for the optimization problem via

$$\min_{a}\{-2\frac{\zeta^T}{\sqrt{n}}a\left(Y_i - \beta_*^T a\right) + 2\frac{\zeta^T}{\sqrt{n}}X_i\left(Y_i - \beta_*^T X_i\right) + \|X_i - a\|_2^2\}.$$

As we can observe in (**??**), the coefficient of second order of $a$ is of order $O(1/\sqrt{n})$ for any fixed $\zeta$, and the coefficients for the last term is always 1, it is easy to observe that, as $n$ large enough, (**??**) has an optimizer in the interior.

We can solve for the optimizer $a = \bar{a}_*(X_i, Y_i, \zeta)$ of the lower bound in (**??**) satisfying the first order optimality condition as

$$\begin{aligned} \bar{a}_*(X_i, Y_i, \zeta) - X_i = &\left(e_i I - \beta_*^T X_i\right)\frac{\zeta}{\sqrt{n}} \\ &+ \left(\beta_*^T\left(\bar{a}_*(X_i, Y_i, \zeta) - X_i\right) I - \left(\bar{a}_*(X_i, Y_i, \zeta) - X_i\right)\beta_*^T\right)\frac{\zeta}{\sqrt{n}}. \end{aligned} \tag{13}$$

Since the optimizer $\bar{a}_*(X_i, Y_i, \zeta)$ is in the interior, it is easy to notice from (13) that $\bar{a}_*(X_i, Y_i, \zeta) - X_i = O\left(\frac{\|\zeta\|_2}{\sqrt{n}}\right)$. Plug in the estimate back into (13) obtain

$$\bar{a}_*(X_i, Y_i, \zeta) = X_i + \left(e_i I - \beta_*^T X_i\right) \frac{\zeta}{\sqrt{n}} + O\big(\frac{\|\zeta\|_2^2}{n}\big). \tag{14}$$

Let us define $a_*(X_i, Y_i, \zeta) = X_i + \left(e_i I - \beta_*^T X_i\right) \frac{\zeta}{\sqrt{n}}$. Using (14), we have

$$\|a_*(X_i, Y_i, \zeta) - \bar{a}_*(X_i, Y_i, \zeta)\|_2 = O\big(\frac{\|\zeta\|_2^2}{n}\big). \tag{15}$$

Then, for the optimal value function of lower bound of the inner optimization problem, we have:

$$\begin{aligned}
&-2\frac{\zeta^T}{\sqrt{n}}\bar{a}_*(X_i, Y_i, \zeta)\left(Y_i - \beta_*^T a\right) + 2\frac{\zeta^T}{\sqrt{n}}X_i\left(Y_i - \beta_*^T X_i\right) + \|X_i - \bar{a}_*(X_i, Y_i, \zeta)\|_2^2 \\
&= -2\frac{\zeta^T}{\sqrt{n}}a_*(X_i, Y_i, \zeta)\left(Y_i - \beta_*^T a\right) + 2\frac{\zeta^T}{\sqrt{n}}X_i\left(Y_i - \beta_*^T X_i\right) + \|X_i - a_*(X_i, Y_i, \zeta)\|_2^2 + O\big(\frac{\|\zeta\|_2^3}{n^{3/2}}\big) \\
&= \frac{\zeta^T V_i \zeta}{n} + O\big(\frac{\|\zeta\|_2^3}{n^{3/2}}\big). \qquad (16)
\end{aligned}$$

For the above equation, first equality is due to (15) and the second equality is by the estimation of $\bar{a}_*(X_i, Y_i, \zeta)$ in (14).

Then for each fixed $i$, let us define a point process

$$N_n^{(i)}(t, \zeta) = \#\left\{X_j : \|X_j - a_*(X_i, Y_i, \zeta)\|_2^2 \le t^{2/d}/n^{2/d}, X_j \neq X_i\right\}.$$

We denote $T_i(n)$ to be the first jump time of $N_n^{(i)}(t, \zeta)$, i.e.

$$T_i(n) = \inf\left\{t \ge 0 : N_n^{(i)}(t, \zeta) \ge 1\right\}.$$

It is easy to observe that, as $n$ goes to infinity, we have

$$N_n^{(i)}(t, \zeta)\,|X_i \Rightarrow Poi\left(\Lambda(X_i, \zeta), t\right),$$

where $Poi\left(\Lambda(X_i, \zeta), t\right)$ denotes a Poisson point process with rate

$$\Lambda(X_i, \zeta) = \gamma f_X\left(X_i + \frac{\zeta}{2\sqrt{\zeta}}\right)\frac{\pi^{d/2}}{\Gamma\left(d/2 + 1\right)}.$$

Then, the conditional survival function for $T_i(n)$, i.e. $P\left(T_i(n) \ge t|X_i\right)$ is

$$P\left(T_i(n) \ge t|X_i\right) = \exp\left(-\Lambda\left(X_i, \zeta\right)t\right)\left(1 + O\left(1/n^{1/d}\right)\right),$$

and we can define $\tau_i$ to be the random variable with survival function being

$$P\left(\tau_i(n) \ge t|X_i\right) = \exp\left(-\Lambda\left(X_i, \zeta\right)t\right).$$

We can also integrate the dependence on $X_i$ and define $\tau$ satisfying

$$P\left(\tau \ge t\right) = E\left[\exp\left(-\Lambda\left(X_1, \zeta\right)t\right)\right].$$

Therefore,for $d = 1$ by the definition of $T_i(n)$ and the estimation in (16), we have the scaled RWP function becomes

$$nR_n(\beta_*) = \max_{\zeta}\Big\{-2\zeta W_n - \frac{1}{n}\sum_{i=1}^{n}\max\left(\zeta^T V_i \zeta - T_i(n)^2/n + O\big(\frac{\|\zeta\|_2^3}{n^{3/2}}\big), 0\right)\Big\}$$

The sequence of global optimizers is tight as $n \to \infty$, because according to Assumption C, $E(V_i)$ is assumed to be strictly positive definite with probability one. In turn, from the previous expression we can apply Lemma 1 in [5] and use the fact that the variable $\zeta$ can be restricted to compact sets for all $n$ sufficiently large. We are then able to concludee

$$nR_n(\beta_*) = \max_{\zeta} \{ -2\zeta^T W_n - E\left[\max\left(\zeta^T V_i \zeta - T_i(n)^2/n, 0\right)\right] \} + o_p(1). \tag{17}$$

When $d = 2$, a similar estimation applies as for the case $d = 1$. the scaled RWP function becomes

$$nR_n(\beta_*) = \max_{\zeta} \{ -2\zeta^T W_n - E\left[\max\left(\zeta^T V_i \zeta - T_i(n)^2, 0\right)\right] \} + o_p(1). \tag{18}$$

For the case when $d \geq 3$, let us define $\zeta = \lambda/(2n^{\frac{3}{2d+2}})$. We follow a similar estimation procedure as in the cases $d = 1, 2$. We also define identical auxiliary Poisson point process, we can write the scaled RWP function to be

$$\begin{aligned} n^{\frac{1}{2}+\frac{3}{2d+2}} R_n(\beta_*) = \max_{\zeta} \{ & -2\zeta^T W_n \\ & - n^{\frac{1}{2}+\frac{3}{2+2d}-\frac{2}{d}} E\left[\max\left(n^{\frac{2}{2}-\frac{6}{2d+2}} \zeta^T V_i \zeta - T_i(n)^{3/d}, 0\right)\right] \} + o_p(1). \end{aligned} \tag{19}$$

This addresses Step 2 and 3 in the proof.

**Step 4:** when $d = 1$, as $n \to \infty$, we have the scaled RWP function given in (17). Let us use $G_1 : \mathbb{R} \to \mathbb{R}$ to denote a deterministic continuous function defined as

$$G_1(\zeta, n) = E\left[\max\left(\zeta^T V_i \zeta - T_i(n)^2/n, 0\right)\right].$$

By Assumption C, we know $EV_i$ is positive, thus $G_1$ as a function of $\zeta$ is strictly convex. Thus the optimizer for the scaled RWP function could be uniquely characterized via the first order optimality condition, which is equivalent to

$$\zeta_n^* = -\frac{W_n}{E[V_i]} + o_p(1), \text{ as } n \to \infty. \tag{20}$$

We plug in (20) into (17) and let $n \to \infty$. Applying the CLT for $W_n$ and the continuous mapping theorem, we have

$$nR_n(\beta_*) = 2W_n^2/E[V_1] - G_1\left(-\frac{W_n}{E[V_1]}, n\right) + o_p(1) \Rightarrow \frac{\tilde{Z}^2}{E[V_1]} = \frac{E\left[X_1^2 e_1^2\right]}{E\left[(e_1 - \beta * X_1)^2\right]} \chi_1^2,$$

where $W_n \Rightarrow \tilde{Z}$ and $\tilde{Z} \sim \mathcal{N}\left(0, E\left[(e_1 - \beta * X_1)^2\right]\right)$.
We conclude the stated convergence for $d = 1$.

**Step 5:** when $d = 2$, as $n \to \infty$, we have the scaled RWP function given in (18). Let us use $G_2 : \mathbb{R} \times \mathbb{N} \to \mathbb{R}$ to denote a deterministic continuous function defined as

$$G_2(\zeta, n) = E\left[\max\left(\zeta^T V_i \zeta - T_i(n)^2, 0\right)\right].$$

Following the same discussion as in Step 4 for the case $d = 1$, we know that the optimizer $\zeta_n^*$ can be uniquely characterized via first order optimality condition given as

$$W_n = -E\left[V_1 I_{(\tau \leq \zeta^T V_1 \zeta)}\right] \zeta + o_p(1), \text{ as } n \to \infty.$$

Since we know that the objective function is strictly convex there exist a continuous mapping, $\tilde{\zeta}:\mathbb{R}^2\to\mathbb{R}^2$, such that $\tilde{\zeta}(W_n)$ is the unique solution to

$$W_n=-E\left[V_1 I_{(\tau\leq\zeta^T V_1\zeta)}\right]\zeta.$$

Then, we can plug-in the first order optimality condition to the value function, and the scaled RWP function becomes,

$$n\mathbb{R}_n\left(\beta_*\right)=2\tilde{\zeta}(W_n)^T W_n-G_2\left(\tilde{\zeta}(W_n),n\right)+o_p(1).$$

Applying Lemma 2 of [5] we can show that as $n\to\infty$,

$$n\mathbb{R}_n\left(\beta_*\right)\Rightarrow 2\tilde{\zeta}(\tilde{Z})^T\tilde{Z}-\tilde{\zeta}\left(\tilde{Z}\right)^T\tilde{G}_2\left(\tilde{\zeta}\left(\tilde{Z}\right)\right)\tilde{\zeta}\left(\tilde{Z}\right)$$

where $\tilde{G}_2:\mathbb{R}^2\to\mathbb{R}^2\times\mathbb{R}^2$ is a continuous mapping defined as

$$\tilde{G}_2\left(\zeta\right)=E\left[V_1\max\left(1-\tau/(\zeta^T V_1\zeta),0\right)\right].$$

This concludes the claim for $d=2$.

**Step 6:** when $d=3$, as $n\to\infty$, we have the scaled RWP function given in (19). Let us write $G_3:\mathbb{R}\times\mathbb{N}\to\mathbb{R}$ to denote a deterministic continuous function defined as

$$G_3\left(\zeta,n\right)=n^{\frac{1}{2}+\frac{3}{2+2d}-\frac{2}{d}}E\left[\max\left(n^{\frac{2}{2}-\frac{6}{2d+2}}\zeta^T V_i\zeta-T_i(n)^{3/d},0\right)\right].$$

Same as discussed in Step 4 and 5, the objective function is strictly convex and the optimizer could be uniquely characterized via first order optimality condition, i.e.

$$W_n=-E\left[V_1\frac{\pi^{d/2}\gamma f_X\left(X_1\right)}{\Gamma\left(d/2+1\right)}\left(\zeta^T V_1\zeta\right)^d\right]\zeta+o_p(1),\ \text{as } n\to\infty.$$

Since we know that the objective function is strictly convex, there exist a continuous mapping, $\tilde{\zeta}:\mathbb{R}^d\to\mathbb{R}^d$, such that $\tilde{\zeta}(W_n)$ is the unique solution to

$$W_n=-E\left[V_1\frac{\pi^{d/2}\gamma f_X\left(X_1\right)}{\Gamma\left(d/2+1\right)}\left(\zeta^T V_1\zeta\right)^d\right]\zeta.$$

Let us plug-in the optimality condition and the scaled RWP function becomes

$$n^{\frac{1}{2}+\frac{3}{2d+2}}R_n\left(\beta_*\right)=-2\tilde{\zeta}(W_n)^T W_n-G_3\left(\tilde{\zeta}(W_n,n)\right)+o_p(1).$$

As $n\to\infty$, we can apply Lemma 2 in [5] to derive estimation for the RWP function and it leads to

$$n^{\frac{1}{2}+\frac{3}{2d+2}}R_n\left(\beta_*\right)\Rightarrow-2\tilde{\zeta}(\tilde{Z})^T\tilde{Z}-\frac{2}{d+2}\tilde{G}_3\left(\tilde{\zeta}(\tilde{Z})\right),$$

where $\tilde{G}_2:\mathbb{R}^d\to\mathbb{R}$ is a deterministic continuous function defined as

$$\tilde{G}_2\left(\zeta\right)=E\left[\frac{\pi^{d/2}\gamma f_X(X_1)}{\Gamma\left(d/2+1\right)}\left(\zeta^T V_1\zeta\right)^{d/2+1}\right].$$

This concludes the case when $d\geq 3$ and for Theorem 1. □

Columbia University, Department of Statistics and Department of Industrial Engineering & Operations Research, 340 S. W. Mudd Building, 500 W. 120 Street, New York, NY 10027, United States.
*E-mail address*: jose.blanchet@columbia.edu

Columbia University, Department of Statistics. 901 SSW, 1255 Amsterdam Ave. New York, NY 10027, United States.
*E-mail address*: yang.kang@columbia.edu