Finite-Sample Guarantees for Wasserstein Distributionally Robust Optimization: Breaking the Curse of Dimensionality

Rui Gao
Department of Information, Risk and Operations Management, University of Texas at Austin, rui.gao@mccombs.utexas.edu

Wasserstein distributionally robust optimization (DRO) aims to find robust and generalizable solutions by hedging against data perturbations in Wasserstein distance. Despite its recent empirical success in operations research and machine learning, existing performance guarantees for generic loss functions are either overly conservative due to the curse of dimensionality, or plausible only in large sample asymptotics. In this paper, we develop a non-asymptotic framework for analyzing the out-of-sample performance for Wasserstein robust learning and the generalization bound for its related Lipschitz and gradient regularization problems. To the best of our knowledge, this gives the first finite-sample guarantee for generic Wasserstein DRO problems without suffering from the curse of dimensionality. Our results highlight the bias-variation trade-off intrinsic in the Wasserstein DRO, which balances between the empirical mean of the loss and the variation of the loss, measured by the Lipschitz norm or the gradient norm of the loss. Our analysis is based on two novel methodological developments that are of independent interest: 1) a new concentration inequality controlling the decay rate of large deviation probabilities by the variation of the loss and, 2) a localized Rademacher complexity theory based on the variation of the loss.

Key words: Distributionally robust optimization, Wasserstein metric, variation regularization, generalization bound, transportation-information inequality

1. Introduction

Distributionally robust optimization (DRO) is an emerging paradigm for statistical learning and decision-making under uncertainty. It aims to provide robust and generalizable solutions by hedging against a set of distributions in the minimax sense. Different choices of distributional uncertainty set have been investigated thoroughly [70, 107, 76, 21, 33, 67, 27, 41, 95, 9, 96, 48, 8, 93, 34, 108]. In this paper, we focus on Wasserstein DRO [96, 34, 108, 17, 38, 98, 25]

$$\inf_{\theta \in \Theta} \sup_{\mathbb{P} : \mathcal{W}_p(\mathbb{P}, \mathbb{P}_n) \leq \rho_n} \mathbb{E}_{\mathbb{P}}[f_{\theta}(z)],$$

which finds a solution \(\theta\) from a space \(\Theta\) so as to minimize the Wasserstein robust loss, defined as the worst-case expectation of the loss function \(f_{\theta}\) among a ball of distributions whose \(p\)-Wasserstein distance \(\mathcal{W}_p\) to the empirical distribution \(\mathbb{P}_n\) of sample size \(n\) is at most \(\rho_n > 0\). Due to its ability to hedge against data perturbations in high dimensions [17, 38] and its regularization effect [71, 72, 14, 37, 87, 5], Wasserstein DRO has recently been studied in many areas in machine learning [71, 14, 13, 80, 77, 23, 39, 87, 54, 31, 62, 55, 2, 78, 28]; as well as other fields, such as automatic control [102, 1, 103, 26], finance [10], energy systems [91, 90, 29], statistics [71, 58, 39, 64, 11, 65], transportation [22]. We refer to [50] for a recent survey.

Among Wasserstein distances of different orders, \(p = 1, 2\) are of particular interest both practically and theoretically. 1-Wasserstein DRO is useful when the loss function is bounded or has linear growth, and often leads to linear programming reformulation when 1-norm or \(\infty\)-norm is used [50, 71, 72]. 2-Wasserstein DRO applies to a larger class of loss functions such as quadratic loss [64, 1, 26]. Efficient gradient-descent algorithms have been developed by virtue of the convex quadratic subproblem associated with the 2-Wasserstein DRO [77, 16, 56, 24]. Moreover, for deep learning problems, the data-dependent gradient regularization induced from the 2-Wasserstein DRO is often more computationally friendly than the Lipschitz regularization induced from the 1-Wasserstein DRO, since the empirical norm of the gradient can be evaluated directly from samples as opposed to the exact computation of Lipschitz norm which is NP hard [86].
Like many other (distributionally) robust optimization frameworks or regularization methods, obtaining a Wasserstein robust solution with good performance guarantees requires a proper hyperparameter tuning, namely, the selection of the radius of the Wasserstein ball \( \rho_n \). On the one hand, the radius \( \rho_n \) cannot be too small since otherwise, the problem behaves like empirical risk minimization or sample average approximation, thus losing the purpose of robustification. On the other hand, the radius \( \rho_n \) cannot be so large that the solution might be overly conservative, which is one of the major criticism faced by traditional robust optimization. Practically, radius selection is often achieved via cross validation. From a statistical point of view, it is crucial to understand what is the correct scaling of the hyperparameter \( \rho_n \) with respect to the sample size \( n \) so as to ensure the robustness and generalization of the solution without sacrificing much out-of-sample performance.

Despite promising applications of Wasserstein DRO, its theoretical performance guarantee is limited. Esfahani and Kuhn [34] provides the first out-of-sample performance guarantee for Wasserstein DRO. Using the concentration of empirical Wasserstein distance [36], they show that if the radius is chosen in the order of \( n^{-1/d} \), where \( d \) is the dimension of the random data \( z \), the underlying data-generating distribution \( \mathbb{P}_{\text{true}} \) is contained in the Wasserstein ball with high probability. Thereby the Wasserstein robust loss of every feasible solution (and in particular the optimal solution) would be an upper bound of its true loss. This provides a finite-sample non-asymptotic guarantee for the Wasserstein robust solution, but unfortunately, such a bound suffers from the curse of dimensionality since the radius shrinks too slow even for problems in moderate dimensions.

To resolve the curse of dimensionality, a series of work by Blanchet et al. [14, 12, 15] consider an approach inspired from the empirical likelihood [51, 30]. Their principle is finding the smallest radius \( \rho_n \) such that with high probability, the Wasserstein ball contains at least one distribution \( \mathbb{P} \) – not necessarily equal to the true data-generating distribution \( \mathbb{P}_{\text{true}} \) – for which there exists an optimal solution to \( \min_{\theta} \mathbb{E}[f_{\theta}] \) that is also optimal to the underlying true problem \( \min_{\theta} \mathbb{E}_{\mathbb{P}_{\text{true}}}[f_{\theta}] \). Through an asymptotic analysis, they show that the radius \( \rho_n \) can be chosen in the square-root order \( 1/\sqrt{n} \) and the constant has a mild dependence on the dimension \( d \) of the random data \( z \). This gives the first radius selection rule that does not suffer from the curse of dimensionality. However, one potential issue with this result is that the bound is valid only in the asymptotic sense, namely, as the sample size \( n \) goes to infinity which is, again, not the typical regime under which robust optimization is applied.

For certain special classes of stochastic optimization problems, the non-asymptotic \( 1/\sqrt{n} \)-rate has been developed. For 1-Wasserstein DRO with certain linear structure, such as linear regression/classification and their kernelization, Shafieezadeh-Abadeh et al. [72] shows that the radius can be chosen as \( \tilde{O}(1/\sqrt{n}) \) to achieve a finite-sample performance guarantee uniformly for all feasible solutions\(^1\). Chen and Paschalidis [23] derives generalization bounds for certain class of 1-Wasserstein DRO problems that are equivalent to norm regularization. Xie et al. [99] provides performance guarantees for stochastic bottleneck problems by relating them to sample average approximations.

Yet, it remains largely unknown whether the non-asymptotic \( 1/\sqrt{n} \)-rate holds for general loss functions. In this paper, we provide an affirmative answer to this open question under reasonable assumptions. Informally, our main result states the following performance guarantees for Wasserstein DRO.

**Theorem (Informal).** Let \( p \in \{1,2\} \). Set \( \rho_n = \rho_0/\sqrt{n} \) for some \( \rho_0 > 0 \). Under appropriate conditions, with high probability,

\[
\mathbb{E}_{\mathbb{P}_{\text{true}}}[f_{\theta}] \leq \sup_{\mathbb{P} : \mathcal{W}_p(\mathbb{P}, \mathbb{P}_{\text{true}}) \leq \rho_n} \mathbb{E}_{\mathbb{P}}[f_{\theta}(z)] + \frac{C}{n}, \quad \forall \theta \in \Theta,
\]

where \( C \) depends on the problem parameters and the confidence level.

\(^1\) We use \( \tilde{O} \) to suppress the logarithmic dependence on \( n \).
This theorem shows that up to a high-order residual, the true loss is upper bounded by the Wasserstein robust loss uniformly for all \( \theta \) and particularly for the robust solution. Minimizing over \( \theta \) yields that the optimal robust loss provides an upper bound for the true optimal loss up to a high-order residual. It justifies the empirical \( 1/\sqrt{n} \) radius selection rule and provides a finite-sample theoretical guarantee for Wasserstein robust loss minimization.

Recall the regularization effect of the Wasserstein DRO \([14, 72, 37]\)^2:

\[
\sup_{P: \mathcal{W}_p(P, P_n) \leq \rho_n} \mathbb{E}_{z \sim P} [f_\theta(z)] = \mathbb{E}_{z \sim P_n} [f_\theta(z)] + \rho_n \cdot \mathcal{V}(f_\theta) + O_p(1/n), \quad \forall \theta \in \Theta,
\]

where \( \mathcal{V}(\cdot) \) represents the variation of the loss, measured by the Lipschitz norm \( \|f\|_{\text{Lip}} \) when \( p = 1 \) or the empirical gradient norm \( \mathbb{E}_{P_n} [\|\nabla_z f_\theta(z)\|^2]^{1/2} \) when \( p = 2 \). Together with this result, our theorem highlights a principled bias-variation trade-off by virtue of the Wasserstein DRO. That is, it ensures robustness and generalization by balancing between the empirical loss \( \mathbb{E}_{P_n} [f_\theta] \) and the variation of the loss \( \mathcal{V}(f_\theta) \) and thereby control the true risk:

\[
\mathbb{E}_{P_{\text{true}}} [f_\theta] \leq \mathbb{E}_{z \sim P_n} [f_\theta(z)] + \frac{\rho_0}{\sqrt{n}} \cdot \mathcal{V}(f_\theta) + O_p(1/n), \quad \forall \theta \in \Theta.
\]

Variation-based regularization has become increasingly popular for many deep learning problems recently. Lipschitz regularization and gradient regularization have shown superior empirical performance for adversarial learning and reinforcement learning \([20, 42, 60, 79, 45, 105, 87, 47, 86, 35, 47, 66, 81, 94, 106]\). Our results also provide statistical guarantees for Lipschitz regularization and gradient regularization.

Below, we briefly describe two methodological advancements that lead to our results. In our analysis, the main object of study is the Wasserstein regularizer:

\[
\mathcal{R}_{Q, \rho; f_\theta} = \sup_{P: \mathcal{W}_p(P, P_q) \leq \rho} \mathbb{E}_{z \sim P} [f_\theta(z)] - \mathbb{E}_{z \sim Q} [f_\theta(z)],
\]

that is, the difference between the Wasserstein robust loss and the nominal loss under some distribution \( Q \) such as \( P_n \) or \( P_{\text{true}} \).

First, in Section 3, leveraging tools from transportation-information inequalities (see, e.g., \([44]\)) in modern probability theory, we derive a new large-deviation type concentration inequality for the empirical loss (Theorem 1). It shows that under proper conditions on the underlying data-generating distribution \( P_{\text{true}} \), the decay rate of the tail probability is upper bounded by the inverse of Wasserstein regularizer \( \mathcal{R}_{P_{\text{true}}, \rho; f_\theta}^{-1} \) as well as the variation of the loss \( \mathcal{V}(f_\theta) \). This result shows that the variation of the loss has a direct control on the deviation of the empirical loss from the ground truth. This is an analog of variance-based control often resulting from Chebyshev’s or Bernstein’s concentration inequalities.

Second, to extend the concentration result above from a single loss function to a family of loss functions, we develop two sets of results in Sections 4.1 and 4.2 respectively, one based on covering number arguments, and the other adapts tools from localized Rademacher complexity theory (see, e.g., \([6, 32]\)). For the latter, we consider subsets of function classes whose variations are controlled, as opposed to usual approaches based on the mean or variance of the loss. These results are demonstrated in Section 5 using various examples, including feature-based newsvendor, linear prediction, portfolio optimization, Lipschitz regularization for kernel classes, and gradient regularization for neural networks.

Overall, we develop a non-asymptotic statistical analysis framework for Wasserstein DRO and its associated variation regularization, and demonstrate the bias-variation trade-off in Wasserstein robust learning, which serves as a counterpart of the well-known bias-variance trade-off theory in statistical learning.

---

\(^2\) We use \( O_p \) for the big O in probability notation.
1.1. Related Work

The generalization bounds for robust optimization dates back to Xu and Mannor [101], which studies generalization of learning algorithms from the viewpoint of robustness. In the introduction, we have elaborated on the literature that provide performance guarantees for Wasserstein DRO [34, 14, 12, 72, 15, 23, 99] and discuss their scopes and limitations. In addition to these literature, motivated by distribution shift in domain adaptation and adversarial learning, Lee and Raginsky [54] and Sinha et al. [77] develop generalization bounds for Wasserstein DRO where the radius is fixed, not varying with the sample size. For divergence DRO and the related variance regularization, Lam [51] studies the calibration of the radius of divergence ball that recovers the best statistical guarantee. Asymptotics and non-asymptotics of divergence DRO and its bias-variance trade-off are investigated in [30, 63, 32]. Besides DRO, Wasserstein distance and transportation-information inequality are also exploited to improve information-theoretic generalization bounds for learning algorithms [100, 57, 69, 92].

The rest of the paper proceeds as follows. In Section 2, we briefly review some results in Wasserstein DRO and its variation regularization effect. We develop a new variation-based concentration inequality in Section 3. Based on these two sections, we derive generalization bounds for variation regularization DRO and its variation regularization effect. We develop a new variation-based concentration inequality are also exploited to improve information-theoretic generalization bounds for learning algorithms [100, 57, 69, 92].

2. Wasserstein DRO and Variation Regularization

In this section, we introduce notations and provide some background on Wasserstein DRO and its variation regularization effect.

**Notation.** Let \( Z \) be a subset of a Banach space equipped with some norm \( \| \cdot \| \) and let \( \| \cdot \|_* \) be its dual norm. Define the diameter of \( Z \) as \( \text{diam}(Z) := \sup_{z \in Z} \| z - z' \| \). Let \( p \in [1, \infty) \) and denote by \( q \) its Hölder conjugate number, i.e., \( \frac{1}{p} + \frac{1}{q} = 1 \). We denote by \( \mathcal{P}_p(Z) \) the set of probability measures \( Z \) with finite \( p \)-th moment, namely, \( Q \in \mathcal{P}_p(Z) \) if and only if \( E_{z \sim Q}[\| z \|^p] < \infty \). The support of a distribution is denoted by \( \text{supp} Q \). The \( \mathcal{L}^p(Q) \)-norm of a \( Q \)-measurable function \( h \) is denoted by \( \| h \|_{Q,p} = E_{z \sim Q}[h(z)^p]^{\frac{1}{p}} \). The sup-norm of a function \( h \) is denoted by \( \| h \|_{\infty} \), and the Lipschitz norm of a Lipschitz continuous function \( h \) is denoted by \( \| h \|_{\text{Lip}} \). We denote \( a \vee b = \max(a, b) \). For the expectation operator \( E_{z \sim Q}[\cdot] \), we often write it as \( E_Q[\cdot] \) provided that the involved random variable is clear from the context.

The Wasserstein distance of order \( p \) between distributions \( P, Q \in \mathcal{P}_p(Z) \) is defined via

\[
W_p(P, Q)^p := \inf_{\gamma \in \mathcal{P}(Z^2)} \left\{ E_{(z,z') \sim \gamma}[\| z - z' \|^p] : \gamma \text{ has marginal distributions } P, Q \right\}.
\]

We denote by \( F := \{ f_\theta : \theta \in \Theta \} \) the class of loss functions. To ease notations, we often suppress the subscript \( \theta \) and use \( f \) to represent a generic loss function from \( F \). Given a nominal distribution \( Q \in \mathcal{P}_p(Z) \) and a radius \( \rho \geq 0 \), the Wasserstein DRO problem is given by

\[
\inf_{f \in F} \sup_{P \in \mathcal{P}_p(Z)} \left\{ E_{z \sim P}[f(z)] : W_p(P, Q) \leq \rho \right\}.
\]

Suppose there exists \( M, L \geq 0 \) such that \( f(z) \leq M + L \| z \|^p \) for all \( z \in Z \), then the inner maximization problem above has a dual problem that always has a minimizer [38]:

\[
\min_{\lambda \geq 0} \left\{ \lambda \rho^p + E_{z \sim Q} \left( \sup_{z \in Z} \{ f(z) - \lambda \| z \|^p \} \right) \right\}.
\]

(1)

In a data-driven problem, the nominal distribution is often chosen as the empirical distribution

\[
P_n = \frac{1}{n} \sum_{i=1}^n \delta_{z_i^n},
\]

constructed from \( n \) i.i.d. samples \( \{ z_i^n \} \) from the underlying true distribution \( P_{\text{true}} \),
where $\delta_z$ denotes the Dirac point mass on $z$. We use $\mathbb{P}_\delta$ or $E_\delta$ to indicate that the probability or expectation is evaluated with respect to the sampling distribution, namely the $n$-fold product distribution $\otimes_{i=1}^n \mathbb{P}_{\text{true}}$ over $Z^n$. We define the Wasserstein regularizer as the difference between the Wasserstein robust loss and the nominal loss:

$$R_{Q,p}(\rho; f) := \sup_{\mathbb{P} \in \mathcal{P}(Z)} \{E_{\mathbb{P}}[f(z)] : W_p(\mathbb{P}, Q) \leq \rho\} - E_Q[f].$$

The connection between Wasserstein DRO and regularization has been established under various settings [34, 71, 14, 72, 37, 87]. The next two results adapted from Gao et al. [37] (see also Bartl et al. [5]) establish connections between the Wasserstein regularizer $R_{Q,p}$ and Lipschitz regularization ($p = 1$) and gradient regularization ($p = 2$) respectively.

**Assumption 1.** Assume the following holds:

(I) There exists $\kappa_1 > 0$ such that for every $f \in \mathcal{F}$,

$$f(\tilde{z}) - f(z) \leq \kappa_1 ||\tilde{z} - z||, \ \forall z, \tilde{z} \in Z.$$

(II) Suppose $\text{diam}(Z) := \sup_{z, \tilde{z} \in Z} ||\tilde{z} - z|| = \infty$, and for every $f \in \mathcal{F}$, there exists $z_0 \in Z$ such that

$$\limsup_{||z - z_0|| \to \infty} \frac{f(\tilde{z}) - f(z_0)}{||z - z_0||} = ||f||_{\text{Lip}}.$$

Assumption (I) means that every $f$ is Lipschitz continuous, and (II) means that the Lipschitz norm is attained at infinity.

**Lemma 1 (Lipschitz regularization).** Let $Q \in \mathcal{P}_1(Z)$ and $\rho \geq 0$. Assume Assumption 1(I) holds, then

$$R_{Q,1}(\rho; f) \leq \rho \cdot ||f||_{\text{Lip}}.$$

Assume, in addition, Assumption 1(II) holds. Then

$$R_{Q,1}(\rho; f) = \rho \cdot ||f||_{\text{Lip}}.$$

**Assumption 2.** Assume every $f \in \mathcal{F}$ is differentiable and there exist $h > 0$ such that

$$||\nabla f(\tilde{z}) - \nabla f(z)||_\ast \leq h||\tilde{z} - z||, \ \forall \tilde{z}, z \in Z, \forall f \in \mathcal{F}.$$

This is a smoothness condition which requires that every $f$ has Lipschitz gradient.

**Lemma 2 (Gradient regularization).** Let $Q \in \mathcal{P}_2(Z)$ and $\rho \geq 0$. Assume Assumption 2 holds. Then

$$|R_{Q,2}(\rho; f) - \rho \cdot ||\nabla f||_\ast| \leq h\rho^2.$$

### 3. Variation-based Concentration Inequality

In this section, we derive a large-deviation type concentration inequality for the empirical mean of a single loss function. We derive an equivalent representation of $R_{Q,1}^{-1}(\rho; f)$ in Section 3.1 and provide a brief overview of transportation-information inequalities in Section 3.2. The new concentration inequality is developed in Section 3.3, whose proof is postponed to Section A.1.
3.1. Inverse of the Wasserstein Regularizer

Fixing \( f \in \mathcal{F} \), we define a function \( \mathcal{I}_p : \mathbb{R}_+ \to \mathbb{R}_+ \cup \{+\infty\} \) via

\[
\mathcal{I}_p(\varepsilon; f) := \sup_{t \geq 0} \left\{ et - \mathbb{E}_{z \sim \mathbb{P}_{\text{true}}} \left[ \sup_{\tilde{z} \in \mathcal{Z}} \left\{ tf(\tilde{z}) - f(z) - \|\tilde{z} - z\|^p \right\} \right] \right\},
\]

which will play a similar role as the rate function in the large deviation principle. The next proposition establishes its connection to the Wasserstein regularizer \( \mathcal{R}_{\mathbb{P}_{\text{true}},p} \), whose proof is given in Appendix A.

**Proposition 1.** Let \( p \in [1, \infty) \) and \( f \in \mathcal{F} \). Suppose there exists \( M, L \geq 0 \) such that \( f(z) \leq M + L\|z\|^p \) for all \( z \in \mathcal{Z} \). Let \( \rho > 0 \) and let \( \lambda_o \) be the dual minimizer of (1). Set \( \underline{\lambda} := \lim_{\|z\| \to \infty} f(z)/\|z\|^p \). Then

\[
\mathcal{I}_p(\mathcal{R}_{\mathbb{P}_{\text{true}},p}(\rho; f); f) \begin{cases} = \rho, & \text{if } \lambda_o > \underline{\lambda}, \\ \geq \rho, & \text{if } \lambda_o = \underline{\lambda}. \end{cases}
\]

Note that the dual optimizer of (1) tends to be large when \( \rho \) is close to zero, in which case \( \lambda_o > \underline{\lambda} \), as observed in [38]. Hence Proposition 1 shows that at least for small \( \rho \), the left inverse of \( \mathcal{R}_{\mathbb{P}_{\text{true}},p}(\cdot; f) \) is \( \mathcal{I}_p(\cdot; f) \).

3.2. Transportation-Information Inequalities

Just like many other results on the concentration of measure, appropriate conditions on the function \( f \) and the distribution \( \mathbb{P}_{\text{true}} \) are required. Since we are dealing with general loss functions that are possibly unbounded, some assumptions on the underlying data-generating distribution are necessary. It turns out for our purpose, it is convenient to work with the transportation-information inequality, a useful condition to establish concentration of measure in modern probability theory.

**Definition 1 (Transportation-information inequality).** Let \( p \in [1, \infty) \). A distribution \( \mathbb{P} \in \mathcal{P}_p(\mathcal{Z}) \) satisfies a transportation-information inequality \( T_p(\tau) \) for some positive constant \( \tau \), if

\[
W_p(Q, \mathbb{P}) \leq \sqrt{\tau H(Q||\mathbb{P})}, \quad \forall Q \in \mathcal{P}_p(\mathcal{Z}),
\]

where \( H(Q||\mathbb{P}) \) denotes the relative entropy \( H(Q||\mathbb{P}) := \int_{\mathcal{Z}} \log(dQ/d\mathbb{P}) \, dQ \), where \( dQ/d\mathbb{P} \) denotes the Radon-Nikodym derivative.

We briefly comment on distributions satisfying transportation-information inequalities, and refer the reader to [44] for a recent survey and [85] for an in-depth discussion. Among different choices of \( p \), \( T_1 \) and \( T_2 \) are of particular interest and have been widely studied in the literature. \( T_1 \) is equivalent to the following condition (see also Lemma 3 in Appendix A): a distribution \( \mathbb{P} \) satisfies \( T_1 \) if and only if there exists \( a > 0 \) such that \( \mathbb{E}[\exp(a\|z\|^2)] < \infty \). In particular, any distribution on a bounded support \( \mathcal{Z} \) with \( \text{diam}(\mathcal{Z}) < \infty \) satisfies \( T_1(2\text{diam}(\mathcal{Z})^2) \). \( T_2 \) is also known as Talagrand’s inequality. It is implied by the logarithmic Sobolev inequality and holds in particular for distributions with a strongly log-concave density function. Note that for \( p_1 \leq p_2 \), \( T_{p_1} \) is weaker than \( T_{p_2} \) since \( W_{p_1} \leq W_{p_2} \). In the sequel, we will focus on the case \( p \in [1, 2] \).

3.3. Concentration for a Single Loss Function

Now we are ready to state our main result in this section.

**Theorem 1 (Variation-based concentration).** Let \( p \in [1, 2] \) and \( f \in \mathcal{F} \). Assume there exist \( M, L > 0 \) such that

\[
f(z) \leq M + L\|z\|^p, \quad \forall z \in \mathcal{Z}.
\]
Assume further that $\mathbb{P}_{\text{true}}$ satisfies $T_\rho(\tau)$ for some $\tau > 0$. Let $\epsilon > 0$. Then

$$\mathbb{P}_{\phi}\{E_n[f] - E_{\text{true}}[f] < -\epsilon\} \leq \exp\left(-nI_\rho(\epsilon; -f)^2/\tau\right).$$

Let $t > 0$. Then with probability at least $1 - e^{-t}$,

$$E_{\text{true}}[f] \leq E_n[f] + R_{\text{true}}(\sqrt{\frac{\epsilon t}{n}}; -f).$$

(2)

Theorem 1 uncovers an interesting connection: the non-asymptotic decay rate of large deviation probabilities is controlled by the inverse of Wasserstein regularizer $I_\rho(\epsilon; -f)$. The negative sign $-f$ appears because here we bound the downside risk, i.e., the probability of empirical loss being smaller than the true loss, whereas $R_{\text{true}}$ is defined via upside excess, i.e., the worst-case loss that is greater than the true loss. As a matter of fact, a similar result holds if we swap the empirical loss and true loss in the theorem:

$$\mathbb{P}_{\phi}\{E_n[f] - E_{\text{true}}[f] > \epsilon\} \leq \exp\left(-nI_\rho(\epsilon; f)^2/\tau\right),$$

and with probability at least $1 - e^{-t}$,

$$E_n[f] \leq E_{\text{true}}[f] + R_{\text{true}}(\sqrt{\frac{\epsilon t}{n}}; f).$$

When $p = 1$, if $f$ is Lipschitz continuous, then by Lemma 1 we have

$$R_{\text{true}}(\sqrt{\frac{\epsilon t}{n}}; -f) \leq \sqrt{\frac{\epsilon t}{n}} \cdot \|f\|_{\text{Lip}} = \sqrt{\frac{\epsilon t}{n}} \cdot \|f\|_{\text{Lip}}.$$When $p = 2$, if $f$ has Lipschitz continuous gradient, then by Lemma 2 we have

$$R_{\text{true},2}(\sqrt{\frac{\epsilon t}{n}}; -f) \leq \sqrt{\frac{\epsilon t}{n}} \cdot \|\nabla f\|_\ast \|f\|_{\text{true},2} + \frac{ht}{n}.$$Substituting these inequalities in Theorem 1 yields following corollary.

**Corollary 1 (Variation regularization).** Let $p \in \{1, 2\}$. When $p = 1$, assume Assumption 1(I) holds; when $p = 2$, assume Assumption 2 holds. Assume further that $\mathbb{P}_{\text{true}}$ satisfies $T_\rho(\tau)$ for some $\tau > 0$. Let $t > 0$. Then with probability at least $1 - e^{-t}$,

$$E_{\text{true}}[f] \leq E_n[f] + \begin{cases} \sqrt{\frac{\epsilon t}{n}} \cdot \|f\|_{\text{Lip}}, & p = 1, \\ \sqrt{\frac{\epsilon t}{n}} \cdot \|\nabla f\|_\ast \|f\|_{\text{true},2} + \frac{ht}{n}, & p = 2. \end{cases}$$

Theorem 1 and Corollary 1 show that the Wasserstein regularizer $R_{\text{true}}(\phi(\sqrt{\frac{\epsilon t}{n}}; -f))$, as well as the variation of the loss, $\|f\|_{\text{Lip}}$ or $\|\nabla f\|_\ast \|f\|_{\text{true},2}$, are natural quantities controlling the deviation of the empirical loss for distributions satisfying a transportation-information inequality. For $p = 1$, thanks to the first part of Lemma 1, the bound in Theorem 1 is tighter than the Lipschitz norm bound in Corollary 1, which was obtained in [18]. Since $\|\nabla f\|_\ast \|f\|_{\text{true},2} \leq \|f\|_{\text{Lip}}$, $p = 2$ suggests a tighter upper bound than $p = 1$, at the cost of a stronger assumption on the underlying distribution.
4. Finite-Sample Guarantees

In the previous section, we have derived a concentration inequality for a single loss function, and the goal of this section is to extend it to a family of loss functions $\mathcal{F}$. In the spirit of [34, 14, 72], we would like to determine a proper scaling of the Wasserstein radius $\rho_n$ with respect to sample size $n$ so that with high probability, the Wasserstein robust loss is an upper bound of the true loss uniformly for all functions in the class $\mathcal{F} = \{ f_\theta : \theta \in \Theta \}$. Whenever this holds, minimizing the Wasserstein robust loss controls the true loss as well.

When $\mathcal{F}$ is a finite set, then a simple application of the union bound to Theorem 1 yields that (2) holds simultaneously for all $f \in \mathcal{F}$ with probability at least $1 - |\mathcal{F}|e^{-t}$, where $| \cdot |$ denotes the cardinality of a set. When $\mathcal{F}$ contains infinitely many functions, some notion of complexity of the function class $\mathcal{F}$ is needed to obtain uniform convergence. In Section 4.1, we prove the result using a standard covering number argument; and in Section 4.2, we adopt techniques from local Rademacher complexity theory [6, 49].

4.1. Covering Number Arguments

Recall that for $\epsilon > 0$, the covering number $\mathcal{N}(\epsilon; \mathcal{H}, \| \cdot \|_\mathcal{H})$ of a set $\mathcal{H}$ with respect to a norm $\| \cdot \|_\mathcal{H}$ is defined as the smallest cardinality of an $\epsilon$-cover of $\mathcal{H}$, where $\mathcal{H}_\epsilon$ is an $\epsilon$-cover of $\mathcal{H}$ if for each $h \in \mathcal{H}$, there exists $\tilde{h} \in \mathcal{H}_\epsilon$ such that $\| \tilde{h} - h \|_\mathcal{H} \leq \epsilon$. Similar to the classic stochastic programming literature (e.g., Shapiro et al. [75, Section 5.3.2]), we can obtain a union bound using the standard covering number argument, whose proof is given in Appendix B.1. Throughout this subsection, we let $\mathcal{F} = \{ f_\theta : \theta \in \Theta \}$ and we impose the following smoothness assumption with respect to the parameter $\theta$.

**Assumption 3.** Assume there exists a measurable function $\kappa : \mathcal{Z} \to \mathbb{R}_+$ with finite moment generating function in a neighborhood of zero such that for all $\tilde{\theta}, \theta \in \Theta$ and $\mathbb{P}_{\text{true}}$-a.e. $z \in \mathcal{Z}$,

$$| f_{\tilde{\theta}}(z) - f_\theta(z) | \leq \kappa(z) | \tilde{\theta} - \theta |_{\Theta}.$$  

**Corollary 2.** Let $p \in [1, 2]$. Assume $\mathbb{P}_{\text{true}}$ satisfies $T_p(\tau)$ for some $\tau > 0$ and Assumption 3 holds. Let $t > 0$. Then we have the following:

1. Assume there exist $M, L > 0$ such that

$$f_\theta(z) \leq M + L \| z \|_p, \quad \forall z \in \mathcal{Z}, \forall \theta \in \Theta.$$  

Then there exists $C > 0$ such that with probability at least $1 - \exp(-Cn) - \mathcal{N}(1/n; \Theta, \| \cdot \|_\Theta) \cdot e^{-t}$,

$$E_{\mathbb{P}_{\text{true}}} [ f_\theta ] \leq E_{\mathbb{P}_n} [ f_\theta ] + \mathcal{R}_{\mathbb{P}_{\text{true}}} (\sqrt{\frac{2t}{n}}, f_\theta) + \frac{3E_{\mathbb{P}_{\text{true}}} [ \kappa ]}{n}, \quad \forall \theta \in \Theta.$$  

2. **(Lipschitz regularization and 1-Wasserstein DRO)** When $p = 1$, assume Assumption 1(I) holds. Then there exists $C > 0$ such that with probability at least $1 - \exp(-Cn) - \mathcal{N}(1/n; \Theta, \| \cdot \|_\Theta) \cdot e^{-t}$,

$$E_{\mathbb{P}_{\text{true}}} [ f_\theta ] \leq E_{\mathbb{P}_n} [ f_\theta ] + \sqrt{\frac{2t}{n}} \cdot \| f_\theta \|_{\text{Lip}} + \frac{3E_{\mathbb{P}_{\text{true}}} [ \kappa ]}{n}, \quad \forall \theta \in \Theta.$$  

Assume, in addition, that Assumption 1(II) holds. Set $\rho_n = \sqrt{\frac{2t}{n}}$. Then there exists $C > 0$ such that with probability at least $1 - \exp(-Cn) - \mathcal{N}(1/n; \Theta, \| \cdot \|_\Theta) \cdot e^{-t}$,

$$E_{\mathbb{P}_{\text{true}}} [ f_\theta ] \leq E_{\mathbb{P}_n} [ f_\theta ] + \mathcal{R}_{\mathbb{P}_n, 1} (\rho_n; f_\theta) + \frac{3E_{\mathbb{P}_{\text{true}}} [ \kappa ]}{n}, \quad \forall \theta \in \Theta.$$  

(III) **(Gradient regularization)** When $p = 2$, assume Assumption 2 holds. Then with probability at least $1 - \mathcal{N}(1/n; \mathcal{F}, \|\cdot\|_\mathcal{F}) \cdot e^{-t}$,

$$
\mathbb{E}_{\mathbb{P}_{\text{true}}} [f_\theta] \leq \mathbb{E}_{\mathbb{P}_n} [f_\theta] + \sqrt{\frac{e}{n}} \cdot \|\nabla f_\theta\|_2 + \frac{h t + 3\mathbb{E}_{\mathbb{P}_{\text{true}}} [\kappa]}{n}, \quad \forall \theta \in \Theta.
$$

In the next result, we provide an empirical counterpart of Corollary 2 for feature-based newsvendor problem in Example 1 and Corollary 3 for linear regularization and show that by choosing the radius $\rho_n = O(1/\sqrt{n})$, the Wasserstein robust loss serves as an upper bound of the true loss for all $f_\theta \in \mathcal{F}$ up to an $O(1/n)$ remainder. In Section 5, we will demonstrate Corollary 2 for feature-based newsvendor problem in Example 1 and Corollary 3 for linear prediction with Lipschitz loss in Example 3.

4.2. Local Rademacher Complexity Arguments

The covering number bound developed in the previous subsection may be loose. To obtain a tighter bound in a more general setting, we derive results using local Rademacher complexity theory.

Let us begin with some technical preparation. Recall the *Rademacher complexity* of a function class $\mathcal{F}$ with respect to a sample $\{z_i^n\}_{i=1}^n$ is defined as

$$
\mathcal{R}_n(\mathcal{F}) := \mathbb{E}_\sigma \left[ \sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^n \sigma_i f(z_i^n) \right],
$$

where $\sigma_i$'s are i.i.d. Rademacher random variables with $\mathbb{P}(\sigma_i = \pm 1) = \frac{1}{2}$. The Rademacher complexity of the function class $\mathcal{F}$ with respect to $\mathbb{P}_{\text{true}}$ for sample size $n$ is defined as $\mathbb{E}_\Theta [\mathcal{R}_n(\mathcal{F})]$. Rademacher complexity plays an important role in bounding the generalization error of statistical learning problems but may be vacuous if $\mathcal{F}$ is too large. The idea of localization is to restrict on a small subset of $\mathcal{F}$ around the optimal solution that often admits low complexity. The *localized Rademacher complexity* [6] at level $r > 0$ is defined as

$$
\mathbb{E}_\Theta \left[ \mathcal{R}_n \left( \{ f \in \mathcal{F} : 0 \leq c \leq 1, T(cf) \leq r \} \right) \right],
$$

where $T : \mathcal{F} \to \mathbb{R}_+$. In our analysis, we choose $T(f) = \|f\|_\text{lip}^2$ when $p = 1$ and $T(f) = \|\nabla f\|_2^2$ when $p = 2$. Part of our techniques below are adapted from the framework developed in [6, 32], which primarily considers $T(f) = \mathbb{E}_{\mathbb{P}_{\text{true}}} [f^2]$. 

**Assumption 4.** Assume there exists a measurable function $\kappa_2 : \mathcal{Z} \to \mathbb{R}_+$ and $a_0 > 0$ such that $\mathbb{E}_{\mathbb{P}_{\text{true}}} [\exp(\kappa_2^2)] < \infty$ for all $|a| < a_0$, and for all $\tilde{\theta}, \theta \in \Theta$ and $\mathbb{P}_{\text{true}}$-a.e. $z \in \mathcal{Z}$,

$$
\|\nabla f_\tilde{\theta}(z) - \nabla f_\theta(z)\|_2 \leq \kappa_2(z) \|\tilde{\theta} - \theta\|_\Theta.
$$

**Corollary 3 (2-Wasserstein DRO).** Assume $\mathbb{P}_{\text{true}}$ satisfies $T_2(\tau)$ for some $\tau > 0$ and Assumptions 2, 3, 4 hold. Assume $\sigma = \sup_{\theta \in \Theta} \mathbb{E}_{\mathbb{P}_{\text{true}}} [\|\nabla f_\theta\|_2^4]^{1/2} / \|\|\nabla f_\theta\|_2^2 \mathbb{P}_{\text{true}} < \infty$. Let $t > 0$ and $n \geq 8\sigma^2 t$. Set $\rho_n = \sqrt{\frac{e}{2n} (1 + \sigma \sqrt{\frac{2t}{n}})}$. Then there exists $C > 0$ such that with probability at least $1 - 2 \exp(-Cn) - 2\mathcal{N}(1/n, \Theta, \|\cdot\|_\Theta) \cdot e^{-t}$,

$$
\mathbb{E}_{\mathbb{P}_{\text{true}}} [f_\theta] \leq \mathbb{E}_{\mathbb{P}_n} [f_\theta] + \rho_n \|\nabla f_\theta\|_2 + h\rho_n^2 + \frac{3\mathbb{E}_{\mathbb{P}_{\text{true}}} [\kappa] + 2\rho_n \kappa_2 \|\mathbb{P}_{\text{true}}\|_{2,\mathbb{P}_{\text{true}}}}{n}, \quad \forall \theta \in \Theta,
$$

and

$$
\mathbb{E}_{\mathbb{P}_{\text{true}}} [f_\theta] \leq \mathbb{E}_{\mathbb{P}_n} [f_\theta] + \mathcal{R}_{\mathbb{P}_n,2}(\rho_n; f_\theta) + 2h\rho_n^2 + \frac{3\mathbb{E}_{\mathbb{P}_{\text{true}}} [\kappa] + 2\rho_n \kappa_2 \|\mathbb{P}_{\text{true}}\|_{2,\mathbb{P}_{\text{true}}}}{n}, \quad \forall \theta \in \Theta.
$$
By choosing a proper level $r_n$, the localized Rademacher complexity of the functions of the subset can be much smaller than the entire family, which enables a better bound. Often, the level $r_n$ is chosen to be the fixed point $r_n^*$ of some function $\psi_n(r)$, which serves as an upper bound on the localized Rademacher complexity at level $r$. A typical assumption imposed on $\psi_n$ is the so-called sub-root condition. A function $\psi: \mathbb{R}_+ \to \mathbb{R}_+$ is sub-root if it is non-constant, non-negative, non-decreasing and the map $r \mapsto \psi(r)/\sqrt{r}$ is non-increasing for all $r > 0$. A sub-root function always has a unique fixed point $r_n^*$ [6]. Similar to the literature, we impose the following assumption.

**Assumption 5 (Sub-root local complexity).** Assume there exists a sub-root function $\psi_n: \mathbb{R}_+ \to \mathbb{R}_+$ such that

$$\psi_n(r) \geq \mathbb{E}_\otimes \left[ \mathcal{R}_n \left( \left\{ c f : f \in \mathcal{F}, \ 0 \leq c \leq 1, T(cf) \leq r \right\} \right) \right].$$

Denote by $r_n^*$ the fixed point of $\psi_n$.

We will verify this assumption for various examples considered in Section 5.

We first study the case of $\rho = 1$. The proof is given in Appendix B.2.2.

**Theorem 2 (Lipschitz regularization).** Assume $\mathbb{P}_{\text{true}}$ satisfies $T_1(\tau)$, Assumption 1(I) holds, and Assumption 5 holds with $T(f) = \|f\|_{\text{Lip}}^2$. Let $t > 0$. Then with probability at least $1 - [\log_2(\kappa_1 \sqrt{tn})] e^{-t}$,

$$\mathbb{E}_{\mathbb{P}_{\text{true}}} \mathbb{E}_{\mathbb{P}_n} [f] \leq \mathbb{E}_{\mathbb{P}_n} [f] + 2 \left( \sqrt{\frac{t}{n}} + 2 \sqrt{\frac{r_n^*}{n}} + \frac{1}{n^{\frac{1}{4}}} \right) \|f\|_{\text{Lip}} + 4 r_n^* + \frac{2}{n}, \ \forall f \in \mathcal{F}.$$

Together with Lemma 1, we obtain the following result.

**Corollary 4 (1-Wasserstein DRO).** Assume $\mathbb{P}_{\text{true}}$ satisfies $T_1(\tau)$, Assumption 1 holds and Assumption 5 holds with $T(f) = \|f\|_{\text{Lip}}^2$. Let $t > 0$. Set

$$\rho_n = 2 \sqrt{\frac{t}{n}} + 4 \sqrt{r_n^*} + \frac{2}{n^{\frac{1}{4}}}.$$

Then with probability at least $1 - [\log_2(\kappa_1 \sqrt{tn})] e^{-t}$,

$$\mathbb{E}_{\mathbb{P}_{\text{true}}} \mathbb{E}_{\mathbb{P}_n} [f] \leq \mathbb{E}_{\mathbb{P}_n} [f] + \mathcal{R}_{\mathbb{P}_n,1} (\rho_n; f) + 4 r_n^* + \frac{2}{n}, \ \forall f \in \mathcal{F}.$$

For many important cases, $r_n^* = O(\frac{1}{\sqrt{n}})$, whence $\rho_n = O(1/\sqrt{n})$. Then Theorem 2 and Corollary 4 show that by choosing a radius in the order of the inverse of $1/\sqrt{n}$, with high probability, the Wasserstein robust loss serves as an upper bound for the true loss up to an $O(1/n)$ gap. Here the probability bound has a $O(\log n)$ term, nearly independent of sample size $n$. By mapping $t$ to $t + \log[\log_2(\kappa_1 \sqrt{tn})]$, one can obtain a probability bound that is independent of sample size, while the radius $\rho_n \sim O(\sqrt{\log n/n})$. In the rest of the paper, we will not make such a transformation, but just keep in mind that these two results are equivalent.

In the next corollary, we consider the loss functions of a composition form $\ell \circ f$, where $\ell: \mathbb{R} \to \mathbb{R}$ is a given Lipschitz function and $f \in \mathcal{F}$, which occurs often in supervised learning. The following result is useful to establish the generalization bound for problems of this type.

**Corollary 5 (Lipschitz composition).** Assume $\mathbb{P}_{\text{true}}$ satisfies $T_1(\tau)$, Assumption 1(I) holds, and Assumption 5 holds with $T(f) = \|f\|_{\text{Lip}}^2$. Let $\ell$ be an $L_\ell$-Lipschitz function and $t > 0$. Then with probability at least $1 - [\log_2(L_\ell \kappa_1 \sqrt{tn})] e^{-t}$,

$$\mathbb{E}_{\mathbb{P}_{\text{true}}} [\ell \circ f] \leq \mathbb{E}_{\mathbb{P}_n} [\ell \circ f] + 2 \left( \sqrt{\frac{t}{n}} L_\ell + 2 L_\ell^2 \sqrt{r_n^*} + \frac{L_\ell}{n^{\frac{1}{4}}} \right) \|f\|_{\text{Lip}} + 4 L_\ell^2 r_n^* + \frac{2 L_\ell}{n}, \ \forall f \in \mathcal{F}.$$
In Section 5, we will illustrate Corollary 5 in supervised learning with linear class (Example 2) and with nonlinear kernel class (Example 5).

The analysis for 2-Wasserstein DRO is aligned with the previous case but requires more care to deal with the data-dependent regularization $\|\|\nabla f\|\|_{p_{n,2}}$; see details in Appendix B.2.3. Define

$$\mathcal{G} := \{ f \in \mathcal{F} : \frac{\|\nabla f\|_{\mathbb{P}_{\text{true}}}^2}{\|\nabla f\|_{\mathbb{P}_{\text{true}}}^2} : f \in \mathcal{F} \}.$$

**Theorem 3 (Gradient regularization).** Assume that $\mathbb{P}_{\text{true}}$ satisfies $T_2(\tau)$, Assumption 2 holds, and Assumption 5 holds with $T(f) = \|\|\nabla f\|\|_{p_{\text{true},2}}^2$. Assume there exists $\kappa_2 > 0$ such that $\|\|\nabla f\|\|_{p_{\text{true},2}} \leq \kappa_2$ for all $f \in \mathcal{F}$. Let $t > 0$. Set

$$\rho_n = 2\sqrt{\frac{d}{n} (1 + 2E_{\mathbb{P}}[\mathcal{R}_{n}(\mathcal{G})]) + 4\sqrt{n} \sqrt{\rho_n} + \frac{2 + \hat{h}t + 2E_{\mathbb{P}}[\mathcal{R}_{n}(\mathcal{G})]}{n^{\sqrt{\rho_n}}},$$

and

$$\epsilon_n = 4\rho_n + \frac{2 + \hat{h}t + 2E_{\mathbb{P}}[\mathcal{R}_{n}(\mathcal{G})]}{n}.$$

Then with probability at least $1 - \lceil \log_2(\kappa_2 \sqrt{\tau n t}) \rceil e^{-t}$,

$$E_{\mathbb{P}_{\text{true}}} [f] \leq E_{\mathbb{P}_{n}} [f] + \rho_n \|\nabla f\|_{p_{\text{true},2}} + \epsilon_n, \quad \forall f \in \mathcal{F}.$$

In the next result, we provide the generalization bound for data-dependent gradient regularization problems and 2-Wasserstein DRO.

**Corollary 6 (2-Wasserstein DRO).** Under the setting in Theorem 3, assume additionally there exists $L > 0$ such that $\|\|\nabla f(z)\|\|_{p_{\text{true},2}} \leq L$ for all $f \in \mathcal{F}$ and $z \in \mathbb{Z}$. Set $\hat{\rho}_n = \rho_n (1 + 2E_{\mathbb{P}}[\mathcal{R}_{n}(\mathcal{G})] + L^2 \sqrt{\frac{d}{2n}})$ and $\tilde{\epsilon}_n = \epsilon_n + \hat{h}^2 \rho_n^2$. Then whenever $2E_{\mathbb{P}}[\mathcal{R}_{n}(\mathcal{G})] + L^2 \sqrt{\frac{d}{2n}} < 1/2$, with probability at least $1 - (\lceil \log_2(\kappa_2 \sqrt{\tau n t}) \rceil + 1)e^{-t}$, for every $f \in \mathcal{F}$,

$$E_{\mathbb{P}_{\text{true}}} [f] \leq E_{\mathbb{P}_{n}} [f] + \hat{\rho}_n \|\nabla f\|_{p_{n,2}} + \epsilon_n,$$

and

$$E_{\mathbb{P}_{\text{true}}} [f] \leq E_{\mathbb{P}_{n}} [f] + \mathcal{R}_{p_{n,2}}(\hat{\rho}_n, f) + \tilde{\epsilon}_n.$$

Whenever $E_{\mathbb{P}}[\mathcal{R}_{n}(\mathcal{G})] = O(1/\sqrt{n})$ and $r_{n^*} = O(1/n)$, which is often the case, Theorem 3 and Corollary 6 show that by choosing a radius in the order of $1/\sqrt{n}$, for every $n \geq O(t)$, with high probability, the Wasserstein robust loss serves as a upper bound for the true loss up to an $O(1/n)$ gap. We will illustrate this result in portfolio optimization (Example 4) and neural networks (Example 6) in Section 5.

5. Examples

In this section, we demonstrate our theoretical results in the context of various applications in operations research and machine learning.

5.1. Big-data Newservendor

We first consider a big-data newservendor problem in the spirit of [3]. In this problem, the decision maker needs to find the optimal ordering quantity for a product with an unknown random demand $y$, subject to holding cost $h > 0$ and backorder cost $b > 0$. In the world of big data, before deciding the ordering quantity, the decision maker observes the a vector of features (such as product information, customer profiles, economic indicators, etc.) and thus can use them make a better ordering decision using these feature information. The vector of features is modeled as a $d$-dimensional random variable.
We consider supervised learning with linear models. Let 
\[
\ell = \mathbb{E}_{\mathbb{P}}\left[ h(\theta^T x - y)_+ + b(y - \theta^Tx)_+ \right],
\]
where \(\mathbb{P}\) is the joint distribution of feature-demand vector \(z = (x, y)\).

**Example 1 (Feature-based newsvendor).** Let \(\mathcal{Z} = (\mathcal{X} \times \mathbb{R}_+)\) be the space of feature-demand vectors. Consider the following distributionally robust feature-based newsvendor problem

\[
\min_{\theta \in \Theta} \sup_{P,W_1(\mathbb{P}_\mathcal{X},\mathbb{P}_\mathcal{Z}) \leq \rho_n} \mathbb{E}_{\mathbb{P}}[h(\theta^T x - y)_+ + b(y - \theta^Tx)_+].
\]

Suppose \(\Theta \subset \{\theta \in \mathbb{R}^d : \|\theta\|_* \leq B\}\) and \(\mathbb{P}_\text{true}\) satisfies \(T_1(\tau)\). Let \(f_\theta(x, y) = h(\theta^T x - y)_+ + b(y - \theta^Tx)_+\) and \(\mathcal{F} = \{f_\theta : \theta \in \Theta\}\).

We have \(\|f_\theta(z) - f_\theta(z)\| = (h \vee b)\|x\|\|\tilde{\theta} - \theta\|\), and thus by Lemma 3 in Appendix A, Assumption 3 holds. Moreover, by [89, Example 5.8], \(\log \mathcal{N}(\varepsilon; \Theta, \|\cdot\|_*) \leq d \log(B(1 + 2/\varepsilon))\). Hence, by Corollary 2(I), there exists \(C > 0\) such that with probability at least \(1 - \exp(-Cn) - \exp(-t + d \log(B(1 + 2n)))\), we have

\[
\mathbb{E}_{\mathbb{P}_\text{true}}[f_\theta] \leq \mathbb{E}_{\mathbb{P}_n}[f_\theta] + \sqrt{\frac{4}{n}}\|f_\theta\|_{\text{Lip}} + \frac{3(h \vee b)\mathbb{E}_{\mathbb{P}_\text{true}}[\|x\|]}{n}, \quad \forall \theta \in \Theta.
\]

By Lemma 15 in Appendix C.1, we have \(\|f_\theta\|_{\text{Lip}} \leq \frac{\sqrt{h \vee b} \mathcal{R}_{\mathbb{P}_n, 1}(f_\theta; \rho_n)}{\sqrt{n}}\). It follows that by setting \(\rho_n = \frac{h \vee b}{h \vee b} \sqrt{\frac{4}{n}}\), with probability at least \(1 - \exp(-Cn) - \exp(-t + d \log(B(1 + 2n)))\),

\[
\mathbb{E}_{\mathbb{P}_\text{true}}[f_\theta] \leq \mathbb{E}_{\mathbb{P}_n}[f_\theta] + \frac{\mathcal{R}_{\mathbb{P}_n, 1}(f_\theta; \rho_n)}{\rho_n} + \frac{3(h \vee b)\mathbb{E}_{\mathbb{P}_\text{true}}[\|x\|]}{n}, \quad \forall \theta \in \Theta.
\]

We remark that in this example, the newsvendor loss function satisfies only Assumption 1(I) but not Assumption 1(II). Hence, we did not directly use Corollary 2(II) to derive the performance guarantee. Instead, Lemma 15 in Appendix C.1 actually shows that \(\mathcal{R}_{\mathbb{P}_n, 1}(f_\theta; \rho_n)\) can achieve a fraction \(\frac{h \vee b}{h \vee b}\) of \(\|f_\theta\|_{\text{Lip}}\) (and thus \(\mathcal{R}_{\mathbb{P}_\text{true}, 1}(f_\theta; \rho_n)\)) uniformly for all \(\theta\), thereby we can still choose \(\rho_n = O(1/\sqrt{n})\) to ensure a good performance guarantee for the Wasserstein robust solution.

### 5.2. Linear Prediction

We consider supervised learning with linear models. Let \(z = (x, y) \in \mathcal{Z} = \mathcal{X} \times \mathcal{Y} \subset \mathbb{R}^d \times \mathbb{R}\). To ease the exposition, we assume \(\|z - \tilde{z}\| = \|x - \tilde{x}\| + \infty \cdot 1\{y \neq \tilde{y}\}\), thereby we only focus on the \(x\)-component when calculating the gradient. Set

\[
l(u, y) := \begin{cases} \ell(u - y), & \text{regression,} \\
\ell(yu), & \text{classification,}
\end{cases}
\]

where \(\ell : \mathbb{R} \to \mathbb{R}\) is \(L\)-Lipschitz continuous, and \(\mathcal{Y} \subset \mathbb{R}\) for regression and \(\mathcal{Y} = \{-1, 1\}\) for classification, and denote \(l \circ f_\theta(z) := l(f_\theta(x), y)\). We denote by \(l'\) the derivative function of \(l\) with respect to its first argument, which is well-defined almost everywhere in \(\mathbb{R}\). Denote by \(\mathbb{P}_\text{true}\) the marginal distribution of \(\mathbb{P}_\text{true}\). The two examples considered in this subsection are on linear predictions for \(p = 1\) and \(p = 2\) respectively.

**Example 2 (Linear class with Lipschitz loss, 1-Wasserstein DRO).** Let \(\Theta \subset \{\theta \in \mathbb{R}^d : \|\theta\|_2 \leq B\}\) for some \(B > 0\). Define \(\mathcal{F} = \{x \mapsto f_\theta(x) = \theta^Tx : \theta \in \Theta\}\).

Consider loss functions of the composite form (3) and let \(f_\theta(x) = \theta^Tx\). Assume \(\mathbb{P}_\text{true}\) is sub-Gaussian, i.e., there exists \(a > 0\) such that \(C := \log \mathbb{E}_{\mathbb{P}_\text{true}}[\exp(a\|x\|^2_2)] < \infty\). Assume \(\ell\) satisfies \(\limsup_{|t| \to \infty} \frac{\ell(t)}{|t|} = L\).
Examples of \( \ell(t) \) include convex losses such as hinge loss \( (1 - t)_+ \), softplus (logistic) loss \( \log(1 + e^t) \), as well as non-convex losses such as inverse S-shaped curve \( \text{sgn}(t) \log\left(\frac{1}{2}(1 + e^t)\right) \).

Let us verify the assumptions in Corollary 5. By Lemma 3 in Appendix A, \( P^x \) satisfies \( T_1(\frac{2}{a}(1 + C)) \). Assumption 1 holds with \( \|f_0\|_{\text{Lip}} = \|\theta\|_2 \leq B =: \kappa_1 \). Furthermore, since

\[
\{ f_0 : \theta \in \Theta, 0 \leq c \leq 1, c^2\|\theta\|_2^2 \leq r \} \subset \{ f_0 : \|\theta\|_2 \leq \sqrt{r} \},
\]

we can set \( \psi_n(r) = \sqrt{n} E^{\text{true}}[\|x\|_2^2]/n \) which, by Lemma 17 in Appendix C.2, is an upper bound of \( E_\Theta[\{ f_0 : \|\theta\|_2 \leq \sqrt{r} \}] \). By Jensen’s inequality, we have

\[
E_{P_{\text{true}}} \left[ \|x\|_2^2 \right] \leq \frac{1}{a} \log E_{P_{\text{true}}} \left[ \exp(a\|x\|_2^2) \right] \leq \frac{C}{a}.
\]

It follows that \( \psi_n(r) = \frac{\sqrt{C}}{an} \) and thus \( r_{n^*} = \frac{C}{an} \).

Let \( t > 0 \). Set

\[
\rho_n = 2L\sqrt{\frac{2(1+C)t}{an}} + 4L^2 \sqrt{\frac{C}{an}} + \frac{2Lt}{\sqrt{Cn/a}}, \quad \epsilon_n = \frac{4L^2C}{an} + \frac{2L}{n}.
\]

Then by Corollary 5 and Lemma 1, with probability at least \( 1 - \left[ (\log_2(LB) + \frac{1}{2} \log_2(\frac{2}{a}(1 + C)tn)) \right] e^{-t} \),

\[
E_{P_{\text{true}}} [l(\theta^T x, y)] - E_{P_{\theta}} [l(\theta^T x, y)] \leq \mathcal{R}_{P_{\theta}}(\rho_n; l \circ f_0) + \epsilon_n = \rho_n \|\theta\|_2 + \epsilon_n, \quad \forall \theta \in \Theta.
\]

The bound obtained in Example 2 is consistent to the existing literature on the generalization bounds for linear predictions. But unlike the typical results (e.g. [74, 23]), we do not impose boundedness assumptions on the loss function \( \ell \) or the domain \( X \). If imposing a positive lower bound on \( \|\theta\|_2 \geq c > 0 \), the bound given in the example further becomes

\[
E_{P_{\text{true}}} [l(\theta^T x, y)] - E_{P_{\theta}} [l(\theta^T x, y)] \leq (\rho_n + \epsilon_n/c) \|\theta\|_2 = \mathcal{R}_{P_{\theta}}(\rho_n + \epsilon_n/c; l \circ f_0).
\]

Thereby one can bound the true loss using only the Wasserstein robust loss with an inflated radius \( \rho_n + \epsilon_n/c \) without having a higher order error term. This bound is of the same form as in Shafieezadeh-Abadeh et al. [72, Theorem 4.6] which has a linear dependence on the dimension \( d \) of \( X \) (albeit under the subgaussian assumption that is slightly weaker than \( T_1 \)); while our bound is independent of \( d \) (at least for the case of 2-norm).

**Example 3 (Linear class with Lipschitz loss, 2-Wasserstein DRO).** Consider a similar setup as in Example 2 but with slightly different notations in order to be consistent with Corollary 3. Let \( \Theta \subset \{ \theta \in \mathbb{R}^d : \|\theta\|_2 \leq B \} \) for some \( B > 0 \). Define

\[
\mathcal{F} = \{ (x, y) \mapsto l(\theta^T x, y) : \theta \in \Theta \},
\]

where \( l \) is defined in (3). Let \( f_0(x) = l(\theta^T x, y) \). Then \( \|\nabla f_0(z)\|_* = \|\nabla x l(\theta^T x, y)\|_* = \|\theta\|_2 \|l'(\theta^T x, y)\| \), recalling \( l' \) denotes the derivative of \( l \) with respect to its first argument. Assume further that \( l \) in (3) has \( h \)-Lipschitz gradient; \( P^x \) satisfies \( T_2(\tau) \) for some \( \tau > 0 \); and \( \inf_{\theta \in \Theta} E_{P_{\text{true}}}[l'(\theta^T x, y)^2] > 0 \). Note that the last condition is mild – indeed, it is satisfied if for every \( \theta \in \Theta \), \( l'(\theta^T ; y) \) is non-zero on some subset of \( X \) with positive \( P^x \)-measure (together with the boundedness assumption on \( \Theta \)).

Let us verify the conditions and compute the constants in Corollary 3. Assumption 2 is satisfied since \( f_0 \) has \( hB^2 \)-Lipschitz gradient, and \( \sigma = \sup_{\theta \in \Theta} E_{P_{\text{true}}}[l'(\theta^T x, y)^4]^{\frac{1}{4}} / E_{P_{\text{true}}}[l'(\theta^T x, y)^2] \leq L^2 / \inf_{\theta \in \Theta} E_{P_{\text{true}}}[l'(\theta^T x, y)^2] < \infty \). We have \( \|f_0(z) - f_0(z)\| \leq L_f \|x\| \|\theta - \theta\|_{\infty} \), and by Lemma 18, \( \|\nabla f_0(z) - \nabla f_0(z)\|_{\infty} \leq (L_f + Bh) \|x\| \|\theta - \theta\|_{\infty} \). Hence, by Lemma 3 in Appendix A, Assumptions 3 and 4 hold. Moreover, by [89, Example 5.8], \( \log \mathcal{N}(\epsilon; \Theta, \|\cdot\|_{\infty}) \leq d \log(B(1+2n)) \). Let \( t > 0 \), \( n \geq 8\sigma^2 t \) and \( \rho_n = \sqrt{\frac{\pi}{n}} (1 + \sqrt{\frac{2t}{n}}) \).
By Corollary 3, there exists $C > 0$ such that with probability at least $1 - 2 \exp(-Cn) - 2 \exp(-t + d \log(B(1 + 2n)))$, for every $\theta \in \Theta$,

$$
\mathbb{E}_{\mathbb{P}_{\text{true}}} [I(\theta^T x, y)] - \mathbb{E}_{\mathbb{P}_n} [I(\theta^T x, y)] \leq 
\rho_n \|\theta\|_2 \cdot \|I'(\theta^T x, y)\|_{\mathbb{P}_n, 2} + 2h^2 \rho_n^2 \frac{3L_f \mathbb{E}_{\mathbb{P}_{\text{true}}} \left[\|x\|\right] + 2\rho_n (L_f + Bh \mathbb{E}_{\mathbb{P}_{\text{true}}} \left[\|x\|^2\right])}{n},
$$

and

$$
\mathbb{E}_{\mathbb{P}_{\text{true}}} [I(\theta^T x, y)] - \mathbb{E}_{\mathbb{P}_n} [I(\theta^T x, y)] \leq 
\mathcal{R}_{\mathbb{P}_n, 2}(\rho_n; f_0) + 2h^2 \rho_n^2 + \frac{3L_f \mathbb{E}_{\mathbb{P}_{\text{true}}} \left[\|x\|\right] + 2\rho_n (L_f + Bh \mathbb{E}_{\mathbb{P}_{\text{true}}} \left[\|x\|^2\right])}{n}.
$$

5.3. Portfolio Optimization

In this subsection, we study the classic Markowitz’s mean-variance portfolio optimization problem. Let $x$ be the vector of random losses of $d$ assets with distribution $\mathbb{P}_{\text{true}}$ and let $w \in \mathcal{W} \subset \mathbb{R}^d$ be the portfolio weights satisfying $w^T 1 = 1$. Note that the variance of a one-dimensional random variable $Y$ has an equivalent representation $\text{Var}[Y] = \min_{u \in \mathbb{R}} \mathbb{E}[(Y - u)^2]$.

**Example 4 (Markowitz model).** Let $\alpha > 0$ and suppose $\mathcal{W} \subset \{w \in \mathbb{R}^d : w^T 1 = 1, \|w\|_2 \leq B\}$, where $B > 0$. Consider the following distributionally robust mean-variance minimization

$$
\min_{w \in \mathcal{W}} \sup_{u \in \mathbb{R}} \mathbb{E}_{\mathbb{P}} \left[(w^T x - u)^2 + \alpha w^T x\right] = \mathbb{E}^* \left[(w^T x - u)^2 + \alpha w^T x\right].
$$

Assume $\mathbb{P}_{\text{true}}$ satisfies $T_2(\tau)$. Then it also satisfies $T_1(\tau)$ and by Corollary 1, for every $w \in \mathbb{R}^d$ with $\|w\|_2 = 1$, $\mathbb{P}_\otimes \{|\mathbb{E}_{\mathbb{P}_{\text{true}}} \left[w^T x\right] - \mathbb{E}_{\mathbb{P}_{\text{true}}} \left[w^T x\right]\| \geq \epsilon\} \leq 2e^{-\epsilon^2/2}$ for all $\epsilon > 0$, thus $\mathbb{P}_{\text{true}}$ is $\tau/\sqrt{2}$-subgaussian. Let $\mu_2 = \mathbb{E}_{\mathbb{P}_{\text{true}}} \left[\|x\|^2\right]^{1/2} < \infty$, $\mu_4 = \mathbb{E}_{\mathbb{P}_{\text{true}}} \left[\|x\|^4\right]^{1/2} < \infty$. Assume there exists $\zeta > 0$ such that $\text{Cov}_{\mathbb{P}_{\text{true}}} \left[x\right] \geq \zeta I$.

Define

$$
\frac{\mu_4}{\mu_2} \in \arg\min_{u \in \mathbb{R}} \sup_{\mathbb{P} : \mathbb{P}_2(\mathbb{P}, \mathbb{P}_{\text{true}}) \leq \rho_n} \mathbb{E}_{\mathbb{P}} \left[(w^T x - u)^2 + \alpha w^T x\right].
$$

We first derive bounds on $u_n$. Let $n \geq t > 0$, then by Lemma 21 in Appendix C.3, with probability at least $1 - e^{-t/(2d\mu_2^2)}$, where $c$ is a universal constant, such that for every $w \in \mathcal{W}$, $u_n(w) \leq 2B^2 (\mu_2^2 + \tau^2 \sqrt{\frac{L}{n}} + \tau^2 \sqrt{2d + \rho_n^2}) =: U_n$. Note that $U := \sup_n U_n$ is finite whenever $\{\rho_n\}_n$ is bounded. Let $\theta = (w, -u + \alpha/2)$, $z = (x, y)$ and $f_0(z) = (\theta^T z)^2 - au - a^2/4$. Thereby we have $f_0(x, 1) = (w^T x - u)^2 + \alpha w^T x$. Set $\Theta = \{(w, u) : w \in \mathcal{W}, |u| \leq U + \alpha/2\}$. Let $\mathbb{P}_{\text{true}} = \mathbb{P}_{\text{true}} \otimes \delta_1$.

Let us verify the assumptions and compute the constants required by Theorem 3. $\mathbb{P}_{\text{true}}$ satisfies $T_2(\tau)$ since any distribution $\mathcal{Q} \in \mathcal{P}_2(\mathcal{Z})$ with finite $H(\mathcal{Q}, \mathbb{P}_{\text{true}})$ has the form $\mathcal{Q}_x \otimes \delta_1$. We have $\nabla f_0(z) = 2(\theta^T z)\theta$. Thus $\nabla^2 f_0(z) = 2\theta \theta^T$, hence Assumption 2 is satisfied with $h = 2(B^2 + U^2)$. To find a sub-root function $\psi_n(r)$ required by Assumption 5, observe that

$$
\|\nabla f_0\|_2 \|\nabla_{\text{true}}\|_2 = 2\|\theta\|_2 \mathbb{E}_{\mathbb{P}_{\text{true}}} \left[\|\theta^T z\|\right]^{1/2} \geq 2\zeta \|\theta\|_2^2.
$$

It follows from Lemma 22 in Appendix C.3 that

$$
\mathbb{E}_\theta \left[\mathcal{R}_n \left\{c f_0 : \theta \in \Theta, 0 \leq c \leq 1, c^2 \|\nabla f_0\|_2 \|\nabla_{\text{true}}\|_2 \leq r\right\}\right] 
\leq \mathbb{E}_\theta \left[\mathcal{R}_n \left\{c f_0 : 0 \leq c \leq 1, \|\theta\|_2 \leq \left(\frac{r}{4\zeta^2\mu_2^2}\right)^{1/4}\right\}\right] 
\leq \frac{\mu_4 \sqrt{r}}{2\zeta^2 \sqrt{n}} := \psi_n(r).
$$
Thus $r_n \leq \frac{\mu_4}{4\xi^2 n}$. Moreover, by Lemma 23 in Appendix C.3,

$$
\mathbb{E}_\Theta \left( \mathbb{R}_n \left( \frac{\|\nabla f_0\|_2^2}{\|\nabla f_0\|_{p_{true,2}}^2} : \theta \in \Theta \right) \right) \leq \frac{\sqrt{\mu_2^2 + 2\mu_2^2 + 1}}{(1 \wedge \xi) \sqrt{n}}.
$$

In addition, $\sup_{\theta \in \Theta} \|\nabla f_0\|_2 \|p_{true,2} \leq 2B(B\mu_2 + U + \alpha/2) =: \kappa_2$.

Let $t > 0$. Then in Theorem 3, we have $\rho_n = O(1/n)$, $\epsilon_n = O(1/n)$, and with probability at least 

$$
1 - [\log_2(\kappa_2 \sqrt{nt})] e^{-t},
$$

$$
\mathbb{E}_{\text{true}} \left[ (w^T x - u)^2 + \alpha w^T x \right] \leq \mathbb{E}_{\text{true}} \left[ (w^T x - u)^2 + \alpha w^T x \right] + \rho_n \|\nabla f_0\|_2 \|p_{true,2} + \epsilon_n,
$$

where $\forall \theta \in \Theta$. By Lemma 25 in Appendix C.3, there exists $\tilde{\rho}_n = \tilde{O}(1/\sqrt{n})$, $\tilde{\epsilon}_n = O(1/n)$ such that for $n \geq O(t)$, with probability at least 

$$
1 - ((\log_2(\kappa_2 \sqrt{nt})) + 2) e^{-t},
$$

$$
\mathbb{E}_{\text{true}} \left[ (w^T x - u)^2 + \alpha w^T x \right] \leq \mathbb{E}_{\text{true}} \left[ (w^T x - u)^2 + \alpha w^T x \right] + \mathbb{R}_{\rho_n,\theta}(\tilde{\rho}_n; f_\theta) + \tilde{\epsilon}_n,
$$

where $\forall (w, u) \in \Theta$. Hence, with probability at least 

$$
1 - ([\log_2(\kappa_2 \sqrt{nt})] + 2) e^{-t} - e^{-\frac{c(t+2d)}{\text{log}^2(\kappa_2 \sqrt{nt})}},
$$

$$
\mathbb{E}_{\text{true}} \left[ (w^T x - u)^2 + \alpha w^T x \right] \leq \mathbb{E}_{\text{true}} \left[ (w^T x - u)^2 + \alpha w^T x \right] + \mathbb{R}_{\rho_n,\theta}(\tilde{\rho}_n; f_\theta) + \tilde{\epsilon}_n.
$$

\hfill \Box

The last equation in Example 4 shows that the true mean-variance of the portfolio $w_n$ is upper bounded by its robust mean-variance up to a higher-order term. We remark that in this example, the parameter space is not bounded as $u \in \mathbb{R}$, which makes it impossible to obtain a bounded complexity for the entire class of loss functions. We circumvent such difficulty by showing that the optimal solution $u_n$ lies in a bounded set $\Theta$ with high probability, thereby it suffices to restrict on $\Theta$. A similar argument applies to the empirical gradient estimate, for which we consider a modified version of Corollary 6 (Lemma 25 in Appendix C.3).

In the next two subsections, we illustrate our results for Lipschitz regularization and gradient regularization for nonlinear class. Similar to Section 5.2, we let $z = (x, y) \in \mathcal{Z} = \mathcal{X} \times \mathcal{Y} \subset \mathbb{R}^d \times \mathbb{R}$ and assume $\|z - \tilde{z}\| = \|x - \tilde{x}\|_2 + \alpha 1\{y \neq \tilde{y}\}$.

### 5.4. Kernel Method

We consider Lipschitz regularization of kernel class (see, e.g., [89, Chapter 12]). Let $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}_+$ be a positive definite kernel on $\mathcal{X} \subset (\mathbb{R}^d, \|\cdot\|_2)$ with $\sigma := (\mathbb{E}_{x - \text{true}} \left[ k(x, x) \right])^{1/2} < \infty$. We can associate $k$ with a feature map $\Phi : \mathcal{X} \to \mathcal{H}$, where $\mathcal{H}$ is a Hilbert space with inner product $\langle \cdot, \cdot \rangle$ and $k(x_1, x_2) = \langle \Phi(x_1), \Phi(x_2) \rangle$. Denote by $\|\cdot\|$ a norm on $\mathcal{H}$. Let $\{x_j\}_{j=1}^m \subset \mathcal{X}$, where $m \in \mathbb{N}_{\geq 1}$. Then we have 

$$
\|\sum_{j=1}^m \theta_j \Phi(x_j)\|_2^2 = \sum_{j,k=1}^m \theta_j \theta_k k(x_j, x_k).
$$

In kernel method, one often consider the following parameterized class 

$$
\mathcal{F} = \left\{ x \mapsto \sum_{j=1}^m \theta_j k(x, x_j) : \sum_{j,k=1}^m \theta_j \theta_k k(x_j, x_k), \ m \in \mathbb{N}_{\geq 1} \leq B \right\},
$$

where $B > 0$. 


Example 5 (Lipschitz Regularization for Kernel Class). Consider loss functions of the form (3) with \( f_\theta \in \mathcal{F} \) defined as above. Assume \( k \) is differentiable and there exists \( \zeta > 0 \) such that
\[
\mathbb{E}_{x \sim p_{\text{true}}} \left[ \sum_{j=1}^m \theta_j \nabla_x k(x, x_j) \right]_{\zeta} \geq \zeta \sum_{j,k=1}^m \theta_j \theta_k k(x, x_k) \text{ for all } \theta \in \Theta,
\]
which can be satisfied when the matrix \( (E_{x \sim p_{\text{true}}}[\nabla_x k(x, x)]^T \nabla_x k(x, x))_{1 \leq j, k \leq m} \) is positive definite. Furthermore, assume \( \kappa_1 = \sup_{\theta \in \Theta, x \in \mathcal{X}} \| \sum_{j=1}^m \theta_j \nabla_x k(x, x_j) \| < \infty \), thus Assumption 1(1) is satisfied.

Let us compute the generalization bound using Corollary 5. To this end, we need to specify \( \psi_n \) and compute its fixed point \( r_n \). Observe that \( c^2 \| f_\theta \|_{\text{Lip}}^2 \leq r \) implies
\[
r \geq c^2 \| f_\theta \|_{\text{Lip}}^2 \geq c^2 \| \nabla f_\theta \|_2^2 \geq \zeta c^2 \| \theta \|_k^2,
\]
hence \( \psi_n(r) \) can be chosen as \( 2\sigma \sqrt{\frac{r}{n\zeta}} \), an upper bound of \( E_{\Theta} [ \# \{ f_\theta : \| \theta \|_k \leq \sqrt{r/\zeta} \} ] \) according to [7, Lemma 22]. Thus \( r_n = \frac{4\sigma^2}{n\zeta} \).

Then by Corollary 5, with probability at least \( 1 - [\log_2 (L\kappa_1 \sqrt{\pi} n) \] e^{-t},
\[
E_{p_{\text{true}}} [ l \circ f_\theta ] \leq E_{p_n} [ l \circ f_\theta ] + r_n \| f_\theta \|_{\text{Lip}} + \epsilon_n, \quad \forall \theta \in \Theta.
\]

This result provides a generalization bound for Lipschitz regularization problems [59, 35, 86] when the loss function class belongs to a kernel class. We remark that the setup in this example is different from Shafieezadeh-Abadeh et al. [72, Section 3.3], in which the distributional uncertainty is imposed on the feature space, while Example 5 considers distributional uncertainty on the original data space.

5.5. Neural Networks

In the last example, we illustrate the generalization bound for gradient regularization for a simple two-layer neural network. Consider
\[
\mathcal{F} = \left\{ (x, y) \mapsto l(W_2 \phi(W_1 x), y) : (W_1, W_2) \in \Theta \right\},
\]
where \( l \) is defined in (3), \( \phi = (\phi_1, \ldots, \phi_d) \) are entry-wise 1-Lipschitz activation functions, and \( \Theta \) is the space of weight matrices
\[
\Theta = \{ \theta = (W_1, W_2) : W_1 \in \mathbb{R}^{d_2 \times d_1}, W_2 \in \mathbb{R}^{1 \times d_2}, W_1 W_1^T = I, \| W_2 \|_2 \leq B \}.
\]

Here the constraint \( W_1 W_1^T = I \) enforces the orthonormal regularization on the weight matrix \([88, 97, 4, 46]\), which is a popular way to ensure the training stability and performance for neural nets. Let \( f_\theta(z) = l(W_2 \phi(W_1 x), y) \).

Example 6 (Gradient Regularization for Neural Networks). Assume \( l \) and \( \phi_i \) has Lipschitz gradient, \( j = 1, \ldots, d_2 \), thereby \( f_\theta \) is smooth and \( \mathcal{F} \) satisfies Assumption 2. Assume \( \eta := \inf_{\theta \in \Theta, x \in \mathcal{X}} l'(W_2 \phi(W_1 x), y) > 0 \), which can be satisfied, for example, when \( \mathcal{Z} \) is bounded and \( l \) is the logistic loss. Furthermore, assume there exists \( \zeta > 0 \) such that \( E_{p_{\text{true}}} [ \| W_2 \phi'(W_1 x) \|_2^2 ] \geq \zeta \| W_2 \|_2^2 \) for every \( (W_1, W_2) \in \Theta \) and \( \sigma = E_{p_{\text{true}}} [ \| x \|_2^2 ]^{1/2} < \infty \).

Let us compute the constants in Corollary 6. We have
\[
\| \| \nabla f \| \|_{p_{\text{true}}, 2} = E_{p_{\text{true}}} [ l'(W_2 \phi(W_1 x), y) W_2 \phi'(W_1 x) W_1 ]^2 \geq \eta^2 \zeta \| W_2 \|_2^2.
\]
Thus, by our assumption, \( c^2 \| \nabla f \|_\ast^2 \leq r \) implies \( \| W_2 \|_2 \leq \frac{\sqrt{r}}{c \eta \sqrt{n}} \). As a result,

\[
\mathbb{E}_\oplus \left[ \mathcal{R}_n \left\{ c f_\theta : \theta \in \Theta, 0 \leq c \leq 1, c^2 \| \nabla f_\theta \|_\ast^2 \leq r \right\} \right] \\
\leq L \mathbb{E}_\oplus \left[ \mathcal{R}_n \left\{ x \mapsto W_2 \phi(W_1 x) : W_1 W_1^T = I, \| W_2 \|_2 \leq \frac{\sqrt{r}}{\eta \sqrt{n}} \right\} \right] \\
\leq \frac{L \sigma \sqrt{2r d_2}}{\eta \sqrt{\zeta n}} =: \psi_n(r),
\]

where the first inequality follows from Lemma 8 in Appendix C, and the second inequality is due to Lemma 27 in Appendix C.4. It follows that \( r^*_n = \frac{2L \sigma^2 d_2}{\eta^2 \zeta n} \). Moreover, by Lemma 28 in Appendix C.4,

\[
\mathbb{E}_\oplus \left[ \mathcal{R}_n(\mathcal{G}) \right] \leq \frac{2L (L \varepsilon + 1) \sigma \sqrt{2d_2}}{\eta^2 \zeta \sqrt{n}}.
\]

With the analysis above, in Corollary 6 it holds that \( \tilde{\rho}_n = O(1/\sqrt{n}), \epsilon_n = O(1/n), \tilde{\epsilon}_n = O(1/n) \), and with probability at least \( 1 - (\log_2(\kappa_2 \sqrt{\eta n})) + 1) \varepsilon^{-1} \), for every \( \theta \in \Theta \),

\[
\mathbb{E}_{p_{\text{true}}} [f_\theta] \leq \mathbb{E}_{p_n} [f_\theta] + \tilde{\rho}_n \mathbb{E}_{p_n} \left[ \| W_2 \phi(W_1 x), y \|^2 \| W_2 \phi'(W_1 x) \|_2^2 \right]^\frac{1}{2} + \epsilon_n,
\]

and

\[
\mathbb{E}_{p_{\text{true}}} [f_\theta] \leq \mathbb{E}_{p_n} [f_\theta] + \mathcal{R}_{p_n,2}(\tilde{\rho}_n; f_\theta) + \tilde{\epsilon}_n.
\]

6. Concluding Remarks

In this paper, we have developed finite-sample performance guarantees for Wasserstein DRO and its associated variation regularization without suffering from the curse of dimensionality. These results help us to understand the empirical success of Wasserstein DRO and/or Lipschitz and gradient regularization. In the meantime, many issues worth investigating are left to future work.

First, we restrict the families of loss functions consistent with Lemma 1 and Lemma 2 that establish the equivalence between Wasserstein DRO and variation regularization. One can extend the results to more general families such as non-smooth losses using the results in [37]. Second, in Section 3, we adopt the widely used transportation inequalities \( T_p \), \( p \in [1, 2] \), which is general enough to cover most subgaussian distributions and works for loss functions of linear and quadratic growth. Nonetheless, one may obtain results for more general distributions and loss functions by considering other families of transportation-information inequalities \([19, 43, 85, 44]\). Third, we focus primarily on the case \( p \in [1, 2] \), but leave the study to future work on the concentration and generalization for another important case \( p = \infty \), which has been widely considered in adversarial robust learning (e.g., \([42, 73, 104]\)). Fourth, in Section 4, we developed a local Rademacher complexity theory based on the variation of the loss. Investigation of these techniques in the context of other problems in statistical learning theory seems interesting, and hopefully would yield new results.

In summary, we hope our results can inspire more fruitful findings for problems in operations research and machine learning in which Wasserstein distributional robustness plays an increasingly prominent role.
Appendix A: Proofs for Section 3

Proof of Proposition 1. Define
\[ \Phi(t) := E_{P_{\text{true}}} \left[ \sup_{\tilde{z} \in Z} \{ t(f(\tilde{z}) - f(z)) - \|\tilde{z} - z\|^p \} \right]. \]

Using the strong dual (1) which is well-defined due to Theorem 1 in [38], we have that
\[ R_{P_{\text{true}}, P}(\rho) = \lambda_o \rho^p + \lambda_o \Phi(1/\lambda_o). \]  

By [38], \( \lambda_o \in [\lambda, \infty) \). We first consider the case \( \lambda_o > \lambda \), in which case the optimizer is in the interior of the domain of the dual objective. The first-order optimality condition of the convex optimization (1) reads
\[ \rho^p + \Phi(1/\lambda_o) \in \frac{1}{\lambda_o} \partial \Phi(1/\lambda_o). \]

Set \( \epsilon = R_{P_{\text{true}}, P}(\rho) \). It follows from the equations above that
\[ \epsilon = \lambda_o \rho^p + \lambda_o \Phi(1/\lambda_o) \in \partial \Phi(1/\lambda_o). \]

But by definition \( I_p(\epsilon; f)^p = \sup_{t \geq 0} \{ \epsilon t - \Phi(t) \} \). This suggests that \( t = 1/\lambda_o \) is an optimizer of \( \sup_{t \geq 0} \{ \epsilon t - \Phi(t) \} \). Hence,
\[ I_p(\epsilon; f)^p = \frac{\epsilon}{\lambda_o} - \Phi(1/\lambda_o) = \lambda_o (\lambda_o \rho^p + \lambda_o \Phi(1/\lambda_o)) - \Phi(1/\lambda_o) = \rho^p. \]

Next, consider the case that the unique dual minimizer \( \lambda_o = \frac{1}{\lambda} \). Taking a feasible solution \( t = 1/\lambda_o \), using (4) we obtain that
\[ I_p(\epsilon; f)^p = \frac{\epsilon}{\lambda_o} - \Phi(1/\lambda_o) = \rho^p. \]

\[ \square \]

The next lemma is mentioned in Section 3.2.

Lemma 3 (Corollary 2.4 in Bolley and Villani [19]). Assume there exists \( a > 0 \) such that \( C := \log E_\mathcal{P}[\exp(a\|Z\|^2)] < \infty \). Then \( \mathcal{P} \) satisfies \( T_1(\tau) \), where
\[ \tau = \inf_{\tilde{z} \in Z, \tilde{z} > 0} \left\{ \frac{2}{a} (1 + \log E_\mathcal{P}[\exp(\tilde{a}\|Z - \tilde{z}\|^2)]) \right\} \leq \frac{2}{a} (1 + C). \]

A.1. Proof of Theorem 1

Our proof is based on Marton’s argument and Herbst’s argument [52, 68]. Let us begin with some definitions and lemmas.

Denote \( Z := (Z^n_1, \ldots, Z^n_m) \). We define a product distance \( d_\rho \) on the space \( Z^n \) as
\[ d_\rho (Z^n, \tilde{Z}^n) := \left( \sum_{i=1}^n \|Z^n_i - \tilde{Z}^n_i\|^p \right)^{1/p}, \]

The \( \rho \)-Wasserstein distance between probability distributions \( \mu \) and \( \mathbb{P}_\circ \) is given by
\[ \mathcal{W}_\rho(\mu, \mathbb{P}_\circ) = \min_{\pi} \left\{ \left( E_{(Z^n, \tilde{Z}^n) \sim \pi} [d_\rho (Z^n, \tilde{Z}^n)^p] \right)^{1/p} : \pi \text{ has marginal distributions } \mu, \mathbb{P}_\circ \right\}. \]

The following tensorization lemma establishes a transportation-information inequality for the product distribution \( \mathbb{P}_\circ \) (see, for example, Proposition 22.5 in [85]).
Lemma 4. Let $p \in [1, 2]$. Suppose $\mathbb{P} \in \mathcal{P}_p(Z)$ satisfies $T_p(\tau)$. Then $\mathbb{P}_\mathbb{Q}$ satisfies $T_p(\tau n^{\frac{2}{p} - 1})$.

Given any function $g : Z \to \mathbb{R}$ which is exponentially integrable with respect to $\nu$, we define a distribution $\nu^{(g)}$, called the $g$-exponential tilting of $\nu$ as (see, e.g., Section 3.1.2 in [68]):

$$
\frac{d\nu^{(g)}}{d\nu} = \frac{\exp(g)}{E_\nu[\exp(g)]}.
$$

It follows that

$$
H(\nu^{(g)}||\nu) = E_\nu^{(g)}[g] - \ln E_\nu[\exp(g)].
$$

We prove below a more general concentration result that applies not only for the empirical mean.

Lemma 5. Let $p \in [1, 2]$. Assume $\mathbb{P}_{\text{true}}$ satisfies $T_p(\tau)$. Let $F : Z^n \to \mathbb{R}$. Assume $E_{\mathbb{Q}}[F] = 0$ and there exist $M, L > 0$ and $z^n_0 \in Z^n$ such that

$$
F(z^n) \leq M + \frac{L}{n} d_p(z^n, z^n_0)^p, \quad \forall z^n \in Z^n.
$$

Define $\mathcal{J}(\cdot : F) : \mathbb{R}_+ \to \mathbb{R}_+$ via

$$
\mathcal{J}(\epsilon : F) := \sup_{t > 0} \left\{ \epsilon t - E_{\mathbb{Q}} \left[ \sup_{z^n \in Z^n} \left\{ t(F(\tilde{z}^n) - F(z^n)) - \frac{1}{n} d_p(\tilde{z}^n, z^n)^p \right\} \right] \right\},
$$

and $\mathcal{R}(\cdot ; F) : \mathbb{R}_+ \to \mathbb{R}_+$ as

$$
\mathcal{R}(\rho ; F) = \inf_{\lambda > 0} \left\{ \lambda \rho^p + E_{\mathbb{Q}} \left[ \sup_{z^n \in Z^n} \left\{ F(\tilde{z}^n) - F(z^n) - \frac{\lambda}{n} d_p(\tilde{z}^n, z^n)^p \right\} \right] \right\}.
$$

Then for any $\epsilon > 0$,

$$
\mathbb{P}_{\mathbb{Q}} \{ F(z^n) > \epsilon \} \leq \exp \left( - n.\mathcal{J}(\epsilon ; F)^2 / \tau \right).
$$

Let $t > 0$. Then with probability at least $1 - e^{-t}$,

$$
F(z^n) \leq \mathcal{R}(\sqrt{\frac{t}{n}} ; F).
$$

Proof of Lemma 5. Define

$$
\Phi(t ; F) := E_{\mathbb{Q}} \left[ \sup_{z^n \in Z^n} \left\{ t(F(\tilde{z}^n) - F(z^n)) - \frac{1}{n} d_p(\tilde{z}^n, z^n)^p \right\} \right],
$$

which is in $[0, \infty)$ for all sufficiently small $t$ because of the growth rate condition on $F$, and thus $
\mathcal{J}(\epsilon ; F)^p = \sup_{t > 0} \{ \epsilon t - \Phi(t ; F) \}$ is finite. Let $s > 0$. Using Lemma 4, for every $\mu \in \mathcal{P}_p(Z^n)$, it holds that

$$
W_p(\mu, \mathbb{P}_{\mathbb{Q}})^p \leq \left( \tau n^{\frac{2}{p} - 1} H(\mu||\mathbb{P}_{\mathbb{Q}}) \right)^{\frac{p}{2}} \left( \left( \frac{2}{p} \right)^{\frac{p}{2} - 1} s^{-\frac{p}{2}} H(\mu||\mathbb{P}_{\mathbb{Q}}) \right)^{\frac{p}{2}} \left( \left( \frac{p_r}{2} \right)^{\frac{p}{2} - 1} s^{-\frac{p}{2}} n^{1 - \frac{p}{2}} \right).
$$

Applying Young’s inequality to the right side yields that

$$
W_p(\mu, \mathbb{P}_{\mathbb{Q}})^p \leq \frac{p}{2} \left( \left( \frac{2}{p} \right)^{\frac{p}{2} - 1} s^{-\frac{p}{2}} H(\mu||\mathbb{P}_{\mathbb{Q}}) \right)^{\frac{p}{2}} + (1 - \frac{p}{2}) \left( \left( \frac{p_r}{2} \right)^{\frac{p}{2} - 1} s^{-\frac{p}{2}} n^{1 - \frac{p}{2}} \right)^{\frac{1}{1 - \frac{p}{2}}}
$$

$$
= s^{\frac{p}{2} - 1} H(\mu||\mathbb{P}_{\mathbb{Q}}) + (1 - \frac{p}{2})(\frac{p_r}{2})^{\frac{p}{2} - 1} s^{\frac{p}{2} - 1} n,
$$

noting that the second term vanishes when $p = 2$. 
Let us assume temporarily that $F$ is bounded from above. Let $t > 0$. Setting $\mu = P(s^{1-\frac{2}{\tau}} t F)$, by (5) we have
\[
W_p(\mu, P_\otimes)^p \leq s^{\frac{2}{\tau} - 1} \left( \mathbb{E}_\mu \left[ s^{1-\frac{2}{\tau}} t F \right] - \ln \mathbb{E}_\otimes \left[ \exp \left( s^{1-\frac{2}{\tau}} t F \right) \right] + (1 - \frac{P}{2}) \left( \frac{P}{2} \right)^{\frac{2}{\tau} - 1} s^{1} \right).
\]
On the other hand, using Kantorovich's duality (see, e.g., Theorem 5.10 in Villani [85]) and the assumption $\mathbb{E}_\otimes[F] = 0$,
\[
W_p(\mu, P_\otimes)^p \geq \mathbb{E}_\mu[t F] + \mathbb{E}_\otimes \left[ \inf_{z^n \in \mathcal{Z}_n} \left\{ \sum_{i=1}^n \|z^n_i - z^n_i\|^p - \inf_{z^n \in \mathcal{Z}_n} \left\{ \sum_{i=1}^n \|z^n_i - z^n_i\|^p - t F(z^n) \right\} \right\} \right] = \mathbb{E}_\mu[t F] - n \Phi \left( \frac{t}{n}; F \right).
\]
Combining the two inequalities above and canceling out the term $\mathbb{E}_\mu[t F]$, we obtain
\[
s^{\frac{2}{\tau} - 1} \ln \mathbb{E}_\otimes[\exp(s^{1-\frac{2}{\tau}} t F)] \leq (1 - \frac{P}{2}) \left( \frac{P}{2} \right)^{\frac{2}{\tau} - 1} s^{1} + n \Phi \left( \frac{t}{n}; F \right).
\]
Using Markov's inequality, for all $s, t > 0$,
\[
P_\otimes \{F(z^n) > \epsilon\} = P_\otimes \left\{ s^{1-\frac{2}{\tau}} t F(z^n) > s^{1-\frac{2}{\tau}} t \epsilon \right\} \leq \mathbb{E}_\otimes \left[ \exp(s^{1-\frac{2}{\tau}} t F) \right] / \exp(s^{1-\frac{2}{\tau}} t \epsilon) \leq \exp \left\{ \left( 1 - \frac{P}{2} \right) \left( \frac{P}{2} \right)^{\frac{2}{\tau} - 1} s^{1} + s^{1-\frac{2}{\tau}} n \Phi \left( \frac{t}{n}; F \right) \right\}.
\]
Mapping $t/n$ to $t$ and minimizing over $s, t > 0$ yields
\[
P_\otimes \{F(z^n) > \epsilon\} \leq \exp \left\{ n \inf_{s, t > 0} \left\{ \left( 1 - \frac{P}{2} \right) \left( \frac{P}{2} \right)^{\frac{2}{\tau} - 1} s^{1} + s^{1-\frac{2}{\tau}} \sup_{t > 0} \{ t \epsilon - \Phi(t; F) \} \right\} \right\} = \exp \left\{ - n \mathcal{J}(\epsilon; F)^2 / \tau \right\}.
\]
Setting $\rho = \sqrt{s \frac{t}{n}}$, $\epsilon = \mathcal{R}(\rho; F)$ and applying Proposition 1 yields the second part of the result, provided that $F$ is bounded from above.

To deal with an unbounded $F$, define $F_k = F \land k$ for $k \in \mathbb{N}_{\geq 1}$. We have proved that the result holds for $F_k$. Observe that for all $z^n, \tilde{z}^n \in \mathcal{Z}$ with $F(z^n) \geq F(\tilde{z}^n)$, it holds that
\[
(F(\tilde{z}^n) \land k) - (F(z^n) \land k) = \begin{cases} 0, & F(z^n) \geq k, \\ k - F(z^n), & F(z^n) < k < F(\tilde{z}), \\ F(\tilde{z}^n) - F(z^n), & F(\tilde{z}) \leq k. \\ \end{cases}
\]
Hence for all $t > 0$ and $k \geq 1$, $\Phi(t; F_k) \leq \Phi(t; F)$, and thus $\mathcal{J}(\epsilon; F_k) \geq \mathcal{J}(\epsilon; F)$. Therefore, by the monotone convergence,
\[
P_\otimes \{F(z^n) > \epsilon\} = \lim_{k \to \infty} P_\otimes \{F_k(z^n) > \epsilon\} \leq \lim_{k \to \infty} \exp \left\{ - n \mathcal{J}(\epsilon; F_k)^2 / \tau \right\} \leq \exp \left\{ - n \mathcal{J}(\epsilon; F)^2 / \tau \right\},
\]
which completes the proof. \qed
Proof of Theorem 1. Set \( F(z^n) = \mathbb{E}_{P_{\text{true}}} [f(z)] - \mathbb{E}_{P_n} [f] \). Then \( F \) satisfies the assumptions in Lemma 5 due to Assumptions 1(I) and 2. Applying Lemma 5 yields that

\[
\Phi(t; F) = \frac{1}{n} \mathbb{E}_\Theta \left[ \sup_{z^n \in Z^n} \left\{ \sum_{i=1}^n \left( t \left( f(z^n_i) - f(\tilde{z}^n_i) \right) - \|\tilde{z}^n_i - z^n_i\| \right) \right\} \right] = \mathbb{E}_{P_{\text{true}}} \left[ \sup_{\tilde{z} \in \mathcal{Z}} \left\{ -t (f(\tilde{z}) - f(z)) - \|\tilde{z} - z\| \right\} \right],
\]

and thus \( \mathcal{F}(\cdot; F) = \mathcal{I}_p(\cdot; -f) \) and \( \mathcal{R}(\cdot; F) = \mathcal{R}_{P_{\text{true}}, p}(\cdot; -f) \), therefore the result follows. \( \square \)

Appendix B: Proofs for Section 4

B.1. Proof for Section 4.1

Proof of Corollary 2. (I) By Assumption 3, it holds that

\[
\mathbb{E}_{P_{\text{true}}} [f_\Theta] - \mathbb{E}_{P_{\text{true}}} [f_\Theta] \leq \mathbb{E}_{P_{\text{true}}} [\kappa] \cdot \|\tilde{\theta} - \theta\|, \\
\mathbb{E}_{P_n} [f_\Theta] - \mathbb{E}_{P_n} [f_\Theta] \leq \mathbb{E}_{P_n} [\kappa] \cdot \|\tilde{\theta} - \theta\|.
\]

By the assumption on \( \kappa \), there exists \( C > 0 \) such that \( \mathbb{P}_\Theta \{ \mathbb{E}_{P_n} [\kappa] > 2\mathbb{E}_{P_{\text{true}}} [\kappa] \} \leq \exp(-Cn) \). Let \( \varepsilon, \delta > 0 \) and let \( \Theta_\delta \) be an \( \varepsilon \)-cover of \( \Theta \). We have that

\[
\mathbb{P}_\Theta \left\{ \exists \theta \in \Theta, s.t. \mathbb{E}_{P_{\text{true}}} [f_\Theta] > \mathbb{E}_{P_n} [f_\Theta] + \varepsilon + 3\varepsilon \mathbb{E}_{P_{\text{true}}} [\kappa] \right\} \\
\leq \exp(-Cn) + \mathbb{P}_\Theta \left\{ \exists \tilde{\theta} \in \Theta_\delta, s.t. \mathbb{E}_{P_{\text{true}}} [f_{\tilde{\theta}}] > \mathbb{E}_{P_n} [f_\Theta] + \varepsilon \right\} \\
\leq \exp(-Cn) + \sum_{\tilde{\theta} \in \Theta_\delta} \mathbb{P}_\Theta \left\{ \mathbb{E}_{P_{\text{true}}} [f_{\tilde{\theta}}] > \mathbb{E}_{P_n} [f_\Theta] + \varepsilon \right\} \\
\leq \exp(-Cn) + \mathcal{N}(\varepsilon; \Theta, \|\cdot\|) \cdot \exp \left( -n\mathcal{I}_p(\varepsilon; -f)^2/\tau \right).
\]

Letting \( \varepsilon = 1/n \) and using Theorem 1 yields the result.

(II)(III) are simple consequences of (I), together with Lemma 1 and Lemma 2. \( \square \)

Proof of Corollary 3. Fix \( f \in \mathcal{F} \). Applying Bennett’s inequality (Lemma 6 below) to \( X_t = -\frac{\|\nabla f(z^n_i)\|^2}{\|\nabla f\|^2\|P_{\text{true}}\|^2} \), \( b = 0 \) and \( v_i = \sigma^2 \), we obtain that

\[
\mathbb{P} \left\{ \mathbb{E}_{P_n} \left[ \frac{\|\nabla f\|^2}{\|\nabla f\|^2\|P_{\text{true}}\|^2} \right] - 1 < -\varepsilon \right\} \leq \exp \left( -\frac{n\varepsilon^2}{2\sigma^2} \right).
\]

Hence, with probability at least \( 1 - e^{-t} \),

\[
\frac{\|\nabla f\|^2\|P_{\text{true}}\|^2}{\|\nabla f\|^2\|P_{\text{true}}\|^2} \geq 1 - \sigma \sqrt{\frac{2t}{n}}.
\]

Thus, for every \( n > 8\sigma^2 t \),

\[
\|\nabla f\|^2\|P_{\text{true}}\|^2 \leq \|\nabla f\|^2\|P_{\text{true}}\|^2 \left( 1 - \sigma \sqrt{\frac{2t}{n}} \right)^{-\frac{1}{2}} \leq \|\nabla f\|^2\|P_{\text{true}}\|^2 \left( 1 + \sigma \sqrt{\frac{2t}{n}} \right),
\]

where the second inequality follows from the simple fact \( 1/\sqrt{1-a} \leq 1 + a \) for \( a \in [0, 1/2] \). Setting \( \rho_n = \sqrt{\frac{2t}{n}} \), by Corollary 1, with probability at least \( 1 - 2e^{-t} \),

\[
\mathbb{E}_{P_{\text{true}}} [f] \leq \mathbb{E}_{P_n} [f] + \rho_n \|\nabla f\|^2\|P_{\text{true}}\|^2 + \frac{ht}{\rho_n^2},
\]

Setting \( \rho_n = \sqrt{\frac{2t}{n}} \), by Corollary 1, with probability at least \( 1 - 2e^{-t} \),

\[
\mathbb{E}_{P_{\text{true}}} [f] \leq \mathbb{E}_{P_n} [f] + \rho_n \|\nabla f\|^2\|P_{\text{true}}\|^2 + \frac{ht}{\rho_n^2},
\]
and
\[ E_{\text{true}} \{ f \} \leq E_{P_n} \{ f \} + R_{P_n,2}(\rho_n; f) + \frac{2htt}{n}. \]

Next we consider a family of losses. By Assumptions 3 and 4, it holds that
\[ E_{P_n} \{ f_\theta \} - E_{\text{true}} \{ f_\theta \} \leq k \cdot ||\bar{\theta} - \theta||_\Theta, \]
\[ E_{P_n} \{ f_\theta \} - E_{P_n} \{ f_\eta \} \leq E_{P_n} \{ k \} \cdot ||\bar{\theta} - \theta||_\Theta, \]
\[ ||\nabla f_\theta||_p \cdot ||P_{n,2} - ||\nabla f_\eta||_p \cdot ||P_{n,2} \leq ||k||_2 \cdot ||P_{n,2} \cdot ||\tilde{\theta} - \theta||_\Theta. \]

By the assumption on \( k \), there exists \( C > 0 \) such that \( P_\oplus \{ E_{P_n} \{ k \} > 2E_{\text{true}} \{ k \} \} \leq \exp(-Cn) \) and \( P_\oplus \{ E_{P_n} \{ k^2 \} > 4E_{\text{true}} \{ k^2 \} \} \leq \exp(-Cn) \). Let \( \epsilon > 0 \) and let \( \Theta_\epsilon \) be an \( \epsilon \)-cover of \( \Theta \). We have that
\[ \begin{align*}
    &P_\oplus \left\{ \exists \theta \in \Theta, s.t. E_{P_{n,2}} \{ f_\theta \} > E_{P_n} \{ f_\theta \} + \rho_n ||\nabla f_\theta||_p \cdot ||P_{n,2} \right. + \frac{hrtt}{n} + (3E_{E_{\text{true}} \{ k \} + 2\rho_n ||k||_2 ||P_{n,2})\epsilon} \\
    &\leq 2 \exp(-Cn) + P_\oplus \left\{ \exists \tilde{\theta} \in \Theta_\epsilon, s.t. E_{P_{n,2}} \{ f_\tilde{\theta} \} > E_{P_n} \{ f_\tilde{\theta} \} + \rho_n ||\nabla f_\tilde{\theta}||_p \cdot ||P_{n,2} + \frac{hrtt}{n} \right. \\
    &\leq 2 \exp(-Cn) + 2\mathcal{N}(\epsilon, \Theta, ||\cdot||_\Theta) \cdot e^{-t}.
\end{align*} \]

Hence the proof is completed by setting \( \epsilon = 1/n \) and invoking Lemma 2 for the second part. \( \square \)

**Lemma 6 (Bennett’s inequality).** Suppose \( X_1, \ldots, X_n \) are independent random variables for which \( X_i \leq b \) and \( E[X_i^2] \leq v_i \) for each \( i \), for nonnegative constants \( b \) and \( v_i \). Let \( W = \sum_{i=1}^{n} v_i \). Then for \( \epsilon \geq 0 \),
\[ P\left\{ \sum_{i=1}^{n} (X_i - E[X_i]) \geq \epsilon \right\} \leq \exp \left( -\frac{\epsilon^2}{2W} \psi_{\text{Benn}} \left( \frac{bx}{W} \right) \right), \]
where \( \psi_{\text{Benn}} \) denotes the function defined on \( [-1, \infty) \) by
\[ \psi_{\text{Benn}}(t) := \frac{(1 + t) \log(1 + t) - t}{t^2/2}, \quad \text{for } t \neq 0, \text{ and } \psi_{\text{Benn}}(0) = 1. \]

**B.2. Proofs for Section 4.2**

**B.2.1. Auxiliary Results** We prepare some auxiliary results that will be used shortly. The following two lemmas are useful properties of Rademacher complexity (see, e.g., [74, Chapter 26]).

**Lemma 7 (Symmetrization).** Let \( \mathcal{H} \) be a family of functions. Then
\[ E_\oplus \left( \sup_{h \in \mathcal{H}} \{ E_{\text{true}} \{ h \} - E_{P_n} \{ h \} \} \right) \leq 2E_\oplus [\mathcal{R}_n(\mathcal{H})]. \]

**Lemma 8 (Contraction).** Let \( \mathcal{H} \) be a family of functions. Let \( \ell : \mathbb{R} \rightarrow \mathbb{R} \) be a Lipschitz function. Denote \( \ell \circ \mathcal{H} = \{ \ell \circ h : h \in \mathcal{H} \} \). Then
\[ E_\oplus [\mathcal{R}_n(\ell \circ \mathcal{H})] \leq ||\ell||_{\text{Lip}} \cdot E_\oplus [\mathcal{R}_n(\mathcal{H})]. \]

Let us define
\[ \mathcal{R}_{\oplus, p}(\rho; \mathcal{F}) := \min_{\lambda \geq 0} \left\{ \lambda \rho^p + E_\oplus \left( \sup_{f \in \mathcal{F}, z^n \in \mathbb{Z}^n} \frac{1}{n} \sum_{i=1}^{n} \left( f(z^n) - f(z^n_i) - \lambda \|z^n - z^n_i\|^p \right) \right) \right\}, \]
and
\[ -\mathcal{F} := \{ -f : f \in \mathcal{F} \}; \]

**Lemma 9.** Let \( p \in 1, 2 \). Assume Assumption 1(I) holds when \( p = 1 \) and Assumptions 2 when \( p = 2 \). Assume \( P_{\text{true}} \) satisfies \( T_p(\tau) \). Let \( t > 0 \). Then with probability at least \( 1 - e^{-t} \), for every \( f \in \mathcal{F} \),
\[ E_{\text{true}} \{ f \} \leq E_{P_n} \{ f \} + \mathcal{R}_{\oplus, p}(\sqrt{\frac{rt}{n}}; -\mathcal{F}) + 2E_\oplus [\mathcal{R}_n(\mathcal{F})]. \]
Proof. Set

\[ F(z^n) = \sup_{f \in \mathcal{F}} \left\{ \mathbb{E}_{\mathbb{P}_{\text{true}}} [f] - \mathbb{E}_{\mathbb{P}_n} [f] \right\} - \mathbb{E}_{\mathbb{Q}} \left[ \sup_{f \in \mathcal{F}} \left\{ \mathbb{E}_{\mathbb{P}_{\text{true}}} [f] - \mathbb{E}_{\mathbb{P}_n} [f] \right\} \right]. \]

Then the assumption on \( f \) implies that \( F \) satisfies the growth assumptions in Lemma 5. Applying Lemma 5 yields that with probability at least \( 1 - e^{-t} \), for every \( f \in \mathcal{F} \),

\[
\mathbb{E}_{\mathbb{P}_{\text{true}}} [f] - \mathbb{E}_{\mathbb{P}_n} [f] - \mathbb{E}_{\mathbb{Q}} \left[ \sup_{f \in \mathcal{F}} \left\{ \mathbb{E}_{\mathbb{P}_{\text{true}}} [f] - \mathbb{E}_{\mathbb{P}_n} [f] \right\} \right] \\
\leq \inf_{\lambda > 0} \left\{ \lambda \left( \frac{\sqrt{t}}{n} \right)^p + \mathbb{E}_{\mathbb{Q}} \left[ \sup_{\bar{z}^n \in \mathbb{Z}^n} \left\{ F(\bar{z}^n) - F(z^n) - \frac{\lambda}{n} d_\rho(\bar{z}^n, z^n)^p \right\} \right] \right\} \\
\leq \inf_{\lambda > 0} \left\{ \lambda \left( \frac{\sqrt{t}}{n} \right)^p + \mathbb{E}_{\mathbb{Q}} \left[ \sup_{\bar{z}^n \in \mathbb{Z}^n} \sup_{f \in \mathcal{F}} \left\{ \frac{1}{n} \sum_{i=1}^n f(\bar{z}^n_i) + f(z^n_i) - \lambda \|\bar{z}^n_i - z^n_i\|^p \right\} \right] \right\} \\
= \mathcal{R}_{\mathbb{Q},p} \left( \sqrt{\frac{t}{n}}; -\mathcal{F} \right).
\]

Thus the result then follows by applying Lemma 7. \( \square \)

The next lemma is used for bounding the fixed point of \( \psi_n \) for local Rademacher complexity.

Lemma 10. Let \( A, B > 0 \) and \( p > 1 \). Let \( r_0 \) be the largest solution to the equation \( Br^\frac{1}{q} + A = r \). Then \( r_0 \leq \frac{q}{q-1} A + B^p \).

Proof. Set \( h(r) := r - Br^\frac{1}{q} - A \). Let \( r_0 = B^p + \frac{q}{q-1} A \). The convexity of \( t \mapsto t^q \) implies that

\[
(r_1 - A)^q = (B^p + \frac{1}{q-1} A)^q \geq B^{pq} + \frac{q}{q-1} AB^{p(q-1)} = B^{pq} + \frac{q}{q-1} AB^q = B^q r_1.
\]

Since \( h \) is convex and \( h(B^p) < 0 \), the inequality above implies that \( h(r) > 0 \) for any \( r > r_0 \). \( \square \)

B.2.2. Proofs for \( p = 1 \) To begin with, an application of Lemma 5 to the loss function \( F = \sup_{f \in \mathcal{F}} \{ \mathbb{P}_{\text{true}} [f] - \mathbb{P}_n [f] \} \) yields the following result.

Lemma 11. Assume \( \mathbb{P}_{\text{true}} \) satisfies \( T_1(\tau) \) and Assumption 1(I) holds. Let \( t > 0 \). Then with probability at least \( 1 - e^{-t} \),

\[
\mathbb{E}_{\mathbb{P}_{\text{true}}} [f] \leq \mathbb{E}_{\mathbb{P}_n} [f] + \sqrt{\frac{t}{n}} \sup_{f \in \mathcal{F}} \|f\|_{\text{Lip}} + 2 \mathbb{E}_{\mathbb{Q}} \left[ \mathcal{R}_n (\mathcal{F}) \right], \quad \forall f \in \mathcal{F}.
\]

Proof of Lemma 11. In view of Lemma 9, it suffices to derive an upper bound on \( \mathcal{R}_{\mathbb{Q},1}(\rho; -\mathcal{F}) \). Assumption 1(I) implies that for any \( \lambda > \sup_{f \in \mathcal{F}} \|f\|_{\text{Lip}} \),

\[
\sup_{f \in (-\mathcal{F}), \bar{z}^n \in \mathbb{Z}^n} \left\{ \frac{1}{n} \sum_{i=1}^n f(\bar{z}^n_i) - f(z^n_i) - \lambda \|\bar{z}^n_i - z^n_i\| \right\} = 0.
\]

Consequently by definition \( \mathcal{R}_{\mathbb{Q},1}(\rho) \leq \rho \sup_{f \in \mathcal{F}} \|f\|_{\text{Lip}} \).

Next, using the peeling technique [82, 40], we can remove the dependence on \( \sup_{f \in \mathcal{F}} \) of the right side of the inequality in Lemma 11.

Lemma 12. Assume \( \mathbb{P}_{\text{true}} \) satisfies \( T_1(\tau) \) and Assumption 1(I) holds. Let \( t > 0 \). Then with probability at least \( 1 - [\log_2(\sqrt{2}n)]/e^{-t} \),

\[
\mathbb{E}_{\mathbb{P}_{\text{true}}} [f] \leq \mathbb{E}_{\mathbb{P}_n} [f] + 2 \sqrt{\frac{t}{n}} \|f\|_{\text{Lip}} + 2 \mathbb{E}_{\mathbb{Q}} \left[ \mathcal{R}_n (\mathcal{F}) \right] + \frac{1}{n}, \quad \forall f \in \mathcal{F}.
\]
**Proof of Lemma 12.** Set \( r = \sup_{f \in \mathcal{F}} ||f||_{\text{Lip}} \leq \kappa_1 \). Let \( K \) be a positive integer whose value will be specified shortly. We define
\[
\mathcal{F}_k := \{ f \in \mathcal{F} : 2^{-k} r < ||f||_{\text{Lip}} \leq 2^{-k+1} r, \quad 1 \leq k \leq K - 1, \}
\]
\[
\mathcal{F}_K := \{ f \in \mathcal{F} : ||f||_{\text{Lip}} \leq 2^{-K} r. \}
\]
Using Lemma 11, for \( k = 1, \ldots, K - 1 \), with probability at least \( 1 - e^{-t} \), for every \( f \in \mathcal{F}_k \),
\[
E_{P_{\text{true}}} [f] - E_{P_n} [f] \leq \sqrt{\frac{t}{n}} 2^{-k+1} r + 2E_\emptyset [\mathcal{R}_n(\mathcal{F}_k)] \leq 2 \sqrt{\frac{t}{n}} ||f||_{\text{Lip}} + 2E_\emptyset [\mathcal{R}_n(\mathcal{F}_k)].
\]
and with probability at least \( 1 - e^{-t} \), for every \( f \in \mathcal{F}_K \),
\[
E_{P_{\text{true}}} [f] - E_{P_n} [f] \leq \sqrt{\frac{t}{n}} 2^{-K} r + 2E_\emptyset [\mathcal{R}_n(\mathcal{F}_K)].
\]
Taking the union bound, with probability at least \( 1 - K e^{-t} \), for every \( f \in \mathcal{F} \),
\[
E_{P_{\text{true}}} [f] \leq E_{P_n} [f] + 2 \sqrt{\frac{t}{n}} ||f||_{\text{Lip}} + 2E_\emptyset [\mathcal{R}_n(\mathcal{F})] + \sqrt{\frac{t}{n}} 2^{-K} r.
\]
Note that \( r \leq \kappa_1 \) by Assumption 1(I). Setting \( K = \lceil \log_2(\kappa_1 \sqrt{n}) \rceil \) yields the result. \( \Box \)

Typically, \( E_\emptyset [\mathcal{R}_n(\mathcal{F})] \) is of the order of \( 1/\sqrt{n} \). By applying Lemma 12 and another peeling argument to a weighted function class \( \{ \sqrt{r} f : f \in \mathcal{F} \} \) and using the sub-root property of \( \psi_n \), we can replace \( E_\emptyset [\mathcal{R}_n(\mathcal{F})] \) with the fixed point \( r_{n*} \) of \( \psi_n \), often in the order of \( 1/n \).

**Proof of Theorem 2.** Let \( r \geq r_{n*} \) whose value will be specified shortly. The sub-root assumption on \( \psi_n \) implies that
\[
\psi_n(r) = \frac{\sqrt{r} \psi_n(r)}{\sqrt{r}} \leq \frac{\sqrt{r} \psi_n(r_{n*})}{\sqrt{r_{n*}}} = \sqrt{r} r_{n*}.
\]
Define
\[
\mathcal{F}_r := \left\{ \frac{\sqrt{r}}{\sqrt{r} \vee ||f||_{\text{Lip}}} f : f \in \mathcal{F} \right\}.
\]
Then \( \mathcal{F}_{k_1} = \mathcal{F} \), \( ||g||_{\text{Lip}} \leq r \) for all \( g \in \mathcal{F}_r \), thus
\[
E_\emptyset [\mathcal{R}_n(\mathcal{F}_r)] \leq E_\emptyset [\mathcal{R} \left( \{ cf : f \in \mathcal{F}, 0 \leq c \leq 1, c^2 ||f||_{\text{Lip}}^2 \leq r \} \right) ] \leq \psi_n(r) \leq \sqrt{r} r_{n*}.
\]
By Lemma 12, with probability at least \( 1 - \lceil \log_2(r \sqrt{n}) \rceil e^{-t} \), for every \( g \in \mathcal{F}_r \),
\[
E_{P_{\text{true}}} [g] \leq E_{P_n} [g] + 2 \sqrt{\frac{t}{n}} ||g||_{\text{Lip}} + 2 \sqrt{r} r_{n*} + \frac{1}{n},
\]
Choose \( r = r_0 \), where \( r_0 \) is the largest solution to \( \frac{1}{n} + 2 \sqrt{r} r_{n*} = r \). By Lemma 10, \( r_{n*} \leq r_0 \leq 4r_{n*} + \frac{2}{n} \).
Let \( \mathcal{F}_{r_0} \ni g = \frac{\sqrt{r_0}}{\sqrt{r_0} \vee ||f||_{\text{Lip}}} f \). If \( ||f||_{\text{Lip}}^2 \leq r_0 \), then \( g = f \), therefore
\[
E_{P_{\text{true}}} [f] \leq E_{P_n} [f] + 2 \sqrt{\frac{t}{n}} ||f||_{\text{Lip}} + 4r_{n*} + \frac{2}{n}.
\]
If \( ||f||_{\text{Lip}}^2 > r_0 \), then
\[
E_{P_{\text{true}}} \left[ \frac{\sqrt{r_0}}{||f||_{\text{Lip}}} f \right] \leq \frac{1}{n} \sum_{i=1}^{n} \sqrt{r_0} f(z_i^n) + 2 \sqrt{\frac{t}{n} \sqrt{r_0}} ||f||_{\text{Lip}} + 4r_{n*} + \frac{2}{n}.
\]
therefore,
\[
E_{P_{\text{true}}} [f] \leq E_{P_n} [f] + 2\sqrt{\frac{t}{n}} \|f\|_{\text{Lip}} + \frac{4r_{n*}^2 + \frac{\sqrt{n}}{\sqrt{t}}}{n} \|f\|_{\text{Lip}}
\]
\[
\leq E_{P_n} [f] + \left(2\sqrt{\frac{t}{n}} + 4\sqrt{r_{n*}} + \frac{2}{n\sqrt{r_{n*}}}\right) \|f\|_{\text{Lip}}.
\]

Combining the two cases gives the result. □

**Proof of Corollary 5.** Define \( \ell \circ \mathcal{F} := \{\ell \circ f : f \in \mathcal{F}\} \). Using Lemma 9, with probability at least \( 1 - e^{-t} \), for every \( f \in \mathcal{F} \),

\[
E_{P_{\text{true}}} [\ell \circ f] \leq E_{P_n} [\ell \circ f] + \mathcal{R}_{\ell,1} \left(\sqrt{\frac{t}{n}}; \ell \circ \mathcal{F}\right) + 2E_\mathcal{F} [\mathcal{R}_n (\ell \circ \mathcal{F})]
\]
\[
\leq E_{P_n} [\ell \circ f] + \mathcal{R}_{\ell,1} \left(\sqrt{\frac{t}{n}}; \ell \circ \mathcal{F}\right) + 2L_{\ell} \mathcal{R}_n (\ell \circ \mathcal{F}),
\]

where we have used Lemma 8 to obtain the second inequality. Using the arguments similar to the proofs of Lemma 11 and Lemma 12, we obtain that with probability at least \( 1 - [\log_2 (\sqrt{t}nL_{\ell}k_1)]e^{-t} \),

\[
E_{P_{\text{true}}} [\ell \circ f] \leq E_{P_n} [\ell \circ f] + \sqrt{\frac{t}{n}} \sup_{f \in \mathcal{F}} \|\ell \circ f\|_{\text{Lip}} + 2E_\mathcal{F} [\mathcal{R}_n (\ell \circ \mathcal{F})]
\]
\[
\leq E_{P_n} [\ell \circ f] + 2\sqrt{\frac{t}{n}} \|\ell \circ f\|_{\text{Lip}} + 2\mathcal{R}_n (\ell \circ \mathcal{F}) + \frac{L_{\ell}}{n}
\]

(6)

Define for any \( r > 0 \) that

\[
\mathcal{F}_r := \left\{\sqrt{\frac{r}{\mathcal{R}_n}} \ell \circ f : f \in \mathcal{F}\right\} \subset \{c\ell \circ f : f \in \mathcal{F}, 0 < c \leq 1, c^2 \|f\|_{\text{Lip}}^2 \leq r\}.
\]

Substituting \( \frac{\sqrt{r}}{\mathcal{R}_n} \ell \) for \( \ell \) in (6), we obtain that with probability at least \( 1 - [\log_2 (\sqrt{t}nL_{\ell}k_1)]e^{-t} \), for every \( \mathcal{F}_r \ni g = \frac{\sqrt{r}}{\mathcal{R}_n} \ell \circ f \),

\[
E_{P_{\text{true}}} [g] \leq E_{P_n} [g] + 2\sqrt{\frac{t}{n}} \frac{\sqrt{r}}{\mathcal{R}_n} L_{\ell} \|f\|_{\text{Lip}} + 2\frac{\sqrt{r}}{\mathcal{R}_n} \|f\|_{\text{Lip}} L_{\ell} \mathcal{R}_n (\ell \circ f) + \frac{L_{\ell}}{n}.
\]

Choose \( r \) to be the largest solution \( r_0 \) to the equation \( r = 2L_{\ell} \mathcal{R}_n \mathcal{R}_n + \frac{L_{\ell}}{n} \). It follows from Lemma 10 that \( r_{n*} \leq r_0 \leq \frac{2L_{\ell}}{n} + 4L_{\ell}^2 r_{n*} \). When \( \|f\|_{\text{Lip}} \leq \mathcal{R}_0 \), we have \( g = \ell \circ f \) and

\[
E_{P_{\text{true}}} [\ell \circ f] \leq E_{P_n} [\ell \circ f] + 2\sqrt{\frac{t}{n}} L_{\ell} \|f\|_{\text{Lip}} + 4L_{\ell}^2 r_{n*} + \frac{2L_{\ell}}{n}.
\]

When \( \|f\|_{\text{Lip}} > \mathcal{R}_0 \), we have \( g = \frac{\sqrt{\mathcal{R}_0}}{\mathcal{R}_n} \ell \circ f \) and

\[
E_{P_{\text{true}}} [\ell \circ f] \leq E_{P_n} [\ell \circ f] + 2\sqrt{\frac{t}{n}} L_{\ell} \|f\|_{\text{Lip}} + (4L_{\ell}^2 r_{n*} + \frac{2L_{\ell}}{n}) \|f\|_{\text{Lip}} \leq \mathcal{R}_0
\]
\[
\leq E_{P_n} [\ell \circ f] + \left(2\sqrt{\frac{t}{n}} L_{\ell} + 4L_{\ell}^2 r_{n*} + \frac{2L_{\ell}}{n\sqrt{r_{n*}}}\right) \|f\|_{\text{Lip}}.
\]

Combining the two cases yields the result. □
B.2.3. Proofs for $p = 2$ \text{ Lemma 13 and 14 below are counterparts of Lemma 11 and Lemma 12.}

**Lemma 13.** Assume $P_{\text{true}}$ satisfies $T_2(\tau)$ and Assumptions 2 holds. Let $t > 0$. Then with probability at least $1 - e^{-t}$,

$$E_{P_{\text{true}}} [f] \leq E_{P_n} [f] + \sqrt{\frac{\pi t}{n}} (1 + 2E_{\mathcal{D}} [R_n(G)]) \sup_{f \in \mathcal{F}} \| \nabla f \|_\ast \| P_{\text{true}, 2} + 2E_{\mathcal{D}} [R_n(F)] + \frac{h \pi t}{n}. $$

**Proof of Lemma 13.** In view of Lemma 9, we derive an upper bound on $\mathcal{R}_{\mathcal{D}, 2}(\rho; -F)$. By Assumption 2, for all $f \in \mathcal{F}$,

$$\mathcal{R}_{\mathcal{D}, 2}(\rho; -F) = \inf_{I} \left\{ \lambda \rho^2 + E_{\mathcal{D}} \left[ \sup_{f \in (-F)} \sup_{z_n \in \mathbb{Z}^n} \left\{ \frac{1}{n} \sum_{i=1}^{n} \| \nabla f(z_i^\alpha) - f(z_i^\alpha) - \frac{\lambda}{n} \| \nabla f(z_i^\alpha) \|^2 \right\} \right] \right\}$$

$$\leq \sup_{I} \left\{ \lambda \rho^2 + E_{\mathcal{D}} \left[ \sup_{f \in (-F)} \sup_{z_n \in \mathbb{Z}^n} \left\{ \frac{1}{n} \sum_{i=1}^{n} \| \nabla f(z_i^\alpha) \|^2 \right\} \right] \right\}$$

$$\leq h \rho^2 + \inf_{\lambda \geq 0} \left\{ \lambda \rho^2 + \frac{1}{4\lambda} E_{\mathcal{D}} \left[ \sup_{f \in (-F)} \frac{1}{n} \sum_{i=1}^{n} \| \nabla f(z_i^\alpha) \|^2 \right] \right\}.$$ 

Observe that by Lemma 7,

$$E_{\mathcal{D}} \left[ \sup_{f \in (-F)} \frac{1}{n} \sum_{i=1}^{n} \| \nabla f(z_i^\alpha) \|^2 \right] \leq E_{\mathcal{D}} \left[ \sup_{f \in (-F)} \frac{1}{n} \sum_{i=1}^{n} \| \nabla f(z_i^\alpha) \|^2 - E_{P_{\text{true}}} \left[ \| \nabla f \|^2 \right] \right] + \sup_{f \in (-F)} E_{P_{\text{true}}} \left[ \| \nabla f \|^2 \right]$$

$$\leq 2E_{\mathcal{D}} [R_n(F)] + \sup_{f \in \mathcal{F}} E_{P_{\text{true}}} \left[ \| \nabla f \|^2 \right].$$

where $\hat{f} := \{ \| \nabla f \|^2 : f \in \mathcal{F} \}$. Hence we have

$$\mathcal{R}_{\mathcal{D}, 2}(\rho; -F) \leq h \rho^2 + \inf_{\lambda \geq 0} \left\{ \lambda \rho^2 + \frac{1}{4\lambda} \left( 2E_{\mathcal{D}} [R_n(F)] + \sup_{f \in \mathcal{F}} E_{P_{\text{true}}} \left[ \| \nabla f \|^2 \right] \right) \right\}. $$

Without loss of generality we assume $\sup_{f \in \mathcal{F}} \| \nabla f \|_\ast P_{\text{true}, 2} > 0$. Picking $\lambda = \frac{\sup_{f \in \mathcal{F}} \| \nabla f \|_\ast P_{\text{true}, 2}}{2\rho}$ yields that

$$\mathcal{R}_{\mathcal{D}, 2}(\rho; -F) \leq h \rho^2 + \rho \sup_{f \in \mathcal{F}} \| \nabla f \|_\ast P_{\text{true}, 2} + \rho \frac{E_{\mathcal{D}} [R_n(F)]}{\sup_{f \in \mathcal{F}} \| \nabla f \|_\ast P_{\text{true}, 2}}.$$ 

Note that

$$\frac{E_{\mathcal{D}} [R_n(F)]}{\sup_{f \in \mathcal{F}} \| \nabla f \|_\ast P_{\text{true}, 2}} \leq \sup_{f \in \mathcal{F}} \| \nabla f \|_\ast P_{\text{true}, 2} \cdot E_{\mathcal{D}} \left[ \sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \| \nabla f(z_i^\alpha) \|^2 - 1 \right]$$

$$\leq \sup_{f \in \mathcal{F}} \| \nabla f \|_\ast P_{\text{true}, 2} \cdot 2E_{\mathcal{D}} [R_n(G)].$$

Therefore, the result follows from Lemma 9 with $\rho = \sqrt{\frac{\pi t}{n}}$. \hfill $\Box$

**Lemma 14.** Assume $P_{\text{true}}$ satisfies $T_2(\tau)$ and Assumption 2 holds. Let $t > 0$. Then with probability at least $1 - [\log_2(\kappa_2 \pi t n)] e^{-t}$,

$$E_{P_{\text{true}}} [f] \leq E_{P_n} [f] + 2\sqrt{\frac{\pi t}{n}} (1 + 2E_{\mathcal{D}} [R_n(G)]) \| \nabla f \|_\ast P_{\text{true}, 2} + 2E_{\mathcal{D}} [R_n(F)] + \frac{h \pi t + 1 + 2E_{\mathcal{D}} [R_n(G)]}{n}. $$
**Proof of Lemma 14.** Set \( r = \sup_{f \in \mathcal{F}} \| \nabla f \|_* \| \mathcal{P}_{\text{true,2}} \leq k_2 \). We define
\[
\mathcal{F}_k := \left\{ f \in \mathcal{F} : 2^{-k} r < \| \mathcal{P}_{\text{true,2}} \leq 2^{-k+1} r \right\}, \quad k = 1, \ldots, K - 1,
\]
\[
\mathcal{F}_K := \left\{ f \in \mathcal{F} : \| \nabla f \|_* \| \mathcal{P}_{\text{true,2}} \leq 2^{-K} r \right\}.
\]
By Lemma 13, for \( k = 1, \ldots, K - 1 \), with probability at least \( 1 - e^{-t} \), for every \( f \in \mathcal{F}_k \),
\[
\mathbb{E}_{\mathcal{P}_{\text{true}}} [f] - \mathbb{E}_{\mathcal{P}_n} [f] \leq \sqrt{\frac{rt}{n} \left( 1 + 2 \mathbb{E}_{\mathcal{P}_n} [\mathcal{R}_n(G_k)] \right)} 2^{-k+1} r + 2 \mathbb{E}_{\mathcal{P}_n} [\mathcal{R}_n(\mathcal{F}_k)] + \frac{hrt}{n},
\]
and with probability at least \( 1 - e^{-t} \), for every \( f \in \mathcal{F}_K \),
\[
\mathbb{E}_{\mathcal{P}_{\text{true}}} [f] - \mathbb{E}_{\mathcal{P}_n} [f] \leq \sqrt{\frac{rt}{n} \left( 1 + 2 \mathbb{E}_{\mathcal{P}_n} [\mathcal{R}_n(G_K)] \right)} 2^{-K} r + 2 \mathbb{E}_{\mathcal{P}_n} [\mathcal{R}_n(\mathcal{F}_K)] + \frac{hrt}{n}.
\]
Taking the union bound, with probability at least \( 1 - Ke^{-t} \), for every \( f \in \mathcal{F} \),
\[
\mathbb{E}_{\mathcal{P}_{\text{true}}} [f] - \mathbb{E}_{\mathcal{P}_n} [f] \leq \sqrt{\frac{rt}{n} \left( 1 + 2 \mathbb{E}_{\mathcal{P}_n} [\mathcal{R}_n(G)] \right)} \| \nabla f \|_* \| \mathcal{P}_{\text{true,2}} + 2 \mathbb{E}_{\mathcal{P}_n} [\mathcal{R}_n(\mathcal{F})] + \frac{hrt}{n} + \frac{1 + 2 \mathbb{E}_{\mathcal{P}_n} [\mathcal{R}_n(G_k)]}{n} 2^{-K} r.
\]
Setting \( K = \lceil \log_2(k_2 \sqrt{nt}) \rceil \) yields the result.

With the two lemmas above, we are ready to prove Theorem 3.

**Proof of Theorem 3.** Let \( r \geq r_{n^*} \) whose value will be specified shortly. The sub-root assumption on \( \psi_n \) implies that
\[
\psi_n(r) = \frac{\sqrt{r} \psi_n(r)}{\sqrt{r}} \leq \frac{\sqrt{r} \psi_n(r_{n^*})}{\sqrt{r}} = \sqrt{r} \psi_n(r_{n^*}).
\]
Define
\[
\mathcal{F}_r := \left\{ \frac{\sqrt{r}}{\sqrt{r} \vee \| \nabla f \|_* \| \mathcal{P}_{\text{true,2}} f \in \mathcal{F} \right\}.
\]
Then \( \mathcal{F}_{\mathcal{F}_r} = \mathcal{F} \), for all \( g \in \mathcal{F} \) it holds that \( \| \| \nabla g \|_* \|^2_{\mathcal{P}_{\text{true,2}}} \leq r \), and
\[
\mathbb{E}_{\mathcal{P}_n} [\mathcal{R}_n(\mathcal{F}_r)] = \mathbb{E}_{\mathcal{P}_n} [\mathcal{R}_n(\{ c f : f \in \mathcal{F}, 0 \leq c \leq 1, c^2 \| \nabla f \|_* \|^2_{\mathcal{P}_{\text{true,2}}} \leq r \})] \leq \psi_n(r) \leq \sqrt{r} \psi_n(r_{n^*}).
\]
By Lemma 14, with probability at least \( 1 - \lceil \log_2(r \sqrt{nt}) \rceil e^{-t} \), for every \( g \in \mathcal{F}_r \),
\[
\mathbb{E}_{\mathcal{P}_{\text{true}}} [g] - \mathbb{E}_{\mathcal{P}_n} [g] \leq \sqrt{\frac{rt}{n} \left( 1 + 2 \mathbb{E}_{\mathcal{P}_n} [\mathcal{R}_n(G_r)] \right)} \| \nabla g \|_* \| \mathcal{P}_{\text{true,2}} + 2 \sqrt{r} \psi_n(r_{n^*}) + \frac{hrt + 1 + 2 \mathbb{E}_{\mathcal{P}_n} [\mathcal{R}_n(G_r)]}{n}.
\]
Choose \( r = r_0 \), where \( r_0 \) is the largest solution to \( \frac{1}{n} + 2 \sqrt{r} \psi_n(r_{n^*}) = r \). By Lemma 10, \( r_{n^*} \leq r_0 \leq 4 r_{n^*} + \frac{2}{n} \). Let \( G_0 \ni g = \frac{r_0}{\sqrt{r} \vee \| \nabla f \|_* \| \mathcal{P}_{\text{true,2}} f \in \mathcal{F} \}. If \( \| \nabla f \|_* \|^2_{\mathcal{P}_{\text{true,2}}} \leq r_0 \), then \( f = g \) and
\[
\mathbb{E}_{\mathcal{P}_{\text{true}}} [f] \leq \mathbb{E}_{\mathcal{P}_n} [f] + 2 \sqrt{\frac{rt}{n} \left( 1 + 2 \mathbb{E}_{\mathcal{P}_n} [\mathcal{R}_n(G)] \right)} \| \nabla f \|_* \| \mathcal{P}_{\text{true,2}} + 4 r_{n^*} \frac{2}{n} + \frac{hrt + 2 \mathbb{E}_{\mathcal{P}_n} [\mathcal{R}_n(G_r)]}{n}.
\]
If \( \|\|\nabla f\|\|_{\mathcal{P}_{\text{true},2}}^2 > r_0 \), then
\[
E_{\mathcal{P}_{\text{true}}} \left[ \frac{\|\|\nabla f\|\|_{\mathcal{P}_{\text{true},2}}^2}{\sqrt{n}} \right] 
\leq E_{\mathcal{P}_n} \left[ \frac{\|\|\nabla f\|\|_{\mathcal{P}_{\text{true},2}}^2}{\sqrt{n}} \right] + 2 \sqrt{\frac{t}{n}} \left( 1 + 2E_\Theta [\mathcal{R}_n(\mathcal{G})] \right) \frac{\|\|\nabla f\|\|_{\mathcal{P}_{\text{true},2}}^2}{\sqrt{n}} + 4r_{n*} + \frac{2}{n} + \frac{h \tau t + 2E_\Theta [\mathcal{R}_n(\mathcal{G})]}{n}.
\]
which implies that
\[
E_{\mathcal{P}_{\text{true}}}[f] 
\leq E_{\mathcal{P}_n}[f] + 2 \sqrt{\frac{t}{n}} \left( 1 + 2E_\Theta [\mathcal{R}_n(\mathcal{G})] \right) \frac{\|\|\nabla f\|\|_{\mathcal{P}_{\text{true},2}}^2}{\sqrt{n}} + 2r_{n*} + \frac{2}{n} + \frac{h \tau t + 2E_\Theta [\mathcal{R}_n(\mathcal{G})]}{n}.
\]
Combining the two cases above gives the desired result. \( \square \)

Proof of Corollary 6. Using McDiarmid’s inequality, with probability at least \( 1 - e^{-t} \), for every \( f \in \mathcal{F} \),
\[
\sup_{f \in \mathcal{F}} E_{\mathcal{P}_n} \left[ \frac{\|\|\nabla f(z)\|\|_{\mathcal{P}_{\text{true},2}}^2}{\|\|\nabla f\|\|_{\mathcal{P}_{\text{true},2}}^2} \right] - 1 \leq E_{\mathcal{P}_n} \left[ \sup_{f \in \mathcal{F}} E_{\mathcal{P}_n} \left[ \frac{\|\|\nabla f(z)\|\|_{\mathcal{P}_{\text{true},2}}^2}{\|\|\nabla f\|\|_{\mathcal{P}_{\text{true},2}}^2} \right] - 1 \right] + L^2 \sqrt{\frac{t}{2n}},
\]
which implies that
\[
\frac{\|\|\nabla f(z)\|\|_{\mathcal{P}_{\text{true},2}}^2}{\|\|\nabla f\|\|_{\mathcal{P}_{\text{true},2}}^2} - 1 \geq -2E_\Theta [\mathcal{R}_n(\mathcal{G})] - L^2 \sqrt{\frac{t}{2n}}.
\]
Thus, whenever \( 2E_\Theta [\mathcal{R}_n(\mathcal{G})] + L^2 \sqrt{\frac{t}{2n}} < 1/2 \), it holds that
\[
\|\|\nabla f\|\|_{\mathcal{P}_{\text{true},2}} \leq \|\|\nabla f\|\|_{\mathcal{P}_{\text{true}}} \left( 1 - 2E_\Theta [\mathcal{R}_n(\mathcal{G})] - L^2 \sqrt{\frac{t}{2n}} \right)^{-1} \leq \|\|\nabla f\|\|_{\mathcal{P}_{\text{true}}} \left( 1 + 2E_\Theta [\mathcal{R}_n(\mathcal{G})] + L^2 \sqrt{\frac{t}{2n}} \right).
\]
Hence, setting \( \tilde{\rho}_n = \rho_n \left( 1 + 2E_\Theta [\mathcal{R}_n(\mathcal{G})] + L^2 \sqrt{\frac{t}{2n}} \right) \) and invoking Theorem 3 and Corollary 1 yields the results. \( \square \)

Appendix C: Proofs for Section 5

C.1. Proofs for Section 5.1

Lemma 15. Under the setting in Example 1, it holds that
\[
\mathcal{R}_{\mathcal{P}_{\text{true}},1}(f_0; \rho_n) \leq \frac{h \vee b}{h \wedge b} \mathcal{R}_{\mathcal{P}_n,1}(f_0; \rho_n).
\]

Proof. By Lemma 1, we have \( \mathcal{R}_{\mathcal{P}_{\text{true}},1}(f_0; \rho_n) \leq \rho_n \|f_0\|_{\text{lip}} \leq \rho_n \max(h, b) \|\|(\theta, -1)\|\|. \) On the other hand, using the duality result (1), we have
\[
\mathcal{R}_{\mathcal{P}_n,1}(f_0; \rho_n) \geq \min_{\lambda \geq 0} \left\{ \lambda \rho_n + E_{\mathcal{P}_n} \left[ \sup_{(x, y) \in Z} \{ \min(h, b) |\theta^T x - y| - \lambda \| (x, y) - (X, Y) \| \} \right] \right\}
\geq \rho_n \min(h, b) \|\|(\theta, -1)\|\|.
\]
Combining the two inequalities yields that
\[
\mathcal{R}_{\mathcal{P}_n,1}(f_0; \rho_n) \geq \frac{h \wedge b}{h \vee b} \mathcal{R}_{\mathcal{P}_{\text{true}},1}(f_0; \rho_n).
\] \( \square \)
Lemma 16. Under the setting in Example 1, it holds that
\[ \mathcal{N}(\epsilon; F, \|\cdot\|_F) \leq \left( \frac{B \max(h, b) \mathbb{E}_{P_{true}}[\|x\|] \vee \mathbb{E}_{P_n}[\|x\|]}{\epsilon} \right)^d. \]

Proof. We have that
\[ \|f_\theta - f_\bar{\theta}\|_{P_{true}, 1} \leq \max(h, b) \mathbb{E}_{P_{true}}[(\bar{\theta} - \theta)^\top x] \leq \max(h, b) \mathbb{E}_{P_{true}}[\|x\|] \|\bar{\theta} - \theta\|, \]
and
\[ \|f_\theta - f_\bar{\theta}\|_{P_n, 1} \leq \max(h, b) \mathbb{E}_{P_n}[(\bar{\theta} - \theta)^\top x] \leq \max(h, b) \mathbb{E}_{P_n}[\|x\|] \|\bar{\theta} - \theta\|. \]

Therefore,
\[ \mathcal{N}(\epsilon; F, \|\cdot\|_F) \leq \mathcal{N}\left( \frac{\epsilon}{\max(h, b) \mathbb{E}_{P_{true}}[\|x\|] \vee \mathbb{E}_{P_n}[\|x\|]}; \Theta, \|\cdot\|_r \right) \leq \left( \frac{B \max(h, b) \mathbb{E}_{P_{true}}[\|x\|] \vee \mathbb{E}_{P_n}[\|x\|]}{\epsilon} \right)^d. \]

C.2. Proofs for Section 5.2

The following lemma is used for Example 2 (see also [74, Lemma 26.10]).

Lemma 17. Assume \( \Theta \subset \{\theta : \|\theta\|_2 \leq B\}. \) Then
\[ \mathbb{E}_{\Theta}[\mathbb{R}_n(\{\theta^\top x : \theta \in \Theta\})] \leq \frac{B}{\sqrt{n}} \mathbb{E}_{P_{true}}[\|x\|_2^2]^{1/2}. \]

Proof. Let \( \sigma_i \) be i.i.d. Rademacher random variables. Using Cauchy-Schwarz inequality and Jensen’s inequality,
\[ \mathbb{E}_{\sigma, \Theta} \left[ \sup_{\theta \in \Theta} \frac{1}{n} \sum_{i=1}^n \sigma_i \langle \theta, x_i^n \rangle \right] = \frac{1}{n} \mathbb{E}_{\sigma, \Theta} \left[ \sup_{\theta \in \Theta} \left( \sum_{i=1}^n \sigma_i x_i^n \right) \right] \leq \frac{1}{n} \mathbb{E}_{\sigma, \Theta} \left[ \sup_{\|\theta\|_2 \leq B} \|\sum_{i=1}^n \sigma_i x_i^n \|_2 \right] \leq \frac{B}{n} \mathbb{E}_{\sigma, \Theta} \left[ \|\sum_{i=1}^n \sigma_i x_i^n \|_2 \right]^{1/2} \leq \frac{B}{n} \left( \mathbb{E}_{\Theta} \left[ \sum_{i=1}^n \|x_i^n\|_2^2 \right] \right)^{1/2} = \frac{B}{\sqrt{n}} \mathbb{E}_{P_{true}}[\|x\|_2^2]^{1/2}. \]

The following lemma is used for Example 3.

Lemma 18. Under the setting in Example 3, it holds that
\[ \|\nabla f_\theta(z)\|_* - \|\nabla f_\bar{\theta}(z)\|_* \leq L_\epsilon \|\bar{\theta} - \theta\|_* + Bh \|x\|_* \|\bar{\theta} - \theta\|_. \]

Proof. We have
\[ \|\nabla f_\theta(z)\|_* - \|\nabla f_\bar{\theta}(z)\|_* = \|\theta\|_* |\ell'((\theta^\top x, y)) - |\bar{\theta}\|_* |\ell'(\bar{\theta}^\top x, y)| \]
\[ = \|\theta\|_* |\ell'((\theta^\top x, y)) - |\bar{\theta}\|_* |\ell'(\bar{\theta}^\top x, y)| + \|\bar{\theta}\|_* |\ell'(\bar{\theta}^\top x, y))| - |\bar{\theta}\|_* |\ell'(\bar{\theta}^\top x, y)| \]
\[ \leq L_\epsilon \|\bar{\theta} - \theta\|_* + \|\bar{\theta}\|_* |\ell'(\bar{\theta}^\top x, y)) - \ell'(\bar{\theta}^\top x, y)| \]
\[ \leq L_\epsilon \|\bar{\theta} - \theta\|_* + Bh \|x\|_* \|\bar{\theta} - \theta\|_. \]
C.3. Proofs for Section 5.3
The following two results on subgaussian distributions are well-known, but for the reader’s convenience, we here provide proofs as well. Recall a $d$-dimensional random variable $X$ is $\sigma$-subgaussian, if for every $w \in \mathbb{R}^d$ with $\|w\|_2 = 1$ and $\epsilon > 0$, $\mathbb{P}\{|w^\top X - \mathbb{E}[w^\top X]| \geq \epsilon\} \leq 2e^{-\frac{\epsilon^2}{2\sigma^2}}$.

**Lemma 19.** Let $X_1, \ldots, X_n$ be independent samples from a one-dimensional $\sigma$-subgaussian distribution. Then
\[
\mathbb{P}_\sigma \left\{ \frac{1}{n} \sum_{i=1}^n X_i^2 - \mathbb{E}[X^2] \geq \epsilon \right\} \leq \exp \left( -cn \min \left( \frac{\epsilon^2}{\sigma^2} , \frac{\epsilon}{\sigma^2} \right) \right),
\]

*Proof.* Denote by $\|X\|_{\psi_2} = \sup_{p \geq 1} \mathbb{E}[|X|^p]^\frac{1}{p}/\sqrt{p}$ and $\|X\|_{\psi_1} = \sup_{p \geq 1} \mathbb{E}[|X|^p]^\frac{1}{p}$ respectively the subgaussian norm and the subexponential norm of a one-dimensional random variable $X$. Since $X$ is subgaussian, $X^2$ is subexponential, and
\[
\|X^2 - \mathbb{E}[X^2]\|_{\psi_1} \leq 2\|X\|_{\psi_1} \leq 4\|X\|_{\psi_2}^2,
\]
where the first inequality is due to the triangle inequality $\|X^2 - \mathbb{E}[X^2]\|_{\psi_1} \leq \|X\|_{\psi_1} + \|\mathbb{E}[X^2]\|_{\psi_1}$ and $\|\mathbb{E}[X^2]\|_{\psi_1} = \mathbb{E}[X^2] \leq \|X\|_{\psi_1}^2$, and the second inequality is due to [83, Lemma 5.14]. Then the result follows from the Bernstein-type inequality of the subexponential distribution [83, Proposition 5.16]. $\square$

**Lemma 20.** Let $x_1^n, \ldots, x_2^n$ be independent samples from $\mathbb{P}_{\text{true}}$. Assume $\mathbb{P}_{\text{true}}$ is $\sigma$-subgaussian. Let $n > t > 0$. Then with probability at least $1 - \epsilon \frac{c(1+2d)}{1+\sqrt{2d}}$,\[
\sup_{w \in \mathbb{R}^d : \|w\|_2 = 1} \left\{ \mathbb{E}_{\mathbb{P}_p}\left[ (w^\top x)^2 \right] - \mathbb{E}_{\mathbb{P}_{\text{true}}}\left[ (w^\top x)^2 \right] \right\} \leq \sigma^2 \left( \sqrt{\frac{t}{n}} + \sqrt{2d} \right),
\]
where $\epsilon$ is a universal constant.

*Proof.* For every $w$ with $\|w\|_2 = 1$, the subgaussian assumption implies that $\mathbb{P}_{\text{true}}\{|w^\top (x^n_i - \mathbb{E}_{\mathbb{P}_{\text{true}}}[x])| \geq \epsilon\} \leq 2e^{-\frac{\epsilon^2}{2\sigma^2}}$ for all $\epsilon > 0$ [89, (2.9)]. Hence, by Lemma 19,
\[
\mathbb{P}_{\text{true}}\left\{ \frac{1}{n} \sum_{i=1}^n (w^\top x_i^n)^2 - \mathbb{E}_{\mathbb{P}_{\text{true}}}\left[ (w^\top x)^2 \right] \geq \epsilon \right\} \leq \exp \left( -cn \min \left( \frac{\epsilon^2}{\sigma^2} , \frac{\epsilon}{\sigma^2} \right) \right),
\]
where $c$ is a universal constant. Let $S_{1/3}$ be a $1/3$-cover of the unit sphere in $\mathbb{R}^d$. By [84, Lemma 2.2],
\[
\sup_{w \in \mathbb{R}^d : \|w\|_2 = 1} \left\{ \frac{1}{n} \sum_{i=1}^n (w^\top x_i^n)^2 - \mathbb{E}_{\mathbb{P}_{\text{true}}}\left[ (w^\top x)^2 \right] \right\} \leq 3 \sup_{w \in S_{1/3}} \frac{1}{n} \left\{ \sum_{i=1}^n (w^\top x_i^n)^2 - \mathbb{E}_{\mathbb{P}_{\text{true}}}\left[ (w^\top x)^2 \right] \right\}.
\]
It follows that
\[
\mathbb{P}_\sigma \left\{ \sup_{w \in \mathbb{R}^d : \|w\|_2 = 1} \left\{ \frac{1}{n} \sum_{i=1}^n (w^\top x_i^n)^2 - \mathbb{E}_{\mathbb{P}_{\text{true}}}\left[ (w^\top x)^2 \right] \right\} > \epsilon \right\}
\]
\[
\leq \mathbb{P}_\sigma \left\{ \sup_{w \in S_{1/3}} \left\{ \frac{1}{n} \sum_{i=1}^n (w^\top x_i^n)^2 - \mathbb{E}_{\mathbb{P}_{\text{true}}}\left[ (w^\top x)^2 \right] \right\} > \epsilon/3 \right\}
\]
\[
\leq |S_{1/3}| \exp \left( -cn \min \left( \frac{\epsilon^2}{\sigma^2} , \frac{\epsilon}{\sigma^2} \right) \right)
\]
\[
\leq 7^d \exp \left( -cn \min \left( \frac{\epsilon^2}{\sigma^2} , \frac{\epsilon}{\sigma^2} \right) \right),
\]
where the second inequality follows from the union bound and the last inequality is due to [53, Lemma 9.5]. Letting $\epsilon = \sigma^2 \left( \sqrt{\frac{t}{n}} + \sqrt{2d} \right)$ and using $n \geq t$ gives
\[
7^d \exp \left( -cn \min \left( \frac{\epsilon^2}{\sigma^2} , \frac{\epsilon}{\sigma^2} \right) \right) \leq \exp \left( -\frac{c}{1+\sqrt{2d}} n \frac{\epsilon^2}{\sigma^2} \right) \leq \exp \left( -\frac{c}{1+\sqrt{2d}} n \left( \frac{t}{n} + 2d \right) \right),
\]
which concludes the result. $\square$
The next lemma provides a probabilistic upper bound on $u_n$ for Example 4.

**Lemma 21.** Let $n \geq t > 0$. Under the setting in Example 4, there exists a universal constant $c > 0$, such that with probability at least $1 - e^{-\frac{c(t+2nd)}{1+2d}}$, for every $w \in \mathcal{W}$,

$$u_n(w) \leq 2B^2(\mu_2^2 + \tau^2 \sqrt{\frac{t}{n}} + \tau^2 \sqrt{2d} + \rho_n^2).$$

**Proof.** Let $u \in \mathbb{R}$ and $\gamma \in \Gamma(\mathbb{P}, \mathbb{P}_n)$. From $(a + b)^2 \leq 2(a^2 + b^2)$, it holds that

$$\mathbb{E}_\mathbb{P}[(w^T \tilde{x} - u)^2] \leq 2\mathbb{E}_{\mathbb{P}_n}[(w^T x - u)^2] + 2\mathbb{E}_\gamma[(w^T (\tilde{x} - x))^2] \leq 2\mathbb{E}_{\mathbb{P}_n}[(w^T x - u)^2] + 2\mathbb{E}_\gamma[\|w\|^2 \|\tilde{x} - x\|^2_2].$$

Minimizing over $u \in \mathbb{R}$ and $\gamma \in \Gamma(\mathbb{P}, \mathbb{P}_n)$ yields that

$$\text{Var}_\mathbb{P}[w^T \tilde{x}] \leq 2\text{Var}_{\mathbb{P}_n}[w^T x] + 2B^2 W_2(\mathbb{P}, \mathbb{P}_n)^2 \leq 2\mathbb{E}_{\mathbb{P}_n}[(w^T x)^2] + 2B^2 \rho_n^2.$$

By Lemma 20, with probability at least $1 - e^{-\frac{c(t+2nd)}{1+2d}}$, for every $w \in \mathcal{W}$,

$$\mathbb{E}_{\mathbb{P}_n}[(w^T x)^2] = \|w\|^2 \mathbb{E}_{\mathbb{P}_n}[(\|w\|^T x)^2] \leq B^2(\mathbb{E}_{\mathbb{P}_\text{true}}[(\|w\|^T x)^2] + \tau^2 (\sqrt{\frac{t}{n}} + \sqrt{2d})) \leq B^2(\mathbb{E}_{\mathbb{P}_\text{true}}[\|x\|^2_2] + \tau^2 (\sqrt{\frac{t}{n}} + \sqrt{2d})) \leq B^2(\mu_2^2 + \tau^2 \sqrt{\frac{t}{n}} + \tau^2 \sqrt{2d} + \rho_n^2).$$

\[\square\]

In the next two results, we compute the Rademacher complexities.

**Lemma 22.** Let $\Theta \subseteq \{\theta \in \mathbb{R}^{d+1} : \|\theta\|_2 \leq B\}$, where $B \geq 0$. Define $\mathcal{F} = \{z \mapsto (\theta^T z)^2 - \alpha u - \alpha^2/4 : \theta \in \Theta\}$. Then

$$\mathbb{E}_{\mathbb{P}_\text{true}}[\mathcal{R}_n(\mathcal{F})] \leq \frac{B^2 \mathbb{E}_{\mathbb{P}_\text{true}}[\|z\|^4_2]^{\frac{1}{2}}}{\sqrt{n}}.$$

**Proof.** Let $\{\sigma_i\}$ be i.i.d. Rademacher random variables and $z_i^n$ be i.i.d samples from $\mathbb{P}_\text{true}$. We have

$$\mathbb{E}_\sigma \left[ \sup_{\theta \in \Theta} \frac{1}{n} \sum_{i=1}^n \sigma_i \left( (\theta^T z_i^n)^2 - \alpha u - \alpha^2/4 \right) \right] = \mathbb{E}_\sigma \left[ \sup_{\theta \in \Theta} \frac{1}{n} \sum_{i=1}^n \sigma_i (\theta^T z_i^n)^2 \right] = \mathbb{E}_\sigma \left[ \frac{1}{n} \sup_{\theta \in \Theta} (\theta^T, \sum_{i=1}^n \sigma_i z_i^n z_i^n^T) \right] \leq \mathbb{E}_\sigma \left[ \frac{1}{n} \sup_{\|\theta\|_2 \leq B} \|\theta^T\|_F \|\sum_{i=1}^n \sigma_i z_i^n z_i^n^T\|_F \right] \leq \frac{B^2}{n} \mathbb{E}_\sigma \left[ \|\sum_{i=1}^n \sigma_i z_i^n z_i^n^T\|_F \right].$$

Hence, the proof is completed by noticing that

$$\mathbb{E}_{\mathbb{P}_\text{true}} \left[ \left\| \sum_{i=1}^n \sigma_i z_i^n z_i^n^T \right\|_F \right] \leq \sqrt{n} \mathbb{E}_{\mathbb{P}_\text{true}} \left[ \|z\|^4_2 \right]^{\frac{1}{2}}.$$

\[\square\]
LEMMA 23. Let $\Theta \subset \mathbb{R}^{d+1}$. Define
\[
G = \left\{ z \mapsto \frac{(\theta^T z)^2}{\mathbb{E}_{\text{true}}[(\theta^T z)^2]} : \theta \in \Theta \right\}.
\]

Then
\[
\mathbb{E}_{\Theta}\left[ \mathcal{R}_n(G) \right] \leq \frac{1}{1 \wedge \zeta} \sqrt{\frac{\mathbb{E}_{\text{true}}[\|z\|^4]}{n}}.
\]

Proof. Let $\{\sigma_i\}_i$ be i.i.d. Rademacher random variables and $z^n_i$ be i.i.d samples from $\mathbb{P}_{\text{true}}$. We have
\[
\mathcal{R}_n(G) = \mathbb{E}_\sigma \left[ \sup_{\theta \in \Theta} \sum_{i=1}^n \sigma_i \frac{(\theta^T z^n_i)^2}{\mathbb{E}_{\text{true}}[(\theta^T z)^2]} \right]
\[
= \mathbb{E}_\sigma \left[ \sup_{\theta \in \Theta} \frac{\theta \theta^T}{\mathbb{E}_{\text{true}}[(\theta^T z)^2]} \sum_{i=1}^n \sigma_i (z^n_i z^n_i^\top) \right]
\[
\leq \mathbb{E}_\sigma \left[ \sup_{\theta \in \Theta} \left\| \frac{\theta \theta^T}{\mathbb{E}_{\text{true}}[(\theta^T z)^2]} \right\|_F \cdot \left\| \sum_{i=1}^n \sigma_i z^n_i z^n_i^\top \right\|_F \right]
\[
\leq \frac{1}{1 \wedge \zeta} \mathbb{E}_\sigma \left[ \left\| \sum_{i=1}^n \sigma_i z^n_i z^n_i^\top \right\|_F \right].
\]

The result follows by noticing that
\[
\mathbb{E}_\Theta,\sigma \left[ \left\| \sum_{i=1}^n \sigma_i z^n_i z^n_i^\top \right\|_F \right] \leq \left( \mathbb{E}_\Theta \left[ \left\| \sum_{i=1}^n z^n_i z^n_i^\top \right\|_F^2 \right] \right)^{\frac{1}{2}} = (n \mathbb{E}_{\text{true}}[\|zz^\top\|_F^2])^{\frac{1}{2}} = \sqrt{n}(\mathbb{E}_{\text{true}}[\|z\|_2^4])^{\frac{1}{2}}.
\]

We next relate the gradient norm $\|\nabla f_0\|_{\text{true},2}$ to its empirical estimate. To this end, we need the following lemma.

LEMMA 24. Let $x^n_1, \ldots, x^n_n$ be i.i.d. samples from $\mathbb{P}_{\text{true}}$. Assume $\mathbb{P}_{\text{true}}$ satisfies $T_1(\tau)$. Then with probability more than $1 - e^{-t}$,
\[
\max_{1 \leq i \leq n} \|x^n_i\|_2 < \mathbb{E}_{\text{true}}[\|x\|_2] + \sqrt{t + \log n}.
\]

Proof. Since $\mathbb{P}_{\text{true}}$ satisfies $T_1(\tau)$, for every $t > 0$, $\mathbb{E}_{\text{true}} \left[ \exp \left( t(\|x\|_2 - \mathbb{E}_{\text{true}}[\|x\|_2]) \right) \right] \leq \exp(\frac{t^2}{2})$, (cf. [89, Theorem 3.19]). By Jensen’s inequality,
\[
\exp \left( t \mathbb{E}_\Theta \left[ \max_{1 \leq i \leq n} \|x^n_i\|_2 - \mathbb{E}_{\text{true}}[\|x\|_2] \right] \right) \leq \exp \left( t \max_{1 \leq i \leq n} \mathbb{E}_{\text{true}}[\|x^n_i\|_2 - \mathbb{E}_{\text{true}}[\|x\|_2]] \right)
\[
\leq \sum_{i=1}^n \mathbb{E}_{\text{true}} \left[ \exp \left( t(\|x^n_i\|_2 - \mathbb{E}_{\text{true}}[\|x\|_2]) \right) \right]
\[
< n \exp(\frac{t^2}{2}).
\]

Thus, using Markov’s inequality,
\[
\mathbb{P}_\Theta \left\{ \max_{1 \leq i \leq n} \|x^n_i\|_2 \geq \mathbb{E}_{\text{true}}[\|x\|_2] + \epsilon \right\} < \inf_{t \in \mathbb{R}} \frac{n \exp(\frac{t^2}{4})}{\exp(\epsilon t)} = n \exp(-\frac{\epsilon^2}{t}),
\]

which completes the proof. \(\square\)
Lemma 25. Under the setting of Example 4, let \( L_n = 1 + \mathbb{E}_{P_{true}}[\|x\|_2] + \sqrt{\tau(t + \log n)} \) and \( \hat{g}_\theta(z) = \frac{1}{2} |\theta^T z| 1 \{ \|z\|_2 \leq L_n \}. \) Then there exists \( C > 0 \) such that with probability at least \( 1 - 2e^{-t}, \)

\[
\|\|\nabla f_\theta\|_2 \|p_{true, 2} \leq C\|\nabla f_\theta\|_2 \|p_{n, 2} \left(1 + 2\mathbb{E}_\theta[\mathcal{R}_n(G)] + L_n^2 \sqrt{\frac{t}{2n}} \right), \quad \forall \theta \in \Theta.
\]

Proof. We have

\[
g_\theta(z) := \frac{\|f_\theta(z)\|_2}{\|\nabla f_\theta\|_2 \|p_{true, 2}} = \frac{|\theta^T z|}{\mathbb{E}_{P_{true}}[(\theta^T z)^2]^{\frac{1}{2}}} \leq \|z\|_2 \\leq 1 \wedge \sqrt{\frac{1}{\lambda}}.
\]

Then by Lemma 24, with probability at least \( 1 - e^{-t}, \) \( \|g_\theta\|_{p_{true, 2}}^2 = \|\hat{g}_\theta\|_{p_{n, 2}}^2. \) According to the proof of Corollary 6 and applying the contraction inequality on the truncation (Lemma 8), with probability at least \( 1 - e^{-t}, \) whenever \( 2\mathbb{E}_\theta[\mathcal{R}_n(G)] + L_n^2 \sqrt{\frac{t}{2n}} < 1/2, \) it holds that

\[
\|\hat{g}_\theta\|_{p_{true, 2}} \leq \|g_\theta\|_{p_{n, 2}} \left(1 + 2\mathbb{E}_\theta[\mathcal{R}_n(G)] + L_n^2 \sqrt{\frac{t}{2n}} \right).
\]

Hence, with probability at least \( 1 - 2e^{-t}, \)

\[
\|\hat{g}_\theta\|_{p_{true, 2}} \leq \|g_\theta\|_{p_{n, 2}} \left(1 + 2\mathbb{E}_\theta[\mathcal{R}_n(G)] + L_n^2 \sqrt{\frac{t}{2n}} \right).
\]

Moreover, since

\[
\sup_{\theta \in \Theta} \mathbb{E}_{P_{true}}[(\theta^T z)^2 1\{\|z\|_2 \leq L_n\}] \leq \mathbb{E}_{P_{true}}[(\theta^T z)^2 1\{\|z\|_2 > L_1\}] \leq \frac{1}{1 \wedge \sqrt{\frac{1}{\lambda}}}
\]

there exists \( c > 0 \) such that

\[
\inf_{\theta \in \Theta} \|\hat{g}_\theta\|_{p_{true, 2}}^2 \geq \frac{\inf_{\theta \in \Theta} \mathbb{E}_{P_{true}}[(\theta^T z)^2 1\{\|z\|_2 \leq L_n\}]}{\sup_{\theta \in \Theta} \mathbb{E}_{P_{true}}[(\theta^T z)^2 1\{\|z\|_2 \leq L_1\}]} \geq 1/c.
\]

\( \square \)

C.4. Proofs for Section 5.5

The next two results are used for Example 6, which relies on the following lemma.

Lemma 26 (Contraction for vector-valued functions). Let \( \mathcal{H} \) be a family of \( m \)-dimensional vector-valued functions on \( X \subset \mathbb{R}^d. \) Let \( \ell : \mathbb{R}^m \to \mathbb{R} \) be an \( L_\ell \)-Lipschitz continuous function. Denote by \( \ell \circ \mathcal{H} := \{x \mapsto \ell(h(x)) : h = (h_1, \ldots, h_m) \in \mathcal{H} \}. \) Then

\[
\mathbb{E}_\theta[\mathcal{R}_n(\ell \circ \mathcal{H})] \leq \frac{\sqrt{2} L_\ell}{\sqrt{m}} \mathbb{E}_{\theta, \sigma} \left[ \sup_{h \in \mathcal{H}} \sum_{i=1}^n \sum_{j=1}^m \sigma_{ij} h_j(x_i^n) \right].
\]

where \( \sigma_{ij} \)'s are i.i.d. Rademacher random variables. In particular, when \( \mathcal{H} = \{x \mapsto Wx : W \in \mathbb{R}^{mxd}, WW^\top = I\}, \) we have

\[
\mathbb{E}_\theta[\mathcal{R}_n(\ell \circ \mathcal{H})] \leq \frac{\sqrt{2} L_\ell \sqrt{m \mathbb{E}_{P_{true}}[\|x\|_2^2]}}{\sqrt{n}}.
\]
Proof. The first part is due to Maurer [61, Corollary 4]. For the second part, denote \( W = (w_1, \ldots, w_m)^T \), where \( w_j \in \mathbb{R}^d \). Using the result in Section 4.2 of Maurer [61], it suffices to compute

\[
\sup_{WW^T = I} \sum_{j=1}^m w_j^T \sum_{i=1}^n \sigma_{ij} x_i^n \leq \sup_{WW^T = I} \sum_{j=1}^m \|w_j\|_2 \cdot \sum_{i=1}^n \sigma_{ij} x_i^n \|2 \leq \left( \sum_{j=1}^m \|\sum_{i=1}^n \sigma_{ij} x_i^n \|_2 \right)^{1/2} \leq \sqrt{m} \sum_{i=1}^n \|x_i^n \|_2^2.
\]

Thereby the second part of the result follows by noticing that

\[
\mathbb{E}_\sigma \left[ \left( \sum_{i=1}^n \|x_i^n \|_2^2 \right)^{1/2} \right] \leq \mathbb{E}_\sigma \left[ \sum_{i=1}^n \|x_i^n \|_2^2 \right]^{1/2} = \sqrt{n} \mathbb{E}_{\text{true}} \left[ \|x \|_2^2 \right]^{1/2}.
\]

\[\square\]

Lemma 27. Under the setting in Example 6, it holds that

\[
\mathbb{E}_\sigma \left[ \mathcal{G}_n \left( \{x \mapsto W_2 \phi(W_1 x) : W_1 W_1^T = I, \|W_2\|_2 \leq B \} \right) \right] \leq B \sqrt{2 d_2} \mathbb{E}_{\text{true}} \left[ \|x \|_2^2 \right]^{1/2}.
\]

Proof. Applying Lemma 26 with \( \mathcal{H} = \{x \mapsto W_1 x : W_1 W_1^T = I\} \) and \( \ell(\cdot) = W_2 \phi(\cdot) \) yields the result. \[\square\]

Lemma 28. Under the setting in Example 6, it holds that

\[
\mathbb{E}_{\sigma, \theta} \left[ \sup_{\theta \in \Theta} \frac{1}{n} \sum_{i=1}^n \sigma_i \left( \frac{\|\nabla_x f_0(x_i^n) \|_2^2}{\ell \|\nabla f_0 \|_2 \|\nabla f_0 \|_2^2} \right) \right] \leq \mathbb{E}_{\sigma, \theta} \left[ \sup_{\theta \in \Theta} \frac{1}{n} \sum_{i=1}^n \sigma_i \left( \frac{\|\nabla_x f_0(x_i^n) \|_2^2}{W_2 \|_2} \right) \right] \cdot \frac{2L(Lh_\phi + 1) \sqrt{2 d_2} \mathbb{E}_{\text{true}} \left[ \|x \|_2^2 \right]^{1/2}}{\eta^2 \zeta \sqrt{n}}.
\]

Proof. Observe that

\[
\mathbb{E}_{\sigma, \theta} \left[ \sup_{\theta \in \Theta} \frac{1}{n} \sum_{i=1}^n \sigma_i \left( \frac{\|\nabla_x f_0(x_i^n) \|_2^2}{\ell \|\nabla f_0 \|_2 \|\nabla f_0 \|_2^2} \right) \right] \leq \mathbb{E}_{\sigma, \theta} \left[ \sup_{\theta \in \Theta} \frac{1}{n} \sum_{i=1}^n \sigma_i \left( \frac{\|\nabla_x f_0(x_i^n) \|_2^2}{W_2 \|_2} \right) \right] \cdot \frac{2L(Lh_\phi + 1) \sqrt{2 d_2} \mathbb{E}_{\text{true}} \left[ \|x \|_2^2 \right]^{1/2}}{\eta^2 \zeta \sqrt{n}}.
\]

Moreover, we have that

\[
\left| l'(W_2 \phi(\bar{r}), y)^2 \frac{\|W_2 \phi'(y) \|_2^2}{W_2 \|_2} - l'(W_2 \phi(t), y)^2 \frac{\|W_2 \phi'(t) \|_2^2}{W_2 \|_2} \right| \leq \left| l'(W_2 \phi(\bar{r}), y)^2 \frac{\|W_2 \phi'(y) \|_2^2}{W_2 \|_2} - l'(W_2 \phi(\bar{r}), y)^2 \frac{\|W_2 \phi'(t) \|_2^2}{W_2 \|_2} \right| \leq L^2 \left( \frac{\|W_2 \phi'(y) \|_2^2}{W_2 \|_2} - \frac{\|W_2 \phi'(t) \|_2^2}{W_2 \|_2} \right) \leq 2L^2 h_\phi \|\bar{r} - t\|_2 + 2L \|\bar{r} - t\|_2.
\]
Thereby, applying Lemma 26 to $\mathcal{H} = \{ x \mapsto W_1 x : W_1 W_1^T = I \}$ and $\ell(\cdot) = l'(W_2 \phi(\cdot))^2 \frac{||W_2 \phi'(\cdot)||_2^2}{||W_2||_2^2}$, we obtain

$$
\mathbb{E}_{\sigma, \phi} \left[ \sup_{\theta \in \Theta} \frac{1}{n} \sum_{i=1}^{n} |\nabla \ell_{\phi}(x^n_i)|^2 \right] \leq 2L(Lh_\phi + 1) \sqrt{\frac{2d_2 \mathbb{E}_{\mathbb{P}_{w2}}[||x||_2^2]}{n}}.
$$

Hence the proof is completed. \qed

References

[1] Abadeh SS, Nguyen VA, Kuhn D, Esfahani PMM (2018) Wasserstein distributionally robust kalman filtering. Advances in Neural Information Processing Systems, 8474–8483.

[2] Abdullah MA, Ren H, Ammar HB, Milenkovic V, Luo R, Zhang M, Wang J (2019) Wasserstein robust reinforcement learning. arXiv preprint arXiv:1907.13196.

[3] Ban GY, Rudin C (2019) The big data newsvendor: Practical insights from machine learning. Operations Research 67(1):90–108.

[4] Bansal N, Chen X, Wang Z (2018) Can we gain more from orthogonality regularizations in training deep networks? Advances in Neural Information Processing Systems, 4261–4271.

[5] Bartl D, Drapeau S, Obloj J, Wiesel J (2020) Robust uncertainty sensitivity analysis. arXiv preprint arXiv:2006.12202.

[6] Bartlett PL, Bousquet O, Mendelson S, et al. (2005) Local rademacher complexities. The Annals of Statistics 33(4):1497–1537.

[7] Bartlett PL, Mendelson S (2002) Rademacher and gaussian complexities: Risk bounds and structural results. Journal of Machine Learning Research 3(Nov):463–482.

[8] Bayraksan G, Love DK (2015) Data-driven stochastic programming using phi-divergences. The Operations Research Revolution, 1–19 (INFORMS).

[9] Ben-Tal A, Den Hertog D, De Waegenaere A, Melenberg B, Rennen G (2013) Robust solutions of optimization problems affected by uncertain probabilities. Management Science 59(2):341–357.

[10] Blanchet J, Chen L, Zhou XY (2018) Distributionally robust mean-variance portfolio selection with wasserstein distances. arXiv preprint arXiv:1802.04885.

[11] Blanchet J, Glynn PW, Yan J, Zhou Z (2019) Multivariate distributionally robust convex regression under absolute error loss. Advances in Neural Information Processing Systems, 11794–11803.

[12] Blanchet J, Kang Y (2016) Sample out-of-sample inference based on wasserstein distance. arXiv preprint arXiv:1605.01340.

[13] Blanchet J, Kang Y (2017) Semi-supervised learning based on distributionally robust optimization. arXiv preprint arXiv:1702.08848.

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

[15] Blanchet J, Murthy K, Si N (2019) Confidence regions in wasserstein distributionally robust estimation. arXiv preprint arXiv:1906.01614.

[16] Blanchet J, Murthy K, Zhang F (2018) Optimal transport based distributionally robust optimization: Structural properties and iterative schemes. arXiv preprint arXiv:1810.02403.

[17] Blanchet J, Murthy KR (2016) Quantifying distributional model risk via optimal transport. arXiv preprint arXiv:1604.01446.

[18] Bobkov SG, Götze F (1999) Exponential integrability and transportation cost related to logarithmic sobolev inequalities. Journal of Functional Analysis 163(1):1–28.

[19] Bolley F, Villani C (2005) Weighted csiszár-kullback-pinsker inequalities and applications to transportation inequalities. Annales de la Faculté des sciences de Toulouse: Mathématiques, volume 14, 331–352.
[20] Bousquet O, Chapelle O, Hein M (2004) Measure based regularization. Advances in Neural Information Processing Systems, 1221–1228.

[21] Calafiore GC, El Ghaoui L (2006) On distributionally robust chance-constrained linear programs. Journal of Optimization Theory and Applications 130(1):1–22.

[22] Carlsson JG, Behroozi M, Mihic K (2018) Wasserstein distance and the distributionally robust tsp. Operations Research 66(6):1603–1624.

[23] Chen R, Paschalidis IC (2018) A robust learning approach for regression models based on distributionally robust optimization. The Journal of Machine Learning Research 19(1):517–564.

[24] Chen Y, Li W (2018) Natural gradient in wasserstein statistical manifold. arXiv preprint arXiv:1809.00210.

[25] Duan C, Fang W, Jiang L, Yao L, Liu J (2018) Distributionally robust chance-constrained approximate ac-opf with wasserstein metric. IEEE Transactions on Power Systems 33(5):4924–4936.

[26] Duchi JC, Hashimoto T, Namkoong H (2018) Distributionally robust losses against mixture covariate shifts. Under review.

[27] Duchi JC, Namkoong H (2019) Variance-based regularization with convex objectives. Journal of Machine Learning Research 20(68):1–65.

[28] Erdoğan E, Iyengar G (2006) Ambiguous chance constrained problems and robust optimization. Mathematical Programming 107(1-2):37–61.

[29] Dizaji, Ghaemian M, Torkamani A, Kuhn D (2018) Distributionally robust optimization with wasserstein distance. arXiv preprint arXiv:1807.00451.

[30] Fournier N, Guillin A (2016) On the rate of convergence in wasserstein distance of the empirical measure. Probability Theory and Related Fields 162(3-4):707–738.

[31] Gao R, Xie L, Xie Y, Tu H (2018) Robust hypothesis testing using wasserstein uncertainty sets. Advances in Neural Information Processing Systems, 7902–7912.

[32] Gozlan N, Léonard C (2007) A large deviation approach to some transportation cost inequalities. Probability Theory and Related Fields 139(1-2):235–283.

[33] Gozlan N, Léonard C (2010) Transport inequalities. a survey. arXiv preprint arXiv:1003.3852.
[45] Gulrajani I, Ahmed F, Arjovsky M, Dumoulin V, Courville A (2017) Improved training of wasserstein gans. arXiv preprint arXiv:1704.00028.

[46] Huang L, Liu X, Lang B, Yu AW, Wang Y, Li B (2018) Orthogonal weight normalization: Solution to optimization over multiple dependent stiefel manifolds in deep neural networks. AAAI.

[47] Jakubovitz D, Giryes R (2018) Improving dnn robustness to adversarial attacks using jacobian regularization. Proceedings of the European Conference on Computer Vision (ECCV), 514–529.

[48] Jiang R, Guan Y (2015) Data-driven chance constrained stochastic program. Mathematical Programming 1–37.

[49] Koltchinskii V, et al. (2006) Local rademacher complexities and oracle inequalities in risk minimization. The Annals of Statistics 34(6):2593–2656.

[50] Kuhn D, Esfahani PM, Nguyen VA, Shafieezadeh-Abadeh S (2019) Wasserstein distributionally robust optimization: Theory and applications in machine learning. Operations Research & Management Science in the Age of Analytics, 130–166 (INFORMS).

[51] Lam H (2016) Recovering best statistical guarantees via the empirical divergence-based distributionally robust optimization. arXiv preprint arXiv:1605.09349.

[52] Ledoux M (2001) The concentration of measure phenomenon. Number 89 (American Mathematical Soc.).

[53] Ledoux M, Talagrand M (2013) Probability in Banach Spaces: isoperimetry and processes (Springer Science & Business Media).

[54] Lee J, Raginsky M (2018) Minimax statistical learning with wasserstein distances. Advances in Neural Information Processing Systems, 2687–2696.

[55] Levine A, Feizi S (2019) Wasserstein smoothing: Certified robustness against wasserstein adversarial attacks. arXiv preprint arXiv:1910.10783.

[56] Li W, Montúfar G (2018) Natural gradient via optimal transport. Information Geometry 1(2):181–214.

[57] Lopez AT, Jog V (2018) Generalization error bounds using wasserstein distances. 2018 IEEE Information Theory Workshop (ITW), 1–5 (IEEE).

[58] Luo F, Mehrorita S (2019) Decomposition algorithm for distributionally robust optimization using wasserstein metric with an application to a class of regression models. European Journal of Operational Research 278(1):20–35.

[59] Luxburg Uv, Bousquet O (2004) Distance-based classification with lipschitz functions. Journal of Machine Learning Research 5(Jun):669–695.

[60] Lyu C, Huang K, Liang HN (2015) A unified gradient regularization family for adversarial examples. 2015 IEEE International Conference on Data Mining, 301–309 (IEEE).

[61] Maurer A (2016) A vector-contraction inequality for rademacher complexities. International Conference on Algorithmic Learning Theory, 3–17 (Springer).

[62] Najafi A, Maeda Si, Koyama M, Miyato T (2019) Robustness to adversarial perturbations in learning from incomplete data. Advances in Neural Information Processing Systems, 5542–5552.

[63] Namkoong H, Duchi JC (2017) Variance-based regularization with convex objectives. Advances in neural information processing systems, 2971–2980.

[64] Nguyen VA, Kuhn D, Esfahani PM (2018) Distributionally robust inverse covariance estimation: The wasserstein shrinkage estimator. arXiv preprint arXiv:1805.07194.

[65] Nguyen VA, Shafieezadeh-Abadeh S, Kuhn D, Esfahani PM (2019) Bridging bayesian and minimax mean square error estimation via wasserstein distributionally robust optimization. arXiv preprint arXiv:1911.03539.

[66] Oberman AM, Calder J (2018) Lipschitz regularized deep neural networks converge and generalize. arXiv preprint arXiv:1808.09540.

[67] Popescu I (2007) Robust mean-covariance solutions for stochastic optimization. Operations Research 55(1):98–112.
[68] Raginsky M, Sason I, et al. (2013) Concentration of measure inequalities in information theory, communications, and coding. *Foundations and Trends in Communications and Information Theory* 10(1-2):1–246.

[69] Russo D, Zou J (2019) How much does your data exploration overfit? controlling bias via information usage. *IEEE Transactions on Information Theory* 66(1):302–323.

[70] Scarf H (1958) A min-max solution of an inventory problem. *Studies in the mathematical theory of inventory and production*.

[71] Shafieezadeh-Abadeh S, Esfahani PM, Kuhn D (2015) Distributionally robust logistic regression. *Advances in Neural Information Processing Systems*, 1576–1584.

[72] Shafieezadeh-Abadeh S, Kuhn D, Esfahani PM (2019) Regularization via mass transportation. *Journal of Machine Learning Research* 20(103):1–68.

[73] Shaham U, Yamada Y, Negahban S (2015) Understanding adversarial training: Increasing local stability of neural nets through robust optimization. *arXiv preprint arXiv:1511.05432*.

[74] Shalev-Shwartz S, Ben-David S (2014) *Understanding machine learning: From theory to algorithms* (Cambridge university press).

[75] Shapiro A, Dentcheva D, et al. (2014) *Lectures on stochastic programming: modeling and theory*, volume 16 (SIAM).

[76] Shapiro A, Kleywegt A (2002) Minimax analysis of stochastic problems. *Optimization Methods and Software* 17(3):523–542.

[77] Sinha A, Namkoong H, Duchi J (2017) Certifying some distributional robustness with principled adversarial training. *arXiv preprint arXiv:1710.10571*.

[78] Smirnova E, Dohmatob E, Mary J (2019) Distributionally robust reinforcement learning. *arXiv preprint arXiv:1902.08708*.

[79] Sokolić J, Giryes R, Sapiro G, Rodrigues MR (2017) Robust large margin deep neural networks. *IEEE Transactions on Signal Processing* 65(16):4265–4280.

[80] Staib M, Jegelka S (2017) Distributionally robust deep learning as a generalization of adversarial training. *NIPS workshop on Machine Learning and Computer Security*.

[81] Terjék D (2019) Virtual adversarial lipschitz regularization. *arXiv preprint arXiv:1907.05681*.

[82] Van De Geer S (1987) A new approach to least-squares estimation, with applications. *The Annals of Statistics* 587–602.

[83] Vershynin R (2010) Introduction to the non-asymptotic analysis of random matrices. *arXiv preprint arXiv:1011.3027*.

[84] Vershynin R (2012) How close is the sample covariance matrix to the actual covariance matrix? *Journal of Theoretical Probability* 25(3):655–686.

[85] Villani C (2008) *Optimal transport: old and new*, volume 338 (Springer Science & Business Media).

[86] Virmaux A, Scaman K (2018) Lipschitz regularity of deep neural networks: analysis and efficient estimation. *Advances in Neural Information Processing Systems*, 3835–3844.

[87] Volpi R, Namkoong H, Sener O, Duchi JC, Murino V, Savarese S (2018) Generalizing to unseen domains via adversarial data augmentation. *Advances in Neural Information Processing Systems*, 5334–5344.

[88] Vorontsov E, Trabelsi C, Kadoury S, Pal C (2017) On orthogonality and learning recurrent networks with long term dependencies. *International Conference on Machine Learning*, 3570–3578.

[89] Wainwright MJ (2019) *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48 (Cambridge University Press).

[90] Wang C, Gao R, Qiu F, Wang J, Xin L (2018) Risk-based distributionally robust optimal power flow with dynamic line rating. *IEEE Transactions on Power Systems* 33(6):6074–6086.

[91] Wang C, Gao R, Wei W, Shafie-khah M, Bi T, Catalao JP (2018) Risk-based distributionally robust optimal gas-power flow with wasserstein distance. *IEEE Transactions on Power Systems* 34(3):2190–2204.
Wang H, Diaz M, Santos Filho JCS, Calmon FP (2019) An information-theoretic view of generalization via Wasserstein distance. 2019 IEEE International Symposium on Information Theory (ISIT), 577–581 (IEEE).

Wang Z, Glynn PW, Ye Y (2016) Likelihood robust optimization for data-driven problems. Computational Management Science 13(2):241–261.

Wei C, Ma T (2019) Data-dependent sample complexity of deep neural networks via Lipschitz augmentation. Advances in Neural Information Processing Systems, 9725–9736.

Wiesemann W, Kuhn D, Sim M (2014) Distributionally robust convex optimization. Operations Research 62(6):1358–1376.

Wozabal D (2014) Robustifying convex risk measures for linear portfolios: A nonparametric approach. Operations Research 62(6):1302–1315.

Xie D, Xiong J, Pu S (2017) All you need is beyond a good init: Exploring better solution for training extremely deep convolutional neural networks with orthonormality and modulation. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 6176–6185.

Xie W (2019) On distributionally robust chance constrained programs with Wasserstein distance. Mathematical Programming 1–41.

Xie W, Zhang J, Ahmed S (2020) Distributionally robust bottleneck combinatorial problems: Uncertainty quantification and robust decision making. arXiv preprint arXiv:2003.00630.

Xu A, Raginsky M (2017) Information-theoretic analysis of generalization capability of learning algorithms. Advances in Neural Information Processing Systems, 2524–2533.

Xu H, Mannor S (2012) Robustness and generalization. Machine learning 86(3):391–423.

Yang I (2017) A convex optimization approach to distributionally robust Markov decision processes with Wasserstein distance. IEEE Control Systems Letters 1(1):164–169.

Yang I (2018) Wasserstein distributionally robust stochastic control: A data-driven approach. arXiv preprint arXiv:1812.09808.

Yin D, Kannan R, Bartlett P (2019) Rademacher complexity for adversarially robust generalization. International Conference on Machine Learning, 7085–7094 (PMLR).

Yoshida Y, Miyato T (2017) Spectral norm regularization for improving the generalizability of deep learning. arXiv preprint arXiv:1705.10041.

Yu T, Thomas G, Yu L, Ermon S, Zou J, Levine S, Finn C, Ma T (2020) Mopo: Model-based offline policy optimization. arXiv preprint arXiv:2005.13239.

Žáčková J (1966) On minimax solutions of stochastic linear programming problems. Časopis pro pěstování matematiky 91(4):423–430.

Zhao C, Guan Y (2018) Data-driven risk-averse stochastic optimization with Wasserstein metric. Operations Research Letters 46(2):262–267.