Wasserstein Regularization for 0-1 Loss

Zhen Yang
Department of Mathematics, The Chinese University of Hong Kong, Hong Kong, zhen.yang@link.cuhk.edu.hk
Rui Gao
Department of Information, Risk and Operations Management, The University of Texas at Austin, rui.gao@mccombs.utexas.edu

Wasserstein distributionally robust optimization (DRO) finds robust solutions by hedging against data perturbation specified by distributions in a Wasserstein ball. The robustness is linked to the regularization effect, which has been studied for continuous losses in various settings. However, existing results cannot be simply applied to the 0-1 loss, which is frequently seen in uncertainty quantification, classification, and chance-constrained programs. In this paper, we relate the Wasserstein DRO with 0-1 loss to a new regularization problem, in which the regularization term is a polynomial of the radius of the Wasserstein ball and the density around the decision boundary. Importantly, this result implies a qualitative difference between 0-1 losses and continuous losses in terms of the radius selection: for most interesting cases, it suffices to choose a radius smaller than the root-n rule. Numerical experiments demonstrate the effectiveness of our implied radius selection rule.

Key words: zero-one loss; distributionally robust optimization; margin regularization; Wasserstein metric

1. Introduction

The emerging distributionally robust optimization (DRO) provides a powerful tool for learning and decision-making when the knowledge of the data-generating model is limited. Among many different modeling choices, distributional uncertainty sets based on Wasserstein distance have raised much interest; we refer to [33, 7] for recent tutorials. Wasserstein DRO has been applied to a variety of practical domains involving 0-1 loss functions, including uncertainty quantification [16, 6, 19], hypothesis testing [49], chance-constrained programs [50, 13, 30, 24, 25, 40, 41, 14], adversarial classification [26, 48], automatic control [27, 39], and power systems [15, 35, 11, 36], etc. In these applications, we are interested in evaluating the worst-case probability of some event E among a family of distributions in the ρ-neighborhood ($\rho > 0$) of some nominal distribution $\mathbb{Q}$ in p-Wasserstein distance, where $p \in [1, \infty]$:

$$\inf_{\mathbb{P}: \mathcal{W}_p(\mathbb{P}, \mathbb{Q}) \le \rho} \mathbb{P}(E). \tag{P}$$

The set E varies among applications. For example, in the uncertainty quantification, E describes the event of interest, and (P) represents the safe probability at a certain confidence level. For another example, in adversarial binary classification, $E = \{x : yF_\theta(x) < 0\}$, where F_θ is a classifier, parameterized by θ, that classifies a sample x to the positive class if $F_\theta(x) > 0$ and the negative class otherwise, and (P) evaluates the worst-case classification accuracy for class $y \in \{\pm 1\}$.

The empirical success of Wasserstein DRO is often credited to its regularization effect. Indeed, studies on Wasserstein regularization not only provide novel interpretations and deepen the theoretical understanding of Wasserstein DRO, but also serve as an indispensable ingredient for establishing out-of-sample performance guarantees for Wasserstein DRO and thereby guide the radius selection [37, 8, 8, 2]. For continuous loss functions, it has been shown that Wasserstein DRO models are closely related to norm regularization [38, 12, 5], Lipschitz regularization [16, 37], gradient regularization [46, 8] and variation regularization [18, 2].

For 0-1 loss functions, however, understanding of the Wasserstein regularization is rather limited. The insights in the continuous setting cannot be translated easily to the 0-1 loss because the proof techniques therein mainly depend on Taylor's expansion of a smooth loss function. In the context of

the chance-constrained program, Xie [50] and Chen et al. [13] show that the robust linear chance constraint can be equivalently viewed as regularized conditional value-at-risk; and in the context of adversarial linear classification, Ho-Nguyen and Wright [26] shows that Wasserstein DRO is equivalent to norm-regularized hinge loss minimization. Beyond these special linear cases, the regularization effect of Wasserstein 0-1 loss remains largely unknown.

To fill in this gap, in this paper, we show a general connection between Wasserstein DRO with 0-1 loss and regularization. A representative case of our result can be stated as follows.

Theorem (Informal) *Let $p \in [1,\infty]$ and $E \subset \mathbb{R}^d$ be a non-empty open set. Let $\mathbb{P}_n$ be an empirical distribution constructed from n i.i.d. samples from some underlying distribution with a positive density on E. Define*

$$\mathfrak{g}(0) \;=\; \lim_{s\downarrow 0} \frac{1}{s}\mathbb{P}_{\mathsf{true}}\left(0 < \mathsf{d}_{E^c}(Z) \le s\right),$$

where $\mathsf{d}_{E^c}(Z)$ denotes the distance of a random sample Z to the complement set of E. Let $\rho_n = \rho_0/n^b$, where $b \in (0,1]$. Then

$$\min_{\mathbb{P}:\mathcal{W}_p(\mathbb{P},\mathbb{P}_n)\le\rho_n} \mathbb{P}(E) \simeq \mathbb{P}_n(E) - (p+1)^{\frac{1}{p+1}}\mathfrak{g}(0)^{\frac{p}{p+1}}\rho_n^{\frac{p}{p+1}}.$$

This theorem shows that the gap between the worst-case loss and the empirical loss is approximately equal to a penalty term proportional to $\mathfrak{g}(0)^{\frac{p}{p+1}}$ and $\rho_n^{\frac{p}{p+1}}$. Two remarks are in order.

(I) The term $\mathfrak{g}(0)$ measures the concentration of the underlying distribution $\mathbb{P}_{\mathsf{true}}$ around the inner boundary of E. Intuitively, to minimize the probability, the worst-case distribution tends to transport samples in E near the inner boundary out of E (see details in Section 2). Hence, a larger density value $\mathfrak{g}(0)$ around the inner boundary would imply that more samples close to the inner boundary of E can be transported out of E at low transport cost, leading to a larger gap between the worst-case loss and the empirical loss.

(II) The penalty term is a polynomial function of the radius ρ_n with an order $\frac{p}{p+1}$, which is in sharp contrast to the case of continuous loss where the penalty has a linear dependence on the radius [18]. As such, when ρ_n approaches zero, the penalty term for 0-1 loss is of a lower order (and thus larger) than the case of continuous loss functions. Intuitively, this is because a perturbation of a sample would lead to a jump in the loss value rather than a continuous change, resulting in a more significant gap between the worst-case loss and the empirical loss.

The result above not only provides a deeper understanding of the regularization effect of Wasserstein DRO for 0-1 loss, but also highlights a qualitative difference in terms of the radius selection compared to the case of continuous loss functions. Indeed, following the principle in [17], we choose ρ_n to be the smallest radius such that the resulting worst-case loss from (P) serves as a high-confidence bound on the true loss $\mathbb{P}_{\mathsf{true}}(E)$. In light of this, if we set $\rho_n = \rho_0/n^{\frac{p+1}{2p}}$, then our result states that

$$\min_{\mathbb{P}:\mathcal{W}_p(\mathbb{P},\mathbb{P}_n)\le\rho_n} \mathbb{P}(E) \;=\; \mathbb{P}_n(E) - (p+1)^{\frac{1}{p+1}}[\mathfrak{g}(0)]^{\frac{p}{p+1}}\rho_0^{\frac{p}{p+1}}/\sqrt{n} \;+\; \mathcal{O}_p(n^{-\frac{3}{4}}),$$

where the gap between the worst-case loss and the nominal loss $\mathbb{P}_n(E)$ is $\mathcal{O}_p(1/\sqrt{n})$. Thereby the worst-case loss can be served as a high-confidence (lower) bound of the true loss up to negligible high-order terms. Particularly, when $p=1$, we have $\rho_n \sim 1/n$. Our result also suggests that the choice $\rho_n \sim 1/\sqrt{n}$ would be overly conservative when $p \in [1,\infty)$, because the resulting confidence interval would have half-length $n^{-\frac{p}{2(p+1)}}$, wider than the standard concentration bounds such as Hoeffding's. In addition to the result presented in the previous informal theorem, we also consider other cases, including those with a fixed nominal distribution, often studied in adversarial learning, and the case where the decision boundary has a positive margin.

Overall, we develop a non-asymptotic statistical analysis framework for the regularization effect of Wasserstein DRO with 0-1 loss, which sheds light on a new radius selection rule that contrasts sharply with the continuous setting. In a nutshell, our main proof techniques are based on the worst-case analysis and tools from order statistics. We make use of the structure of the worst-case distribution (see [19] and Lemma 1 in Section 2) to show that the gap between the worst-case loss and the nominal loss depends mainly on (i) the value of the optimal dual Lagrangian multiplier when the nominal distribution is fixed, and (ii) the number of the points transported by the worst-case distribution when the nominal distribution is a (random) empirical distribution. Thereby, to bound the gap between robust loss and nominal loss, it amounts to bound the optimal dual Lagrangian multiplier or the number of the points transported. In our proof, these quantities are bounded using a careful analysis of the worst-case scenarios, leveraging tools from the concentration of measure for order statistics.

1.1. Related Literature

For Wasserstein DRO with 0-1 loss functions, general reformulations for uncertainty quantification problems are derived in [16, 6, 19] using duality arguments. Based on these dual representations, deterministic reformulations have been provided in several settings, such as chance-constrained programs (CCP) including linear CPP [50, 13, 30, 24, 25, 40, 41], nonlinear CCP [22, 27], CPP with binary decisions [50, 13, 30, 41]), CPP with the special nominal distributions [40, 42]; and machine learning problems including adversarial classification [26, 44], hypothesis testing [20, 49], safe reinforcement learning [32]; as well as application domains in energy management [15, 35, 11] and automatic control [39], etc. These works have been focused on developing efficient computational tools but not on the regularization effect of Wasserstein DRO.

Recently there has been an emerging interest in studying the connection between Wasserstein DRO and regularization. For specific problem settings, existing works have established the equivalence between Wasserstein DRO and regularization, such as logistic regression [38, 5], piecewise-linear convex losses [16], linear regression and classification and their kernelization [12, 37], support vector machines [5], etc. For general settings, a connection between Wasserstein DRO and gradient regularization for smooth loss functions was established in [18, 46], which were generalized by [4, 2] under weaker assumptions; a finer analysis of the asymptotic equivalence for 2-Wasserstein DRO was established in [8]. As mentioned earlier, all these results rely on the smoothness of the loss functions, which enables the Taylor expansion on each data point.

For the 0-1 loss, regularized reformulations developed in [50, 13, 24, 26] exploit the linearity in the description of the set E. Their equivalence is exact, but the gap depends on the optimal value of the dual Lagrangian multiplier, whose dependence on the sample size is unclear. In contrast, we develop a probabilistic bound on the gap between the robust loss and the empirical loss as a function of the radius and sample size for a general measurable set E.

We also remark that for other distributional uncertainty sets, the robust 0-1 loss can sometimes be interpreted as regularization in a broad sense as well. For example, for certain moment-based sets [21, 9, 47, 52, 53, 23, 51, 54, 34] or Wasserstein set centered at a Gaussian distribution [40], the distributionally robust (generalized) linear chance constraint can be equivalently represented as regularized mean using standard deviation. For ϕ-divergence-based sets, the distributionally robust chance constraint is equivalent to a nominal chance constraint with an adjusted safety level [28, 29, 31], which can also be viewed as regularization.

The paper proceeds as follows. In Section 2, we review the duality of Wasserstein DRO for 0-1 loss and the structure of the worst-case distribution. In Section 3, we study the regularization effect when the nominal distribution is fixed, which is a comparatively simple case yet provides essential insights. Next, in Section 4, we study the case when the nominal distribution is an empirical distribution constructed from i.i.d. samples. In Section 5, we discuss some applications of our results, especially their implications on the radius selection. We conclude the paper in Section 6. Proofs and auxiliary results are deferred to the Appendix.

2. Duality and Worst-case Distribution

In this section, we introduce notations and present the strong duality and the structure of worst-case distribution for (P).

Throughout this paper, let $p \in [1,\infty]$ and denote by $(\mathcal{Z}, \|\cdot\|) \subset \mathbb{R}^d$ a normed space and by $\mathcal{B}(\mathcal{Z})$ its Borel σ-algebra. Let $\|\cdot\|_*$ denote the dual norm. The interior of a set A is denoted by $\mathrm{int}(A)$. The support of a distribution $\mathbb{Q}$ is denoted by $\mathrm{supp}\,\mathbb{Q}$. We use $\mathcal{O}$ and $\mathcal{O}_p$ to represent the big O and the big O in probability notations, respectively. Let $\mathcal{P}(\mathcal{Z})$ denote the set of all Borel probability measures on $\mathcal{Z}$. We denote random variables with capital letters and the realization with small letters, e.g. Z and z. The Wasserstein distance of order p between distributions $\mathbb{P}, \mathbb{Q} \in \mathcal{P}(\mathcal{Z})$ is defined via

$$\mathcal{W}_p(\mathbb{P},\mathbb{Q})^p := \begin{cases} \inf_{\pi\in\mathcal{P}(\mathcal{Z}^2)} \left\{ \mathbb{E}_{(\tilde{Z},Z)\sim\pi}\left[\|\tilde{Z}-Z\|^p\right] : \pi \text{ has marginals } \mathbb{P},\mathbb{Q} \right\}, & p\in[1,\infty), \\ \inf_{\pi\in\mathcal{P}(\mathcal{Z}^2)} \left\{ \pi\text{-}\operatorname*{ess\,sup}_{\mathcal{Z}\times\mathcal{Z}} \|\tilde{z}-z\| : \pi \text{ has marginals } \mathbb{P},\mathbb{Q} \right\}, & p=\infty. \end{cases}$$

Let $\mathbb{Q}$ be a Borel probability measure on $\mathcal{Z}$, $\rho > 0$, and E be an open set in $\mathcal{Z}$. By strong duality for Wasserstein DRO (Lemma 2 in Appendix A), the problem (P) admits a strong dual problem

$$\begin{cases} \sup_{\lambda\geq 0}\left\{ -\lambda\rho^p + \mathbb{E}_{\mathbb{Q}}\left[\mathbf{1}_E(Z)\cdot\min(1,\lambda \mathrm{d}_{E^c}(Z)^p)\right]\right\}, & p\in[1,\infty), \\ \mathbb{Q}\left(\mathrm{d}_{E^c}(Z)>\rho\right), & p=\infty, \end{cases} \tag{D}$$

where $\mathrm{d}_{E^c}(\cdot)$ denotes the distance to the complement of the set E. When $p \in [1,\infty)$, the dual maximizer, denoted by λ_*, exists. The following proposition shows that the assumption of an open set E does not lose the generality, whose proof is given in Appendix EC.1.

Proposition 1. *Let $p \in [1,\infty]$. For any measurable subset $E \subset \mathcal{Z}$, it holds that*

$$\inf_{\mathbb{P}\in\mathcal{P}(\mathcal{Z}):\mathcal{W}_p(\mathbb{P},\mathbb{Q})\leq\rho} \mathbb{P}(E) = \min_{\mathbb{P}\in\mathcal{P}(\mathcal{Z}):\mathcal{W}_p(\mathbb{P},\mathbb{Q})\leq\rho} \mathbb{P}(\mathrm{int}(E)).$$

Throughout the paper, we assume $\mathbb{Q}(E) > 0$ without loss of generality; otherwise, the worst-case loss would be zero.

To describe the worst-case distribution, let us define the sets

$$\begin{aligned} \mathcal{Z}_* &:= \begin{cases} \left\{ z\in E\cap\mathrm{supp}\ \mathbb{Q} : 0<\mathrm{d}_{E^c}(z)\leq\lambda_*^{-1/p}\right\}, & p\in[1,\infty), \\ \{z\in E\cap\mathrm{supp}\ \mathbb{Q} : 0<\mathrm{d}_{E^c}(z)\leq\rho\}, & p=\infty. \end{cases} \\ \mathcal{Z}_= &:= \begin{cases} \left\{ z\in E\cap\mathrm{supp}\ \mathbb{Q} : \mathrm{d}_{E^c}(z)=\lambda_*^{-1/p}\right\}, & p\in[1,\infty), \\ \varnothing, & p=\infty, \end{cases} \end{aligned} \tag{Z}$$

which represent the sets of sample points transported in the worst case. Define two set-valued maps $\mathcal{T}_*, \mathcal{T}_- : \mathcal{Z} \rightrightarrows \mathcal{Z}$ as

$$\mathcal{T}_*(z) = \begin{cases} z, & z\in\mathcal{Z}_*^c, \\ \arg\min_{\tilde{z}\in E^c}\|\tilde{z}-z\|, & z\in\mathcal{Z}_*\setminus\mathcal{Z}_=, \\ z\cup\arg\min_{\tilde{z}\in E^c}\|\tilde{z}-z\|, & z\in\mathcal{Z}_=, \end{cases}$$

and

$$\mathcal{T}_-(z) = \begin{cases} z, & z\in\mathcal{Z}_*^c\cup\mathcal{Z}_=, \\ \arg\min_{\tilde{z}\in E^c}\|\tilde{z}-z\|, & z\in\mathcal{Z}_*\setminus\mathcal{Z}_=. \end{cases}$$

Then it follows from Lemma 3 in Appendix A that there exist $\mathbb{Q}$-measurable transport maps $T_*, T_- : \mathcal{Z}\to\mathcal{Z}$ that are measurable selections of $\mathcal{T}_*$ and $\mathcal{T}_-$ respectively. Recall that the pushforward $T_\#\mathbb{P}$ of a probability distribution $\mathbb{P}$ under a measurable map $T : \mathcal{Z}\to\mathcal{Z}$ is defined as $T_\#\mathbb{P}(A) := \mathbb{P}\{z\in\mathcal{Z} : T(z)\in A\}$, for every measurable set $A\subset\mathcal{Z}$. The following lemma characterizes the structure of the worst-case distribution, whose proof is given in Appendix EC.1.

LEMMA 1 **(Worst-case distribution)**.

(I) *When $p = \infty$ and when $p \in [1, \infty)$ with a dual optimizer $\lambda_* = 0$, let T_* be a measurable selection of $\mathcal{T}_*$. Then $\mathbb{P}_* := T_{*\#}\mathbb{Q}$ is a worst-case distribution with probability $\mathbb{P}_*(E) = \mathbb{Q}(E \setminus \mathcal{Z}_*)$.*

(II) *When $p \in [1, \infty)$ and all dual optimizers $\lambda_* > 0$, any worst-case transport plan γ_* satisfies $\rho^p = \mathbb{E}_{(Z,\tilde{Z})\sim\gamma_*}[\|\tilde{Z} - Z\|^p]$ and*

$$\{(z, \mathcal{T}_-(z)) : z \in \mathcal{Z}_*\} \subset \operatorname{supp} \gamma_* \subset \{(z, \mathcal{T}_*(z)) : z \in \mathcal{Z}_*\}.$$

Moreover, there exist $t^* \in [0, 1]$ and measurable selections T_* of $\mathcal{T}_*$ and T_- of $\mathcal{T}_-$ such that

$$\mathbb{P}_* := t_* T_{*\#}\mathbb{Q} + (1 - t_*) T_{-\#}\mathbb{Q}$$

is a worst-case distribution with probability $\mathbb{P}_*(E) = \mathbb{Q}(E \setminus \mathcal{Z}_*) + (1 - t_*)\mathbb{Q}(\mathcal{Z}_=)$.

(III) *Let $p \in [1, \infty)$ and set $\rho_{\max} := \left(\mathbb{E}_{\mathbb{Q}}\left[\mathbf{1}_E(Z)\mathrm{d}_{E^c}(Z)^p\right]\right)^{\frac{1}{p}}$. If $\rho < \rho_{\max}$, then every dual optimizer $\lambda_* > 0$; if $\rho \geq \rho_{\max}$, then $\lambda_* = 0$ is a dual optimizer.*

Lemma 1(I) and (II) separate two different situations: one corresponds to ∞-Wasserstein DRO or p-Wasserstein distance with an unbinding Wasserstein constraint, and the other corresponds to p-Wasserstein distance with a binding Wasserstein constraint. The former does not have mass splitting, whereas the latter may have mass splitting. Lemma 1(III) describes the condition for a binding Wasserstein constraint, i.e., $\lambda_* > 0$.

We illustrate this result in Figure 1. The support of the nominal distribution $\mathbb{Q}$ is demonstrated by diamond dots and shaded regions, representing point mass and continuous density, respectively. In the case descrbied in Lemma 1(I), the worst-case distribution $\mathbb{P}_*$ transports all points in $\mathcal{Z}_*$ to their respective nearest boundary points. In the case descrbied in Lemma 1(II), the worst-case distribution $\mathbb{P}_*$ transports all points in $\mathcal{Z}_* \setminus \mathcal{Z}_=$ with full probability to E^c and splits t_*-fraction of points in $\mathcal{Z}_=$ (as illustrated by $\widehat{\xi}$) to their respective nearest boundary points (as illustrated by ξ_*). If $t_* = 1$ or $\mathbb{Q}(\mathcal{Z}_=) = 0$, then the splitting at $\mathcal{Z}_=$ does not occur, and all points in $\mathcal{Z}_*$ are transported out in full probability to E^c.

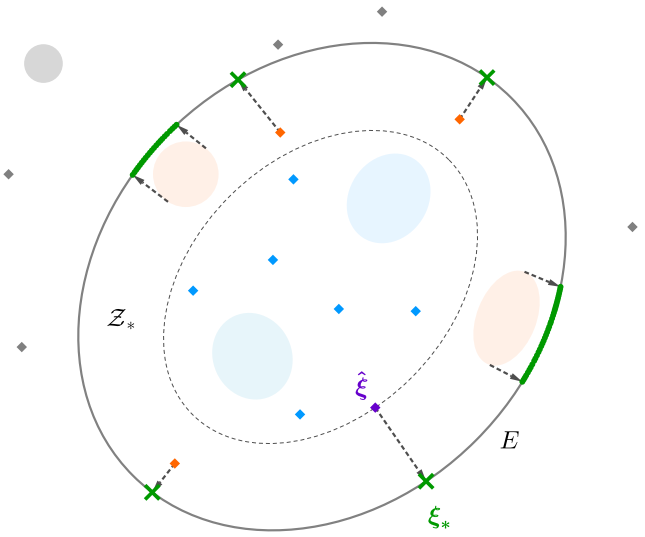

Figure 1 The worst-case distribution transports points (diamonds and shaded region) in $\mathcal{Z}_*$ to their nearest boundary points (crosses and curves), with the possible splitting of probability mass at $\mathcal{Z}_=$ (purple diamond).

3. Regularization Effect when the Nominal Distribution is Fixed

Define the *Wasserstein regularizer*

$$\mathcal{R}_{\mathbb{Q},p}(\rho) := \mathbb{Q}(E) - \min_{\mathbb{P}\in\mathcal{P}(\mathcal{Z}):\mathcal{W}_p(\mathbb{P},\mathbb{Q})\le\rho} \mathbb{P}(E),$$

that is, the difference between the nominal probability and the Wasserstein robust probability. In this section, we study the regularization effect of Wasserstein 0-1 loss when $\mathbb{Q}$ is a fixed distribution. This is commonly seen in *adversarial learning*, where $\mathbb{Q}$ represents the data distribution in the training environment, and the testing environment can be adversarially different from the training environment, captured by the worst-case distribution in the Wasserstein ball.

We start by defining the following function that plays a crucial role in our results. Let G be the conditional cumulative distribution function of the random variable $\mathsf{d}_{E^c}(Z)$ given that $\mathsf{d}_{E^c}(Z) > 0$, where Z is sampled from $\mathbb{Q}$:

$$G(s) := \mathbb{Q}\left(\mathsf{d}_{E^c}(Z) \le s \mid \mathsf{d}_{E^c}(Z) > 0\right) = \frac{\mathbb{Q}\left(0 < \mathsf{d}_{E^c}(Z) \le s\right)}{\mathbb{Q}(E)},$$

which describes how the probability in $\operatorname{supp}\mathbb{Q}\cap E$ is distributed in an s-margin from the boundary of E. Note that $G(s)$ is well-defined due to our assumptions that E is open and $\mathbb{Q}(E) > 0$.

3.1. ∞-Wasserstein Regularization

In this subsection, we study $p = \infty$. Using the definition of G, the Wasserstein regularizer $\mathcal{R}_{\mathbb{Q},\infty}(\rho)$ admits a simple expression by virtue of the structure of the worst-case distribution (Lemma 1(I)).

THEOREM 1. *Let $p = \infty$. Then*

$$\mathcal{R}_{\mathbb{Q},\infty}(\rho) = G(\rho)\mathbb{Q}(E).$$

Proof. Recall that $\mathcal{Z}_* = \{z \in E : \mathsf{d}_{E^c}(z) \le \rho\}$ as defined in (Z). According to Lemma 1(I), the worst-case probability is $\mathbb{Q}(E \setminus \mathcal{Z}_*)$. Then

$$\mathcal{R}_{\mathbb{Q},\infty}(\rho) = \mathbb{Q}(E) - \mathbb{Q}(E \setminus \mathcal{Z}_*) = \mathbb{Q}(\mathcal{Z}_*) = G(\rho)\mathbb{Q}(E),$$

where the last equality follows from the definition of the conditional distribution. □

In view of Theorem 1, to obtain a Taylor expansion of $\mathcal{R}_{\mathbb{Q},\infty}(\rho)$ with respect to ρ around zero, we need some kind of differentiability of the function G in a neighborhood of 0. In the sequel, whenever the derivative G' exists, we denote

$$g(s) := G'(s), \quad \mathfrak{g}(0) := g(0)\mathbb{Q}(E).$$

We impose the following assumption.

ASSUMPTION 1. *There exist constants $\delta > 0$ and $L_g > 0$ such that G is differentiable on $[0,\delta]$, $G(\delta) > 0$ and*

$$|g(x) - g(y)| \le L_g|x - y|, \quad \forall x, y \in [0,\delta].$$

The differentiability at 0 and δ refers to the one-sided differentiability and the condition $G(\delta) > 0$ excludes the case that G is constantly zero on $[0,\delta]$. Under Assumption 1, G admits a second order approximation $|G(y) - G(x) - g(x)(y-x)| \le \frac{L_g}{2}(x-y)^2$ (Lemma EC.1 in Appendix EC.2). It follows that

$$|\mathcal{R}_{\mathbb{Q},\infty}(\rho) - \mathfrak{g}(0)\cdot\rho| \le \frac{L_g}{2}\rho^2. \tag{1}$$

3.2. p-Wasserstein Regularization ($p \in [1,\infty)$)

In this subsection, we consider $p \in [1,\infty)$, which turns out to be qualitatively different from $p=\infty$. By (D) we have

$$\begin{aligned}\mathcal{R}_{\mathbb{Q},p}(\rho) &= \inf_{\lambda\geq 0}\left\{\lambda\rho^p + \mathbb{E}_{\mathbb{Q}}[\mathbf{1}_E(Z)\max(0,1-\lambda \mathsf{d}_{E^c}(Z)^p)]\right\}\\ &= \inf_{\lambda\geq 0}\left\{\lambda\rho^p + \mathbb{Q}(E)\int_0^{\lambda^{-1/p}}(1-\lambda s^p)dG(s)\right\}.\end{aligned}$$

The worst-case behavior depends crucially on the distance between $\operatorname{supp}\mathbb{Q}\cap E$ and the boundary of E. More specifically, define the *margin*

$$\underline{s} = \inf\{s>0:\ G(s)>0\}, \tag{2}$$

which represents the least distance between $\operatorname{supp}\mathbb{Q}\cap E$ and the boundary of E. By Assumption 1, the following quantity is well-defined:

$$\mathfrak{g}(\underline{s}) \ :=\ \lim_{s\downarrow\underline{s}}\frac{\mathbb{Q}\left(\underline{s}\leq \mathsf{d}_{E^c}(Z)\leq s\right)}{s-\underline{s}} = g(\underline{s})\mathbb{Q}(E),$$

representing the density of the random variable $\mathsf{d}_{E^c}(Z)$ at $\underline{s}$ (from right), where Z is distributed as $\mathbb{Q}$. Note that this definition is consistent with $\mathfrak{g}(0)$ in (1).

3.2.1. Positive Margin According to Lemma 1, for small ρ, when $\lambda_*^{-1/p} > \underline{s}$,

$$\rho^p \geq \mathbb{E}_{\mathbb{Q}}\left[\mathsf{d}_{E^c}(Z)^p\mathbf{1}_{\{\mathcal{Z}_*\setminus\mathcal{Z}_=\}}(Z)\right] \geq \underline{s}^p\mathbb{E}_{\mathbb{Q}}\left[1_{\{\mathcal{Z}_*\setminus\mathcal{Z}_=\}}(Z)\right] \geq \mathbb{Q}(E)\cdot\underline{s}^p(G(\lambda_*^{-1/p}-)-G(\underline{s})). \tag{3}$$

This indicates that when $\underline{s}>0$, $G(\lambda_*^{-1/p}-)\to G(\underline{s})$ as $\rho\to 0$ (and thus λ_* is away from zero as $\rho\to 0$). By the structure of the worst-case distribution (Lemma 1(II)), this means that if $\operatorname{supp}\mathbb{Q}\cap E$ has a positive margin $\underline{s}$, then as the radius ρ becomes small, only points near the $\underline{s}$-margin are transported in the worst case. Thereby

$$\mathcal{R}_{\mathbb{Q},p}(\rho) \simeq \mathbb{Q}(\mathcal{Z}_*) \simeq \mathbb{Q}(E)\cdot(G(\lambda_*^{-1/p}-)-G(\underline{s})) \overset{(3)}{\simeq} \underline{s}^{-p}\rho^p.$$

Note that for $\rho<\underline{s}$, $\underline{s}^{-p}\rho^p$ tends to 0 as p approaches to ∞, which is consistent with (1) as $\mathfrak{g}(0)=0$.

The discussion above is formalized in the following result under suitable smoothness conditions. We consider three cases: (I) G has point mass at $\underline{s}$, (II) G has positive density at $\underline{s}$, and (III) G has zero density at $\underline{s}$, respectively. The detailed proof is given in Appendix EC.2.1.

THEOREM 2. *Let $p\in[1,\infty)$. Suppose $\underline{s}>0$, and when $G(\underline{s})=0$, assume Assumption 1 holds.*

(I) *When $G(\underline{s})>0$, for any $\rho<[G(\underline{s})\mathbb{Q}(E)]^{\frac{1}{p}}\underline{s}$, it holds that*

$$\mathcal{R}_{\mathbb{Q},p}(\rho) = \underline{s}^{-p}\rho^p.$$

(II) *When $G(\underline{s})=0$ and $g(\underline{s})>0$, there exists a positive constant $\bar\rho$ dependent on $(E,\mathbb{Q},p)$ such that for any $\rho<\bar\rho$, it holds that*

$$\underline{s}^{-p}\rho^p - \frac{2p}{\underline{s}^{2p+1}\mathfrak{g}(\underline{s})}\rho^{2p} \ \leq\ \mathcal{R}_{\mathbb{Q},p}(\rho)\ \leq\ \underline{s}^{-p}\rho^p.$$

(III) *When $G(\underline{s})=0$ and $g(\underline{s})=0$, assume there exist constants $\tau\geq 1, c_{\underline{s}}>0$ such that $\liminf_{x\to 0_+}\frac{G(\underline{s}+x)}{x^\tau}=c_{\underline{s}}$. Then there exists a positive constant $\bar\rho$ dependent on $(E,\mathbb{Q},p)$ such that for any $\rho<\bar\rho$, it holds that*

$$\rho^p\underline{s}^{-p} - \frac{2^{1/\tau}p}{c_{\underline{s}}^{1/\tau}\underline{s}^{(1+1/\tau)p+1}\mathbb{Q}(E)^{1/\tau}}\rho^{(1+1/\tau)p} \ \leq\ \mathcal{R}_{\mathbb{Q},p}(\rho)\ \leq\ \underline{s}^{-p}\rho^p.$$

All three cases show that $\mathcal{R}_{\mathbb{Q},p}(\rho) \simeq \underline{s}^{-p}\rho^p$ when $\underline{s} > 0$.They differ mainly in the remainder term. In (I), for sufficiently small radius, by Lemma 1 a fraction of the point mass in $\mathcal{Z}_* = \mathcal{Z}_= = \{z \in E : \mathrm{d}_{E^c}(z) = \underline{s}\}$ is split and transported to its nearest boundary, resulting a linear decrease of probability in the worst case. In (II) and (III), when G is differentiable at $\underline{s}$, the smoothness conditions in the neighborhood of $\underline{s}$ ensure there is sufficient probability around $\underline{s}$ to be transported so that $\lambda_* \geq \underline{s}^{-p} - \mathcal{O}(\rho^a)$ for some $a > 0$.

3.2.2. Zero Margin Next, we consider the other case $\underline{s} = 0$. Consider the following motivating example.

EXAMPLE 1. Consider $\mathcal{Z} = ([-1,1], |\cdot|)$ and $E = (0, \frac{1}{2})$. Let $p \in [1,\infty)$. Suppose $\mathbb{Q}$ is uniform distribution on $[0,1]$. Then $\mathbb{Q}(E) = \frac{1}{2}$ and $G(s) = \mathbb{P}\left(Z \leq s \mid 0 < Z < \frac{1}{2}\right) = 2s$ for $0 \leq s < \frac{1}{2}$. It follows from (D) that

$$\begin{aligned}
\mathcal{R}_{\mathbb{Q},p}(\rho) &= \inf_{\lambda \geq 0} \left\{\lambda\rho^p + \mathbb{E}_{\mathbb{Q}}[\mathbf{1}_E(Z)\max(0, 1 - \lambda \mathrm{d}_{E^c}(Z)^p)]\right\} \\
&= \begin{cases} \frac{1}{2}, & \text{if } \lambda_* = 0, \\ \inf_{\lambda>0}\left\{\lambda\rho^p + \mathbb{Q}(E)\int_0^{\lambda^{-1/p}}(1-\lambda s^p)g(s)ds\right\}, & \text{if } \lambda_* > 0, \end{cases} \\
&= \begin{cases} \frac{1}{2}, & \text{if } \lambda_* = 0, \\ \inf_{\lambda>0}\left\{\lambda\rho^p + \int_0^{\lambda^{-1/p}}(1-\lambda s^p)ds\right\}, & \text{if } \lambda_* > 0, \end{cases} \\
&= \begin{cases} \frac{1}{2}, & \text{if } \lambda_* = 0, \\ (p+1)^{\frac{1}{p+1}}\rho^{\frac{p}{p+1}}, & \text{if } \lambda_* > 0. \end{cases}
\end{aligned}$$

Note that for any fixed ρ, $\mathcal{R}_{\mathbb{Q},p}(\rho)$ is monotonically decreasing with respect to p; see Lemma EC.3 in Appendix EC.2.2 for details. ♣

Example 1 shows that when $\rho \to 0$, the Wasserstein regularizer $\mathcal{R}_{\mathbb{Q},p}(\rho)$ is $\mathcal{O}(\rho^{\frac{p}{p+1}})$, which is smaller than the $\mathcal{O}(\rho)$-order for loss functions with certain modulus of continuity [18]. This is largely because, in the latter case, perturbation of data leads to a linear change in the value of the loss function, but in the case of 0-1 loss, the change is discrete, and thus Taylor expansion of the loss function used in the continuous losses does not apply to the 0-1 loss. As such, for 0-1 loss, $\mathcal{R}_{\mathbb{Q},p}(\rho)$ cannot be Lipschitz continuous at 0. Nonetheless, if the nominal distribution $\mathbb{Q}$ has certain smoothness property, $\mathcal{R}_{\mathbb{Q},p}(\rho)$ can be Hölder smooth at 0 as in Example 1. This intuition extends to more general cases, as in the following theorem, whose proof can be found in Appendix EC.2.3.

THEOREM 3. *Let $p \in [1,\infty)$. Suppose $\underline{s} = 0$. Assume Assumption 1 holds.*

(I) *When $g(0) > 0$, there exists a constant $\bar{\rho}$ dependent on $(E, \mathbb{Q}, p)$ such that for any $\rho < \bar{\rho}$,*

$$\left|\mathcal{R}_{\mathbb{Q},p}(\rho) - (p+1)^{\frac{1}{p+1}}\mathfrak{g}(0)^{\frac{p}{p+1}}\rho^{\frac{p}{p+1}}\right| \leq \frac{p}{p+2}2^{\frac{2}{p}-1}L_g(p+1)^{\frac{2}{p+1}}g(0)^{-\frac{2}{p+1}}\mathbb{Q}(E)^{\frac{p-1}{p+1}}\rho^{\frac{2p}{p+1}}.$$

(II) *When $g(0) = 0$, there exists a constant $\bar{\rho}$ dependent on $(E, \mathbb{Q}, p)$ such that for any $\rho < \bar{\rho}$,*

$$0 \leq \mathcal{R}_{\mathbb{Q},p}(\rho) \leq \left(\tfrac{p+2}{2}\right)^{\frac{2}{p+2}}(L_g\mathbb{Q}(E))^{\frac{p}{p+2}}\rho^{\frac{2p}{p+2}}.$$

Theorem 3 shows that when the margin $\underline{s} = 0$,

$$\mathcal{R}_{\mathbb{Q},p}(\rho) \simeq (p+1)^{\frac{1}{p+1}}\mathfrak{g}(0)^{\frac{p}{p+1}}\rho^{\frac{p}{p+1}},$$

The Hölder order $\frac{p}{p+1}$ increases as p increases and converges to 1 as $p \to \infty$, consistent with (1). The constant $\mathfrak{g}(0)$ indicates that the more probability in $\operatorname{supp}\mathbb{Q} \cap E$ concentrates around the boundary of E, the larger the Wasserstein regularizer is. This makes intuitive sense, because the perturbation of

points near the boundary leads to a jump in the value of the 0-1 loss, and more concentration on the boundary makes it easier to change the loss with a tiny perturbation.

Unlike Example 1 with a uniform nominal distribution $\mathbb{Q}$, in general cases, the dual optimizer λ_* cannot be solved explicitly. As such, the pivotal proof technique for Theorem 3 is to develop upper and lower bounds on the dual optimizer λ_*. To this end, in the proof we exploit the relationship between ρ and λ_*, Lemma 1 and (3). Using the approximation $\mathcal{R}_{\mathbb{Q},p}(\rho) \simeq \mathbb{Q}(\mathcal{Z}_*) \simeq \mathbb{Q}(E) \cdot G(\lambda_*^{-1/p} -)$ from Lemma 1, we can establish the lower and upper bounds on $\mathcal{R}_{\mathbb{Q},p}(\rho)$.

We close this section by summarizing our main findings on the approximation of $\mathcal{R}_{\mathbb{Q},p}(\rho)$ in Table 1. Note that the limiting cases as $p \to \infty$ in the last column are consistent with the case $p = \infty$.

Table 1 Expansion of $\mathcal{R}_{\mathbb{Q},p}(\rho)$ for a fixed nominal distribution $\mathbb{Q}$

	$p = \infty$	$p \in [1, \infty)$
Positive margin $\underline{s}$	$0 + \mathcal{O}(\rho^2)$ (Equation (1) with $\mathfrak{g}(0) = 0$)	$\underline{s}^{-p}\rho^p + \mathcal{O}(\rho^{(1+1/\tau)p})$ (Theorem 2)
Zero margin	$\mathfrak{g}(0)\rho + \mathcal{O}(\rho^2)$ (Equation (1))	$(p+1)^{\frac{1}{p+1}}\mathfrak{g}(0)^{\frac{p}{p+1}}\rho^{\frac{p}{p+1}} + \mathcal{O}(\rho^{\frac{2p}{p+1}})$ (Theorem 3(I))

4. Regularization Effect when the Nominal Distribution is Empirical

In this section, we consider the case where $\mathbb{Q}$ is an empirical distribution.

$$\mathbb{P}_n = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{\delta}_{z_i^n},$$

where z_i^n, $i = 1, \ldots, n$, are n i.i.d. samples from some underlying distribution $\mathbb{P}_{\mathsf{true}}$ and $\boldsymbol{\delta}_z$ denotes the Dirac point mass on z. In this case, we would choose a sample-size-dependent radius $\rho = \rho_n$. As we will see, intuitions in the case of a fixed nominal distribution are mostly extended to the case of an empirical distribution. Nonetheless, the randomness of the empirical distribution $\mathbb{P}_n$ makes the analysis much more complicated than the case of a fixed nominal distribution.

In light of the structure of the worst-case distribution (Lemma 1), let us assume without loss of generality that z_i^n, $i = 1, \ldots, n$, are ordered in a way such that $\mathsf{d}_{E^c}(z_1^n) \le \mathsf{d}_{E^c}(z_2^n) \le \cdots \le \mathsf{d}_{E^c}(z_n^n)$. Let I be the smallest index such that $\mathsf{d}_{E^c}(z_I^n) > 0$. Setting $N = n - I + 1$, then we have $N \sim \mathit{Binomial}(n, \mathbb{P}_{\mathsf{true}}(E))$. For $1 \le j \le N$, define $S_j := \mathsf{d}_{E^c}(z_{I+j-1}^n)$ and adopt the convention that $S_0 = 0$. By properties of order statistics, conditioning on N, $\{S_j\}_{j=1}^N$ are independent. With slight abuse of notation, let

$$G(s) := \mathbb{P}_{\mathsf{true}}\left(\mathsf{d}_{E^c}(Z) \le s \mid \mathsf{d}_{E^c}(Z) > 0\right),$$

where Z is sampled from $\mathbb{P}_{\mathsf{true}}$. Set

$$\mathfrak{g}(\underline{s}) \;=\; \lim_{s \downarrow \underline{s}} \frac{\mathbb{P}_{\mathsf{true}}\left(\underline{s} \le \mathsf{d}_{E^c}(Z) \le s\right)}{s - \underline{s}} = g(\underline{s})\mathbb{P}_{\mathsf{true}}(E),$$

where $\underline{s}$ is defined in a same way as in (2) and recall that $g = G'$ wherever G is differentiable.

In Section 4.1 and Section 4.2 below, we study cases of $p = \infty$ and $1 \le p < \infty$, respectively.

4.1. ∞-Wasserstein Regularization

In view of Lemma 1, define $0 \le J \le N$ to be such that

$$S_J \le \rho_n < S_{J+1}.$$

Namely, J is the number of sample points perturbed by the worst-case distribution, as specified by the set $\mathcal{Z}_*$ defined in (Z). Note that $\mathcal{Z}_*$ has been exploited in the proof of Theorem 1, the case of a fixed nominal distribution. We have that $G(S_J) \le G(\rho_n) < G(S_{J+1})$ and $J \sim Binomial\,(n, \mathbb{P}_{\mathsf{true}}(E)G(\rho_n))$. Using Bernstein's inequality, we can obtain

$$J = n\mathbb{P}_{\mathsf{true}}(E)G(\rho_n) + \mathcal{O}_p(n^{\frac{1}{4}}).$$

Meanwhile, by Lemma 1,

$$\frac{J}{n} \le \mathcal{R}_{\mathbb{P}_n,\infty}(\rho_n) < \frac{J+1}{n}.$$

Then the confidence interval of J implies the confidence interval of $\mathcal{R}_{\mathbb{P}_n,\infty}(\rho_n)$. The discussion above is formalized in the following result; see Appendix EC.3.1 for a detailed proof.

THEOREM 4. *Let $\rho_n = \rho_0/\sqrt{n}$ with $\rho_0 > 0$. Assume Assumption 1 holds. Let $t > 0$. Then there exist constants n_0 and C both dependent on $(E, \mathbb{P}_{\mathsf{true}}, \rho_0, t)$ such that for every $n > n_0$, it holds with probability at least $1 - 2\exp\left(-\frac{t^2}{2\mathfrak{g}(0)\rho_0+1}\right)$ that*

$$\left|\mathcal{R}_{\mathbb{P}_n,\infty}(\rho_n) - \mathfrak{g}(0)\rho_n\right| \le \frac{C}{n^{3/4}}.$$

Moreover, when $g(0) = 0$, there exist constants n_0 and C dependent on $(E, \mathbb{P}_{\mathsf{true}}, \rho_0, t)$ such that for any $n > n_0$, it holds with probability at least $1 - 2\exp\left(-\frac{3t^2}{3\mathbb{P}_{\mathsf{true}}(E)L_g\rho_0^2+2t}\right)$ that

$$\mathcal{R}_{\mathbb{P}_n,\infty}(\rho_n) \le \frac{C}{n}.$$

The first-order approximation $\mathfrak{g}(0)\rho_n$ has the same form as in (1) for a fixed nominal distribution. When $\mathfrak{g}(0) = 0$, the remainders in (1) and Theorem 4 are both $\mathcal{O}(\rho_n^2)$. However, when $\mathfrak{g}(0) > 0$, the remainder in Theorem 4 is $\mathcal{O}(\rho_n^{\frac{3}{2}})$, which is larger than $\mathcal{O}(\rho_n^2)$ remainder in (1). This is because randomness on the empirical distribution $\mathbb{P}_n$ leads to randomness on the number J of points being perturbed by the worst-case distribution, and the standard deviation of J results in an extra $\mathcal{O}(n^{-3/4})$ term in the remainder.

The following result shows that the order of the remainder is tight, whose detailed proof can be found in Appendix EC.3.2.

PROPOSITION 2. *Let $p = \infty$ and $\rho_n = \rho_0/\sqrt{n}$ with $\rho_0 > 0$. Assume that Assumption 1 holds with $g(0) > 0$. Let $t > 0$. Then there exists a constant C dependent on $(E, \mathbb{P}_{\mathsf{true}}, \rho_0, t)$ such that it holds with probability at least $\frac{2\sqrt{2}}{\sqrt{\pi}(\sqrt{t^2+4}+t)}e^{-\frac{t^2}{2}}$ that*

$$\lim_{n\to\infty} \left|\mathcal{R}_{\mathbb{P}_n,\infty}(\rho_n) - \mathfrak{g}(0)\rho_n\right| - \frac{C}{n^{3/4}} \ge 0.$$

This result shows that with a positive probability, the remainder is $O(n^{-3/4})$, and thereby the bound in Theorem 1 cannot be improved in general.

4.2. p-Wasserstein Regularization ($p \in [1,\infty)$)

The analysis for the case of $p \in [1,\infty)$ is more involved. Recall from Section 3.2 that in the case of a fixed nominal distribution, the crux of our analysis was to develop lower and upper bounds on the dual optimizer λ_* that specifies $\mathcal{Z}_*$ as defined in (Z). When the nominal distribution becomes an empirical distribution $\mathbb{P}_n$ subjective to randomness due to sampling, the dual optimizer λ_* becomes random as well. This creates additional challenges in deriving bounds for λ_*.

Define

$$U_0 = 0, \quad U_j = G(S_j),\ j = 1,\dots,N.$$

Throughout this subsection, we assume that G is a continuous function; note that this is a weaker condition than assuming G to be the cumulative distribution function of a continuous random variable. By [10, Theorem 2.1.10], $U_j \sim Beta(j, N+1-j)$, $j = 1,\dots,N$. Set $0 \le J \le N$ to be such that

$$\frac{1}{n}\sum_{j=0}^{J} S_j^p \le \rho_n^p < \frac{1}{n}\sum_{j=0}^{J+1} S_j^p. \tag{4}$$

Here the random variable J specifies the points in $\operatorname{supp}\mathbb{P}_n$ that are transported by the worst-case distribution and thereby specifies $\mathcal{Z}_*$ as defined in (Z). From this aspect, it plays a similar role as the dual optimizer λ_* as in our analysis in the previous section. By Lemma 1, it follows that

$$\frac{J}{n} \le \mathcal{R}_{\mathbb{P}_n,p}(\rho_n) < \frac{J+1}{n}. \tag{5}$$

Recall from (2) and our definition of G in this section that

$$\underline{s} = \inf\{s>0:\ G(s)>0\} = \inf\{s>0:\ \mathbb{P}_{\mathsf{true}}(\mathsf{d}_{E^c}(Z)\le s \mid \mathsf{d}_{E^c}(Z)>0)>0\}.$$

Similar to the case of a fixed nominal distribution, we separate the cases $\underline{s}>0$ and $\underline{s}=0$. We will consider a general radius selection rule $\rho_n = \rho_0 n^{-b}$ for some $\rho_0>0$ and $0<b\le 1$ in this subsection.

4.2.1. Positive Margin We first consider the case $\underline{s}>0$. The idea of our analysis is sketched below.

When $G(\underline{s})>0$, there exist $\mathcal{O}_p(n)$ samples whose distance to the boundary of E is $\underline{s}$. So it would cost $\mathcal{O}(1)$ distance to transport all these points to the boundary of E. As such, for $\rho_n = O(n^{-b})$, $b\in(0,1]$, with high probability, all points that are transported in the worst-case distribution would have $\underline{s}$-distance away from the boundary of E, namely, $S_j = \underline{s}$ for $0\le j\le J$. Thereby, using (4) and(5), we have

$$\mathcal{R}_{\mathbb{P}_n,p}(\rho_n) \simeq \underline{s}^{-p}\rho_n^p.$$

When $G(\underline{s})=0$, using $U_j\sim Beta(j,N+1-j)$, the confidence interval of $U_j = G(S_j)$ has the form $\frac{j}{N+1} \pm t\sqrt{\frac{j(N+1-j)}{(N+1)^2(N+2)}}$. Under smoothness assumption on G^{-1}, we can derive the confidence interval of S_j as

$$\begin{aligned} G^{-1}\left(\tfrac{j}{N+1} \pm t\sqrt{\tfrac{j(N+1-j)}{(N+1)^2(N+2)}}\right) &\simeq G^{-1}\left(\tfrac{j}{N+1}\right) \pm \mathcal{O}_p\left(t\sqrt{\tfrac{j(N+1-j)}{(N+1)^2(N+2)}}\right) \\ &\simeq \underline{s} + \mathcal{O}_p(\tfrac{j}{N+1}) \pm \mathcal{O}_p\left(t\sqrt{\tfrac{j(N+1-j)}{(N+1)^2(N+2)}}\right). \end{aligned} \tag{6}$$

It follows that

$$\begin{aligned} \frac{1}{n}\sum_{j=0}^{J} S_j^p &\simeq \frac{1}{n}\sum_{j=0}^{J}\left(\underline{s} + \mathcal{O}_p(\tfrac{j}{N+1}) \pm \mathcal{O}_p\left(t\sqrt{\tfrac{j(N+1-j)}{(N+1)^2(N+2)}}\right)\right)^p \\ &= \frac{J}{n}\underline{s}^p + \mathcal{O}_p(1). \end{aligned} \tag{7}$$

Thereby, using (4) we can derive a confidence bound on J, which, by (5), further renders a confidence bound on $\mathcal{R}_{\mathbb{P}_n,p}(\rho_n)$

$$\mathcal{R}_{\mathbb{P}_n,p}(\rho_n) \simeq \underline{s}^{-p}\rho_n^p.$$

The analysis above is rigorously stated in the following result, whose proof is given in Appendix EC.4.1.

THEOREM 5. *Let $p \in [1,\infty)$ and $\rho_n = \rho_0/n^b$ with $\rho_0 > 0$, where $0 < b \le 1$. Suppose $\underline{s} > 0$.*

(I) *When $G(\underline{s}) > 0$, let $t > 0$. Then there exists a constant n_0 dependent on $(E, \mathbb{P}_{\mathsf{true}}, p, b, t)$ such that for any $n \ge n_0$, with probability at least $1 - \exp\left(-\frac{t^2}{2\mathbb{P}_{\mathsf{true}}(E)G(\underline{s})(1-\mathbb{P}_{\mathsf{true}}(E)G(\underline{s}))+\frac{2}{3}}\right)$, it holds that*

$$\left|\mathcal{R}_{\mathbb{P}_n,p}(\rho_n) - \underline{s}^{-p}\rho_n^p\right| < \frac{1}{n}.$$

(II) *When $G(\underline{s}) = 0$ and $g(\underline{s}) > 0$, assume Assumption 1 holds. Let $t_1, t_2 > 0$. Then there exist constants n_0 and C both dependent on $(E, \mathbb{P}_{\mathsf{true}}, p, \rho_0, b, t_1, t_2)$ such that for any $n \ge n_0$, with probability at least $1 - n\exp\left(-\frac{t_1^2}{2+4t_1}\right) - 2\exp(-2t_2^2)$, it holds that*

$$\left|\mathcal{R}_{\mathbb{P}_n,p}(\rho_n) - \underline{s}^{-p}\rho_n^p\right| < \frac{C}{n^{\min\left(1,2bp,\frac{3bp+1}{2}\right)}}.$$

(III) *When $G(\underline{s}) = 0$ and $g(\underline{s}) = 0$, assume there exist constants $\tau, c_{\underline{s}} > 0$ such that $\liminf_{x\to 0_+} \frac{G(\underline{s}+x)}{x^\tau} = c_{\underline{s}}$. Then there exist constants n_0 and C both dependent on $(E, \mathbb{P}_{\mathsf{true}}, p, \rho_0, b, t_1, t_2)$ such that for any $n \ge n_0$, with probability at least $1 - n\exp\left(-\frac{t_1^2}{2+4t_1}\right) - 2\exp(-2t_2^2)$, it holds that*

$$\left|\mathcal{R}_{\mathbb{P}_n,p}(\rho_n) - \underline{s}^{-p}\rho_n^p\right| < \frac{C}{n^{\min\left(1,(1+\frac{1}{\tau})bp,\frac{1}{\tau}+bp,\frac{1}{2}+bp\right)}}.$$

Compared to Theorem 2 for a fixed nominal distribution, the dominating term $\underline{s}^{-p}\rho_n^p$ in Theorem 5 is of the same form, yet the remainders are not entirely the same. For all three cases (I)(II)(III) of Theorem 5, the $1/n$ term in the remainder—which does not exist in Theorem 2—comes from the discrete structure of the worst-case distribution; see Lemma 1 and (4) (5). For the case (II) of Theorem 5, the remainder term $n^{-2bp} = \mathcal{O}(\rho_n^{2p})$ is consistent with the remainder term ρ^{2p} in Theorem 2(II), and the extra remainder term $n^{-\frac{3bp+1}{2}}$ comes from the higher-order expansion in (7), as a result of the randomness in U_j due to sampling. For the case (III) of Theorem 5, the remainder term $n^{-(1+\frac{1}{\tau})bp} = \mathcal{O}(\rho_n^{(1+\frac{1}{\tau})p})$ is consistent with the remainder term $\rho^{(1+\frac{1}{\tau})p}$ in Theorem 2(III), and the extra remainder term also results from the randomness of U_j due to sampling.

4.2.2. Zero Margin Next, we consider the case $\underline{s} = 0$. Similar to (6), we derive the confidence interval of S_j as

$$\begin{aligned} G^{-1}\left(\tfrac{j}{N+1} \pm t\sqrt{\tfrac{j(N+1-j)}{(N+1)^2(N+2)}}\right) &\simeq G^{-1}\left(\tfrac{j}{N+1}\right) \pm \mathcal{O}_p\left(t\sqrt{\tfrac{j(N+1-j)}{(N+1)^2(N+2)}}\right) \\ &\simeq \tfrac{j}{g(0)(N+1)} + \mathcal{O}_p\left(\tfrac{j^2}{(N+1)^2}\right) + \mathcal{O}_p\left(t\sqrt{\tfrac{j(N+1-j)}{(N+1)^2(N+2)}}\right). \end{aligned}$$

Then it follows that

$$\begin{aligned} \frac{1}{n}\sum_{j=0}^{J} S_j^p &\simeq \frac{1}{n}\sum_{j=0}^{J}\left(\tfrac{j}{g(0)(N+1)} + \mathcal{O}_p(\tfrac{j^2}{(N+1)^2}) \pm \mathcal{O}_p\left(t\sqrt{\tfrac{j(N+1-j)}{(N+1)^2(N+2)}}\right)\right)^p \\ &\simeq \frac{J^{p+1}}{g(0)^p n(N+1)^p} + o_p(1) \\ &\simeq \frac{J^{p+1}}{(p+1)\left(g(0)\mathbb{P}_{\mathsf{true}}(E)\right)^p n^{p+1}} + o_p(1). \end{aligned} \tag{8}$$

Thereby, using (4) we can derive a confidence bound on J which, by (5), further renders a confidence bound on $\mathcal{R}_{\mathbb{P}_n,p}(\rho_n)$

$$\mathcal{R}_{\mathbb{P}_n,p}(\rho_n) \simeq (p+1)^{\frac{1}{p+1}}\mathfrak{g}(0)^{\frac{p}{p+1}}\rho_n^{\frac{p}{p+1}}.$$

The following theorem formalizes the result above; see Appendix EC.4.2 for a detailed proof.

THEOREM 6. *Let $p \in [1,\infty)$. Suppose $\underline{s} = 0$ and $\rho_n = \rho_0/n^b$ for some $\rho_0 > 0$ and $0 < b \le 1$. Assume Assumption 1 holds. Let $t_1, t_2 > 0$. Then with probability at least $1 - n\exp(-\frac{t_1^2}{2+4t_1}) - 2\exp(-2t_2^2)$, the following holds:*

(I) *When $g(0) > 0$, there exist constants n_0 and C both dependent on $(E, \mathbb{P}_{\mathsf{true}}, p, \rho_0, t_1, t_2)$ such that for any $n > n_0$,*

$$\left| \mathcal{R}_{\mathbb{P}_n,p}(\rho_n) - (p+1)^{\frac{1}{p+1}} \mathfrak{g}(0)^{\frac{p}{p+1}} \rho_n^{\frac{p}{p+1}} \right| \le \frac{C}{n^{\min\left(1, \frac{2bp}{p+1}, \frac{p+bp+1}{2(p+1)}\right)}}.$$

(II) *When $g(0) = 0$ and $0 < b < \frac{p+2}{2p}$, there exists a positive integer n_0 dependent on $(E, \mathbb{P}_{\mathsf{true}}, p, \rho_0, b, t_1, t_2)$ such that for any $n > n_0$,*

$$\mathcal{R}_{\mathbb{P}_n,p}(\rho_n) < \left(\tfrac{p+2}{2}\right)^{\frac{2}{p+2}} \left(\mathbb{P}_{\mathsf{true}}(E) L_{G'} \rho_0^2\right)^{\frac{p}{p+2}} \frac{1}{n^{\frac{2bp}{p+2}}} + \frac{1}{n}.$$

The regularization term $(p+1)^{\frac{1}{p+1}} [\mathfrak{g}(0)]^{\frac{p}{p+1}} \rho_n^{\frac{p}{p+1}}$ is consistent with Theorem 3 for a fixed nominal distribution. Moverover, the remainder terms $n^{-\frac{2bp}{p+1}} = \mathcal{O}(\rho_n^{\frac{2p}{p+1}})$ in (I) and $n^{-\frac{2bp}{p+1}} = \mathcal{O}(\rho_n^{\frac{2p}{p+1}})$ in (II) of Theorem 6 are also consistent with the remainders in Theorem 3(I) and (II), respectively. The extra remainder term $1/n$ in both (I) and (II) comes from the discrete structure of the worst-case distribution (Lemma 1 and (4)(5)), similar to Theorem 5. The extra remainder term $n^{-\frac{p+bp+1}{2(p+1)}}$ is from the higher-order remainder in (8), coming from the randomness of U_j due to sampling.

In both Theorems 5 and 6, we did not specify the scaling of the radius ρ_n, but allowed b to be a free parameter. In Section 5.2 below, we will discuss the implication of these results on the radius selection.

We close this section by summarizing our main findings on the approximation of $\mathcal{R}_{\mathbb{P}_n,p}(\rho_n)$ in Table 2, where the scaling factor b of the radius ρ_n is chosen as the values specified in Section 5.2. The dominating terms are consistent with those in Table 1, whereas the remainders have a lower order, resulting from the sampling randomness of the empirical nominal distribution.

Table 2 Expansion of $\mathcal{R}_{\mathbb{P}_n,p}(\rho_n)$

	$p = \infty$	$p \in [1,\infty)$
Positive margin $\underline{s}$	$0 + \mathcal{O}(n^{-\frac{3}{4}})$ (Theorem 4 with $\mathfrak{g}(0) = 0$)	$\underline{s}^{-p}\rho_n^p + \mathcal{O}(n^{-\min(1,\frac{1}{2}+\frac{1}{2\tau})})$ (Theorem 5 with $\rho_n = n^{-\frac{1}{2p}}$ and $\rho_n = n^{-1/2}$)
Zero margin	$\mathfrak{g}(0)\rho_n + \mathcal{O}(n^{-\frac{3}{4}})$ (Theorem 4 with $\rho_n = n^{-1/2}$)	$(p+1)^{\frac{1}{p+1}}\mathfrak{g}(0)^{\frac{p}{p+1}}\rho_n^{\frac{p}{p+1}} + \mathcal{O}(n^{-\frac{3}{4}})$ (Theorem 6 with $\rho_n = n^{-\frac{p+1}{2p}}$)

5. Applications

In this section, we discuss two applications of our theoretical results. In Section 5.1, we exemplify our results for linear classification. In Section 5.2, we highlight an implication of our result on choosing the radius for the Wasserstein ball.

5.1. Classification and Margin Regularization

In this subsection, we illustrate our result in the context of binary classification. Let $p \in [1,\infty)$, $\mathcal{Z} = \mathcal{X} \times \{\pm 1\}$, where $\mathcal{X} \subset \mathbb{R}^d$, and $\mathbb{P}_n = \frac{1}{n}\sum_{i=1}^n \boldsymbol{\delta}_{z_i^n}$, where the samples $z_i^n = (x_i^n, y_i^n)$, $i = 1, \cdots, n$, are generated from some underlying distribution $\mathbb{P}_{\mathsf{true}}$. Consider a robust binary classification problem that minimizes the worst-case mis-classification error

$$\min_{\theta\in\Theta} \sup_{\mathbb{P}:\mathcal{W}_p(\mathbb{P},\mathbb{P}_n)\le\rho_n} \mathbb{P}(Y\cdot F_\theta(X)\le 0),$$

or equivalently,

$$\min_{\theta\in\Theta}\left\{1 - \inf_{\mathbb{P}:\mathcal{W}_p(\mathbb{P},\mathbb{P}_n)\le\rho_n} \mathbb{P}(Y\cdot F_\theta(X)>0)\right\}.$$

Suppose there is a norm on $\mathcal{Z}$ defined as $\|z-\widehat{z}\| = \|x-\widehat{x}\|_2 + \infty\cdot|y-\widehat{y}|$, where $z=(x,y), \widehat{z}=(\widehat{x},\widehat{y}) \in \mathcal{Z}$, that is, there is no label uncertainty. Then any distribution $\mathbb{P}$ in the Wasserstein ball satisfies $\mathbb{P}(Y = y) = \mathbb{P}_n(Y=y) := \alpha_{n,y}$, $y = \pm 1$. In addition, let $\mathbb{P}_{X|Y}$ be the conditional distribution of X given Y, then

$$\mathcal{W}_p(\mathbb{P},\mathbb{P}_n)^p = \alpha_{n,1}\mathcal{W}_p(\mathbb{P}_{X|Y=1},\mathbb{P}_{n,X|Y=1})^p + \alpha_{n,-1}\mathcal{W}_p(\mathbb{P}_{X|Y=1},\mathbb{P}_{n,X|Y=-1})^p.$$

It follows that the classification accuracy equals

$$\begin{aligned}
&\inf_{\mathbb{P}:\mathcal{W}_p(\mathbb{P},\mathbb{P}_n)\le\rho_n} \mathbb{P}\big(Y\cdot F_\theta(X)>0\big)\\
&= \inf_{\rho_{n,1}^p+\rho_{n,-1}^p\le\rho_n^p} \sum_{y=\pm1} \alpha_{n,y}\cdot \inf_{\mathbb{P}:\mathcal{W}_p(\mathbb{P}_{X|Y=y},\mathbb{P}_{n,X|Y=y})\le\rho_{n,y}} \mathbb{P}\big(Y\cdot F_\theta(X)>0 \mid Y=y\big).
\end{aligned}$$

Define a set $E_{\theta,y} = \{x\in\mathcal{X}:\, yF_\theta(x)>0\}$. Suppose F_θ is not a constant function. $\mathbb{P}_{\mathsf{true},X|Y=y}$ has positive density everywhere on $\mathcal{X}$, then $\underline{s}=0$. Using Theorem 6, we obtain that

$$\inf_{\mathbb{P}:\mathcal{W}_p(\mathbb{P}_{X|Y=y},\mathbb{P}_{n,X|Y=y})\le\rho_{n,y}} \mathbb{P}\big(Y\cdot F_\theta(X)>0\mid Y=y\big) \simeq \mathbb{P}_n\big(Y\cdot F_\theta(X)>0\mid Y=y\big) - (p+1)^{\frac{1}{p+1}}\mathfrak{g}_{\theta,y}(0)^{\frac{p}{p+1}}\rho_{n,y}^{\frac{p}{p+1}},$$

where

$$\mathfrak{g}_{\theta,y}(0) := \lim_{s\downarrow 0}\frac{\mathbb{P}_{\mathsf{true},X|Y=y}\big(X\in E_{\theta,y},\, 0<\mathsf{d}_{E_{\theta,y}^c}(X)\le s\big)}{s}.$$

As a result,

$$\begin{aligned}
&\sup_{\mathbb{P}:\mathcal{W}_p(\mathbb{P},\mathbb{P}_n)\le\rho_n} \mathbb{P}\big(Y\cdot F_\theta(X)\le 0\big)\\
&\simeq \mathbb{P}_n\big(Y\cdot F_\theta(X)\le 0\big) + (p+1)^{\frac{1}{p+1}} \sup_{\rho_{n,1}^p+\rho_{n,-1}^p\le\rho_n^p}\left\{\alpha_{n,1}\mathfrak{g}_{\theta,1}(0)^{\frac{p}{p+1}}\rho_{n,1}^{\frac{p}{p+1}} + \alpha_{n,-1}\mathfrak{g}_{\theta,-1}(0)^{\frac{p}{p+1}}\rho_{n,-1}^{\frac{p}{p+1}}\right\}\\
&= \mathbb{P}_n\big(Y\cdot F_\theta(X)\le 0\big) + (p+1)^{\frac{1}{p+1}}\left(\mathfrak{g}_{\theta,1}(0)\alpha_{n,1}^{\frac{p+1}{p}} + \mathfrak{g}_{\theta,-1}(0)\alpha_{n,-1}^{\frac{p+1}{p}}\right)^{\frac{p}{p+1}}\rho_n^{\frac{p}{p+1}},
\end{aligned}$$

where the last equality follows from Hölder's inequality. Therefore, we arrive at the following regularization form

$$\begin{aligned}
&\inf_{\theta\in\Theta}\sup_{\mathbb{P}:\mathcal{W}_p(\mathbb{P},\mathbb{P}_n)\le\rho_n} \mathbb{P}\big(Y\cdot F_\theta(X)\le 0\big)\\
&\simeq \inf_{\theta\in\Theta}\left\{\mathbb{P}_n\big(Y\cdot F_\theta(X)\le 0\big) + (p+1)^{\frac{1}{p+1}}\left(\mathfrak{g}_{\theta,1}(0)\alpha_{n,1}^{\frac{p+1}{p}} + \mathfrak{g}_{\theta,-1}(0)\alpha_{n,-1}^{\frac{p+1}{p}}\right)^{\frac{p}{p+1}}\rho_n^{\frac{p}{p+1}}\right\}.
\end{aligned}$$

From the perspective of regularization, a large density $\mathfrak{g}_{\theta,y}(0)$ implies a small empirical margin around the decision boundary, and therefore the corresponding Wasserstein robust decision optimization would penalize decisions with a small margin.

Next, we provide further analysis in the case of linear classification,

$$F_\theta(x) = w^\top x + r,\ x \in \mathcal{X},\ \text{where } \theta = (w,r) \in \Theta = \mathbb{R}^{d+1}. \tag{9}$$

When $w = 0$, since there is no label uncertainty it holds that

$$\sup_{\mathbb{P}:\mathcal{W}_p(\mathbb{P},\mathbb{P}_n)\le\rho_n} \mathbb{P}\big(Y\cdot r \le 0\big) \simeq \mathbb{P}_n\big(Y\cdot r \le 0\big).$$

Next, we consider the case $w \neq 0$. We introduce the orthogonal decomposition $\mathbb{R}^d = \text{span}\{w\} \oplus V$, where V is the orthogonal complement of $\text{span}\{w\}$. Let $(e_1, \cdots, e_{d-1})$ be an orthonormal basis of V. Then for any $x \in \mathbb{R}^d$, there exist constants $t, v = (v_1, v_2, \cdots, v_{d-1})$ such that

$$x = t\frac{w}{\|w\|_2} + \sum_{k=1}^{d-1} v_k e_k,$$

where $t = w^\top x/\|w\|_2$ and $v_i = e_k^\top x$ for $k = 1, 2, \cdots, d-1$. Denote the density of $\mathbb{P}_{\text{true},X|Y=y}$ as f_y, $y = \pm 1$. With some analysis as detailed in Appendix EC.5, we have the following approximation of the problem (9)

$$\begin{aligned}
&\inf_{(w,r)\in\mathbb{R}^{d+1}} \sup_{\mathbb{P}:\mathcal{W}_p(\mathbb{P},\mathbb{P}_n)\le\rho_n} \mathbb{P}\big(Y\cdot(w^\top X + r) \le 0\big) \\
&\simeq \inf_{(w,r)\in\mathbb{R}^{d+1}} \left\{ \mathbb{P}_n\big(Y\cdot(w^\top X + r) \le 0\big) + (p+1)^{\frac{1}{p+1}} \left(\mathfrak{g}_{w,r,1}(0)\alpha_{n,1}^{\frac{p+1}{p}} + \mathfrak{g}_{w,r,-1}(0)\alpha_{n,-1}^{\frac{p+1}{p}} \right)^{\frac{p}{p+1}} \rho_n^{\frac{p}{p+1}} \right\}.
\end{aligned} \tag{10}$$

where we have $\mathfrak{g}_{0,r,y}(0) = 0$, and

$$\mathfrak{g}_{w,r,y}(0) = \int_{\mathbb{R}^{d-1}} f_y\Big(-rw/\|w\|_2^2 + \sum_{k=1}^{d-1} v_k e_k \Big) dv, \quad w \neq 0.$$

When $r \neq 0$, since f_y is a density function, $f_y(-\infty) = 0$, by dominated convergence, we have that

$$\lim_{\|w\|\to 0} \mathfrak{g}_{w,r,y}(0) = \int_{v\in\mathbb{R}^{d-1}} \lim_{\|w\|\to 0} f_y\Big(-r\frac{w}{\|w\|_2}\frac{1}{\|w\|_2} + \sum_{k=1}^{d-1} v_k e_k \Big) dv = 0.$$

Hence, the regularization term $\mathfrak{g}_{w,r,y}(0)$ encourages a solution with small norm. In [26] it is shown that the problem (9) is equivalent to empirical hinge loss minimization with norm regularization. Different from this result, (10) directly compares the worst-case classification error and the empirical classification error.

5.2. Implication on the Radius Selection

In this subsection, we discuss the implication of our results in Section 4 on the choice of radius ρ_n in a data-driven setting. Particularly, we will focus on the convergence rate of ρ_n with respect to the sample size n, highlighting a qualitative difference in radius selection between 0-1 loss and continuous loss functions.

Our principle for radius selection is consistent with [17], namely, choosing the smallest radius such that the resulting worst-case loss from (P) serves as a high-confidence bound on the true loss $\mathbb{P}_{\text{true}}(E)$. By standard concentration bounds such as Hoeffding's inequality, the difference between the true loss $\mathbb{P}_{\text{true}}(E)$ and the nominal loss $\mathbb{P}_n(E)$ is $\mathcal{O}_p(1/\sqrt{n})$. As a result, if we choose ρ_n such that $\mathcal{R}(\rho_n)$ is of the order $1/\sqrt{n}$, then Theorems 5 and 6 would ensure the worst-case loss to be a high-confidence bound of the true loss up to negligible high-order terms. We have the following results.

COROLLARY 1. *Let $p \in [1,\infty)$. Under the setup of Theorem 5, suppose $b = \frac{1}{2p}$. Then*

$$\left| \mathcal{R}_{\mathbb{P}_n,p}(\rho_n) - \underline{s}^{-p} \frac{\rho_0^p}{\sqrt{n}} \right| \leq \frac{C}{n^\kappa},$$

where $\kappa = 1$ in the settings of Theorem 5(I)(II), and $\kappa = \min(1, \frac{1}{2} + \frac{1}{2\tau})$ in the setting of Theorem 5(III).

COROLLARY 2. *Let $p \in [1,\infty)$. Under the setup of Theorem 6, suppose $b = \frac{p+1}{2p}$. Then*

$$\left| \mathcal{R}_{\mathbb{P}_n,p}(\rho_n) - (p+1)^{\frac{1}{p+1}} [\mathfrak{g}(0)]^{\frac{p}{p+1}} \frac{\rho_0^{\frac{p}{p+1}}}{\sqrt{n}} \right| \leq \frac{C}{n^{\frac{3}{4}}}.$$

Note that for $p = \infty$, the expressions above are also consistent with Theorem 4. When $\underline{s} > 0$, Corollary 1 shows that $\rho_n = \rho_0 / n^{\frac{1}{2p}}$, which is slower than $1/\sqrt{n}$ for all $p \in (1,\infty)$; and when $\underline{s} = 0$, Corollary 2 shows that $\rho_n = \rho_0 / n^{\frac{p+1}{2p}}$, which is faster than $1/\sqrt{n}$ for all $p \in [1,\infty)$. This is in sharp contrast to the $1/\sqrt{n}$ radius selection rule for continuous loss functions [17]. For 0-1 loss functions, if we still adopt the $1/\sqrt{n}$ radius, than the worst-case loss would be overly conservative for all $p \in [1,\infty)$ when $\underline{s} = 0$, and overly optimistic for $p \in (1,\infty)$ when $\underline{s} > 0$.

We would like to remark that, in the results above, we only focus on the order of the Wasserstein radius ρ_n so that it yields a tight confidence interval with half-length of the order $1/\sqrt{n}$, but do not specify the constant upfront. The latter requires new concentration inequalities that bound the difference between empirical loss and true loss using the Wasserstein regularizer. Furthermore, we study the selection rule when the set E is fixed, although it can be dependent on the decision for some applications. Study on these topics is beyond the scope of this paper.

In the following, we demonstrate our results in Corollary 2 numerically. Recall that in Example 1, we have $\mathcal{Z} = ([-1,1], |\cdot|)$, $E = (0, \frac{1}{2})$ and $p = 1$. Suppose $\mathbb{P}_{\mathsf{true}}$ is a uniform distribution on $[0,1]$. Then $\mathbb{P}_{\mathsf{true}}(E) = \frac{1}{2}$. Let $\mathbb{P}_n$ be the empirical distribution formed by n i.i.d. samples from $\mathbb{P}_{\mathsf{true}}(E)$, where $n \in \{10^1, 10^2, \ldots, 10^6\}$. Given $\mathbb{P}_n$, we evaluate the worst-case loss, examine whether it serves as a lower bound of $\mathbb{P}_{\mathsf{true}}(E)$, and compute its gap from $\mathbb{P}_{\mathsf{true}}(E)$. Ideally, we want this gap to be minor, suggesting a tight lower bound. We repeat this process for 1000 independent trials. Based on these trials, we evaluate the probability that the worst-case loss fails to be a lower bound of the true loss and compute the average gap over the trials in which the worst-case loss is a lower bound on the true loss. We compare the radius selection rule $\rho_n = \rho_0/n$ as suggested by Corollary 2 and the standard $1/\sqrt{n}$-rule.

The results are shown in Figure 2, where we vary $\rho_0 \in \{0.3, 3, 30\}$ for $\rho_n = \rho_0/n$ and $\rho_0 \in \{0.01, 0.1, 1\}$ for $\rho_n = \rho_0/\sqrt{n}$. As shown in Fig. 2(b), the failure probabilities are well-controlled for most cases, and hence the robust formulation (P) with both scaling schemes yield high-confidence bounds for the true loss. On the other hand, as shown in the Fig. 2(a), the average gap when $\rho_n \sim 1/n$ converges to zero at a faster rate than the $1/\sqrt{n}$-rule. This verifies that Corollary 2 suggests a tighter confidence bound for the true loss and the $1/\sqrt{n}$-rule yields a more conservative bound. As a robustness check, we perform a similar experiment under a different setup. Additional numerical results are provided in Appendix EC.7.

6. Concluding Remarks

In this paper, we have established a general connection between Wasserstein distributionally robust 0-1 loss and regularization, which complements the literature on Wasserstein DRO and regularization for continuous loss functions. These results unveil a qualitative difference in the radius selection for Wasserstein 0-1 loss compared to the case of continuous loss functions. For the future work, it would be interesting to establish a concentration bound using the Wasserstein regularizer explicitly and develop the generalization bounds for Wasserstein distributionally robust problems involving 0-1 loss functions.

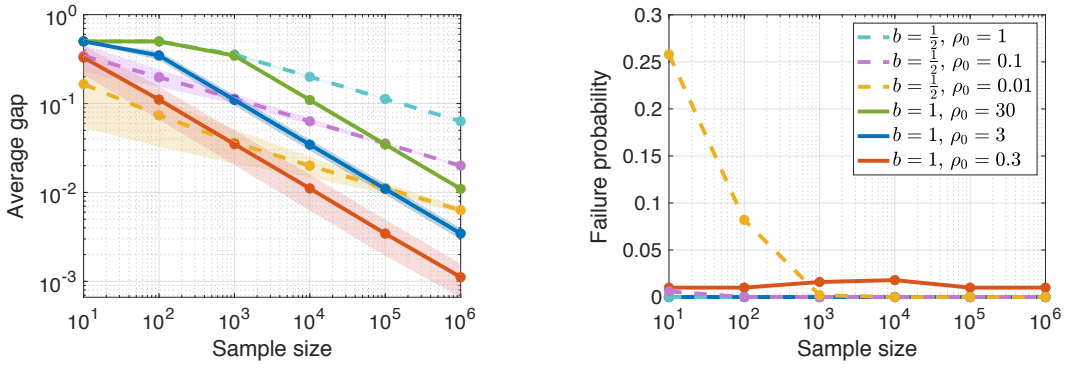

Figure 2 Comparisons between two scaling schemes $\rho_n \sim 1/n$ and $\rho_n \sim 1/\sqrt{n}$ for 1-Wasserstein DRO, represented by solid lines and dashed lines, respectively.

Appendix A: Duality and worst-case distribution for general DRO problems

We present the duality and worst-case distributions for DRO problems with general loss functions on a Polish space $(\mathcal{Z}, \mathrm{d})$. Consider the problem

$$v_P := \sup_{\mathbb{P}\in\mathcal{P}(\mathcal{Z}):\mathcal{W}_p(\mathbb{P},\mathbb{Q})\le\rho} \mathbb{E}_{\mathbb{P}}[f(Z)].$$

Its dual problem is defined as

$$v_D := \begin{cases} \min_{\lambda\ge 0}\left\{\lambda\rho^p + \mathbb{E}_{\mathbb{Q}}\left[\sup_{\tilde{z}\in\mathcal{Z}}\{f(\tilde{z}) - \lambda \mathrm{d}(\tilde{z},Z)^p\}\right]\right\}, & p\in[1,\infty), \\ \mathbb{E}_{\mathbb{Q}}\left[\sup_{\tilde{z}\in\mathcal{Z}}\{f(\tilde{z}) : \mathrm{d}(\tilde{z},Z)\le\rho\}\right], & p=\infty. \end{cases}$$

The following strong duality holds due to [19, 18].

Lemma 2. *Let $\mathbb{Q}\in\mathcal{P}(\mathcal{Z})$, $p\in[1,\infty]$, $\rho>0$, and f be a Borel measurable function on $\mathcal{Z}$. Then we have the following:*

(I) $v_P = v_D$.

(II) *When $p\in[1,\infty)$, suppose $v_D<\infty$. Then the dual minimizer λ_* exists.*

Particularly, when $f=-\mathbf{1}_E$, we have

$$\sup_{\tilde{z}\in\mathcal{Z}}\{-\mathbf{1}_E(\tilde{z}) - \lambda\mathrm{d}(\tilde{z},Z)^p\} = -\inf_{\tilde{z}\in\mathcal{Z}}\{\mathbf{1}_E(\tilde{z}) + \lambda\mathrm{d}(\tilde{z},Z)^p\} = -\mathbf{1}_E(Z)\cdot\min(1,\lambda\mathrm{d}_{E^c}(Z)^p),$$

and

$$\sup_{\tilde{z}\in\mathcal{Z}}\{-\mathbf{1}_E(\tilde{z}) : \mathrm{d}(\tilde{z},Z)\le\rho\} = -\inf_{\tilde{z}\in\mathcal{Z}}\{\mathbf{1}_E(\tilde{z}) : \mathrm{d}(\tilde{z},Z)\le\rho\} = -\mathbf{1}\{\mathrm{d}_E^c(Z)>\rho\}.$$

thereby we obtain (D).

The existence of a worst-case distribution is established in the following result, whose proof is given in Appendix EC.6.

Lemma 3. *Suppose that bounded subsets of $(\mathcal{Z},\mathrm{d})$ are totally bounded. Assume f is upper semi-continuous on $\mathcal{Z}$ and $\inf\left\{\lambda\ge 0 : \mathbb{E}_{\mathbb{Q}}\left[\sup_{\tilde{z}\in\mathcal{Z}}\{f(\tilde{z})-\lambda\mathrm{d}(\tilde{z},Z)^p\}\right]<\infty\right\}<\infty$. Then a worst-case distribution $\mathbb{P}_*$ exists and is given as follows:*

(I) *When $p=\infty$, there exists a $\mathbb{Q}$-measurable map $T_*:\mathcal{Z}\to\mathcal{Z}$ satisfying*

$$T_*(z)\in\arg\max_{\tilde{z}\in\mathcal{Z}}\{f(\tilde{z}) : \mathrm{d}(\tilde{z},z)\le\rho\}, \quad \mathbb{Q}-a.e.$$

Then $\mathbb{P}_* := T_{*\#}\mathbb{Q}$ is a worst-case distribution.

(II) *When* $p \in [1,\infty)$ *and* $\lambda_* = 0$*, there exists a* $\mathbb{Q}$*-measurable map satisfying*

$$T_*(z) \in \operatorname*{arg\,min}_{\tilde{z}\in\mathcal{Z}} \left\{ \mathsf{d}(z,\tilde{z}) : \ \tilde{z} \in \operatorname*{arg\,max}_{z\in\mathcal{Z}} \ f(z') \right\}, \quad \mathbb{Q}-a.e.$$

Then $\mathbb{P}_* := T_{*\#}\mathbb{Q}$ *is a worst-case distribution.*

(III) *When* $p \in [1,\infty)$ *and* $\lambda_* > 0$*, there exist* $\mathbb{Q}$*-measurable maps* $T_*, T_- : \mathcal{Z} \to \mathcal{Z}$ *satisfying*

$$T_*(z) \in \operatorname*{arg\,max}_{\tilde{z}\in\mathcal{Z}} \left\{ \mathsf{d}(z,\tilde{z}) : \ \tilde{z} \in \operatorname*{arg\,max}_{z'\in\mathcal{Z}} \{ f(z') - \lambda_* \mathsf{d}(z',z)^p \} \right\},$$

$$T_-(z) \in \operatorname*{arg\,min}_{\tilde{z}\in\mathcal{Z}} \left\{ \mathsf{d}(z,\tilde{z}) : \ \tilde{z} \in \operatorname*{arg\,max}_{z'\in\mathcal{Z}} \{ f(z') - \lambda_* \mathsf{d}(z',z)^p \} \right\}.$$

Let t_* *be the largest number in* $[0,1]$ *such that*

$$\rho^p = t_* \mathbb{E}_{\mathbb{Q}} \left[\mathsf{d}(T_*(Z),Z)^p \right] + (1-t_*) \mathbb{E}_{\mathbb{Q}} \left[\mathsf{d}(T_-(Z),Z)^p \right].$$

Then $\mathbb{P}_* := t_* T_{*\#}\mathbb{Q} + (1-t_*) T_{-\#}\mathbb{Q}$ *is a worst-case distribution.*

References

[1] Aliprantis CD, Border KC (2006) *Infinite Dimensional Analysis: a Hitchhiker's Guide* (Springer).

[2] An Y, Gao R (2021) Generalization bounds for (Wasserstein) robust optimization. *Advances in Neural Information Processing Systems* 34:10382–10392.

[3] Baricz Á (2008) Mills' ratio: Monotonicity patterns and functional inequalities. *Journal of Mathematical Analysis and Applications* 340(2):1362–1370.

[4] Bartl D, Drapeau S, Obłój J, Wiesel J (2021) Sensitivity analysis of Wasserstein distributionally robust optimization problems. *Proceedings of the Royal Society A* 477(2256):20210176.

[5] Blanchet J, Kang Y, Murthy K (2019) Robust Wasserstein profile inference and applications to machine learning. *Journal of Applied Probability* 56(3):830–857.

[6] Blanchet J, Murthy K (2019) Quantifying distributional model risk via optimal transport. *Mathematics of Operations Research* 44(2):565–600.

[7] Blanchet J, Murthy K, Nguyen VA (2021) Statistical analysis of Wasserstein distributionally robust estimators. *Tutorials in Operations Research: Emerging Optimization Methods and Modeling Techniques with Applications*, 227–254 (INFORMS).

[8] Blanchet J, Murthy K, Si N (2022) Confidence regions in Wasserstein distributionally robust estimation. *Biometrika* 109(2):295–315.

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

[10] Casella G, Berger RL (2002) *Statistical Inference*, volume 2 (Duxbury Pacific Grove, CA).

[11] Chen G, Zhang H, Hui H, Song Y (2021) Fast Wasserstein-distance-based distributionally robust chance-constrained power dispatch for multi-zone hvac systems. *IEEE Transactions on Smart Grid* 12(5):4016–4028.

[12] Chen R, Paschalidis IC (2018) A robust learning approach for regression models based on distributionally robust optimization. *Journal of Machine Learning Research* 19(13).

[13] Chen Z, Kuhn D, Wiesemann W (2022) Data-driven chance constrained programs over Wasserstein balls. *Operations Research*. Articles in Advance.

[14] Chen Z, Kuhn D, Wiesemann W (2022) On approximations of data-driven chance constrained programs over Wasserstein balls. *arXiv preprint arXiv:2206.00231* .

[15] 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.

[16] Esfahani PM, Kuhn D (2018) Data-driven distributionally robust optimization using the Wasserstein metric: Performance guarantees and tractable reformulations. *Mathematical Programming* 171(1):115–166.

[17] Gao R (2022) Finite-sample guarantees for Wasserstein distributionally robust optimization: Breaking the curse of dimensionality. *Operations Research*. Articles in Advance.

[18] Gao R, Chen X, Kleywegt AJ (2022) Wasserstein distributionally robust optimization and variation regularization. *Operations Research*. Articles in Advance.

[19] Gao R, Kleywegt AJ (2022) Distributionally robust stochastic optimization with Wasserstein distance. *Mathematics of Operations Research*. Articles in Advance.

[20] Gao R, Xie L, Xie Y, Xu H (2018) Robust hypothesis testing using Wasserstein uncertainty sets. *Advances in Neural Information Processing Systems* 31:7913–7923.

[21] Ghaoui LE, Oks M, Oustry F (2003) Worst-case value-at-risk and robust portfolio optimization: A conic programming approach. *Operations research* 51(4):543–556.

[22] Gu Y, Wang Y (2021) Distributionally robust chance-constrained programmings for non-linear uncertainties with Wasserstein distance. *arXiv preprint arXiv:2103.04790* .

[23] Hanasusanto GA, Roitch V, Kuhn D, Wiesemann W (2017) Ambiguous joint chance constraints under mean and dispersion information. *Operations Research* 65(3):751–767.

[24] Ho-Nguyen N, Kılınç-Karzan F, Küçükyavuz S, Lee D (2021) Distributionally robust chance-constrained programs with right-hand side uncertainty under Wasserstein ambiguity. *Mathematical Programming* 1–32.

[25] Ho-Nguyen N, Kılınç-Karzan F, Küçükyavuz S, Lee D (2022) Strong formulations for distributionally robust chance-constrained programs with left-hand side uncertainty under Wasserstein ambiguity. *INFORMS Journal on Optimization*. Articles in Advance.

[26] Ho-Nguyen N, Wright SJ (2022) Adversarial classification via distributional robustness with Wasserstein ambiguity. *Mathematical Programming* 1–37.

[27] Hota AR, Cherukuri A, Lygeros J (2019) Data-driven chance constrained optimization under Wasserstein ambiguity sets. *2019 American Control Conference (ACC)*, 1501–1506 (IEEE).

[28] Hu Z, Hong JL (2012) Kullback-Leibler divergence constrained distributionally robust optimization. *Available at Optimization Online* .

[29] Hu Z, Hong LJ, So AMC (2013) Ambiguous probabilistic programs. *Available at Optimization Online* .

[30] Ji R, Lejeune MA (2021) Data-driven distributionally robust chance-constrained optimization with Wasserstein metric. *Journal of Global Optimization* 79(4):779–811.

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

[32] Kandel A, Moura SJ (2020) Safe Wasserstein constrained deep Q-learning. *arXiv preprint arXiv:2002.03016* .

[33] 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).

[34] Li B, Jiang R, Mathieu JL (2019) Ambiguous risk constraints with moment and unimodality information. *Mathematical Programming* 173(1):151–192.

[35] Li W, Liu Y, Liang H, Man Y, Li F (2021) Distributed tracking-admm approach for chance-constrained energy management with stochastic wind power in smart grid. *CSEE Journal of Power and Energy Systems* 1–11.

[36] Ordoudis C, Nguyen VA, Kuhn D, Pinson P (2021) Energy and reserve dispatch with distributionally robust joint chance constraints. *Operations Research Letters* 49(3):291–299.

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

[38] Shafieezadeh Abadeh S, Mohajerin Esfahani PM, Kuhn D (2015) Distributionally robust logistic regression. *Advances in Neural Information Processing Systems* 28:1576–1584.

[39] Shang C, Ding SX, Ye H (2021) Distributionally robust fault detection design and assessment for dynamical systems. *Automatica* 125:109434.

[40] Shen H, Jiang R (2021) Convex chance-constrained programs with Wasserstein ambiguity. *arXiv preprint arXiv:2111.02486* .

[41] Shen H, Jiang R (2022) Chance-constrained set covering with Wasserstein ambiguity. *Mathematical Programming* 1–54.

[42] Shen H, Jiang R (2022) Wasserstein two-sided chance constraints with an application to optimal power flow. *arXiv preprint arXiv:2204.00191* .

[43] Skorski M (2021) Bernstein-type bounds for Beta distribution. *arXiv preprint arXiv:2101.02094* .

[44] Taskesen B, Nguyen VA, Kuhn D, Blanchet J (2020) A distributionally robust approach to fair classification. *arXiv preprint arXiv:2007.09530* .

[45] Vershynin R (2018) *High-dimensional probability: An introduction with applications in data science*, volume 47 (Cambridge university press).

[46] 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*, volume 31.

[47] Wagner MR (2008) Stochastic 0–1 linear programming under limited distributional information. *Operations Research Letters* 36(2):150–156.

[48] Wang Y, Nguyen VA, Hanasusanto GA (2021) Wasserstein robust classification with fairness constraints. *arXiv preprint arXiv:2103.06828* .

[49] Xie L, Gao R, Xie Y (2021) Robust hypothesis testing with Wasserstein uncertainty sets. *arXiv preprint arXiv:2105.14348* .

[50] Xie W (2021) On distributionally robust chance constrained programs with Wasserstein distance. *Mathematical Programming* 186(1):115–155.

[51] Xie W, Ahmed S (2018) On deterministic reformulations of distributionally robust joint chance constrained optimization problems. *SIAM Journal on Optimization* 28(2):1151–1182.

[52] Xu H, Caramanis C, Mannor S (2012) Optimization under probabilistic envelope constraints. *Operations Research* 60(3):682–699.

[53] Yang W, Xu H (2016) Distributionally robust chance constraints for non-linear uncertainties. *Mathematical Programming* 155(1):231–265.

[54] Zhang Y, Jiang R, Shen S (2018) Ambiguous chance-constrained binary programs under mean-covariance information. *SIAM Journal on Optimization* 28(4):2922–2944.

Supplementary Materials

Appendix EC.1: Proofs for Section 2

Let $\Gamma(\mathbb{P},\mathbb{Q})$ denote the set of all Borel probability distributions on $\mathcal{Z}\times\mathcal{Z}$ with marginal distributions $\mathbb{P}$ and $\mathbb{Q}$.

Proof of Proposition 1. If $E=\mathcal{Z}$, then both sides equal 1. Below we assume $\mathbb{E}\subsetneq\mathcal{Z}$. The case of $p\in[1,\infty)$ has been shown in [19, Proposition 3]. In the following, we prove for $p=\infty$. Consider $\inf_{\mathbb{P}\in\mathcal{P}(\mathcal{Z})}\left\{\mathbb{P}(\mathrm{int}(E)):\mathcal{W}_p(\mathbb{P},\mathbb{Q})\le\rho\right\}$. By Lemma 3, a worst-case distribution $\mathbb{P}_*$ exists.

If $\mathbb{P}_*(\mathrm{int}(E))=1$, then

$$1=\mathbb{P}_*(\mathrm{int}(E))=\inf_{\mathbb{P}\in\mathcal{P}(\mathcal{Z}):\mathcal{W}_p(\mathbb{P},\mathbb{Q})\le\rho}\mathbb{P}(\mathrm{int}(E))\le\inf_{\mathbb{P}\in\mathcal{P}(\mathcal{Z}):\mathcal{W}_p(\mathbb{P},\mathbb{Q})\le\rho}\mathbb{P}(E)\le\mathbb{P}_*(E)\le 1.$$

It follows that $\inf_{\mathbb{P}\in\mathcal{P}(\mathcal{Z})}\{\mathbb{P}(E):\mathcal{W}_\infty(\mathbb{P},\mathbb{Q})\le\rho\}=\inf_{\mathbb{P}\in\mathcal{P}(\mathcal{Z})}\{\mathbb{P}(\mathrm{int}(E)):\mathcal{W}_\infty(\mathbb{P},\mathbb{Q})\le\rho\}=1$.

Otherwise, consider $\mathbb{P}_*(\mathrm{int}(E))<1$. Let $\epsilon>0$. For any $z\in E\setminus\mathrm{int}(E)$, the set $\{\tilde z\in E^c:0<\|z-\tilde z\|<\epsilon\}$ is non-empty and the graph

$$G_\epsilon:=\{(z,\tilde z)\in(E\setminus\mathrm{int}(E))\times E^c:0<\|z-\tilde z\|<\epsilon\},$$

is measurable in $\mathcal{Z}\times\mathcal{Z}$. By Aumann's measurable selection theorem [1, Theorem 18.26], there exists a measurable map $T^\epsilon:\mathcal{Z}\to\mathcal{Z}$ such that $T^\epsilon(z)=z$ for all $z\in\mathrm{int}(E)\cup E^c$, and $T^\epsilon(z)\in\{\tilde z\in E^c:0<\|z-\tilde z\|<\epsilon\}$ for all $z\in E\setminus\mathrm{int}(E)$. Define

$$\mathbb{P}^\epsilon:=(1-t^\epsilon)\,T^\epsilon_\#\mathbb{P}_*+t^\epsilon\mathbb{Q},$$

where $t^\epsilon\in(0,1)$ is to be determined shortly. Let $\pi_1^\epsilon\in\arg\min_{\pi\in\Gamma(T^\epsilon_\#\mathbb{P}_*,\mathbb{Q})}\pi\text{-}\operatorname*{ess\,sup}_{\mathcal{Z}\times\mathcal{Z}}\|\tilde z-z\|$. Let $\pi_2\in\Gamma(\mathcal{Z}\times\mathcal{Z})$ be given by $\pi_2(B):=\mathbb{Q}\{z\in\mathcal{Z}:(z,z)\in B\}$. For $i\in\{1,2\}$, define the canonical projection $\mathrm{proj}_i:\mathcal{Z}\times\mathcal{Z}\to\mathcal{Z}$ as $\mathrm{proj}_i(z_1,z_2)=z_i$. Note that $\mathrm{proj}_{1\#}\pi_2=\mathrm{proj}_{2\#}\pi_2=\mathbb{Q}$. Define $\pi^\epsilon:=(1-t^\epsilon)\,\pi_1^\epsilon+t^\epsilon\pi_2$. Observe that

$$\begin{aligned}\mathrm{proj}_{1\#}\pi^\epsilon&=(1-t^\epsilon)\,\mathrm{proj}_{1\#}\pi_1^\epsilon+t^\epsilon\,\mathrm{proj}_{1\#}\pi_2=(1-t^\epsilon)\,T^\epsilon_\#\mathbb{P}_*+t^\epsilon\mathbb{Q}=\mathbb{P}^\epsilon;\\ \mathrm{proj}_{2\#}\pi^\epsilon&=(1-t^\epsilon)\,\mathrm{proj}_{2\#}\pi_1^\epsilon+t^\epsilon\,\mathrm{proj}_{2\#}\pi_2=(1-t^\epsilon)\,\mathbb{Q}+t^\epsilon\mathbb{Q}=\mathbb{Q}.\end{aligned}$$

Then

$$\begin{aligned}\mathcal{W}_\infty(\mathbb{P}^\epsilon,\mathbb{Q})&\le\pi^\epsilon\text{-}\operatorname*{ess\,sup}_{\mathcal{Z}\times\mathcal{Z}}\|\tilde z-z\|\\&=\left((1-t^\epsilon)\pi_1^\epsilon+t^\epsilon\pi_2\right)\text{-}\operatorname*{ess\,sup}_{\mathcal{Z}\times\mathcal{Z}}\|\tilde z-z\|\\&=(1-t^\epsilon)\,\pi_1^\epsilon\text{-}\operatorname*{ess\,sup}_{\mathcal{Z}\times\mathcal{Z}}\|\tilde z-z\|\\&=(1-t^\epsilon)\,\mathcal{W}_\infty(T^\epsilon_\#\mathbb{P}_*,\mathbb{Q}).\end{aligned}$$

Next, define $\pi^*\in\arg\min_{\pi\in\Gamma(\mathbb{P}_*,\mathbb{Q})}\pi-\operatorname*{ess\,sup}_{\mathcal{Z}\times\mathcal{Z}}\|\tilde z-z\|$. Consider $\pi^T\in\Gamma(\mathcal{Z}\times\mathcal{Z})$ given by $\pi^T(B):=\pi^*(\{(z,\tilde z):(T^\epsilon(z),\tilde z)\in B\subset\mathcal{Z}\times\mathcal{Z}\})$. Note that for any $A\in\mathcal{B}(\mathcal{Z})$,

$$\begin{aligned}\mathrm{proj}_{1\#}\pi^T(A)&=\pi^T(A\times\mathcal{Z})=\pi^*(\{(z,\tilde z):T^\epsilon(z)\in A,\tilde z\in\mathcal{Z}\})=\pi^*\left((T^\epsilon)^{-1}(A)\times\mathcal{Z}\right)\\&=\mathrm{proj}_{1\#}\pi^*\left((T^\epsilon)^{-1}(A)\right)=\mathbb{P}_*\left((T^\epsilon)^{-1}(A)\right)=T^\epsilon_\#\mathbb{P}_*(A),\\ \mathrm{proj}_{2\#}\pi^T(A)&=\pi^T(\mathcal{Z}\times A)=\pi^*(\{(z,\tilde z):T^\epsilon(z)\in\mathcal{Z},\tilde z\in A\})=\pi^*(\mathcal{Z}\times A)\\&=\mathrm{proj}_{2\#}\pi^*(A)=\mathbb{Q}(A).\end{aligned}$$

Thus $\mathrm{proj}_{1\#}\pi^T = T^\epsilon_\# \mathbb{P}_*$ and $\mathrm{proj}_{2\#}\pi^T = \mathbb{Q}$. As a result,

$$\begin{aligned}
\mathcal{W}_\infty(T^\epsilon_\#\mathbb{P}_*, \mathbb{Q}) &= \inf_{\pi\in\Gamma(T^\epsilon_\#\mathbb{P}_*,\mathbb{Q})} \pi\text{-ess}\sup_{\mathcal{Z}\times\mathcal{Z}} \|\tilde{z}-z\| \\
&\le \pi^T\text{-ess}\sup_{\mathcal{Z}\times\mathcal{Z}} \|\tilde{z}-z\| \\
&= \pi^*\text{-ess}\sup_{\mathcal{Z}\times\mathcal{Z}} \|T^\epsilon(\tilde{z})-z\| \\
&\le \pi^*\text{-ess}\sup_{\mathcal{Z}\times\mathcal{Z}} (\|T^\epsilon(\tilde{z})-\tilde{z}\| + \|\tilde{z}-z\|) \\
&\le \pi^*\text{-ess}\sup_{\mathcal{Z}\times\mathcal{Z}} \|\tilde{z}-z\| + \epsilon \\
&= \mathcal{W}_\infty(\mathbb{P}_*, \mathbb{Q}) + \epsilon.
\end{aligned}$$

Therefore, $\mathcal{W}_\infty(\mathbb{P}^\epsilon, \mathbb{Q}) \le (1-t^\epsilon)\,\mathcal{W}_\infty(T^\epsilon_\#\mathbb{P}_*, \mathbb{Q}) \le (1-t^\epsilon)(\rho+\epsilon)$. Pick $t^\epsilon = \epsilon/(\rho+\epsilon)$. Then, $\mathcal{W}_\infty(T^\epsilon_\#\mathbb{P}_*, \mathbb{Q}) \le \rho$ and

$$\begin{aligned}
\inf_{\mathbb{P}\in\mathcal{P}(\mathcal{Z}):\mathcal{W}_p(\mathbb{P},\mathbb{Q})\le\rho} \mathbb{P}(\mathrm{int}(E)) &\le \inf_{\mathbb{P}\in\mathcal{P}(\mathcal{Z}):\mathcal{W}_p(\mathbb{P},\mathbb{Q})\le\rho} \mathbb{P}(E) \\
&\le \mathbb{P}^\epsilon(E) \\
&\le (1-t^\epsilon)\, T^\epsilon_\#\mathbb{P}_*(E) + t^\epsilon \mathbb{Q}(E) \\
&= (1-t^\epsilon)\, \mathbb{P}_*(T^\epsilon(E)) + t^\epsilon \mathbb{Q}(E) \\
&= (1-t^\epsilon)\, \mathbb{P}_*(\mathrm{int}(E)) + t^\epsilon \mathbb{Q}(E) \\
&\le (1-t^\epsilon) \inf_{\mathbb{P}\in\mathcal{P}(\mathcal{Z}):\mathcal{W}_p(\mathbb{P},\mathbb{Q})\le\rho} \mathbb{P}(\mathrm{int}(E)) + t^\epsilon.
\end{aligned}$$

Letting $\epsilon \downarrow 0$ yields the result. □

Proof of Lemma 1. (III) Let $\mathbb{P}_*$ be a worst-case distribution. If $\rho < \rho_{\max}$, from the definition of $\rho_{\max}$, not all probability in E can be transported out to E^c, and thus the worst-case probability $\mathbb{P}_*(E) > 0$. Suppose $\lambda_* = 0$, then by (D) we have $\mathbb{P}_*(E) = -\lambda_*\rho^p + \mathbb{E}_{\mathbb{Q}}\left[\mathbf{1}_E(Z)\cdot\min(1, \lambda_* \mathrm{d}_{E^c}(Z)^p)\right] = 0$, contradictory to $\mathbb{P}_*(E) > 0$. Hence $\lambda_* > 0$ when $\rho < \rho_{\max}$.

If $\rho \ge \rho_{\max}$, then all probability in E are transported out to E^c, and thus $\mathbb{P}_*(E) = 0$. In the previous case, we have shown that when $\lambda = 0$, $\mathbb{P}_*(E) = 0$. Therefore, $\lambda_* = 0$ is a dual optimizer.

(II) Note that for every $z \in \mathcal{Z}$, $\mathcal{T}_*(z) = \arg\min_{\tilde{z}\in\mathcal{Z}}\{\mathbf{1}_E(\tilde{z}) + \lambda_* \mathrm{d}(\tilde{z}, z)\}$. Hence by [6, Theorem 1] or [19, Corollary 1],

$$\{(z, \mathcal{T}_-(z)) : z \in \mathcal{Z}_*\} \subset \mathrm{supp}\,\gamma_* \subset \{(z, \mathcal{T}_*(z)) : z \in \mathcal{Z}_*\}.$$

From [19, Lemma 3 and Corollary 1], there exist two $\mathbb{Q}$-measurable transport maps $T_*, T_- : \mathcal{Z} \to \mathcal{Z}$ that

$$T_*(z) \begin{cases} = z, & z \in \mathcal{Z}_*^c, \\ \in \arg\min_{\tilde{z}\in E^c} \mathrm{d}(\tilde{z}, z), & z \in \mathcal{Z}_*. \end{cases}$$

$$T_-(z) \begin{cases} = z, & z \in \mathcal{Z}_*^c \cup \mathcal{Z}_=, \\ \in \arg\min_{\tilde{z}\in E^c} \mathrm{d}(\tilde{z}, z), & z \in \mathcal{Z}_* \setminus \mathcal{Z}_=. \end{cases}$$

Then T_* and T_- are measurable selections of $\mathcal{T}_*$ and $\mathcal{T}_-$ respectively. Since $-\mathbf{1}_E(\cdot)$ is upper semi-continuous on $\mathcal{Z}$, from Lemma 3, there exists $t_* \in [0,1]$ such that $\mathbb{P}_*$ defined in (II) is a worst-case distribution. We have

$$\mathbf{1}_E(T_*(z)) = \begin{cases} 1, & z \in E\setminus\mathcal{Z}_*, \\ 0, & z \in E^c \cup \mathcal{Z}_*, \end{cases} \qquad \mathbf{1}_E(T_-(z)) = \begin{cases} 1, & z \in (E\setminus\mathcal{Z}_*) \cup \mathcal{Z}_=, \\ 0, & z \in E^c \cup (\mathcal{Z}_*\setminus\mathcal{Z}_=). \end{cases}$$

Therefore,

$$\begin{aligned}\mathbb{P}_*(E) = \mathbb{E}_{\mathbb{P}_*}[\mathbf{1}_E] &= t_* \int_{\mathcal{Z}} \mathbf{1}_E(T_*(z))\mathbb{Q}(dz) + (1-t_*)\int_{\mathcal{Z}} \mathbf{1}_E(T_-(z))\mathbb{Q}(dz)\\ &= t_*\mathbb{Q}(E\setminus \mathcal{Z}_*) + (1-t_*)\mathbb{Q}((E\setminus\mathcal{Z}_*)\cup \mathcal{Z}_=)\\ &= \mathbb{Q}(E\setminus\mathcal{Z}_*) + (1-t_*)\mathbb{Q}(\mathcal{Z}_=).\end{aligned}$$

(I) Define the transport map T_* the same as in the proof of (II). With the similar reasoning in the proof of (II), $\mathbb{P}_*$ defined in (I) is a worst-case distribution. We have

$$\mathbf{1}_E(T_*(z)) = \begin{cases} 1, & z \in E\setminus\mathcal{Z}_*,\\ 0, & z \in E^c \cup \mathcal{Z}_*.\end{cases}$$

Thus $\mathbb{P}_*(E) = \mathbb{Q}(E\setminus\mathcal{Z}_*)$. The proof is completed. □

Appendix EC.2: Proofs for Section 3

From Assumption 1, we can derive a second-order approximation of G.

Lemma EC.1. *Assume Assumption 1 holds. Then for all $s_1, s_2 \in [0,\delta]$,*

$$|G(s_2) - G(s_1) - g(s_1)(s_2-s_1)| \le \frac{L_g}{2}(s_1-s_2)^2.$$

Proof of Lemma EC.1. By the mean value theorem,

$$G(s_2) = G(s_1) + g(s_1)(s_2-s_1) + \int_0^1 \left(g(s_1+t(s_2-s_1)) - g(s_1)\right)(s_2-s_1)dt.$$

Then from Assumption 1, it follows that

$$|G(s_2) - G(s_1) - g(s_1)(s_2-s_1)| \le \int_0^1 L_g t(s_1-s_2)^2 dt = \frac{L_g}{2}(s_1-s_2)^2,$$

which completes the proof. □

EC.2.1. Proof of Theorem 2

We first derive rough yet simple bounds for $\mathcal{R}_{\mathbb{Q},p}(\rho)$.

Lemma EC.2. *Let $p \in [1,\infty)$ and E be some open and measurable set. Assume Assumption 1 holds and $G(\underline{s}) = 0$. Let $\underline{s} > 0$. For $\rho < \left(\mathbb{Q}(E)G\left((\underline{s}+\delta)/2\right)\right)^{1/p}\underline{s}$, it holds that $\underline{s} \le \lambda_*^{-1/p} < \delta$ and*

$$\lambda_*\rho^p \le \mathcal{R}_{\mathbb{Q},p}(\rho) \le \underline{s}^{-p}\rho^p.$$

Proof. From Lemma 1, it holds that $\underline{s} \le \lambda_*^{-1/p}$, since otherwise, $\mathbb{Q}(\mathcal{Z}_*) = 0$ and then $\rho^p = \mathbb{E}_{(Z,\tilde{Z})\sim\gamma_*}[\|Z-\tilde{Z}\|^p] = 0$. Since $G(\underline{s}) = 0$, by Assumption 1, $\underline{s} < \delta$ and $G((\underline{s}+\delta)/2) > 0$. We claim that $\lambda_*^{-1/p} < \delta$. Suppose on the contrary that $\lambda_*^{-1/p} \ge \delta$. From Lemma 1 and (3), it follows that

$$\rho^p \ge \mathbb{Q}(E)\cdot\underline{s}^p\left(G(\lambda_*^{-1/p}-) - G(\underline{s})\right) \ge \mathbb{Q}(E)\cdot\underline{s}^p G\left(\frac{\underline{s}+\delta}{2}\right),$$

which contradicts with $\rho < \left(\mathbb{Q}(E)G\left(\frac{\underline{s}+\delta}{2}\right)\right)^{1/p}\underline{s}$.

Recall $\mathcal{Z}_*$ and $\mathcal{Z}_=$ defined in (Z). Let $s := \mathsf{d}_{E^c}(z)$. Since $\underline{s} \le \lambda_*^{-1/p} < \delta$ and G is differentiable on $[0,\delta]$, we have $\mathbb{Q}(\mathcal{Z}_=) = 0$. Then by Lemma 1,

$$\rho^p = \mathbb{E}_{\mathbb{Q}}[\mathbf{1}_{\mathcal{Z}_*}(Z)\mathsf{d}^p_{E^c}(Z)] = \int_{0+}^{\lambda_*^{-1/p}} s^p \mathbb{Q}(E) dG(s) = \mathbb{Q}(E)\int_{\underline{s}}^{\lambda_*^{-1/p}} s^p dG(s) = \mathbb{Q}(E)\int_{\underline{s}}^{\lambda_*^{-1/p}} s^p dG(s), \tag{EC.1}$$

and

$$\mathcal{R}_{\mathbb{Q},p}(\rho) = \mathbb{E}_{\mathbb{Q}}[\mathbf{1}_{\mathcal{Z}_*}] = \mathbb{Q}(E)\int_{\underline{s}}^{\lambda_*^{-1/p}} 1 dG(s) = \mathbb{Q}(E)[G(\lambda_*^{-1/p}) - G(\underline{s})] = \mathbb{Q}(E)G(\lambda_*^{-1/p}). \tag{EC.2}$$

With (EC.1) and (EC.2), it follows that

$$\underline{s}^p \mathbb{Q}(E)[G(\lambda_*^{-1/p}) - G(\underline{s})] \le \rho^p = \mathbb{Q}(E)\int_{\underline{s}}^{\lambda_*^{-1/p}} s^p dG(s) \le \lambda_*^{-1}\mathbb{Q}(E)[G(\lambda_*^{-1/p}) - G(\underline{s})],$$

which indicates that

$$\lambda_* \rho^p \le \mathcal{R}_{\mathbb{Q},p}(\rho) \le \underline{s}^{-p}\rho^p.$$

Hence the proof is completed. □

Based on the rough bounds above, next, we derive finer bounds.

Proof of Theorem 2. (I) When $G(\underline{s}) > 0$, consider the set $\mathcal{Z}_{\underline{s}} = \{z \in \operatorname{supp}\mathbb{Q} \cap E : \mathsf{d}_{E^c}(z) = \underline{s}\}$. By definition $\mathbb{Q}(\mathcal{Z}_{\underline{s}}) = G(\underline{s})\mathbb{Q}(E)$. Since $\rho^p < G(\underline{s})\mathbb{Q}(E)\underline{s}^p = \mathbb{Q}(\mathcal{Z}_{\underline{s}})\underline{s}^p$, by the structure of the worst-case distribution (Lemma 1), the worst-case distribution transports a fraction $t_* = \frac{\rho^p}{\mathbb{Q}(\mathcal{Z}_{\underline{s}})\underline{s}^p}$ of the set $\mathcal{Z}_= = \mathcal{Z}_{\underline{s}}$ to the boundary of E, resulting a change

$$\mathcal{R}_{\mathbb{Q},p}(\rho) = t_*\mathbb{Q}(\mathcal{Z}_{\underline{s}}) = \frac{\rho^p}{\mathbb{Q}(\mathcal{Z}_{\underline{s}})\underline{s}^p}\mathbb{Q}(\mathcal{Z}_{\underline{s}}) = \underline{s}^{-p}\rho^p.$$

(II) When $G(\underline{s}) = 0$, $g(\underline{s}) > 0$ and $\rho < \left(\mathbb{Q}(E)G\left((\underline{s}+\delta)/2\right)\right)^{1/p}\underline{s}$, by Lemma EC.2, $\underline{s} \le \lambda_*^{-1/p} < \delta$. Define $\epsilon = \lambda_*^{-1/p} - \underline{s}$. We are to derive an upper bound for ϵ, which is equivalent to a lower bound for λ_*. By Assumption 1, it follows that

$$\rho^p = \mathbb{Q}(E)\int_{\underline{s}}^{\underline{s}+\epsilon} s^p dG(s) \ge \mathbb{Q}(E)\int_0^{\epsilon} g(\underline{s}+s)\underline{s}^p ds \ge \mathbb{Q}(E)\int_0^{\epsilon}\left(g(\underline{s}) - L_g s\right)\underline{s}^p ds = \mathbb{Q}(E)\left(g(\underline{s})\epsilon - \frac{1}{2}L_g\epsilon^2\right)\underline{s}^p,$$

which implies

$$\frac{1}{2}L_g\mathbb{Q}(E)\epsilon^2 - g(\underline{s})\mathbb{Q}(E)\epsilon + \underline{s}^{-p}\rho^p \ge 0.$$

When $\rho < (2L_g)^{-\frac{1}{p}} g(\underline{s})^{\frac{2}{p}}\mathbb{Q}(E)^{\frac{1}{p}}\underline{s}$, this is equivalent to either

$$\epsilon \ge \frac{g(\underline{s})\mathbb{Q}(E) + \sqrt{\left(g(\underline{s})\mathbb{Q}(E)\right)^2 - 2L_g\mathbb{Q}(E)\underline{s}^{-p}\rho^p}}{L_g\mathbb{Q}(E)},$$

or

$$\epsilon \le \frac{g(\underline{s})\mathbb{Q}(E) - \sqrt{\left(g(\underline{s})\mathbb{Q}(E)\right)^2 - 2L_g\mathbb{Q}(E)\underline{s}^{-p}\rho^p}}{L_g\mathbb{Q}(E)}.$$

The former one cannot hold because otherwise.

$$\rho^p = \int_{\underline{s}}^{\underline{s}+\epsilon} s^p\mathbb{Q}(E)dG(s) \ge \int_{\underline{s}}^{\underline{s}+g(\underline{s})L_g^{-1}} s^p\mathbb{Q}(E)dG(s) \ge \underline{s}^p\mathbb{Q}(E)G\left(\underline{s}+g(\underline{s})L_g^{-1}\right),$$

contradicts that $\rho < \left(\underline{s}^p \mathbb{Q}(E) G\left(\underline{s} + g(\underline{s}) L_g^{-1}\right)\right)^{\frac{1}{p}}$. Hence,

$$\epsilon \leq \frac{g(\underline{s})\mathbb{Q}(E) - \sqrt{\left(g(\underline{s})\mathbb{Q}(E)\right)^2 - 2L_g \mathbb{Q}(E)\underline{s}^{-p}\rho^p}}{L_g \mathbb{Q}(E)} = \frac{2\underline{s}^{-p}\rho^p}{g(\underline{s})\mathbb{Q}(E) + \sqrt{\left(g(\underline{s})\mathbb{Q}(E)\right)^2 - 2L_g \mathbb{Q}(E)\underline{s}^{-p}\rho^p}} \leq \frac{2\underline{s}^{-p}\rho^p}{g(\underline{s})\mathbb{Q}(E)}.$$

Using the elementary inequality $(1+x)^{-p} \geq 1 - px$ for $x \geq 0$ and $p \geq 1$, it follows that

$$\begin{aligned}
\mathcal{R}_{\mathbb{Q},p}(\rho) &\geq (\underline{s} + \epsilon)^{-p}\rho^p \\
&\geq \rho^p \left(\underline{s} + \frac{2\underline{s}^{-p}\rho^p}{g(\underline{s})\mathbb{Q}(E)}\right)^{-p} \\
&= \rho^p \underline{s}^{-p} \left(1 + \frac{2\underline{s}^{-p-1}\rho^p}{g(\underline{s})\mathbb{Q}(E)}\right)^{-p} \\
&\geq \rho^p \underline{s}^{-p} \left(1 - p\frac{2\underline{s}^{-p-1}\rho^p}{g(\underline{s})\mathbb{Q}(E)}\right) \\
&= \rho^p \underline{s}^{-p} - 2p\left(g(\underline{s})\mathbb{Q}(E)\right)^{-1} \underline{s}^{-2p-1}\rho^{2p} \\
&= \rho^p \underline{s}^{-p} - 2p\mathfrak{g}(\underline{s})^{-1} \underline{s}^{-2p-1}\rho^{2p}.
\end{aligned}$$

where the last equality holds since $\mathfrak{g}(\underline{s}) = g(\underline{s})\mathbb{Q}(E)$. Hence we obtain the desired result by setting

$$\bar{\rho} = \min\left(\left(\mathbb{Q}(E) G\left((\underline{s} + \delta)/2\right)\right)^{1/p} \underline{s},\ (2L_g)^{-\frac{1}{p}} g(\underline{s})^{\frac{2}{p}} \mathbb{Q}(E)^{\frac{1}{p}} \underline{s},\ \left(\underline{s}^p \mathbb{Q}(E) G(\underline{s} + g(\underline{s}) L_g^{-1})\right)^{\frac{1}{p}}\right).$$

(III) By our assumption on G, there exists a constant $0 < \Delta_G < \delta - \underline{s}$ such that $G(\underline{s} + x) \geq \frac{1}{2} c_{\underline{s}} x^\tau$ for $x \in (0, \Delta_G]$. Define $\epsilon = \lambda_*^{-1/p} - \underline{s}$. It can be claimed that $\epsilon \leq \Delta_G$. Indeed, if $\epsilon > \Delta_G$, using a similar reasoning as in Lemma EC.2 where we replace $(\underline{s} + \delta)/2$ with $\underline{s} + \Delta_G$ yields that

$$\rho^p \geq \mathbb{Q}(E)\underline{s}^p G(\underline{s} + \Delta_G) > 0.$$

which contradicts that $\rho < \left(G(\underline{s} + \Delta_G)\mathbb{Q}(E)\right)^{\frac{1}{p}} \underline{s}$. Thus by (EC.1), we have

$$\rho^p \geq \mathbb{Q}(E)\underline{s}^p G(\underline{s} + \epsilon) \geq \mathbb{Q}(E)\underline{s}^p \cdot \tfrac{1}{2} c_{\underline{s}} \epsilon^\tau,$$

which is equivalent to

$$\epsilon \leq 2^{\frac{1}{\tau}} c_{\underline{s}}^{-\frac{1}{\tau}} \underline{s}^{-\frac{p}{\tau}} \mathbb{Q}(E)^{-\frac{1}{\tau}} \rho^{\frac{p}{\tau}}.$$

Accordingly, it holds that

$$\begin{aligned}
\mathcal{R}_{\mathbb{Q},p}(\rho) \geq (\underline{s} + \epsilon)^{-p}\rho^p &\geq \rho^p \left(\underline{s} + 2^{\frac{1}{\tau}} c_{\underline{s}}^{-\frac{1}{\tau}} \underline{s}^{-\frac{p}{\tau}} \mathbb{Q}(E)^{-\frac{1}{\tau}} \rho^{\frac{p}{\tau}}\right)^{-p} \\
&= \rho^p \underline{s}^{-p} \left(1 + 2^{\frac{1}{\tau}} c_{\underline{s}}^{-\frac{1}{\tau}} \underline{s}^{-\frac{p}{\tau}-1} \mathbb{Q}(E)^{-\frac{1}{\tau}} \rho^{\frac{p}{\tau}}\right)^{-p} \\
&\geq \rho^p \underline{s}^{-p} \left(1 - p2^{\frac{1}{\tau}} c_{\underline{s}}^{-\frac{1}{\tau}} \underline{s}^{-\frac{p}{\tau}-1} \mathbb{Q}(E)^{-\frac{1}{\tau}} \rho^{\frac{p}{\tau}}\right) \\
&= \rho^p \underline{s}^{-p} - 2^{\frac{1}{\tau}} p c_{\underline{s}}^{-\frac{1}{\tau}} \underline{s}^{-\frac{p}{\tau}-p-1} \mathbb{Q}(E)^{-\frac{1}{\tau}} \rho^{\frac{p}{\tau}+p},
\end{aligned}$$

where the second inequality comes from that $(1+x)^{-p} \geq 1 - px$ for $p \geq 1$ and $x \geq 0$. Letting $\bar{\rho} = \left(G(\underline{s} + \Delta_G)\mathbb{Q}(E)\right)^{\frac{1}{p}} \underline{s}$ yields the result. □

EC.2.2. Proof for Example 1

LEMMA EC.3. *Let $p \in [1,\infty)$. In Example 1, $\mathcal{R}_{\mathbb{Q},p}(\rho)$ is monotonically decreasing with respect to p.*

Proof of Lemma EC.3 . For every $z \in E = (0, \frac{1}{2})$, it holds that $\mathrm{d}_{E^c}(z) = \min\left(z, \frac{1}{2} - z\right) \le \frac{1}{4}$. Hence we have

$$\rho_{\max} = \left(\mathbb{E}_{\mathbb{Q}}\left[\mathbf{1}_E(Z)\mathrm{d}_{E^c}(Z)^p\right]\right)^{1/p} = \left(\int_0^{\frac{1}{4}} s^p ds\right)^{1/p} = \left(\frac{1}{4}\right)^{1+\frac{1}{p}} (p+1)^{-1/p},$$

and note that $\frac{1}{32} \le \left(\frac{1}{4}\right)^{1+\frac{1}{p}} (p+1)^{-1/p} < \frac{1}{4}$ for all $p \ge 1$. Define $F(p) := \ln\left((p+1)^{\frac{1}{p+1}} \rho^{\frac{p}{p+1}}\right)$ and then $F'(p) = (1 - \ln(p+1) + \ln\rho)(p+1)^{-2}$.

Fixing $\rho > 0$, consider the following three cases:

(I) $0 < \rho \le \frac{1}{32}$.

In this case, $\lambda_* > 0$ and $\mathcal{R}_{\mathbb{Q},p}(\rho) = (p+1)^{\frac{1}{p+1}} \rho^{\frac{p}{p+1}}$. Observe that for all $p \ge 1$,

$$F'(p) = (1 - \ln(p+1) + \ln\rho)\,(p+1)^{-2} < 0.$$

Hence $\mathcal{R}_{\mathbb{Q},p}(\rho)$ is decreasing as p increases.

(II) $\rho \ge \frac{1}{4}$.

In this case $\lambda_* = 0$ and $\mathcal{R}_{\mathbb{Q},p}(\rho) = 1$ for all $p \ge 1$.

(III) $\frac{1}{32} < \rho < \frac{1}{4}$.

In this case, it follows from the strict monotonicity of the function $p \mapsto \left(\frac{1}{4}\right)^{1+\frac{1}{p}} (p+1)^{-1/p}$ on $p \ge 1$ that there exists a unique $p_* > 1$ such that $\left(\frac{1}{4}\right)^{1+\frac{1}{p}} (p+1)^{-1/p} \le \rho$ when $p \in [1, p_*]$ and $\left(\frac{1}{4}\right)^{1+\frac{1}{p}} (p+1)^{-1/p} > \rho$ when $p \in (p_*, \infty)$. Then, $\mathcal{R}_{\mathbb{Q},p}(\rho) = 1$ for $p \in [1, p_*]$ and $\mathcal{R}_{\mathbb{Q},p}(\rho) = (p+1)^{\frac{1}{p+1}} \rho^{\frac{p}{p+1}}$ for $p \in (p_*, \infty)$. Observe that when $p \in (p_*, \infty)$,

$$\begin{aligned} F'(p) = (1 - \ln(p+1) + \ln\rho)\,(p+1)^{-2} &< \left(1 - \ln(p+1) + \ln\left(4^{-1-\frac{1}{p}}(p+1)^{-\frac{1}{p}}\right)\right)(p+1)^{-2} \\ &< \left(1 - \frac{p+1}{p}\ln(p+1)\right)(p+1)^{-2} \\ &\le 0, \end{aligned}$$

where the last inequality comes from the fact that $\ln(1+x) \ge \frac{x}{x+1}$ for $x > 0$. Thus, it implies that $\mathcal{R}_{\mathbb{Q},p}(\rho)$ is decreasing as p increases.

In all, $\mathcal{R}_{\mathbb{Q},p}(\rho)$ is monotonically decreasing with respect to p. □

EC.2.3. Proof of Theorem 3

(I) We first derive the upper bound of $\mathcal{R}_{\mathbb{Q},p}(\rho)$. It follows from (3.2) that

$$\begin{aligned} \mathcal{R}_{\mathbb{Q},p}(\rho) &\le \inf_{\lambda>0}\left\{\lambda\rho^p + \mathbb{Q}(E)\int_0^{\lambda^{-1/p}} (1 - \lambda s^p) dG(s)\right\} \\ &\overset{\tau=\lambda^{-1/p}}{=} \inf_{\tau>0}\left\{\tau^{-p}\rho^p + \mathbb{Q}(E)\int_0^{\tau} (1 - \tau^{-p}s^p) dG(s)\right\} \\ &\le \inf_{0<\tau<\delta}\left\{\tau^{-p}\rho^p + \mathbb{Q}(E)\int_0^{\tau} (1 - \tau^{-p}s^p) g(s) ds\right\}, \end{aligned}$$

where δ is from Assumption 1. By Assumption 1,

$$\int_0^{\tau} (1 - \tau^{-p}s^p) g(s) ds \le \int_0^{\tau} (1 - \tau^{-p}s^p)(g(0) + L_g s) ds = \frac{p}{p+1} g(0)\tau + \frac{p}{2(p+2)} L_g \tau^2.$$

It follows that

$$\mathcal{R}_{\mathbb{Q},p}(\rho) \le \inf_{0<\tau<\delta}\left\{\tau^{-p}\rho^p + \frac{p}{p+1}\mathbb{Q}(E)g(0)\tau + \frac{p}{2(p+2)}\mathbb{Q}(E)L_g\tau^2\right\}. \tag{EC.3}$$

When $g(0)>0$ and $\rho \le (p+1)^{-\frac{1}{p}}\left(g(0)\mathbb{Q}(E)\right)^{\frac{1}{p}}\delta^{\frac{p+1}{p}}$, we have $\bar{\tau} := (p+1)^{\frac{1}{p+1}}\left(g(0)\mathbb{Q}(E)\right)^{-\frac{1}{p+1}}\rho^{\frac{p}{p+1}} < \delta$, and

$$\begin{aligned}
\mathcal{R}_{\mathbb{Q},p}(\rho) &\le \inf_{0<\tau\le\bar{\tau}}\left\{\tau^{-p}\rho^p + \frac{p}{p+1}\mathbb{Q}(E)g(0)\tau + \frac{p}{2(p+2)}L_g\mathbb{Q}(E)\tau^2\right\}\\
&\le \inf_{0<\tau\le\bar{\tau}}\left\{\tau^{-p}\rho^p + \frac{pg(0)}{p+1}\mathbb{Q}(E)\tau + \frac{pL_g}{2(p+2)}(p+1)^{\frac{2}{p+1}}g(0)^{-\frac{2}{p+1}}\mathbb{Q}(E)^{\frac{p-1}{p+1}}\rho^{\frac{2p}{p+1}}\right\}\\
&= (p+1)^{\frac{1}{p+1}}\left(g(0)\mathbb{Q}(E)\right)^{\frac{p}{p+1}}\rho^{\frac{p}{p+1}} + \frac{pL_g}{2(p+2)}(p+1)^{\frac{2}{p+1}}g(0)^{-\frac{2}{p+1}}\mathbb{Q}(E)^{\frac{p-1}{p+1}}\rho^{\frac{2p}{p+1}}.
\end{aligned}$$

Next, we derive a lower bound of $\mathcal{R}_{\mathbb{Q},p}(\rho)$. Notice that when $g(0)>0$, under Assumption 1, there exist constants $0<\delta_1<\delta$ and $0<C_1\le C_2<\infty$ such that

$$0 < C_1 \le g(x) \le C_2 < \infty,$$

for any $x\in[0,\delta_1], y\in[0,\delta_1]$. Hence, $g(x) > C_1$ on $[0,\delta_1]$. Let $\rho < \left(\frac{C_1\mathbb{Q}(E)}{p+1}\right)^{\frac{1}{p}}\delta_1^{\frac{p+1}{p}}$. We claim that $\lambda_*^{-1/p} \le \delta_1$. Indeed, suppose on the contrary that $\lambda_*^{-1/p} > \delta_1$. Then it follows from Lemma 1 that

$$\rho^p \ge \mathbb{E}_{\mathbb{Q}}\left[\mathsf{d}_{E^c}(Z)^p\mathbf{1}_{\{\mathcal{Z}_*\setminus\mathcal{Z}_=\}}(Z)\right] \ge \mathbb{Q}(E)\int_0^{\delta_1} s^p dG(s) \ge C_1\mathbb{Q}(E)\int_0^{\delta_1} s^p ds \quad = \frac{C_1\mathbb{Q}(E)}{p+1}\delta_1^{p+1} > \rho^p.$$

Contradiction arises. Hence, $\lambda_*^{-1/p} \le \delta_1$. Then it follows from Assumption 1 that

$$\begin{aligned}
\rho^p \ge F(\lambda_*) := \mathbb{Q}(E)\int_0^{\lambda_*^{-1/p}} s^p dG(s) &\ge \mathbb{Q}(E)\int_0^{\lambda_*^{-1/p}} s^p[g(0)-L_g s]ds\\
&= \frac{g(0)\mathbb{Q}(E)}{p+1}\lambda_*^{-\frac{p+1}{p}} - \frac{L_g\mathbb{Q}(E)}{p+2}\lambda_*^{-\frac{p+2}{p}}\\
&=: \underline{F}(\lambda_*).
\end{aligned}$$

Let $\rho < 2^{-\frac{p+1}{p^2}}\left(g(0)\mathbb{Q}(E)\right)^{\frac{1}{p}}(p+1)^{-\frac{1}{p}}\delta_1^{\frac{p+1}{p}}$. Then $\underline{\lambda} := \frac{1}{2}\left(g(0)\mathbb{Q}(E)\right)^{\frac{p}{p+1}}(p+1)^{-\frac{p}{p+1}}\rho^{-\frac{p^2}{p+1}} > \delta_1^{-p}$. When $\rho < \left(2^{-\frac{1}{p}} - 2^{-\frac{p+2}{p}}\right)^{\frac{p+1}{p}}(p+1)^{-\frac{1}{p}}g(0)^{\frac{p+2}{p}}\mathbb{Q}(E)^{\frac{1}{p}}L_g^{-\frac{p+1}{p}}$, it holds that

$$\begin{aligned}
F\left(\underline{\lambda}\right) \ge \underline{F}\left(\underline{\lambda}\right) &= 2^{\frac{p+1}{p}}\rho^p - \frac{2^{\frac{p+2}{p}}(p+1)^{\frac{p+2}{p+1}}}{(p+2)}L_g g(0)^{-\frac{p+2}{p+1}}\mathbb{Q}(E)^{-\frac{1}{p+1}}\rho^{\frac{p(p+2)}{p+1}}\\
&\ge 2^{\frac{p+1}{p}}\rho^p - 2^{\frac{p+2}{p}}(p+1)^{\frac{1}{p+1}}L_g g(0)^{-\frac{p+2}{p+1}}\mathbb{Q}(E)^{-\frac{1}{p+1}}\rho^{\frac{p(p+2)}{p+1}}\\
&> \rho^p,
\end{aligned}$$

Since $F(\lambda)$ is strictly decreasing on $(\delta_1^{-p},+\infty)$, we have $\lambda_* > \underline{\lambda}$. It follows that

$$\begin{aligned}
\mathcal{R}_{\mathbb{Q},p}(\rho) &= \inf_{\lambda>\underline{\lambda}} \left\{ \lambda\rho^p + \int_0^{\lambda^{-1/p}} (1-\lambda s^p)\mathbb{Q}(E)dG(s) \right\} \\
&= \inf_{0<\tau<\underline{\lambda}^{-\frac{1}{p}}} \left\{ \tau^{-p}\rho^p + \mathbb{Q}(E)\int_0^{\tau} (1-\tau^{-p}s^p)g(s)ds \right\} \\
&\geq \inf_{0<\tau<\underline{\lambda}^{-\frac{1}{p}}} \left\{ \tau^{-p}\rho^p + \mathbb{Q}(E)\int_0^{\tau} (1-\tau^{-p}s^p)[g(0)-L_g s]ds \right\} \\
&= \inf_{0<\tau<\underline{\lambda}^{-\frac{1}{p}}} \left\{ \tau^{-p}\rho^p + \frac{p}{p+1}g(0)\mathbb{Q}(E)\tau - \frac{p\mathbb{Q}(E)}{2(p+2)}L_g\tau^2 \right\} \\
&\geq \inf_{0<\tau<\underline{\lambda}^{-\frac{1}{p}}} \left\{ \tau^{-p}\rho^p + \frac{pg(0)\mathbb{Q}(E)}{p+1}\tau \right\} - \frac{p\mathbb{Q}(E)}{2(p+2)}L_g\underline{\lambda}^{-\frac{2}{p}} \\
&= (p+1)^{\frac{1}{p+1}}\left(g(0)\mathbb{Q}(E)\right)^{\frac{p}{p+1}}\rho^{\frac{p}{p+1}} - \frac{2^{\frac{2-p}{p}}pL_g}{(p+2)}(p+1)^{\frac{2}{p+1}}g(0)^{-\frac{2}{p+1}}\mathbb{Q}(E)^{\frac{p-1}{p+1}}\rho^{\frac{2p}{p+1}}.
\end{aligned}$$

The result is proved by letting

$$\bar{\rho} = \min\left((p+1)^{-\frac{1}{p}}(g(0)\mathbb{Q}(E))^{\frac{1}{p}}\delta^{\frac{p+1}{p}}, \left(\frac{C_1\mathbb{Q}(E)}{p+1}\right)^{\frac{1}{p}}\delta_1^{\frac{p+1}{p}},\ 2^{-\frac{p+1}{p^2}}(p+1)^{-\frac{1}{p}}(g(0)\mathbb{Q}(E))^{\frac{1}{p}}\delta^{\frac{p+1}{p}}, \right.$$
$$\left. \left(2^{\frac{-1}{p}}-2^{-\frac{p+2}{p}}\right)^{\frac{p+1}{p}}(p+1)^{-\frac{1}{p}}g(0)^{\frac{p+2}{p}}\mathbb{Q}(E)^{\frac{1}{p}}L_g^{-\frac{p+1}{p}} \right).$$

(II) Let $\rho < (p+2)^{-\frac{1}{p}}L_g^{\frac{1}{p}}\mathbb{Q}(E)^{\frac{1}{p}}\delta^{\frac{p+2}{p}}$. Then $(p+2)^{\frac{1}{p+2}}L_g^{-\frac{1}{p+2}}\mathbb{Q}(E)^{-\frac{1}{p+2}}\rho^{\frac{p}{p+2}} < \delta$ and from (EC.3), it holds that

$$\begin{aligned}
\mathcal{R}_{\mathbb{Q},p}(\rho) &\leq \inf_{0<\tau<\delta} \left\{ \tau^{-p}\rho^p + \frac{p}{2(p+2)}\mathbb{Q}(E)L_g\tau^2 \right\} \\
&= \frac{1}{2}(p+2)^{\frac{2}{p+2}}L_g^{\frac{p}{p+2}}\mathbb{Q}(E)^{\frac{p}{p+2}}\rho^{\frac{2p}{p+2}} \\
&\leq \left(\frac{p+2}{2}\right)^{\frac{2}{p+2}}L_g^{\frac{2}{p+2}}\mathbb{Q}(E)^{\frac{2}{p+2}}\rho^{\frac{2p}{p+2}}.
\end{aligned}$$

The proof is completed by setting $\bar{\rho} = (p+2)^{-\frac{1}{p}}L_g^{\frac{1}{p}}\mathbb{Q}(E)^{\frac{1}{p}}\delta^{\frac{p+2}{p}}$.

Appendix EC.3: Proof of Section 4.1

To ease the notation, denote $\alpha = \mathbb{P}_{\mathsf{true}}(E)$ and $\alpha_n = \alpha G(\rho_n)$.

EC.3.1. Proof of Theorem 4.

By Bernstein's inequality (e.g., [45]), it follows that

$$\mathbb{P}\left(|J - n\alpha_n| > tn^{\frac{1}{4}}\right) \leq 2\exp\left(-\frac{\frac{1}{2}t^2\sqrt{n}}{n\alpha_n(1-\alpha_n)+\frac{1}{3}tn^{\frac{1}{4}}}\right).$$

Notice that

$$\frac{\frac{1}{2}t^2\sqrt{n}}{n\alpha_n(1-\alpha_n)+\frac{1}{3}tn^{\frac{1}{4}}} \geq \frac{\frac{1}{2}t^2\sqrt{n}}{n\alpha_n+\frac{1}{3}tn^{\frac{1}{4}}} \geq \frac{3t^2\sqrt{n}}{6n\alpha(g(0)\rho_n+\frac{1}{2}L_g\rho_n^2)+2tn^{\frac{1}{4}}} \geq \frac{3t^2}{6\alpha g(0)\rho_0+3\alpha L_g\frac{\rho_0^2}{\sqrt{n}}+3\frac{t}{n^{\frac{1}{4}}}}$$
$$\geq \frac{t^2}{2\alpha g(0)\rho_0+1},$$

where the second inequality follows the Lemma EC.1 when $n \geq \frac{\rho_0^2}{\delta^2}$, and the last inequality holds since $3\alpha L_g\frac{\rho_0^2}{\sqrt{n}}+3tn^{-\frac{1}{4}} \leq 3$ when $n \geq (\alpha L_g\rho_0^2+t)^4$. Below we let $n \geq \max(\frac{\rho_0^2}{\delta^2}, (\alpha L_g\rho_0^2+t)^4)$. It follows that

$$\mathbb{P}\left(|J-n\alpha_n| > tn^{\frac{1}{4}}\right) \leq 2\exp\left(-\frac{t^2}{2\alpha g(0)\rho_0+1}\right),$$

which is equivalent to

$$\mathbb{P}\left(n\alpha_n - tn^{\frac{1}{4}} \leq J \leq n\alpha_n + tn^{\frac{1}{4}}\right) \geq 1-2\exp\left(-\frac{t^2}{2\alpha g(0)\rho_0+1}\right).$$

Therefore, whenever the event above holds, by Lemma 1 and Lemma EC.1,

$$\mathcal{R}_{\mathbb{P}_n,\infty}(\rho_n) < \frac{J+1}{n} \leq \frac{n\alpha G(\rho_n)}{n}+\frac{tn^{\frac{1}{4}}}{n} \leq \alpha g(0)\rho_n+\frac{L_g}{2}\frac{\alpha\rho_0^2}{n}+\frac{1}{n}+\frac{t}{n^{3/4}},$$

and

$$\mathcal{R}_{\mathbb{P}_n,\infty}(\rho_n) \geq \frac{J}{n} \geq \frac{n\alpha G(\rho_n)}{n}-\frac{tn^{\frac{1}{4}}}{n} = \alpha g(0)\rho_n-\frac{L_g}{2}\frac{\alpha\rho_0^2}{n}-\frac{t}{n^{3/4}}.$$

When $g(0)=0$, by Bernstein's inequality (e.g., [45]),

$$\mathbb{P}\left(|J-n\alpha_n| > t\right) \leq 2\exp\left(-\frac{\frac{1}{2}t^2}{n\alpha_n(1-\alpha_n)+\frac{1}{3}t}\right).$$

Notice that

$$\frac{\frac{1}{2}t^2}{n\alpha_n(1-\alpha_n)+\frac{1}{3}t} \geq \frac{3t^2}{6n\alpha_n+2t} \quad \geq \frac{3t^2}{6n\alpha(\frac{1}{2}L_g\rho_n^2)+2t} = \frac{3t^2}{3\alpha L_g\rho_0^2+2t}.$$

Then Bernstein's inequality above implies that

$$\mathbb{P}\left(|J-n\alpha_n| > t\right) \leq 2\exp\left(-\frac{3t^2}{3\alpha L_g\rho_0^2+2t}\right).$$

Therefore

$$\mathcal{R}_{\mathbb{P}_n,\infty}(\rho_n) < \frac{J+1}{n} \leq \frac{n\alpha G(\rho_n)+t+1}{n} \leq \alpha g(0)\rho_n+\frac{L_g}{2}\frac{\alpha\rho_0^2}{n}+\frac{1+t}{n} = \frac{L_g\alpha\rho_0^2+2t+2}{2n},$$

and

$$\mathcal{R}_{\mathbb{P}_n,\infty}(\rho_n) \geq \frac{J}{n} \geq \frac{n\alpha G(\rho_n)}{n}-\frac{t}{n} = \alpha g(0)\rho_n-\frac{L_g}{2}\frac{\alpha\rho_0^2}{n}-\frac{t}{n} = -\frac{L_g\alpha\rho_0^2+t}{2n},$$

The proof is completed by setting $C = \frac{1}{2}L_g\alpha\rho_0^2+2t+2$ and $n_0 = \max\left(\frac{\rho_0^2}{\delta^2}, (\alpha L_g\rho_0^2+s)^4\right)$.

EC.3.2. Proof of Proposition 2.

Recall that $J \sim B(n,\alpha_n)$, $\sigma_n^2 = n\alpha_n(1-\alpha_n)$. We first derive an asymptotic anti-concentration inequality for J.

LEMMA EC.4. *Under the setting in Proposition 2, let $t>0$. Then*

$$\lim_{n\to\infty} \mathbb{P}\Big(|J-n\alpha_n| \ge t\sigma_n \Big) \ge \frac{2\sqrt{2}}{\sqrt{\pi}(\sqrt{t^2+4}+t)} e^{-\frac{t^2}{2}}.$$

Proof. For two sequences of positive real numbers $\{a_n\}_n$, $\{b_n\}_n$, we use $a_n \lesssim b_n$ to represent $\lim_{n\to\infty} \frac{a_n}{b_n} \le 1$, and $a_n \simeq b_n$ to represent $\lim_{n\to\infty} \frac{a_n}{b_n} = 1$. By Stirling's formula, we have that

$$\begin{aligned}
\mathbb{P}\Big(n\alpha_n - t\sigma_n < J < n\alpha_n + t\sigma_n\Big) &\le \sum_{-t\sigma_n < k-n\alpha_n < t\sigma_n} \binom{n}{k} \alpha_n^k (1-\alpha_n)^{n-k} \\
&\simeq \sum_{-t\sigma_n < k-n\alpha_n < t\sigma_n} \frac{1}{\sqrt{2\pi}} \sqrt{\frac{n}{k(n-k)}} \frac{n^n}{k^k (n-k)^{n-k}} \alpha_n^k (1-\alpha_n)^{n-k} \\
&= \sum_{-t\sigma_n < k-n\alpha_n < t\sigma_n} \frac{1}{\sqrt{2\pi}} \sqrt{\frac{n}{k(n-k)}} \left(\frac{n\alpha_n}{k}\right)^k \left(\frac{n(1-\alpha_n)}{n-k}\right)^{n-k}.
\end{aligned}$$

Let $x = \frac{k-n\alpha_n}{\sigma_n} = \frac{k-n\alpha_n}{\sqrt{n\alpha_n(1-\alpha_n)}}$, then

$$\begin{aligned}
&\left(\frac{n\alpha_n}{k}\right)^k \left(\frac{n(1-\alpha_n)}{n-k}\right)^{n-k} \\
&= \exp\left\{ \ln\left(\left(\frac{n\alpha_n}{k}\right)^k \right) + \ln\left(\left(\frac{n-n\alpha_n}{n-k}\right)^{n-k} \right) \right\} \\
&= \exp\left\{ -k\ln\left(\frac{k}{n\alpha_n}\right) + (k-n)\ln\left(\frac{n-k}{n-n\alpha_n}\right) \right\} \\
&= \exp\left\{ -k\ln\left(\frac{n\alpha_n + x\sqrt{n\alpha_n(1-\alpha_n)}}{n\alpha_n}\right) + (k-n)\ln\left(\frac{n-n\alpha_n - x\sqrt{n\alpha_n(1-\alpha_n)}}{n(1-\alpha_n)}\right) \right\} \\
&= \exp\left\{ -k\ln\left(1 + x\sqrt{\frac{1-\alpha_n}{n\alpha_n}}\right) + (k-n)\ln\left(1 - x\sqrt{\frac{\alpha_n}{n(1-\alpha_n)}}\right) \right\} \\
&\simeq \exp\left\{ -k\left(x\sqrt{\frac{1-\alpha_n}{n\alpha_n}} - \frac{x^2(1-\alpha_n)}{2n\alpha_n}\right) + (k-n)\left(-x\sqrt{\frac{\alpha_n}{n(1-\alpha_n)}} - \frac{x^2\alpha_n}{2n(1-\alpha_n)}\right) \right\} \\
&= \exp\left\{ (-n\alpha_n - x\sqrt{n\alpha_n(1-\alpha_n)}) \left(x\sqrt{\frac{1-\alpha_n}{n\alpha_n}} - \frac{x^2(1-\alpha_n)}{2n\alpha_n}\right) \right. \\
&\qquad \left. + (n\alpha_n + x\sqrt{n\alpha_n(1-\alpha_n)} - n)\left(-x\sqrt{\frac{\alpha_n}{n(1-\alpha_n)}} - \frac{x^2\alpha_n}{2n(1-\alpha_n)}\right) \right\} \\
&\simeq \exp\left\{ \left(-x\sqrt{n\alpha_n(1-\alpha_n)} + \frac{1}{2}x^2(1-\alpha_n) - x^2(1-\alpha_n)\right) + \left(x\sqrt{n\alpha_n(1-\alpha_n)} + \frac{1}{2}x^2\alpha_n - x^2\alpha_n\right) \right\} \\
&= \exp\left\{ -\frac{1}{2}x^2(1-\alpha_n) - \frac{1}{2}x^2\alpha_n \right\} \\
&= \exp\left\{ -\frac{1}{2}x^2 \right\} \\
&= e^{-\frac{(k-n\alpha_n)^2}{2\sigma_n^2}},
\end{aligned}$$

where the first $\simeq$ comes from Taylor's expansion with the fact that $n\alpha_n \geq \alpha g(0)\rho_0\sqrt{n} - \frac{\alpha}{2}L_g\rho_0^2 \to \infty$ as $n \to \infty$. It follows that

$$\mathbb{P}\Big(n\alpha_n - t\sigma_n < J < n\alpha_n + t\sigma_n\Big) \lesssim \sum_{-t\sigma_n < k - n\alpha_n < t\sigma_n} \frac{1}{\sqrt{2\pi}}\sqrt{\frac{n}{k(n-k)}}e^{-\frac{(k-n\alpha_n)^2}{2\sigma_n^2}},$$

Notice that

$$\begin{aligned}
\mathbb{P}\Big(n\alpha_n - t\sigma_n < J < n\alpha_n + t\sigma_n\Big) &\lesssim \sum_{-t\sigma_n < k - n\alpha_n < t\sigma_n} \frac{1}{\sqrt{2\pi}}\sqrt{\frac{n}{k(n-k)}}e^{-\frac{(k-n\alpha_n)^2}{2\sigma_n^2}} \\
&\lesssim \sum_{-t\sigma_n < k - n\alpha_n < t\sigma_n} \frac{1}{\sqrt{2\pi}}\sqrt{\frac{n}{(n\alpha_n - t\sigma_n)(n - n\alpha_n - t\sigma_n)}}e^{-\frac{(k-n\alpha_n)^2}{2\sigma_n^2}} \\
&= \sum_{-t < \frac{k-n\alpha_n}{\sigma_n} < t} \frac{1}{\sqrt{2\pi}}\sigma_n\sqrt{\frac{n}{(n\alpha_n - t\sigma_n)(n - n\alpha_n - t\sigma_n)}}e^{-\frac{(k-n\alpha_n)^2}{2\sigma_n^2}}\frac{1}{\sigma_n} \\
&\lesssim \sqrt{\frac{n^2\alpha_n(1-\alpha_n)}{(n\alpha_n - t\sigma_n)(n - n\alpha_n + t\sigma_n)}}\int_{-t}^{t}\frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}dx \\
&\simeq 1 - 2\int_t^{\infty}\frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}dx \\
&\leq 1 - \frac{2\sqrt{2}}{\sqrt{\pi}(\sqrt{t^2+4}+t)}e^{-\frac{t^2}{2}}.
\end{aligned}$$

where the second inequality is based on the fact $n > n\alpha_n + t\sqrt{n\alpha_n(1-\alpha_n)}$ for sufficiently large n, and the last inequality comes from the inequality in [3] that $\frac{1}{\sqrt{2\pi}}\int_t^{\infty}e^{-\frac{x^2}{2}}dx \geq \frac{1}{\sqrt{2\pi}}\frac{2}{(\sqrt{t^2+4}+t)}e^{-t^2/2}$. Therefore,

$$\begin{aligned}
\mathbb{P}\Big(\,|J - n\alpha_n| \geq t\sigma_n\Big) = 1 - \mathbb{P}\Big(n\alpha_n - t\sigma_n < J < n\alpha_n + t\sigma_n\Big) &\gtrsim 1 - \left(1 - \frac{2\sqrt{2}}{\sqrt{\pi}(\sqrt{t^2+4}+t)}e^{-\frac{t^2}{2}}\right) \\
&= \frac{2\sqrt{2}}{\sqrt{\pi}(\sqrt{t^2+4}+t)}e^{-\frac{t^2}{2}}.
\end{aligned}$$

Hence the proof is completed. □

Proof of Proposition 2. When $g(0) > 0$, it holds that $\alpha g(0)\rho_n - \frac{\alpha}{2}L_g\rho_n^2 \leq \alpha_n \leq \alpha g(0)\rho_n + \frac{\alpha}{2}L_g\rho_n^2$. Then,

$$\sigma_n = \sqrt{n\alpha_n(1-\alpha_n)} \geq \sqrt{n\Big(\alpha g(0)\rho_n - \frac{\alpha}{2}L_g\rho_n^2\Big)} \geq \sqrt{\frac{1}{2}\alpha g(0)\rho_0\sqrt{n}} \geq \sqrt{\frac{1}{2}\alpha g(0)\rho_0}\cdot n^{\frac{1}{4}},$$

where the second inequality is due to the fact that $\frac{\alpha}{2}g(0)\rho_n \geq \frac{\alpha}{2}L_g\rho_n^2$ when $n \geq L_g^2\rho_0^2 g(0)^{-2}$. Applying Lemma EC.4 gives

$$\mathbb{P}\Big(\,|J - n\alpha_n| \geq t\sqrt{\frac{1}{2}\alpha g(0)\rho_0}\cdot n^{\frac{1}{4}}\Big) \geq \mathbb{P}\Big(\,|J - n\alpha_n| \geq t\sigma_n\Big) \gtrsim \frac{2\sqrt{2}}{\sqrt{\pi}(\sqrt{t^2+4}+t)}e^{-\frac{t^2}{2}}.$$

Hence, when $n > \left(\frac{L_g \alpha \rho_0^2 + 2}{t\sqrt{\frac{1}{2}\alpha g(0)\rho_0}}\right)^4$, by the Lemma EC.1, at least one of

$$\begin{aligned}\mathcal{R}_{\mathbb{P}_n,\infty}(\rho_n) < \frac{J+1}{n} &\le \frac{n\alpha G(\rho_n)}{n} + \frac{-t\sqrt{\frac{1}{2}\alpha g(0)\rho_0}n^{\frac{1}{4}} + 1}{n} \\ &\le \alpha g(0)\rho_n + \frac{L_g}{2}\frac{\alpha\rho_0^2}{n} + \frac{1}{n} - \frac{t\sqrt{\frac{1}{2}\alpha g(0)\rho_0}}{n^{3/4}} \\ &\le \alpha g(0)\rho_n - \frac{t\sqrt{\frac{1}{2}\alpha g(0)\rho_0}}{2n^{3/4}},\end{aligned}$$

and

$$\begin{aligned}\mathcal{R}_{\mathbb{P}_n,\infty}(\rho_n) \ge \frac{J}{n} \ge \frac{n\alpha G(\rho_n)}{n} + \frac{t\sqrt{\frac{1}{2}\alpha g(0)\rho_0}n^{\frac{1}{4}}}{n} &= \alpha g(0)\rho_n - \frac{L_g}{2}\frac{\alpha\rho_0^2}{n} + \frac{t\sqrt{\frac{1}{2}\alpha g(0)\rho_0}}{n^{3/4}} \\ &\ge \alpha g(0)\rho_n + \frac{t\sqrt{\frac{1}{2}\alpha g(0)\rho_0}}{2n^{3/4}},\end{aligned}$$

holds with probability at least $\frac{2\sqrt{2}}{\sqrt{\pi}(\sqrt{t^2+4}+t)}e^{-\frac{t^2}{2}}$ which implies that

$$\lim_{n\to+\infty}\left|\mathcal{R}_{\mathbb{P}_n,\infty}(\rho_n) - \alpha g(0)\rho_n\right| - \frac{t\sqrt{2\alpha g(0)\rho_0}}{4n^{3/4}} \ge 0,$$

The proof is completed by setting $C = \frac{t\sqrt{2\alpha g(0)\rho_0}}{4}$. □

Appendix EC.4: Proofs for Section 4.2

EC.4.1. Proof of Theorem 5

We first state several useful preparatory results, whose proofs are postponed to Section EC.4.1.4. Recall that $U_j \sim Beta(j, N+1-j)$, $j = 1,\dots,N$. Define $\sigma_j = \frac{\sqrt{j(N+1-j)}}{(N+1)\sqrt{N+2}}$, $j = 1,\dots,N$, namely the standard deviation of U_j.

Lemma EC.5. *Let $t_1 > 0$ and $t_2 > 0$. It holds with the probability at least $1 - n\exp\left(-\frac{t_1^2}{2+t_1}\right) - 2\exp(-2t_2^2)$ that for any $1 \le j \le N$,*

$$\mathbb{E}[U_j] - t_1\sigma_j \le U_j(\omega) \le \mathbb{E}[U_j] + t_1\sigma_j,$$

and $|N(\omega) - n\alpha| \le t_2\sqrt{n}$.

By the definition of $\underline{s}$, it holds that with probability one, $U_j \ge \underline{s}$ for all $1 \le j \le N$. Then by Lemma EC.5, there exists a subset $\Omega_0 \subset \Omega$ with $\mathbb{P}(\Omega_0) \ge 1 - n\exp\left(-\frac{t_1^2}{2+t_1}\right) - 2\exp(-2t_2^2)$, such that for all $\omega \in \Omega_0$, $\max\left(G(\underline{s}), \mathbb{E}[U_j] - t_1\sigma_j\right) \le U_j(\omega) \le \mathbb{E}[U_j] + t_1\sigma_j$ for $1 \le j \le N(\omega)$ and $|N(\omega) - n\alpha| \le t_2\sqrt{n}$. Define

$$\begin{aligned} n_1 &:= \tfrac{4t_2^2}{\alpha^2}, \\ n_2 &:= \left(2\alpha^{-1}G(\delta)^{-1}\left(\rho_0^p\underline{s}^{-p} + t_1\sqrt{\rho_0^p\underline{s}^{-p}}\right)\right)^{(\min\{bp,(bp+1)/2\})^{-1}}. \end{aligned} \tag{EC.4}$$

Lemma EC.6. *Suppose $G(\delta) > 0$ and $n > \lceil\max(n_1, n_2)\rceil$. Then for all $\omega \in \Omega_0$, it holds that*

$$|N(\omega) - n\alpha| \le t_2\sqrt{n}, \quad G(\underline{s}) \le U_j(\omega) \le \mathbb{E}[U_j] + t_1\sigma_j < G(\delta), \quad \forall 1 \le j \le N(\omega).$$

LEMMA EC.7. *Assume Assumption 1 holds with $G(\underline{s}) = 0$ and $g(\underline{s}) > 0$. Then there exist constants $\delta_1, M, L_{G^{-1}} > 0$ such that $G^{-1}(x)$ exists on $[0, G(\delta_1)]$, and for all $x_1, x_2 \in [0, G(\delta_1)]$,*

$$M|x_1 - x_2| \le |G^{-1}(x_1) - G^{-1}(x_2)| \le L_{G^{-1}}|x_1 - x_2|.$$

Note that here the generalized inverse of G at 0 is $[0, \underline{s}]$; yet to ease the presentation, we set $G^{-1}(0) = \underline{s}$.

LEMMA EC.8. *Suppose $G(\delta) > 0$ and $G(\underline{s}) = 0$. Let $n > \lceil \max(n_1, n_2) \rceil$. Then for all $\omega \in \Omega_0$, it holds that*

$$n\rho_n^p \delta^{-p} - 1 < J(\omega) \le n\rho_n^p \underline{s}^{-p}.$$

EC.4.1.1. Proof of Theorem 5(I) Define

$$n_3 := 2^{\frac{1}{bp}} \alpha_{\underline{s}}^{-\frac{1}{bp}} \underline{s}^{-\frac{1}{b}} \rho_0^{\frac{1}{b}},$$
$$n_4 := 4t^2/\alpha_{\underline{s}}^2.$$

where $\alpha_{\underline{s}} := \alpha G(\underline{s})$.

Proof of Theorem 5 (I). Since $G(\underline{s}) > 0$, it holds with probability 1 that for all $1 \le j \le N$, $S_j \ge \underline{s}$. Define $N_{\underline{s}} = \max\{0 \le j \le N : S_j = \underline{s}\}$. Then $N_{\underline{s}} \sim Binomial\left(n, \alpha G(\underline{s})\right)$. Applying Bernstein's inequality (e.g., [45]) gives that

$$\mathbb{P}\left(N_{\underline{s}} - n\alpha_{\underline{s}} > t\sqrt{n}\right) \le \exp\left(-\frac{t^2}{2\alpha_{\underline{s}}(1-\alpha_{\underline{s}}) + \frac{2}{3}\frac{t}{\sqrt{n}}}\right) \le \exp\left(-\frac{t^2}{2\alpha_{\underline{s}}(1-\alpha_{\underline{s}}) + \frac{2}{3}}\right),$$

where the second inequality holds for $n \ge t^2$ when $n > n_1$. Observe that when $n \ge \max(n_3, n_4)$, it holds that $\left(n\alpha_{\underline{s}} - t\sqrt{n}\right) \cdot \frac{1}{n}\underline{s}^p \ge \frac{1}{2}\alpha_{\underline{s}}\underline{s}^p \ge \rho_n^p$, hence the worst-case distribution transports only points in $\mathcal{Z}_{N_{\underline{s}}}$. Then by (4) we have

$$J\underline{s}^p \le n\rho_n^p < (J+1)\underline{s}^p,$$

which is equivalent to

$$\rho_n^p \underline{s}^{-p} - 1 < J \le \rho_n^p \underline{s}^{-p}.$$

Therefore, the result follows from (5), by setting $n_0 = \lceil \max(n_1, n_3, n_4) \rceil$.

EC.4.1.2. Proof of Theorem 5(II)
Based on Lemma EC.8, we can derive a finer bounds for $J(\omega)$.
Define $\delta_3 := \min\{s > 0 : G(s) = G(\delta_1)/2\}$, and

$$n_5 := \left(2\alpha^{-1}G(\delta_1)^{-1}\left(\rho_0^p \underline{s}^{-p} + t_1\sqrt{\rho_0^p \underline{s}^{-p}}\right)\right)^{(\min\{bp,(bp+1)/2\})^{-1}},$$
$$n_6 := 2(\zeta_1 + \zeta_2 + \zeta_3),$$
$$n_7 := 2^{\frac{1}{bp}}(\zeta_1 + \zeta_2 + \zeta_3)^{\frac{1}{bp}},$$

where

$$\zeta_1 = 2p\alpha^{-1}\rho_0^p \underline{s}^{-(p+1)} L_{G^{-1}},$$
$$\zeta_2 = \tfrac{8}{3}t_1 p\alpha^{-1}\rho_0^{\frac{p}{2}} \underline{s}^{-(\frac{p}{2}+1)} L_{G^{-1}},$$
$$\zeta_3 = (2 + \tfrac{8}{3}t_1)p\alpha^{-1}\underline{s}^{-1} L_{G^{-1}}.$$

Note that due to $\delta_1 < \delta$, we have $n_4 > n_2$ and $n_5 > n_3$.

LEMMA EC.9. *When $G(\underline{s})=0$ with $g(\underline{s})>0$, assume Assumption 1 holds and $n>\max(n_1,n_5,n_6,n_7)$. Let $t_1>0$ and $t_2>0$. Then for any $\omega\in\Omega_0$, it holds that*

$$\left(1+\frac{\eta_1}{n^{bp}}+\frac{\eta_2}{n^{\frac{bp+1}{2}}}+\frac{\eta_3}{n}\right)^{-1} n\rho_n^p\underline{s}^{-p}-1<J(\omega)\le\left(1-\frac{\zeta_1}{n^{bp}}-\frac{\zeta_2}{n^{\frac{bp+1}{2}}}-\frac{\zeta_3}{n}\right)^{-1} n\rho_n^p\underline{s}^{-p},$$

where

$$\begin{aligned}
\eta_1&=2p\alpha^{-1}\rho_0^p\underline{s}^{-(p+1)}L_{G^{-1}}\left(1+(1+t_1)\underline{s}^{-1}L_{G^{-1}}\right)^{p-1},\\
\eta_2&=\tfrac{8}{3}pt_1\alpha^{-1}\rho_0^{\frac{p}{2}}\underline{s}^{-(\frac{p}{2}+1)}L_{G^{-1}}\left(1+(1+t_1)\underline{s}^{-1}L_{G^{-1}}\right)^{p-1},\\
\eta_3&=\left(4+\tfrac{8\sqrt{2}}{3}t_1\right)p\alpha^{-1}\underline{s}^{-1}L_{G^{-1}}\left(1+(1+t_1)\underline{s}^{-1}L_{G^{-1}}\right)^{p-1}.
\end{aligned}$$

Proof of Lemma EC.9. Let $\omega\in\Omega_0$. Since $n>\max(n_1,n_5)$, applying Lemma EC.6 with $\delta=\delta_1$ yields that for $1\le j\le J(\omega)$,

$$U_j(\omega)\le\mathbb{E}[U_j]+t_1\sigma_j<G(\delta_1).$$

By Lemma EC.7, for every $1\le j\le J(\omega)$,

$$\begin{aligned}
S_j(\omega)=G^{-1}(U_j(\omega))&\le G^{-1}(E[U_j]+t_1\sigma_j)\le G^{-1}(0)+L_{G^{-1}}(E[U_j]+t_1\sigma_j)\\
&=\underline{s}+L_{G^{-1}}(E[U_j]+t_1\sigma_j),
\end{aligned}\tag{EC.5}$$

and

$$\begin{aligned}
S_j(\omega)=G^{-1}(U_j(\omega))&\ge G^{-1}(E[U_j])-L_{G^{-1}}|U_j(\omega)-E[U_j]|\\
&\ge G^{-1}(E[U_j])-L_{G^{-1}}t_1\sigma_j\\
&\ge G^{-1}(0)-L_{G^{-1}}E[U_j]-L_{G^{-1}}t_1\sigma_j\\
&=\underline{s}-L_{G^{-1}}(E[U_j]+t_1\sigma_j),
\end{aligned}\tag{EC.6}$$

We first derive the lower bound for $J(\omega)$. From (4) and (EC.5), it holds that

$$\begin{aligned}
n\rho_n^p&<\sum_{j=1}^{J(\omega)+1}S_j(\omega)^p\\
&\le\sum_{j=1}^{J(\omega)+1}\left(\underline{s}+L_{G^{-1}}(E[U_j]+t_1\sigma_j)\right)^p\\
&\le\sum_{j=1}^{J(\omega)+1}\underline{s}^p\left(1+\underline{s}^{-1}L_{G^{-1}}\left(\frac{j}{N(\omega)+1}+t_1\frac{\sqrt{j}}{N(\omega)+1}\right)\right)^p.
\end{aligned}$$

Due to the inequality that $\left(1+\underline{s}^{-1}L_{G^{-1}}x\right)^p \le 1+p\underline{s}^{-1}L_{G^{-1}}\left(1+(1+t_1)\underline{s}^{-1}L_{G^{-1}}\right)^{p-1}x$ for $p\ge 1$ and $0\le x\le 1+t_1$, it follows that

$$\begin{aligned}
n\rho_n^p &< \sum_{j=1}^{J(\omega)+1}\underline{s}^p\left(1+p\underline{s}^{-1}L_{G^{-1}}\left(1+(1+t_1)\underline{s}^{-1}L_{G^{-1}}\right)^{p-1}\left(\frac{j}{N(\omega)+1}+t_1\frac{\sqrt{j}}{N(\omega)+1}\right)\right)\\
&= \underline{s}^p(J(\omega)+1)+\sum_{j=1}^{J(\omega)+1}p\underline{s}^{p-1}L_{G^{-1}}\left(1+(1+t_1)\underline{s}^{-1}L_{G^{-1}}\right)^{p-1}\left(\frac{j}{N(\omega)+1}+t_1\frac{\sqrt{j}}{N(\omega)+1}\right)\\
&\le \underline{s}^p\left(J(\omega)+1+p\underline{s}^{-1}L_{G^{-1}}\left(1+(1+t_1)\underline{s}^{-1}L_{G^{-1}}\right)^{p-1}\int_0^{J(\omega)+2}\left(\frac{x}{N(\omega)+1}+t_1\frac{\sqrt{x}}{N(\omega)+1}\right)dx\right)\\
&\le \underline{s}^p(J(\omega)+1)\left(1+\frac{p\underline{s}^{-1}L_{G^{-1}}}{N(\omega)+1}\left(1+(1+t_1)\underline{s}^{-1}L_{G^{-1}}\right)^{p-1}\left(\frac{(J(\omega)+2)^2}{2(J(\omega)+1)}+t_1\frac{2(J(\omega)+2)^{\frac{3}{2}}}{3(J(\omega)+1)}\right)\right)\\
&\le \underline{s}^p(J(\omega)+1)\left(1+p\underline{s}^{-1}L_{G^{-1}}\left(1+(1+t_1)\underline{s}^{-1}L_{G^{-1}}\right)^{p-1}\left(\frac{J(\omega)+2}{(N(\omega)+1)}+t_1\frac{4\sqrt{J(\omega)+2}}{3(N(\omega)+1)}\right)\right),
\end{aligned}$$

where the last inequality holds since $2(J(\omega)+1)\ge J(\omega)+2$. Using Lemma EC.8 and $n>n_1$ which implies $|N(\omega)-n\alpha|>t_2\sqrt{n}$,

$$\frac{J(\omega)+2}{N(\omega)+1}\le\frac{n\rho_n^p\underline{s}^{-p}+2}{N(\omega)}\le\frac{\rho_0^p\underline{s}^{-p}n^{1-bp}+2}{\frac{1}{2}\alpha n}=\frac{2\rho_0^p\underline{s}^{-p}}{\alpha n^{bp}}+\frac{4}{\alpha n},\tag{EC.7}$$

and

$$\frac{\sqrt{J(\omega)+2}}{N(\omega)+1}\le\frac{\sqrt{J(\omega)}+\sqrt{2}}{N(\omega)}\le\frac{\sqrt{n\rho_n^p\underline{s}^{-p}}+\sqrt{2}}{\frac{1}{2}\alpha n}=\frac{2\rho_0^{\frac{p}{2}}\underline{s}^{-\frac{p}{2}}}{\alpha n^{\frac{bp+1}{2}}}+\frac{2\sqrt{2}}{\alpha n},\tag{EC.8}$$

where the first inequality comes from the fact that $\sqrt{a+b}\le\sqrt{a}+\sqrt{b}$ for $a,b\ge 0$. Then it follows that

$$n\rho_n^p<\underline{s}^p(J(\omega)+1)\left(1+\frac{\eta_1}{n^{bp}}+\frac{\eta_2}{n^{\frac{bp+1}{2}}}+\frac{\eta_3}{n}\right),$$

which implies that

$$J(\omega)>\left(1+\frac{\eta_1}{n^{bp}}+\frac{\eta_2}{n^{\frac{bp+2}{2}}}+\frac{\eta_3}{n}\right)^{-1}n\rho_n^p\underline{s}^{-p}-1.$$

Next, we derive the upper bound for $J(\omega)$, $\omega\in\Omega_0$. Using (4) and (EC.6), it holds that

$$\begin{aligned}
n\rho_n^p &\ge \sum_{j=0}^{J(\omega)}S_j(\omega)^p\\
&\ge \sum_{j=0}^{J(\omega)}\left(\underline{s}-L_{G^{-1}}(E[U_j]+t_1\sigma_j)\right)_+^p\\
&\ge \sum_{j=0}^{J(\omega)}\underline{s}^p\left(1-\underline{s}^{-1}L_{G^{-1}}\left(\frac{j}{N(\omega)+1}+t_1\frac{\sqrt{j}}{N(\omega)+1}\right)\right)_+^p\\
&\ge \sum_{j=0}^{J(\omega)}\underline{s}^p\left(1-p\underline{s}^{-1}L_{G^{-1}}\left(\frac{j}{N(\omega)+1}+t_1\frac{\sqrt{j}}{N(\omega)+1}\right)\right),
\end{aligned}$$

where the last inequality follows from the inequality $(1-x)_+^p \geq 1-px$ for $p \geq 1$ and $x \geq 0$. If $J(\omega)=0$, then $J(\omega)$ is upper bounded by zero. If $J(\omega) \geq 1$, it holds that

$$\begin{aligned}
n\rho_n^p \underline{s}^{-p} &\geq \sum_{j=1}^{J(\omega)} \left(1 - p\underline{s}^{-1} L_{G^{-1}}\left(\frac{j}{N(\omega)+1} + t_1 \frac{\sqrt{j}}{N(\omega)+1}\right)\right) \\
&= J(\omega) - \sum_{j=0}^{J(\omega)} p\underline{s}^{-1} L_{G^{-1}}\left(\frac{j}{N(\omega)+1} + t_1 \frac{\sqrt{j}}{N(\omega)+1}\right) \\
&\geq J(\omega) - p\underline{s}^{-1} L_{G^{-1}} \int_0^{J(\omega)+1} \left(\frac{x}{N(\omega)+1} + t_1 \frac{\sqrt{x}}{N(\omega)+1}\right) dx \\
&= J(\omega)\left(1 - p\underline{s}^{-1} L_{G^{-1}}\left(\frac{(J(\omega)+1)^2}{2(N(\omega)+1)J(\omega)} + t_1 \frac{2(J(\omega)+1)^{\frac{3}{2}}}{3(N(\omega)+1)J(\omega)}\right)\right) \\
&\geq J(\omega)\left(1 - p\underline{s}^{-1} L_{G^{-1}}\left(\frac{J(\omega)+1}{N(\omega)+1} + t_1 \frac{4\sqrt{J(\omega)+1}}{3(N(\omega)+1)}\right)\right) \\
&\geq J(\omega)\left(1 - p\underline{s}^{-1} L_{G^{-1}}\left(\frac{J(\omega)+2}{N(\omega)+1} + t_1 \frac{4\sqrt{J(\omega)+2}}{3(N(\omega)+1)}\right)\right),
\end{aligned}$$

where the second last inequality holds since $2J(\omega) \geq J(\omega)+1$. Using (EC.7) and (EC.8), we obtain that

$$\begin{aligned}
n\rho_n^p \underline{s}^{-p} &\geq J(\omega)\left(1 - p\underline{s}^{-1} L_{G^{-1}}\left(\frac{2\rho_0^p \underline{s}^{-p}}{\alpha n^{bp}} + \frac{2}{\alpha n} + t_1 \frac{8\rho_0^{\frac{p}{2}} \underline{s}^{-\frac{p}{2}}}{3\alpha n^{\frac{bp+1}{2}}} + t_1 \frac{8}{3\alpha n}\right)\right) \\
&= J(\omega)\left(1 - \frac{\zeta_1}{n^{bp}} - \frac{\zeta_2}{n^{\frac{bp+1}{2}}} - \frac{\zeta_3}{n}\right).
\end{aligned}$$

Since $\frac{\zeta_1}{n^{bp}} + \frac{\zeta_2}{n^{\frac{bp+1}{2}}} + \frac{\zeta_3}{n} < \frac{1}{2}$ when $n > \max(n_6, n_7)$, this implies that

$$J(\omega) \leq \left(1 - \frac{\zeta_1}{n^{bp}} - \frac{\zeta_2}{n^{\frac{bp+1}{2}}} - \frac{\zeta_3}{n}\right)^{-1} n\rho_n^p \underline{s}^{-p}.$$

□

Proof of Theorem 5 (II). Recall the bound on $\mathcal{R}_{\mathbb{P}_n,p}(\rho_n)$ in (5) that

$$\frac{J(\omega)}{n} \leq \mathcal{R}_{\mathbb{P}_n,p}(\rho_n) < \frac{J(\omega)+1}{n}.$$

Applying Lemma EC.9 gives that

$$\begin{aligned}
\mathcal{R}_{\mathbb{P}_n,p}(\rho_n) \geq \frac{J(\omega)}{n} &> \left(1 + \frac{\eta_1}{n^{bp}} + \frac{\eta_2}{n^{\frac{bp+1}{2}}} + \frac{\eta_3}{n}\right)^{-1} \rho_n^p \underline{s}^{-p} - \frac{1}{n} \\
&\geq \left(1 - \frac{\eta_1}{n^{bp}} - \frac{\eta_2}{n^{\frac{bp+1}{2}}} - \frac{\eta_3}{n}\right) \rho_n^p \underline{s}^{-p} - \frac{1}{n} \\
&= \rho_n^p \underline{s}^{-p} - \left(\frac{\eta_1}{n^{2bp}} + \frac{\eta_2}{n^{\frac{3bp+1}{2}}} + \frac{\eta_3}{n^{1+bp}}\right) \rho_0^p \underline{s}^{-p} - \frac{1}{n},
\end{aligned}$$

where the last inequality comes from the inequality that $(1+x)^{-1} \geq 1-x$ for $x \geq 0$. Similarly,

$$\begin{aligned}\mathcal{R}_{\mathbb{P}_n,p}(\rho_n) < \frac{J(\omega)+1}{n} &\leq \left(1-\frac{\zeta_1}{n^{bp}}-\frac{\zeta_2}{n^{\frac{bp+1}{2}}}-\frac{\zeta_3}{n}\right)^{-1}\rho_n^p \underline{s}^{-p}+\frac{1}{n}\\ &\leq \left(1+4\frac{\zeta_1}{n^{bp}}+4\frac{\zeta_2}{n^{\frac{bp+1}{2}}}+4\frac{\zeta_3}{n}\right)\rho_n^p \underline{s}^{-p}+\frac{1}{n}\\ &= \rho_n^p \underline{s}^{-p}+4\left(\frac{\zeta_1}{n^{2bp}}+\frac{\zeta_2}{n^{\frac{3bp+1}{2}}}+\frac{\zeta_3}{n^{1+bp}}\right)\rho_0^p \underline{s}^{-p}+\frac{1}{n},\end{aligned}$$

where the last inequality is due to facts that $\frac{\zeta_1}{n^{bp}}+\frac{\zeta_2}{n^{\frac{bp+1}{2}}}+\frac{\zeta_3}{n} < \frac{1}{2}$ when $n > \max(n_6, n_7)$ and $(1-x)^{-1} \leq 1+4x$ for $0 \leq x \leq \frac{1}{2}$. Therefore, the result is proved by setting $n_0 = \lceil \max\{n_1, n_5, n_6, n_7\} \rceil$ and $C = \max\left((\eta_1+\eta_2+\eta_3)\rho_0^p \underline{s}^{-p}+1,\ 4(\zeta_1+\zeta_2+\zeta_3)\rho_0^p \underline{s}^{-p}+1\right)$. □

EC.4.1.3. Proof of Theorem 5(III)
Using (4), for $\omega \in \Omega_0$, it holds that

$$n\rho_n^p \geq \sum_{j=0}^{J(\omega)} S_j(\omega)^p \geq J(\omega)\underline{s}^p,$$

which implies that

$$J(\omega) \leq n\underline{s}^{-p}\rho_n^p.$$

Hence, for $\omega \in \Omega_0$,

$$\mathcal{R}_{\mathbb{P}_n,p}(\rho_n) < \frac{J(\omega)+1}{n} \leq \underline{s}^{-p}\rho_n^p + \frac{1}{n}.$$

To deive a lower bound for $J(\omega)$, by our assumption on G, there exists a constant $\Delta_G > 0$ such that $G(\underline{s}+x) \geq \frac{1}{2}c_{\underline{s}}x^\tau$ for $x \in (0, \Delta_G]$. Define

$$n_8 := \left(2\alpha^{-1}G(\underline{s}+\Delta_G)^{-1}\left(\rho_0^p \underline{s}^{-p}+t_1\sqrt{\rho_0^p \underline{s}^{-p}}\right)\right)^{\frac{1}{\min(1,bp)}},$$

When $n > \max(n_1, n_8)$, applying Lemma EC.6 with $\delta = \underline{s}+\Delta_G$ yields that for $1 \leq j \leq J(\omega)$,

$$U_j(\omega) < G(\underline{s}+\Delta_G).$$

As a result, $S_j(\omega)-\underline{s} \leq \Delta_G$, and thus

$$U_j(\omega) \geq \frac{1}{2}c_{\underline{s}}(S_j(\omega)-\underline{s})^\tau,$$

which implies that

$$S_j(\omega) \leq \left(\frac{2}{c_{\underline{s}}}U_j(\omega)\right)^{\frac{1}{\tau}}+\underline{s}.$$

Using (4) and Lemma EC.5, it holds that

$$\begin{aligned}n\rho_n^p < \sum_{j=0}^{J(\omega)+1}\left(\left(\frac{2}{c_{\underline{s}}}U_j(\omega)\right)^{\frac{1}{\tau}}+\underline{s}\right)^p &= \sum_{j=0}^{J(\omega)+1}\underline{s}^p\left(1+\underline{s}^{-1}\left(\frac{2}{c_{\underline{s}}}U_j(\omega)\right)^{\frac{1}{\tau}}\right)^p\\ &\leq \sum_{j=0}^{J(\omega)+1}\underline{s}^p\left(1+2^{\frac{1}{\tau}}c_{\underline{s}}^{-\frac{1}{\tau}}\underline{s}^{-1}\left(E[U_j]+t_1\sigma_j\right)^{\frac{1}{\tau}}\right)^p.\end{aligned}$$

Due to the elementary inequalities

$$\left(1+2^{\frac{1}{\tau}}\underline{s}^{-1}c_{\underline{s}}^{-\frac{1}{\tau}}x\right)^{p} \le 1+p2^{\frac{1}{\tau}}c_{\underline{s}}^{-\frac{1}{\tau}}\underline{s}^{-1}\left(1+2^{\frac{1}{\tau}}c_{\underline{s}}^{-\frac{1}{\tau}}\underline{s}^{-1}(1+t_1)^{\frac{1}{\tau}}\right)^{p-1}x, \quad \forall 1\le x\le (1+t_1)^{\frac{1}{\tau}},$$
$$(1+x)^{\frac{1}{\tau}} \le 1+\left(1+\tau^{-1}(1+t_1)^{\frac{1}{\tau}-1}\right)x, \qquad \forall 0\le x\le t_1,$$

it follows that

$$\begin{aligned}
&n\rho_n^p\\
&< (J(\omega)+1)\,\underline{s}^p + \sum_{j=1}^{J(\omega)+1} p2^{\frac{1}{\tau}}c_{\underline{s}}^{-\frac{1}{\tau}}\underline{s}^{p-1}\left(1+2^{\frac{1}{\tau}}c_{\underline{s}}^{-\frac{1}{\tau}}\underline{s}^{-1}(1+t_1)^{\frac{1}{\tau}}\right)^{p-1}\left(\tfrac{j}{N(\omega)+1}+t_1\tfrac{\sqrt{j}}{N(\omega)+1}\right)^{\frac{1}{\tau}}\\
&= (J(\omega)+1)\underline{s}^p + \sum_{j=1}^{J(\omega)+1} p2^{\frac{1}{\tau}}c_{\underline{s}}^{-\frac{1}{\tau}}\underline{s}^{p-1}\left(1+2^{\frac{1}{\tau}}c_{\underline{s}}^{-\frac{1}{\tau}}\underline{s}^{-1}(1+t_1)^{\frac{1}{\tau}}\right)^{p-1}\left(\tfrac{j}{N(\omega)+1}\right)^{\frac{1}{\tau}}\left(1+\tfrac{t_1}{\sqrt{j}}\right)^{\frac{1}{\tau}}\\
&\le (J(\omega)+1)\underline{s}^p+\\
&\quad \sum_{j=1}^{J(\omega)+1} p\left(\tfrac{2}{c_{\underline{s}}}\right)^{\frac{1}{\tau}}\underline{s}^{p-1}\left(1+2^{\frac{1}{\tau}}c_{\underline{s}}^{-\frac{1}{\tau}}\underline{s}^{-1}(1+t_1)^{\frac{1}{\tau}}\right)^{p-1}\left(\left(\tfrac{j}{N(\omega)+1}\right)^{\frac{1}{\tau}}+\left(1+\tau^{-1}(1+t_1)^{\frac{1}{\tau}-1}\right)\tfrac{t_1 j^{\frac{1}{\tau}-\frac{1}{2}}}{(N(\omega)+1)^{\frac{1}{\tau}}}\right)\\
&\le (J(\omega)+1)\underline{s}^p+\\
&\quad \int_{j=0}^{J(\omega)+2} p\left(\tfrac{2}{c_{\underline{s}}}\right)^{\frac{1}{\tau}}\underline{s}^{p-1}\left(1+\left(\tfrac{2}{c_{\underline{s}}}\right)^{\frac{1}{\tau}}\underline{s}^{-1}(1+t_1)^{\frac{1}{\tau}}\right)^{p-1}\left(\left(\tfrac{x}{N(\omega)+1}\right)^{\frac{1}{\tau}}+\left(\tfrac{\tau+(1+t_1)^{\frac{1}{\tau}-1}}{\tau}\right)\tfrac{t_1 x^{\frac{1}{\tau}-\frac{1}{2}}}{(N(\omega)+1)^{\frac{1}{\tau}}}\right)dx\\
&= (J(\omega)+1)\underline{s}^p\\
&\quad + p\left(\tfrac{2}{c_{\underline{s}}}\right)^{\frac{1}{\tau}}\underline{s}^{p-1}\left(1+\left(\tfrac{2}{c_{\underline{s}}}\right)^{\frac{1}{\tau}}\underline{s}^{-1}(1+t_1)^{\frac{1}{\tau}}\right)^{p-1}\left(\tfrac{\tau J(\omega+2)^{\frac{1}{\tau}+1}}{(\tau+1)(N(\omega)+1)^{\frac{1}{\tau}}}+2\left(\tfrac{\tau+(1+t_1)^{\frac{1}{\tau}-1}}{\tau+2}\right)\tfrac{t_1 J(\omega+2)^{\frac{\tau+2}{2\tau}}}{(N(\omega)+1)^{\frac{1}{\tau}}}\right)\\
&\le (J(\omega)+1)\underline{s}^p\left(1+\tfrac{p}{\underline{s}}\left(\tfrac{2}{c_{\underline{s}}}\right)^{\frac{1}{\tau}}\left(1+\left(\tfrac{2}{c_{\underline{s}}}\right)^{\frac{1}{\tau}}\tfrac{(1+t_1)^{\frac{1}{\tau}}}{\underline{s}}\right)^{p-1}\left(\tfrac{2J(\omega+2)^{\frac{1}{\tau}}}{(N(\omega)+1)^{\frac{1}{\tau}}}+4\tfrac{\tau+(1+t_1)^{\frac{1}{\tau}-1}}{\tau+2}\tfrac{t_1 J(\omega+2)^{\frac{2-\tau}{2\tau}}}{(N(\omega)+1)^{\frac{1}{\tau}}}\right)\right).
\end{aligned}$$

Notice that from (EC.7),

$$\left(\frac{J(\omega)+2}{N(\omega)+1}\right)^{\frac{1}{\tau}} \le \left(\frac{2\rho_0^p\underline{s}^{-p}}{\alpha n^{bp}}+\frac{4}{\alpha n}\right)^{\frac{1}{\tau}} \le \left(\frac{2\rho_0^p\underline{s}^{-p}+4}{\alpha}\right)^{\frac{1}{\tau}}\frac{1}{n^{\min\{\frac{bp}{\tau},\frac{1}{\tau}\}}},$$

and

$$\begin{aligned}
\frac{(J(\omega+2)^{\frac{2-\tau}{2\tau}}}{(N(\omega)+1)^{\frac{1}{\tau}}} \le \frac{(n\rho_n^p\underline{s}^{-p}+2)^{\frac{1}{\tau}-\frac{1}{2}}}{(\frac{1}{2}\alpha n)^{\frac{1}{\tau}}} &\le \frac{\left((\rho_0^p\underline{s}^{-p}+2)n^{\max(1-bp,0)}\right)^{\frac{1}{\tau}-\frac{1}{2}}}{(\frac{1}{2}\alpha n)^{\frac{1}{\tau}}}\\
&\le \left(\tfrac{2}{\alpha}\right)^{\frac{1}{\tau}}\left(\rho_0^p\underline{s}^{-p}+2\right)^{\frac{2-\tau}{2\tau}}\frac{1}{n^{\frac{1}{\tau}-\max(1-bp,0)\cdot(\frac{1}{\tau}-\frac{1}{2})}}\\
&\le \left(\tfrac{2}{\alpha}\right)^{\frac{1}{\tau}}\left(\rho_0^p\underline{s}^{-p}+2\right)^{\frac{2-\tau}{2\tau}}\frac{1}{n^{\min(\frac{1}{\tau},\frac{1}{2})}},
\end{aligned}$$

where the last inequality holds since $\frac{1}{\tau}-\max(1-bp,0)\cdot(\frac{1}{\tau}-\frac{1}{2})\geq\frac{1}{\tau}-(\frac{1}{\tau}-\frac{1}{2})\geq\frac{1}{2}$ when $0<\tau<2$ and $\frac{1}{\tau}-\max(1-bp,0)\cdot(\frac{1}{\tau}-\frac{1}{2})\geq\frac{1}{\tau}$ when $\tau\geq 2$. Set

$$\mu_1 := 2p2^{\frac{1}{\tau}}c_{\underline{s}}^{-\frac{1}{\tau}}\underline{s}^{-1}\left(1+2^{\frac{1}{\tau}}c_{\underline{s}}^{-\frac{1}{\tau}}\underline{s}^{-1}(1+t_1)^{\frac{1}{\tau}}\right)^{p-1}\left(\frac{2\rho_0^p\underline{s}^{-p}+4}{\alpha}\right)^{\frac{1}{\tau}},$$

$$\mu_2 := t_1p2^{\frac{1}{\tau}}c_{\underline{s}}^{-\frac{1}{\tau}}\underline{s}^{-1}\left(1+2^{\frac{1}{\tau}}c_{\underline{s}}^{-\frac{1}{\tau}}\underline{s}^{-1}(1+t_1)^{\frac{1}{\tau}}\right)^{p-1}4\left(\frac{\tau+(1+t_1)^{\frac{1}{\tau}-1}}{\tau+2}\right)\left(\frac{2}{\alpha}\right)^{\frac{1}{\tau}}\left(\rho_0^p\underline{s}^{-p}+2\right)^{\frac{2-\tau}{2\tau}}.$$

Then we have that

$$n\rho_n^p < (J(\omega)+1)\underline{s}^p\left(1+\frac{\mu_1}{n^{\min\{\frac{bp}{\tau},\frac{1}{\tau}\}}}+\frac{\mu_2}{n^{\min(\frac{1}{\tau},\frac{1}{2})}}\right),$$

which implies that

$$J(\omega) > n\rho_n^p\underline{s}^{-p}\left(1+\frac{\mu_1}{n^{\min\{\frac{bp}{\tau},\frac{1}{\tau}\}}}+\frac{\mu_2}{n^{\min(\frac{1}{\tau},\frac{1}{2})}}\right)^{-1}-1.$$

Therefore, it follows from (5) that

$$\begin{aligned}
\mathcal{R}_{\mathbb{P}_n,p}(\rho_n) &\geq \frac{J(\omega)}{n}\\
&> \left(1+\frac{\mu_1}{n^{\min\{\frac{bp}{\tau},\frac{1}{\tau}\}}}+\frac{\mu_2}{n^{\min(\frac{1}{\tau},\frac{1}{2})}}\right)^{-1}\rho_n^p\underline{s}^{-p}-\frac{1}{n}\\
&\geq \left(1-\frac{\mu_1}{n^{\min(\frac{bp}{\tau},\frac{1}{\tau})}}-\frac{\mu_2}{n^{\min(\frac{1}{\tau},\frac{1}{2})}}\right)\rho_n^p\underline{s}^{-p}-\frac{1}{n}\\
&= \rho_n^p\underline{s}^{-p}-\left(\frac{\mu_1}{n^{\min\{(1+\frac{1}{\tau})bp,\frac{1}{\tau}+bp\}}}+\frac{\mu_2}{n^{\min(\frac{1}{\tau}+bp,\frac{1}{2}+bp)}}\right)\rho_0^p\underline{s}^{-p}-\frac{1}{n},
\end{aligned}$$

where the last inequality comes from the inequality $(1+x)^{-1}\geq 1-x$ for $x\geq 0$. Thus, we obtain the desired result by setting $n_0=\lceil\max(n_1,n_8)\rceil$ and $C=(\mu_1+\mu_2)\rho_0^p\underline{s}^{-p}+1$. □

EC.4.1.4. Proofs for Preparatory Results

Proof of Lemma EC.5. Applying Hoeffding's inequality on N gives that

$$\mathbb{P}\left(|N(\omega)-n\alpha|>t_2\sqrt{n}\right)\leq 2\exp(-2t_2^2).$$

By Bernstein's inequality for Beta distribution (e.g., [43]), it holds that

$$\mathbb{P}\left(|U_j-\mathbb{E}[U_j]|>t_1\sigma_j\mid N\right)\leq\exp\left(-\frac{t_1^2\sigma_j^2}{2\sigma_j^2+2c_jt_1\sigma_j}\right)=\exp\left(-\frac{t_1^2}{2+2t_1c_j/\sigma_j}\right),\quad\forall 1\leq j\leq N,$$

where $c_j:=\max\left(\sqrt{\frac{j(N-j+1)}{(N+3)(N+1)^2}},\frac{|N-2j+1|}{(N+3)(N+1)}\right)$. Next we derive an upper bound for c_j/σ_j. Recalling $\sigma_j=\frac{\sqrt{j(N+1-j)}}{(N+1)\sqrt{N+2}}$, $j=1,\ldots,N$, we compute that

$$\sqrt{\frac{j(N-j+1)}{(N+3)(N+1)^2}}\Big/\sigma_j=\sqrt{\frac{N+2}{N+3}}\leq 1,$$

and

$$\frac{|N-2j+1|}{(N+3)(N+1)}\Big/\sigma_j=\frac{|N-2j+1|\sqrt{N+2}}{\sqrt{j(N-j+1)}(N+3)}\leq\frac{N+1}{\sqrt{j(N-j+1)}\sqrt{N+2}}\leq\frac{N+1}{\sqrt{N}\sqrt{N+2}}\leq 2.$$

Thus, $\frac{c_j}{\sigma_j} \le 2$. It implies that

$$\mathbb{P}\Big(|U_j(\omega) - \mathbb{E}[U_j]| > t_1\sigma_j \mid N\Big) \le \exp\left(-\tfrac{t_1^2}{2+4t_1}\right), \quad \forall 1 \le j \le N.$$

To facilitate the notation, let us define $U_j = 0$ for $j = N+1, \dots, n$. Then it follows that

$$\mathbb{P}\Big(|U_j(\omega) - \mathbb{E}[U_j]| > t_1\sigma_j\Big) \le \exp\left(-\frac{t_1^2}{2+4t_1}\right), \quad \forall 1 \le j \le n.$$

Then the result follows from the union bound

$$\begin{aligned}
&\mathbb{P}\Big(|N - n\alpha| > t_2\sqrt{n},\ \exists 1 \le j \le N,\ s.t.\ |U_j(\omega) - \mathbb{E}[U_j]| > t_1\sigma_j\Big)\\
&\le \mathbb{P}\Big(|N - n\alpha| > t_2\sqrt{n}\Big) + \mathbb{P}\Big(\exists 1 \le j \le N,\ s.t.\ |U_j(\omega) - \mathbb{E}[U_j]| > t_1\sigma_j\Big)\\
&\le \mathbb{P}\Big(|N - n\alpha| > t_2\sqrt{n}\Big) + \mathbb{P}\Big(\exists 1 \le j \le n,\ s.t.\ |U_j(\omega) - \mathbb{E}[U_j]| > t_1\sigma_j\Big)\\
&\le 2\exp(-2t_2^2) + n\exp\left(-\tfrac{t_1^2}{2+4t_1}\right).
\end{aligned}$$

□

Proof of Lemma EC.6. Let $\omega \in \Omega_0$. By the definition of $\underline{s}$, it follows from (4) that

$$J(\omega)\underline{s}^p \le \sum_{j=0}^{J(\omega)} S_j(\omega)^p \le n\rho_n^p,$$

which implies that

$$J(\omega) \le n\rho_n^p \underline{s}^{-p}.$$

Using Lemma EC.5, the fact that $|N(\omega) - n\alpha| \le t_2\sqrt{n}$ when $n > n_1$, and $n > n_2$, we have

$$U_{J(\omega)}(\omega) \le \mathbb{E}[U_J] + t_1\sigma_{J(\omega)} \le \tfrac{J(\omega)}{N(\omega)+1} + \tfrac{t_1\sqrt{J(\omega)}}{N(\omega)+1} \le \tfrac{n\rho_n^p\underline{s}^{-p} + t_1\sqrt{n\rho_n^p\underline{s}^{-p}}}{\frac{1}{2}n\alpha} = \tfrac{2\rho_0^p\underline{s}^{-p}}{\alpha n^{bp}} + \tfrac{2t_1\sqrt{\rho_0^p\underline{s}^{-p}}}{\alpha n^{\frac{bp+1}{2}}} < G(\delta).$$

□

Proof of Lemma EC.7. Notice that when $g(\underline{s}) > 0$, under Assumption 1, there exist constants $\delta_1 \in (\underline{s}, \delta)$ and $C_1, C_2 \in (0, \infty)$ such that

$$0 < C_1 \le g(s) \le C_2 < \infty, \quad \forall x \in [0, \delta_1].$$

Hence, G is strictly monotone and continuous on $(\underline{s}, \delta_1]$. Applying the inverse function theorem gives that $G^{-1}(x)$ is continuous on $(0, G(\delta_1)]$ and $\frac{1}{C_2} \le (G^{-1})'(s) \le \frac{1}{C_1}$ for all $x \in (0, G(\delta_1)]$. By the mean value theorem, for all $x_1, x_2 \in [0, G(\delta_1)]$ with $x_1 < x_2$, there exists $x_0 \in (x_1, x_2)$ such that

$$|G^{-1}(x_1) - G^{-1}(x_2)| = (G^{-1})'(x_0)|x_1 - x_2|,$$

which implies the result with $M = \frac{1}{C_2}$ and $L_{G^{-1}} = \frac{1}{C_1}$. □

Proof of Lemma EC.8. Let $\omega \in \Omega_0$. Since $n > \max(n_1, n_2)$, by Lemma EC.6, for all $\omega \in \Omega_0$,

$$G(\underline{s}) \le U_j(\omega) < G(\delta), \quad \forall 1 \le j \le J(\omega),$$

Taking G^{-1} on each term yields

$$\underline{s} \le S_j(\omega) \le \delta, \quad \forall 1 \le j \le J(\omega).$$

From (4), it holds that

$$J\underline{s}^p \le \sum_{j=0}^{J} S_j(\omega)^p \le n\rho_n^p < \sum_{j=0}^{J+1} S_j(\omega)^p \le (J+1)\delta^p,$$

which is equivalent to

$$n\rho_n^p\delta^{-p} - 1 < J \le n\rho_n^p\underline{s}^{-p},$$

which completes the proof. □

EC.4.2. Proof of Theorem 6

We first state several useful preparatory results, whose proofs are postponed to Section EC.4.2.3.

Recall in Lemma EC.7 that when $G(\delta) > 0$ and $g(0) > 0$, there exists a $0 < \delta_1 < \delta$ such that $g(s) > 0$ on $[0, \delta_1]$ and $G^{-1}(x)$ exists on $[0, G(\delta_1)]$. Define $\delta_2 := \inf\{s > 0 : G(s) = G(\delta_1)/2\}$, and

$$n_9 := \left(t_1\sqrt{\tfrac{1}{4} + \tfrac{t_1^2}{8}} + \tfrac{\sqrt{2}t_1^2}{4}\right)^{\frac{2}{1-2bp}} \left(\frac{\delta_2}{\rho_0}\right)^{\frac{2p}{1-2bp}} \alpha^{\frac{1}{1-2bp}},$$

$$n_{10} := \left(\tfrac{2}{\sqrt{\alpha}G(\delta_1)}\right)^{(\min\{bp,\frac{1}{2}\})^{-1}} \left(\left(t_1\sqrt{1+\tfrac{t_1^2}{2}} + \tfrac{\sqrt{2}t_1^2}{2} + \tfrac{4\rho_0^p}{\delta_2^p\sqrt{\alpha}}\right) + \left(1 + t_1\sqrt{2+t_1^2} + t_1^2 + \tfrac{4\rho_0^p}{\delta_2^p\alpha}\right)^{\frac{1}{2}}\right)^{(\min\{bp,\frac{1}{2}\})^{-1}}.$$

Note that n_9 is only used for the case of $bp < \frac{1}{2}$.

Recall Ω_0 defined in Lemma EC.5: for all $\omega \in \Omega_0$, $\max\left(\underline{s}, \mathbb{E}[U_j] - t_1\sigma_j\right) \le U_j(\omega) \le \mathbb{E}[U_j] + t_1\sigma_j$ for $1 \le j \le N(\omega)$ and $|N(\omega) - n\alpha| \le t_2\sqrt{n}$.

Lemma EC.10. *Suppose $G(\delta) > 0$ and $g(0) > 0$. Let $n > \lceil \max(n_1, n_9, n_{10})\rceil$. Then for all $\omega \in \Omega_0$, it holds that*

$$|N(\omega) - n\alpha| \le t_2\sqrt{n}, \quad G(\underline{s}) \le U_j(\omega) \le \mathbb{E}[U_j] + t_1\sigma_j < G(\delta_1), \quad \forall 1 \le j \le N(\omega).$$

Define

$$n_{11} := \left(\frac{2}{\alpha\rho_0}\right)^{\frac{p}{p+1-bp}} (p+1)^{-\frac{1}{p+1-bp}} L_{G^{-1}}^{\frac{p}{p+1-bp}} (1+t_1)^{\frac{p}{p+1-bp}} \left(4(p+1)^2\frac{t_1^2 L_{G^{-1}}^2}{M^2} + 4\right)^{\frac{p+1}{p+1-bp}}.$$

Lemma EC.11. *Assume Assumption 1 holds with $g(0) > 0$ and $n > \lceil \max(n_1, n_9, n_{10}, n_{11})\rceil$. Let $s > 0$ and $t > 0$. Then for any $\omega \in \Omega_0$, it holds that*

$$2 \le L_{G^{-1}}^{-\frac{p}{p+1}} (1+s)^{-\frac{p}{p+1}} c_n^{\frac{1}{p+1}} - 2 < J(\omega) \le 2^{\frac{1}{p+1}} M^{-\frac{p}{p+1}} c_n^{\frac{1}{p+1}} + 1.$$

where $c_n = n\rho_n^p (N(\omega)+1)^p (1+p)$.

Lemma EC.12. *Assume that Assumption 1 holds with $g(0) > 0$. Let $h = \frac{L_g}{C_1^3}$. Then for any $x, y \in [0, G(\delta_1)]$,*

$$|(G^{-1})'(x) - (G^{-1})'(y)| \le h|x - y|.$$

EC.4.2.1. Proof of Theorem 6 (I)

In light of Lemma EC.11, now we derive a finer bound on $J(\omega)$. Define

$$n_{12} := \left(\frac{2p\beta_1 + 2\sqrt{2}p\beta_2}{\gamma}\right)^{\max\left(\frac{2(p+1)}{p-bp+1}, \frac{p+1}{bp}\right)},$$

$$n_{13} := \left(\frac{\beta_1 + \beta_2}{p+1}\right)^{\max\left(\frac{2(p+1)}{p-bp+1}, \frac{p+1}{bp}\right)},$$

where

$$\gamma = p\left(1+\frac{h}{(G^{-1})'(0)}+t_1\frac{L_{G^{-1}}}{(G^{-1})'(0)}\right)^{p-1},$$
$$\beta_1 = \frac{h\gamma}{(G^{-1})'(0)}\left(\frac{12\rho_0^{\frac{p}{p+1}}(1+p)^{\frac{1}{p+1}}}{M^{\frac{p}{p+1}}\alpha}\right),$$
$$\beta_2 = \frac{(p+1)\gamma t_1 L_{G^{-1}}}{(p+\frac{1}{2})(G^{-1})'(0)}L_{G^{-1}}^{\frac{p}{2(p+1)}}(1+t_1)^{\frac{p}{2(p+1)}}\left(\frac{1}{2}\alpha\rho_0\right)^{-\frac{p}{2(p+1)}}(1+p)^{-\frac{1}{2(p+1)}}.$$

Lemma EC.13. *Assume Assumption 1 holds with $g(0)>0$ and $n>\lceil\max(n_1,n_9,n_{10},n_{11},n_{12})\rceil$. Let $t_1,t>0$. Then for any $\omega\in\Omega_0$, it holds that*

$$\left(1+\frac{\beta_1}{n^{\frac{p}{2(p+1)}}}+\frac{\beta_2}{n^{\frac{p+2}{4(p+1)}}}\right)^{-\frac{1}{p+1}}\left(\frac{c_n^{\frac{1}{p+1}}}{(G^{-1})'(0)^p}\right)^{\frac{1}{p+1}}-2\le J(\omega)\le\left(1-\frac{p\beta_1}{\gamma n^{\frac{p}{2(p+1)}}}-\frac{\sqrt{2}p\beta_2}{\gamma n^{\frac{p+2}{4(p+1)}}}\right)^{-\frac{1}{p+1}}\left(\frac{c_n^{\frac{1}{p+1}}}{(G^{-1})'(0)^p}\right)^{\frac{1}{p+1}}+1.$$

Proof of Lemma EC.13. Let $\omega\in\Omega_0$. Due to Lemma EC.11, below we consider $J(\omega)\ge J_1(\omega)\ge 2$.

We first derive the lower bound. Recall from the Lemma EC.10 that when $n>\max(n_1,n_9,n_{10})$, $U_j(\omega)\le G(\delta_1)$ for $1\le j\le J(\omega)$. It follows from Lemma EC.7 that

$$G^{-1}(\mathbb{E}[U_j])-L_{G^{-1}}t_1\sigma_j\le G^{-1}(U_j(\omega))\le G^{-1}(\mathbb{E}[U_j])+L_{G^{-1}}t_1\sigma_j. \tag{EC.9}$$

Moreover, thanks to Lemma EC.10, $U_j(\omega)<G(\delta_1)$, $\forall 1\le j\le N(\omega)$. Using Lemma EC.12, we have

$$G^{-1}(\mathbb{E}[U_j])\le(G^{-1})'(0)\mathbb{E}[U_j]+h(\mathbb{E}[U_j])^2, \tag{EC.10}$$

and

$$G^{-1}(\mathbb{E}[U_j])\ge(G^{-1})'(0)\mathbb{E}[U_j]-h(\mathbb{E}[U_j])^2. \tag{EC.11}$$

It follows from (4) and (EC.10) that

$$\begin{aligned}
n\rho_n^p &< \sum_{j=0}^{J(\omega)+1}G^{-1}\left(U_j(\omega)\right)^p\\
&\le \sum_{j=1}^{J(\omega)+1}\left(G^{-1}(\mathbb{E}[U_j])+L_{G^{-1}}t_1\sigma_j\right)^p\\
&\le \sum_{j=1}^{J(\omega)+1}\left((G^{-1})'(0)\mathbb{E}[U_j]+h(\mathbb{E}[U_j])^2+L_{G^{-1}}t_1\sigma_j\right)^p\\
&= \sum_{j=1}^{J(\omega)+1}\left((G^{-1})'(0)\frac{j}{N(\omega)+1}+h\frac{j^2}{(N(\omega)+1)^2}+t_1L_{G^{-1}}\frac{\sqrt{j}}{N(\omega)+1}\right)^p\\
&= \frac{(G^{-1})'(0)^p}{(N(\omega)+1)^p}\sum_{j=1}^{J(\omega)+1}j^p\left(1+h\frac{j}{(G^{-1})'(0)(N(\omega)+1)}+t_1\frac{L_{G^{-1}}}{(G^{-1})'(0)}\frac{1}{\sqrt{j}}\right)^p.
\end{aligned} \tag{EC.12}$$

Notice that

$$0\le h\frac{j}{(G^{-1})'(0)(N(\omega)+1)}+t_1\frac{L_{G^{-1}}}{(G^{-1})'(0)}\frac{1}{\sqrt{j}}\le\frac{h}{(G^{-1})'(0)}+t_1\frac{L_{G^{-1}}}{(G^{-1})'(0)}.$$

Using the inequality that $(1+x)^p \le 1+p\left(1+\frac{h}{(G^{-1})'(0)}+t_1\frac{L_{G^{-1}}}{(G^{-1})'(0)}\right)^{p-1}x$ for $0\le x\le \frac{h}{(G^{-1})'(0)}+t_1\frac{L_{G^{-1}}}{(G^{-1})'(0)}$, it follows from (EC.12) and the definition of γ that

$$\begin{aligned}
n\rho_n^p &< \frac{(G^{-1})'(0)^p}{(N(\omega)+1)^p}\sum_{j=1}^{J(\omega)+1} j^p\left(1+h\gamma\frac{j}{(G^{-1})'(0)(N(\omega)+1)}+\gamma\frac{L_{G^{-1}}t_1}{(G^{-1})'(0)}\frac{1}{\sqrt{j}}\right)\\
&\le \frac{(G^{-1})'(0)^p}{(N(\omega)+1)^p}\int_{j=0}^{J(\omega)+2} x^p\left(1+h\gamma\frac{x}{(G^{-1})'(0)(N(\omega)+1)}+\gamma\frac{L_{G^{-1}}t_1}{(G^{-1})'(0)}\frac{1}{\sqrt{x}}\right)dx\\
&= \frac{(G^{-1})'(0)^p}{(N(\omega)+1)^p}\left(\frac{(J(\omega)+2)^{p+1}}{p+1}+\frac{h\gamma(J(\omega)+2)^{p+2}}{(p+2)(G^{-1})'(0)(N(\omega)+1)}+\gamma\frac{L_{G^{-1}}t_1(J(\omega)+2)^{p+\frac{1}{2}}}{(p+\frac{1}{2})(G^{-1})'(0)}\right),
\end{aligned}$$

which implies that

$$\begin{aligned}
\frac{c_n}{(G^{-1})'(0)^p} &\le (J(\omega)+2)^{p+1}+\frac{(p+1)h\gamma(J(\omega)+2)^{p+2}}{(p+2)(G^{-1})'(0)(N+1)}+\frac{(p+1)\gamma t_1 L_{G^{-1}}(J(\omega)+2)^{p+\frac{1}{2}}}{(p+\frac{1}{2})(G^{-1})'(0)}\\
&= (J(\omega)+2)^{p+1}\left(1+\frac{(p+1)h\gamma}{(p+2)(G^{-1})'(0)}\left(\frac{J(\omega)+2}{N(\omega)+1}\right)+\frac{(p+1)\gamma t_1 L_{G^{-1}}}{(p+\frac{1}{2})(G^{-1})'(0)}\frac{1}{\sqrt{J(\omega)+2}}\right)\\
&\le (J(\omega)+2)^{p+1}\left(1+\frac{h\gamma}{(G^{-1})'(0)}\left(\frac{J_2(\omega)+2}{N(\omega)+1}\right)+2\frac{\gamma t_1 L_{G^{-1}}}{(G^{-1})'(0)}\frac{1}{\sqrt{J_1(\omega)+2}}\right),
\end{aligned}\tag{EC.13}$$

where c_n is defined in Lemma EC.11, and

$$\begin{aligned}
J_1(\omega) &:= L_{G^{-1}}^{-\frac{p}{p+1}}(1+s)^{-\frac{p}{p+1}}c_n^{\frac{1}{p+1}}-2,\\
J_2(\omega) &:= 2^{\frac{1}{p+1}}M^{-\frac{p}{p+1}}c_n^{\frac{1}{p+1}}+1.
\end{aligned}$$

To bound the right-hand side, using Lemma EC.11 we have that

$$\begin{aligned}
\frac{J_2(\omega)+2}{N(\omega)+1}\le\frac{3J_2(\omega)}{N(\omega)+1}\le\frac{6J_2(\omega)}{n\alpha}\le 6\frac{2^{\frac{1}{p+1}}c_n^{\frac{1}{p+1}}}{M^{\frac{p}{p+1}}n\alpha} &= 6\frac{2^{\frac{1}{p+1}}\left(n\rho_n^p(N(\omega)+1)^p(1+p)\right)^{\frac{1}{p+1}}}{M^{\frac{p}{p+1}}n\alpha}\\
&\le 6\frac{2^{\frac{1}{p+1}}\left(n\rho_n^p(2n)^p(1+p)\right)^{\frac{1}{p+1}}}{M^{\frac{p}{p+1}}n\alpha}\\
&= \frac{12\rho_0^{\frac{p}{p+1}}(1+p)^{\frac{1}{p+1}}}{M^{\frac{p}{p+1}}\alpha n^{\frac{bp}{p+1}}}.
\end{aligned}$$

As a result,

$$\frac{h\gamma}{(G^{-1})'(0)}\left(\frac{J_2(\omega)+2}{N(\omega)+1}\right)\le\frac{h\gamma}{(G^{-1})'(0)}\left(\frac{12\rho_0^{\frac{p}{p+1}}(1+p)^{\frac{1}{p+1}}}{M^{\frac{p}{p+1}}\alpha n^{\frac{p}{2(p+1)}}}\right)=\frac{\beta_1}{n^{\frac{bp}{p+1}}}.$$

Similar to the reasoning above, applying Lemma EC.11 gives that

$$\begin{aligned}\frac{1}{\sqrt{J_1(\omega)+2}} &= L_{G^{-1}}^{\frac{p}{2(p+1)}}(1+t_1)^{\frac{p}{2(p+1)}}c_n^{-\frac{1}{2(p+1)}}\\ &\le L_{G^{-1}}^{\frac{p}{2(p+1)}}(1+t_1)^{\frac{p}{2(p+1)}}\left(n\rho_n^p(N(\omega))^p(1+p)\right)^{-\frac{1}{2(p+1)}}\\ &\le L_{G^{-1}}^{\frac{p}{2(p+1)}}(1+t_1)^{\frac{p}{2(p+1)}}\left(nn^{-bp}\rho_0^p\left(\frac{1}{2}\alpha n\right)^p(1+p)\right)^{-\frac{1}{2(p+1)}}\\ &= L_{G^{-1}}^{\frac{p}{2(p+1)}}(1+t_1)^{\frac{p}{2(p+1)}}\left(\frac{1}{2}\alpha\rho_0\right)^{-\frac{p}{2(p+1)}}(1+p)^{-\frac{1}{2(p+1)}}\frac{1}{n^{\frac{p+1-bp}{2(p+1)}}},\end{aligned}$$

which implies that

$$\begin{aligned}2\frac{\gamma t_1 L_{G^{-1}}}{(G^{-1})'(0)}\frac{1}{\sqrt{J_1(\omega)+1}} &\le 2\frac{\gamma t_1 L_{G^{-1}}}{(G^{-1})'(0)}L_{G^{-1}}^{\frac{p}{2(p+1)}}(1+t_1)^{\frac{p}{2(p+1)}}\left(\frac{1}{2}\alpha\rho_0\right)^{-\frac{p}{2(p+1)}}(1+p)^{-\frac{1}{2(p+1)}}\frac{1}{n^{\frac{p+1-bp}{2(p+1)}}}\\ &= \frac{\beta_2}{n^{\frac{p+1-bp}{2(p+1)}}}.\end{aligned} \tag{EC.14}$$

Hence, from (EC.13), it holds that

$$J(\omega)\ge\left(1+\frac{\beta_1}{n^{\frac{bp}{p+1}}}+\frac{\beta_2}{n^{\frac{p+1-bp}{2(p+1)}}}\right)^{-\frac{1}{p+1}}(G^{-1})'(0)^{-\frac{p}{p+1}}c_n^{\frac{1}{p+1}}-2.$$

Next, we drive the upper bound. By (EC.18) and (EC.11),

$$\begin{aligned}n\rho_n^p &\ge \sum_{j=1}^{J(\omega)}\left(G^{-1}(\mathbb{E}[U_j])-L_{G^{-1}}s\sigma_j\right)_+^p\\ &\ge \sum_{j=1}^{J(\omega)}\left((G^{-1})'(0)\mathbb{E}[U_j)]-h(\mathbb{E}[U_j])^2-t_1L_{G^{-1}}\sigma_j\right)_+^p\\ &\ge \sum_{j=1}^{J(\omega)}\left((G^{-1})'(0)\frac{j}{N(\omega)+1}-h\frac{j^2}{(N(\omega)+1)^2}-t_1L_{G^{-1}}\frac{\sqrt{j}}{N(\omega)+1}\right)_+^p\\ &= \frac{(G^{-1})'(0)^p}{(N(\omega)+1)^p}\sum_{j=1}^{J(\omega)}j^p\left(1-h\frac{j}{(G^{-1})'(0)(N(\omega)+1)}-\frac{t_1L_{G^{-1}}}{(G^{-1})'(0)}\frac{1}{\sqrt{j}}\right)_+^p\\ &\ge \frac{(G^{-1})'(0)^p}{(N(\omega)+1)^p}\int_0^{J(\omega)-1}x^p\left(1-h\frac{x}{(G^{-1})'(0)(N(\omega)+1)}-\frac{t_1L_{G^{-1}}}{(G^{-1})'(0)}\frac{1}{\sqrt{x}}\right)_+^p dx.\end{aligned}$$

Using the inequality $(1-x)_+^p\ge 1-px$ for $p\ge 1$ and $x\ge 0$, we obtain that

$$\begin{aligned}n\rho_n^p &\ge \frac{(G^{-1})'(0)^p}{(N(\omega)+1)^p}\int_0^{J(\omega)-1}x^p\left(1-ph\frac{x}{(G^{-1})'(0)(N(\omega)+1)}-p\frac{t_1L_{G^{-1}}}{(G^{-1})'(0)}\frac{1}{\sqrt{x}}\right)dx\\ &= \frac{(G^{-1})'(0)^p(J(\omega)-1)^{p+1}}{(N(\omega)+1)^p(p+1)}\left(1-\frac{(p+1)ph(J(\omega)-1)}{(p+2)(G^{-1})'(0)(N(\omega)+1)}-\frac{(p+1)pt_1L_{G^{-1}}}{\left(p+\frac{1}{2}\right)(G^{-1})'(0)}\frac{1}{\sqrt{J(\omega)-1}}\right)\\ &\ge \frac{(G^{-1})'(0)^p(J(\omega)-1)^{p+1}}{(N(\omega)+1)^p(p+1)}\left(1-\frac{ph(J(\omega)-1)}{(G^{-1})'(0)(N(\omega)+1)}-\frac{2pt_1L_{G^{-1}}}{(G^{-1})'(0)}\frac{1}{\sqrt{J_1(\omega)-1}}\right),\end{aligned}$$

which implies that

$$\frac{c_n}{(G^{-1})'(0)^p} \geq (J(\omega)-1)^{p+1}\left(1-\frac{ph(J(\omega)-1)}{(G^{-1})'(0)(N(\omega)+1)}-\frac{2pt_1L_{G^{-1}}}{(G^{-1})'(0)}\frac{1}{\sqrt{J_1(\omega)-1}}\right). \tag{EC.15}$$

Accordingly, by (EC.14), we have

$$\frac{ph}{(G^{-1})'(0)}\left(\frac{J_2(\omega)-1}{N+1}\right) \leq \frac{ph}{(G^{-1})'(0)}\left(\frac{J_2(\omega)+2}{N(\omega)+1}\right) \leq \frac{p\beta_1}{\gamma n^{\frac{bp}{p+1}}},$$

and due to $|N(\omega)-n\alpha| \leq t_2\sqrt{n}$ with $n > n_1$, we have

$$\begin{aligned}
\frac{1}{\sqrt{J_1(\omega)-1}} &= \left(L_{G^{-1}}^{-\frac{p}{p+1}}(1+t_1)^{-\frac{p}{p+1}}c_n^{\frac{1}{p+1}}-3\right)^{-\frac{1}{2}} \\
&\leq \left(L_{G^{-1}}^{-\frac{p}{p+1}}(1+t_1)^{-\frac{p}{p+1}}\left[nn^{-bp}\rho_0^p\left(\frac{1}{2}\alpha n\right)^p(1+p)\right]^{\frac{1}{p+1}}-3\right)^{-\frac{1}{2}} \\
&= \left(L_{G^{-1}}^{-\frac{p}{p+1}}(1+t_1)^{-\frac{p}{p+1}}\left(\frac{1}{2}\alpha\rho_0\right)^{\frac{p}{p+1}}(1+p)^{\frac{1}{p+1}}n^{\frac{p-bp+1}{p+1}}-3\right)^{-\frac{1}{2}} \\
&\leq \left(\frac{1}{2}L_{G^{-1}}^{-\frac{p}{p+1}}(1+t_1)^{-\frac{p}{p+1}}\left(\frac{1}{2}\alpha\rho_0\right)^{\frac{p}{p+1}}(1+p)^{\frac{1}{p+1}}n^{\frac{p-bp+1}{p+1}}\right)^{-\frac{1}{2}} \\
&= \sqrt{2}L_{G^{-1}}^{\frac{p}{2(p+1)}}(1+t_1)^{\frac{p}{2(p+1)}}\left(\frac{1}{2}\alpha\rho_0\right)^{-\frac{p}{2(p+1)}}(1+p)^{-\frac{1}{2(p+1)}}\frac{1}{n^{\frac{p+1-bp}{2(p+1)}}},
\end{aligned}$$

where the last inequality is due to that $n \geq n_{11} \geq 6^{\frac{p+1}{p-bp+1}}L_{G^{-1}}^{\frac{p}{p-bp+1}}(1+t_1)^{\frac{p}{p-bp+1}}\left(\frac{1}{2}\alpha\rho_0\right)^{-\frac{p}{p-bp+1}}(1+p)^{-\frac{1}{p-bp+1}}$. It follows from (EC.14) that

$$\begin{aligned}
\frac{2pt_1L_{G^{-1}}}{(p+2)(G^{-1})'(0)}\frac{1}{\sqrt{J_1(\omega)-1}} &\leq \frac{2pt_1L_{G^{-1}}}{(p+2)(G^{-1})'(0)}\sqrt{2}L_{G^{-1}}^{\frac{p}{2(p+1)}}(1+t_1)^{\frac{p}{2(p+1)}}\left(\frac{1}{2}\alpha\rho_0\right)^{-\frac{p}{2(p+1)}}(1+p)^{-\frac{1}{2(p+1)}}\frac{1}{n^{\frac{p+1-bp}{2(p+1)}}} \\
&= \frac{\sqrt{2}p}{\gamma}\frac{\beta_2}{n^{\frac{p+1-bp}{2(p+1)}}}.
\end{aligned}$$

Thus, using (EC.15), we obtain that

$$\begin{aligned}
\frac{c_n}{(G^{-1})'(0)^p} &\geq (J(\omega)-1)^{p+1}\left(1-\frac{ph(J(\omega)-1)}{(G^{-1})'(0)(N(\omega)+1)}-\frac{2pt_1L_{G^{-1}}}{(G^{-1})'(0)}\frac{1}{\sqrt{J_1(\omega)-1}}\right) \\
&\geq (J(\omega)-1)^{p+1}\left(1-\frac{p\beta_1}{\gamma n^{\frac{bp}{p+1}}}-\frac{\sqrt{2}p}{\gamma}\frac{\beta_2}{n^{\frac{p+1-bp}{2(p+1)}}}\right).
\end{aligned}$$

Since $\frac{p\beta_1}{\gamma n^{\frac{bp}{p+1}}}+\frac{\sqrt{2}p}{\gamma}\frac{\beta_2}{n^{\frac{p+1-bp}{2(p+1)}}} < \frac{1}{2}$ when $n > n_{12}$, the above implies that

$$J(\omega) \leq \left(1-\frac{p\beta_1}{\gamma n^{\frac{bp}{p+1}}}-\frac{\sqrt{2}p}{\gamma}\frac{\beta_2}{n^{\frac{p+1-bp}{2(p+1)}}}\right)^{-\frac{1}{p+1}}(G^{-1})'(0)^{-\frac{p}{p+1}}c_n^{\frac{1}{p+1}}+1.$$

Therefore the proof is completed. □

Proof of Theorem 6(I). Using (5), Lemma EC.13 and the inequality that $(1+x)^{-\frac{1}{p+1}} \geq 1-\frac{1}{p+1}x$ for $p \geq 1$ and $x \geq 0$, it follows that

$$\begin{aligned}
\mathcal{R}_{\mathbb{P}_n,p}(\rho_n) &\geq \frac{J(\omega)-1}{n} \\
&\geq \left(1+\frac{\beta_1}{n^{\frac{bp}{p+1}}}+\frac{\beta_2}{n^{\frac{p+1-bp}{2(p+1)}}}\right)^{-\frac{1}{p+1}} (G^{-1})'(0)^{-\frac{p}{p+1}} c_n^{\frac{1}{p+1}} - \frac{3}{n} \\
&\geq \left(1-\frac{1}{p+1}\frac{\beta_1}{n^{\frac{bp}{p+1}}}-\frac{1}{p+1}\frac{\beta_2}{n^{\frac{p+1-bp}{2(p+1)}}}\right)(G^{-1})'(0)^{-\frac{p}{p+1}} c_n^{\frac{1}{p+1}} - \frac{3}{n} \\
&= \left(1-\frac{1}{p+1}\left(\frac{\beta_1}{n^{\frac{bp}{p+1}}}+\frac{\beta_2}{n^{\frac{p+1-bp}{2(p+1)}}}\right)\right)(G^{-1})'(0)^{-\frac{p}{p+1}} \rho_n^{\frac{p}{p+1}} (1+p)^{\frac{1}{p+1}} \left(\frac{N(\omega)+1}{n}\right)^{\frac{p}{p+1}} - \frac{3}{n}.
\end{aligned}$$

Then, since $\frac{1}{p+1}\left(\frac{\beta_1}{n^{\frac{bp}{p+1}}}+\frac{\beta_2}{n^{\frac{p+1-bp}{2(p+1)}}}\right) < 1$ when $n > n_{13}$, using $|N(\omega)-n\alpha| \leq t_2\sqrt{n}$ with $n > n_1$ and $(1-x)^{\frac{p}{p+1}} \geq 1-2^{\frac{1}{p+1}}\frac{p}{p+1}x$ for $p \geq 1$ and $0 \leq x \leq \frac{1}{2}$, it holds that

$$\begin{aligned}
&\mathcal{R}_{\mathbb{P}_n,p}(\rho_n) \\
&\geq \left(1-\frac{1}{p+1}\left(\frac{\beta_1}{n^{\frac{bp}{p+1}}}+\frac{\beta_2}{n^{\frac{p+1-bp}{2(p+1)}}}\right)\right)(G^{-1})'(0)^{-\frac{p}{p+1}} \rho_n^{\frac{p}{p+1}} (1+p)^{\frac{1}{p+1}} \left(\frac{n\alpha-t_2\sqrt{n}}{n}\right)^{\frac{p}{p+1}} - \frac{3}{n} \\
&= \left(1-\frac{1}{p+1}\left(\frac{\beta_1}{n^{\frac{bp}{p+1}}}+\frac{\beta_2}{n^{\frac{p+1-bp}{2(p+1)}}}\right)\right)(G^{-1})'(0)^{-\frac{p}{p+1}} (\alpha\rho_n)^{\frac{p}{p+1}} (1+p)^{\frac{1}{p+1}} \left(1-\frac{t_2}{\alpha\sqrt{n}}\right)^{\frac{p}{p+1}} - \frac{3}{n} \\
&\geq \left(1-\frac{1}{p+1}\left(\frac{\beta_1}{n^{\frac{bp}{p+1}}}+\frac{\beta_2}{n^{\frac{p+1-bp}{2(p+1)}}}\right)\right)(G^{-1})'(0)^{-\frac{p}{p+1}} (\alpha\rho_n)^{\frac{p}{p+1}} (1+p)^{\frac{1}{p+1}} \left(1-\frac{2^{\frac{1}{p+1}}p}{p+1}\frac{t_2}{\alpha\sqrt{n}}\right) - \frac{3}{n} \\
&= (G^{-1})'(0)^{-\frac{p}{p+1}} (\alpha\rho_n)^{\frac{p}{p+1}} (1+p)^{\frac{1}{p+1}} - 2^{\frac{1}{p+1}} p t_2 \left((G^{-1})'(0)(1+p)\right)^{-\frac{p}{p+1}} \alpha^{-\frac{1}{p+1}} \rho_0^{\frac{p}{p+1}} \frac{1}{n^{\frac{2bp+p+1}{2(p+1)}}} \\
&\quad - \frac{1}{p+1}\left(\frac{\beta_1}{n^{\frac{2bp}{(p+1)}}}+\frac{\beta_2}{n^{\frac{p+bp+1}{2(p+1)}}}\right)(G^{-1})'(0)^{-\frac{p}{p+1}} (\alpha\rho_0)^{\frac{p}{p+1}} (1+p)^{\frac{1}{p+1}} \left(1-\frac{2^{\frac{1}{p+1}}p}{p+1}\frac{t_2}{\alpha\sqrt{n}}\right) - \frac{3}{n} \\
&\geq (G^{-1})'(0)^{-\frac{p}{p+1}} (\alpha\rho_n)^{\frac{p}{p+1}} (1+p)^{\frac{1}{p+1}} - \frac{\overline{C}_1}{n^{\min\left\{1,\frac{2bp}{(p+1)},\frac{p+bp+1}{2(p+1)}\right\}}},
\end{aligned}$$

where $\overline{C}_1 := 2^{\frac{1}{p+1}} p t_2 \left((G^{-1})'(0)(1+p)\right)^{-\frac{p}{p+1}} \alpha^{-\frac{1}{p+1}} \rho_0^{\frac{p}{p+1}} + (\beta_1+\beta_2)(G^{-1})'(0)^{-\frac{p}{p+1}} (\alpha\rho_0)^{\frac{p}{p+1}} (1+p)^{-\frac{p}{p+1}} + 3$.

On the other hand, using (5) and Lemma EC.13, we can obtain that

$$\mathcal{R}_{\mathbb{P}_n,p}(\rho_n) < \frac{J(\omega)}{n} \leq \left(1-\frac{p\beta_1}{\gamma n^{\frac{bp}{p+1}}}-\frac{\sqrt{2}p}{\gamma}\frac{\beta_2}{n^{\frac{p+1-bp}{2(p+1)}}}\right)^{-\frac{1}{p+1}} c_n^{\frac{1}{p+1}} + \frac{2}{n}.$$

Since $\frac{p\beta_1}{\gamma n^{\frac{bp}{p+1}}}+\frac{\sqrt{2}p}{\gamma}\frac{\beta_2}{n^{\frac{p+1-bp}{2(p+1)}}}<\frac{1}{2}$ when $n>n_{12}$ and $(1-x)^{-\frac{1}{p+1}}\le 1+\frac{2^{\frac{p+2}{p+1}}}{p+1}x$ for $p\ge 1$ and $0<x\le\frac{1}{2}$, it follows that

$$\begin{aligned}\mathcal{R}_{\mathbb{P}_n,p}(\rho_n)&\le\left(1+\frac{2^{\frac{p+2}{p+1}}}{p+1}\left(\frac{p\beta_1}{\gamma n^{\frac{bp}{p+1}}}+\frac{\sqrt{2}p}{\gamma}\frac{\beta_2}{n^{\frac{p+1-bp}{2(p+1)}}}\right)\right)(G^{-1})'(0)^{-\frac{p}{p+1}}c_n^{\frac{1}{p+1}}+\frac{2}{n}\\&=\left(1+\frac{2^{\frac{p+2}{p+1}}}{p+1}\left(\frac{p\beta_1}{\gamma n^{\frac{bp}{p+1}}}+\frac{\sqrt{2}p}{\gamma}\frac{\beta_2}{n^{\frac{p+1-bp}{2(p+1)}}}\right)\right)(G^{-1})'(0)^{-\frac{p}{p+1}}\rho_n^{\frac{p}{p+1}}(1+p)^{\frac{1}{p+1}}\left(\frac{N(\omega)+1}{n}\right)^{\frac{p}{p+1}}+\frac{2}{n}.\end{aligned}$$

Since $|N(\omega)-n\alpha|<t_2\sqrt{n}$ implied by $n>n_1$, we have

$$\begin{aligned}&\mathcal{R}_{\mathbb{P}_n,p}(\rho_n)\\&\le\left(1+\frac{2^{\frac{p+2}{p+1}}}{p+1}\left(\frac{p\beta_1}{\gamma n^{\frac{bp}{p+1}}}+\frac{\sqrt{2}p}{\gamma}\frac{\beta_2}{n^{\frac{p+1-bp}{2(p+1)}}}\right)\right)(G^{-1})'(0)^{-\frac{p}{p+1}}\rho_n^{\frac{p}{p+1}}(1+p)^{\frac{1}{p+1}}\left(\frac{n\alpha+t_2\sqrt{n}+1}{n}\right)^{\frac{p}{p+1}}+\frac{2}{n}\\&\le\left(1+\frac{2^{\frac{p+2}{p+1}}}{p+1}\left(\frac{p\beta_1}{\gamma n^{\frac{bp}{p+1}}}+\frac{\sqrt{2}p}{\gamma}\frac{\beta_2}{n^{\frac{p+1-bp}{2(p+1)}}}\right)\right)(G^{-1})'(0)^{-\frac{p}{p+1}}(\alpha\rho_n)^{\frac{p}{p+1}}(1+p)^{\frac{1}{p+1}}\left(1+\frac{t_2+1}{\alpha\sqrt{n}}\right)^{\frac{p}{p+1}}+\frac{2}{n}\\&\le\left(1+\frac{2^{\frac{p+2}{p+1}}}{p+1}\left(\frac{p\beta_1}{\gamma n^{\frac{bp}{p+1}}}+\frac{\sqrt{2}p}{\gamma}\frac{\beta_2}{n^{\frac{p+1-bp}{2(p+1)}}}\right)\right)(G^{-1})'(0)^{-\frac{p}{p+1}}(\alpha\rho_n)^{\frac{p}{p+1}}(1+p)^{\frac{1}{p+1}}\left(1+\frac{t_2+1}{\alpha\sqrt{n}}\right)+\frac{2}{n}\\&=(G^{-1})'(0)^{-\frac{p}{p+1}}(\alpha\rho_n)^{\frac{p}{p+1}}(1+p)^{\frac{1}{p+1}}+(t_2+1)(G^{-1})'(0)^{-\frac{p}{p+1}}\alpha^{-\frac{1}{p+1}}\rho_0^{\frac{p}{p+1}}(1+p)^{\frac{1}{p+1}}\frac{1}{n^{\frac{2bp+p+1}{2(p+1)}}}\\&\quad+\frac{2^{\frac{p+2}{p+1}}}{p+1}\left(\frac{p\beta_1}{\gamma n^{\frac{2bp}{(p+1)}}}+\frac{\sqrt{2}p}{\gamma}\frac{\beta_2}{n^{\frac{p+bp+1}{2(p+1)}}}\right)(G^{-1})'(0)^{-\frac{p}{p+1}}(\alpha\rho_0)^{\frac{p}{p+1}}(1+p)^{\frac{1}{p+1}}\left(1+\frac{t_2+1}{\alpha\sqrt{n}}\right)+\frac{2}{n}\\&\ge(G^{-1})'(0)^{-\frac{p}{p+1}}(\alpha\rho_n)^{\frac{p}{p+1}}(1+p)^{\frac{1}{p+1}}+\frac{\overline{C}_2}{n^{\min\left(1,\frac{2bp}{(p+1)},\frac{p+bp+1}{2(p+1)}\right)}},\end{aligned}$$

where $\overline{C}_2=(t_2+1)(G^{-1})'(0)^{-\frac{p}{p+1}}\alpha^{-\frac{1}{p+1}}\rho_0^{\frac{p}{p+1}}(1+p)^{\frac{1}{p+1}}+2^{\frac{p+2}{p+1}}\left(\frac{p\beta_1+\sqrt{2}p\beta_2}{\gamma}\right)\left(\frac{\alpha\rho_0}{(G^{-1})'(0)(1+p)}\right)^{\frac{p}{p+1}}\left(\frac{3}{2}+\frac{1}{\alpha}\right)+2$.
Therefore, the proof is completed by setting $n_0=\lceil\max\{n_1,n_9,n_{10},n_{11},n_{12},n_{13}\}\rceil$ and $C=\max\left(\overline{C}_1,\overline{C}_2\right)$. □

EC.4.2.2. Proof of Theorem 6 (II)
Define $\inf_3:=\min\{s>0:G(s)=G(\delta)/2\}$, and

$$n_{14}:=\left(t_1\sqrt{\frac{1}{4}+\frac{t_1^2}{8}}+\frac{\sqrt{2}t_1^2}{4}\right)^{\frac{2}{1-2bp}}\left(\frac{\delta_3}{\rho_0}\right)^{\frac{2p}{1-2bp}}\alpha^{\frac{1}{1-2bp}},$$

$$n_{15}:=\left(\frac{2}{\sqrt{\alpha}G(\delta 3)}\right)^{\frac{1}{\min(bp,1/2)}}\left(\left(t_1\sqrt{1+\frac{t_1^2}{2}}+\frac{\sqrt{2}t_1^2}{2}+\frac{4\rho_0^p}{\delta_3^p\sqrt{\alpha}}\right)+\left(1+t_1\sqrt{2+t_1^2}+t_1^2+\frac{4\rho_0^p}{\delta_3^p\alpha}\right)^{\frac{1}{2}}\right)^{\frac{1}{\min(bp,1/2)}},$$

$$n_{16}:=2^{\frac{2(p+1)}{p+2-2bp}}(\alpha L_g\rho_0^2)^{-\frac{p}{p+2-2bp}}(p+2)^{-\frac{2}{p+2-2bp}}\left((\frac{1}{2})^{p+2}(t_1+1)^{p+2}(p+2)^{p+2}+2^{p+2}(t_1+1)^{p+2}\right)^{\frac{2}{p+2-2bp}},$$

$$n_{17}:=\frac{2}{\alpha},\qquad n_{18}:=2^{\frac{2(p+3)}{p+2-2bp}}(p+2)^{-\frac{2}{p+2-2bp}}(t_1+1)^{\frac{p+2}{p+2-2bp}}\left(\alpha L_g\rho_0^2\right)^{-\frac{p}{p+2-2bp}}.$$

Proof of Theorem 6 (II). When $n > \max(n_1, n_{14}, n_{15})$,applying Lemma EC.10 with $\delta_1 = \delta$ and $\delta_2 = \delta_3$ yields that $S_j < \delta$ for all $1 \le j \le N(\omega)$. By Lemma EC.1,

$$S_j(\omega) \ge \sqrt{\frac{2}{L_g} U_j(\omega)}.$$

Taking this bound into (4) gives that

$$n\rho_n^p \ge \sum_{j=0}^{J(\omega)} \left(\frac{2}{L_g} U_j(\omega)\right)^{\frac{p}{2}} \ge \sum_{j=0}^{J(\omega)} \left(\frac{2}{L_g}\right)^{\frac{p}{2}} \left(\mathbb{E}[U_j] - t_1\sigma_j\right)_+^{\frac{p}{2}} \quad \ge \sum_{j=0}^{J(\omega)} \left(\frac{2}{L_g}\right)^{\frac{p}{2}} \left(\frac{j}{N(\omega)+1} - t_1 \frac{\sqrt{j}}{N(\omega)+1}\right)_+^{\frac{p}{2}},$$

which implies that

$$\left(\frac{L_g}{2}\right)^{\frac{p}{2}} n\rho_n^p (N(\omega)+1)^{\frac{p}{2}} > \sum_{j=1}^{J(\omega)} (j - t_1\sqrt{j})_+^{\frac{p}{2}}.$$

Without loss of generality, assume $J(\omega) \ge 1$, otherwise the bound holds trivially $\mathcal{R}_{\mathbb{P}_n,p} \le 1/n$. Using the inequality $(1-x)^{\frac{p}{2}} \ge 1 - px$ for $p \ge 1$ and $0 \le x \le \frac{1}{2}$, we have that

$$\begin{aligned}
\left(\frac{L_g}{2}\right)^{\frac{p}{2}} n\rho_n^p (N(\omega)+1)^{\frac{p}{2}} &\ge \sum_{j=0}^{J(\omega)} (j - t_1\sqrt{j})_+^{\frac{p}{2}} \\
&= \sum_{j=0}^{J(\omega)} j^{\frac{p}{2}} \left(1 - \frac{t_1}{\sqrt{j}}\right)_+^{\frac{p}{2}} \\
&\ge \int_0^{J(\omega)-1} x^{\frac{p}{2}} \left(1 - \frac{t_1}{\sqrt{x}}\right)_+^{\frac{p}{2}} dx \\
&\ge \int_{4t_1^2}^{J(\omega)-1} x^{\frac{p}{2}} \left(1 - \frac{p}{2}\frac{t_1}{\sqrt{x}}\right) dx \\
&= \frac{2}{p+2}(J(\omega)-1)^{\frac{p+2}{2}} - \frac{pt_1}{(p+1)}(J(\omega)-1)^{\frac{p+1}{2}} - \frac{2^{p+3}t_1^{p+2}}{p+2} + \frac{p2^{p+1}t_1^{p+2}}{(p+1)} \\
&\ge \frac{2}{p+2} J(\omega)^{\frac{p+2}{2}} (1 - J(\omega)^{-1})^{\frac{p+2}{2}} - t_1 (J(\omega)-1)^{\frac{p+1}{2}} - \frac{2^{p+3}t_1^{p+2}}{p+2} \\
&\ge \frac{2}{p+2} J(\omega)^{\frac{p+2}{2}} - J(\omega)^{\frac{p}{2}} - t_1 J(\omega)^{\frac{p+1}{2}} - \frac{2^{p+3}t_1^{p+2}}{p+2} \\
&\ge \frac{2}{p+2} J(\omega)^{\frac{p+2}{2}} - (t_1+1) J(\omega)^{\frac{p+1}{2}} - \frac{2^{p+3}(t_1+1)^{p+2}}{p+2},
\end{aligned}$$

which implies that

$$l_n > J(\omega)^{\frac{p+2}{2}} - \frac{(t_1+1)(p+2)}{2} J(\omega)^{\frac{p+1}{2}}. \tag{EC.16}$$

where $l_n = \frac{p+2}{2}\left(\frac{L_g}{2}\right)^{\frac{p}{2}} n\rho_n^p (N+1)^{\frac{p}{2}} - 2^{p+2}(t_1+1)^{p+2}$.

Define $f(x) = x^{\frac{p+2}{2}} - \frac{(t_1+1)(p+2)}{2} x^{\frac{p+1}{2}}$. The roots of $f'(x) = 0$ are $x_1 = 0$ and $x_2 = \frac{1}{4}(t_1+1)^2(p+2)^2$. Hence, $f(x)$ is decreasing on $[0, x_2]$ and increasing on $(x_2, +\infty)$. Since $l_n \ge \frac{p+2}{2}\left(\frac{L_g}{2}\right)^{\frac{p}{2}} n\rho_n^p (\frac{1}{2}\alpha n)^{\frac{p}{2}} - 2^{p+2}(t_1+1)^{p+2} \ge (\frac{1}{2})^{p+2}(t_1+1)^{p+2}(p+2)^{p+2}$ and $l_n^{\frac{2}{p+2}} \ge x_2$ when $n \ge n_{16}$,

$$f(2^{\frac{2}{p+2}} l_n^{\frac{2}{p+2}}) = 2l_n - \frac{(t_1+1)(p+2)}{2^{\frac{1}{p+2}}} l_n^{\frac{p+1}{p+2}} \ge l_n,$$

which, by the monotonicity of $f(x)$ and (EC.16), implies that

$$J(\omega) \le l_n^{\frac{2}{p+2}} = \left(\frac{p+2}{2} \left(\frac{L_g}{2} \right)^{\frac{p}{2}} n\rho_n^p (N(\omega)+1)^{\frac{p}{2}} - 2^{p+2}(t_1+1)^{p+2} \right)^{\frac{2}{p+2}}.$$

Since $|N(\omega) - n\alpha| \le t_2\sqrt{n}$ implied by $n > n_1$, it holds that

$$\begin{aligned}
J(\omega) &\le \left(\frac{p+2}{2} \left(\frac{L_g}{2} \right)^{\frac{p}{2}} n\rho_n^p \left(\frac{3}{2}\alpha n + 1 \right)^{\frac{p}{2}} - 2^{p+2}(t_1+1)^{p+2} \right)^{\frac{2}{p+2}} \\
&\le \left(\frac{p+2}{2} \left(\frac{L_g}{2} \right)^{\frac{p}{2}} n\rho_n^p (2\alpha n)^{\frac{p}{2}} - 2^{p+2}(t_1+1)^{p+2} \right)^{\frac{2}{p+2}} \\
&= \left(\frac{p+2}{2} (\alpha L_g)^{\frac{p}{2}} \rho_0^p n^{\frac{p+2-2bp}{2}} - 2^{p+2}(t_1+1)^{p+2} \right)^{\frac{2}{p+2}} + 1 \\
&= \left(\frac{p+2}{2} \right)^{\frac{2}{p+2}} \left(\alpha L_g \rho_0^2 \right)^{\frac{p}{p+2}} n^{\frac{p+2-2bp}{p+2}} \left(1 - \frac{2^{p+3}}{p+2} \left(\alpha L_g \rho_0^2 \right)^{-\frac{p}{2}} (t_1+1)^{p+2} n^{-\frac{p+2-2bp}{2}} \right)^{\frac{2}{p+2}},
\end{aligned}$$

where the second inequality holds since $n \ge n_{17}$. Observe that $\frac{2^{p+3}}{p+2} \left(\alpha L_g \rho_0^2 \right)^{-\frac{p}{2}} (t_1+1)^{p+2} n^{-\frac{p+2-2bp}{2}} \le 1$ when $n \ge n_{18}$. Using the inequality that $(1+x)^{\frac{2}{p+2}} \le 1 + \frac{2}{p+2}x$ for $p \ge 1$ and $x \ge -1$, we have

$$J(\omega) \le \left(\frac{p+2}{2} \right)^{\frac{2}{p+2}} \left(\alpha L_g \rho_0^2 \right)^{\frac{p}{p+2}} n^{\frac{p+2-2bp}{p+2}} \left(1 - \frac{2}{p+2} \frac{2^{p+3}}{p+2} \left(\alpha L_g \rho_0^2 \right)^{-\frac{p}{2}} (t_1+1)^{p+2} n^{-\frac{p+2-2bp}{2}} \right).$$

Therefore, for $n > n_0 = \lceil \max\{n_1, n_{14}, n_{15}, n_{16}, n_{17}, n_{18}\} \rceil$,

$$\mathcal{R}_{\mathbb{P}_n,p}(\rho_n) < \frac{J(\omega)+1}{n} \le \left(\frac{p+2}{2} \right)^{\frac{2}{p+2}} \left(L_g \rho_0^2 \right)^{\frac{p}{p+2}} \frac{1}{n^{\frac{2bp}{p+2}}} + \frac{1}{n}.$$

□

EC.4.2.3. Proofs of Preparatory Results

Proof of Lemma EC.10. Define $\underline{J}(\omega) := \max\{1 \le j \le N(\omega) : U_j(\omega) \le G(\delta_1)/2\}$ and $\delta_2 = \inf\{s > 0 : G(s) = G(\delta_1)/2\}$, recalling δ is defined in Assumption 1. Then $\delta_2 > 0$ since G is right-continuous and $G(\delta_2) = G(\delta_1)/2 > 0$. When $J(\omega) \le \underline{J}(\omega)$, $G(\underline{s}) \le U_j(\omega) \le G(\delta_1)/2$ for all $j \le J(\omega)$. When $J(\omega) > \underline{J}(\omega)$, for all $\underline{J}(\omega) < j \le J(\omega)$ we have $U_j(\omega) > G(\delta_1)/2 = G(\delta_2)$, and thus $S_j \ge \delta_2$. Then from (4), we have

$$\rho_n^p \ge \frac{1}{n} \sum_{j=1}^{J(\omega)} S_j(\omega)^p \ge \frac{1}{n} \sum_{j=\underline{J}(\omega)+1}^{J(\omega)} S_j(\omega)^p \ge \frac{1}{n} \sum_{j=\underline{J}(\omega)+1}^{J(\omega)} \delta_2^p,$$

which implies that

$$J(\omega) \le \underline{J}(\omega) + \frac{n\rho_n^p}{\delta_2^p}.$$

By Lemma EC.5, for $\omega \in \Omega_0$,

$$\mathbb{E}[U_{\underline{J}}] - t_1 \sigma_{\underline{J}(\omega)} \le U_{\underline{J}(\omega)} \le \frac{G(\delta_1)}{2},$$

which yields that

$$\frac{\underline{J}(\omega)}{N(\omega)+1} - t_1 \frac{\sqrt{\underline{J}(\omega)}}{N(\omega)+1} \le \frac{G(\delta_1)}{2}.$$

Solving out the inequality yields that

$$0 \le \underline{J}(\omega) \le \frac{G(\delta_1)}{2}(N(\omega)+1) + t_1\sqrt{\frac{G(\delta_1)}{2}(N(\omega)+1) + \frac{t_1^2}{4}} + \frac{t_1^2}{2}.$$

It follows that

$$\begin{aligned}
U_{J(\omega)}(\omega) &\le \mathbb{E}[U_J] + t_1\sigma_{J(\omega)} \\
&\le \frac{J(\omega)}{N(\omega)+1} + t_1\frac{\sqrt{J(\omega)}}{N(\omega)+1} \\
&\le \frac{\underline{J}(\omega) + n\rho_n^p\delta_2^{-p}}{N(\omega)+1} + t_1\frac{\sqrt{\underline{J}(\omega) + n\rho_n^p\delta_2^{-p}}}{N(\omega)+1} \\
&\le \frac{G(\delta_1)}{2} + \left(t_1\sqrt{\frac{G(\delta_1)}{2(N(\omega)+1)} + \frac{t_1^2}{4(N(\omega)+1)^2}} + \frac{t_1^2}{2(N(\omega)+1)} + \frac{n\rho_n^p}{\delta_2^p(N(\omega)+1)}\right) \\
&\quad + t_1\left(\frac{G(\delta_1)}{2}(N(\omega)+1) + t_1\sqrt{\frac{G(\delta_1)}{2}(N(\omega)+1) + \frac{t_1^2}{4}} + \frac{t_1^2}{2} + \frac{n\rho_n^p}{\delta_2^p}\right)^{\frac{1}{2}} /(N(\omega)+1).
\end{aligned}$$

Denote by K_2 and K_3 the second and the third term of the last inequality, respectively. Then, for K_2, it follows from inequalities that $|N(\omega) - n\alpha| \le t_2\sqrt{n}$ and $n > n_1$ that

$$\begin{aligned}
K_2 &= \frac{1}{\sqrt{N(\omega)+1}}\left(t_1\sqrt{\frac{G(\delta_1)}{2} + \frac{t_1^2}{4(N(\omega)+1)}} + \frac{t_1^2}{2\sqrt{N(\omega)+1}} + \frac{n^{1-bp}\rho_0^p}{\delta_2^p\sqrt{N(\omega)+1}}\right) \\
&\le \frac{1}{\sqrt{n\alpha - t_2\sqrt{n}}}\left(t_1\sqrt{\frac{G(\delta_1)}{2} + \frac{t_1^2}{4(N(\omega)+1)}} + \frac{t_1^2}{2\sqrt{N(\omega)+1}} + \frac{n^{1-bp}\rho_0^p}{\delta_2^p\sqrt{n\alpha - t_2\sqrt{n}}}\right) \\
&\le \frac{1}{\sqrt{\frac{1}{2}n\alpha}}\left(t_1\sqrt{\frac{1}{2} + \frac{t_1^2}{4}} + \frac{t_1^2}{2} + \frac{\sqrt{2}\rho_0^p n^{\frac{1}{2}-bp}}{\delta_2^p\sqrt{\alpha}}\right).
\end{aligned}$$

In the last inequality above, observe that when $0 < bp < \frac{1}{2}$ and $n \ge n_9$, it holds that $t_1\sqrt{\frac{1}{2} + \frac{t_1^2}{4}} + \frac{t_1^2}{2} \le \frac{\sqrt{2}\rho_0^p n^{\frac{1}{2}-bp}}{\delta_2^p\sqrt{\alpha}}$. Then we can derive that

$$\begin{aligned}
K_2 &\le \begin{cases} \frac{4\rho_0^p}{\delta_2^p\alpha n^{bp}}, \text{ if } 0 < bp < \frac{1}{2}, \\ \frac{1}{\sqrt{\frac{1}{2}n\alpha}}\left(t_1\sqrt{\frac{1}{2} + \frac{t_1^2}{4}} + \frac{t_1^2}{2} + \frac{\sqrt{2}\rho_0^p}{\delta_2^p\sqrt{\alpha}}\right), \text{ if } bp \ge \frac{1}{2}, \end{cases} \\
&\le \frac{1}{\sqrt{\alpha}n^{\min(bp,\frac{1}{2})}}\left(t_1\sqrt{1 + \frac{t_1^2}{2}} + \frac{\sqrt{2}t_1^2}{2} + \frac{4\rho_0^p}{\delta_2^p\sqrt{\alpha}}\right).
\end{aligned}$$

For K_3, due to $|N(\omega)-n\alpha| \le t_2\sqrt{n}$ and $n > n_1$, it holds that

$$\begin{aligned}
K_3 &= \frac{1}{\sqrt{N(\omega)+1}}\left(\frac{G(\delta_1)}{2} + t_1\sqrt{\frac{G(\delta_1)}{2(N(\omega)+1)} + \frac{t_1^2}{4(N(\omega)+1)^2}} + \frac{t_1^2}{2(N(\omega)+1)} + \frac{n^{1-bp}\rho_0^p}{\delta_2^p(N(\omega)+1)}\right)^{\frac{1}{2}} \\
&\le \frac{1}{\sqrt{n\alpha - t\sqrt{n}}}\left(\frac{1}{2} + t_1\sqrt{\frac{1}{2(N(\omega)+1)} + \frac{t_1^2}{4(N(\omega)+1)^2}} + \frac{t_1^2}{2(N(\omega)+1)} + \frac{n^{1-bp}\rho_0^p}{\delta_2^p(n\alpha - t\sqrt{n})}\right)^{\frac{1}{2}} \\
&\le \frac{1}{\sqrt{\frac{1}{2}n\alpha}}\left(\frac{1}{2} + t_1\sqrt{\frac{1}{2} + \frac{t_1^2}{4}} + \frac{t_1^2}{2} + \frac{2n^{-bp}\rho_0^p}{\delta_2^p\alpha}\right)^{\frac{1}{2}} \\
&\le \frac{1}{\sqrt{\alpha n}}\left(1 + t_1\sqrt{2+t_1^2} + t_1^2 + \frac{4\rho_0^p}{\delta_2^p\alpha}\right)^{\frac{1}{2}}.
\end{aligned}$$

Finally, since $n > n_{10}$, we have

$$K_2 + K_3 \le \frac{1}{\sqrt{\alpha}n^{\min(bp,\frac{1}{2})}}\left(t_1\sqrt{1+\frac{t_1^2}{2}} + \frac{\sqrt{2}t_1^2}{2} + \frac{4\rho_0^p}{\delta_2^p\sqrt{\alpha}}\right) + \frac{1}{\sqrt{\alpha n}}\left(1 + t_1\sqrt{2+t_1^2} + t_1^2 + \frac{4\rho_0^p}{\delta_2^p\alpha}\right)^{\frac{1}{2}} < \frac{G(\delta_1)}{2}.$$

Therefore, we have shown that $U_j(\omega) \le \mathbb{E}[U_j] + t_1\sigma_j < G(\delta_1)$ for any $1 \le j \le J(\omega)$ and $\omega \in \Omega_0$, which completes the proof. □

Proof of Lemma EC.11. Recall from the Lemma EC.10 that when $n > \max(n_1, n_9, n_{10})$, $U_j(\omega) \le G(\delta_1)$ for $1 \le j \le J(\omega)$. It follows from Lemma EC.7 that

$$M\mathbb{E}[U_j] \le G^{-1}(\mathbb{E}[U_j]) \le L_{G^{-1}}\mathbb{E}[U_j]. \tag{EC.17}$$

Using (4) and $S_0 = 0$, it follows that

$$\begin{aligned}
n\rho_n^p &< \sum_{j=1}^{J(\omega)+1} G^{-1}\left(U_j(\omega)\right)^p \\
&\le \sum_{j=1}^{J(\omega)+1}\left(G^{-1}(\mathbb{E}[U_j]) + L_{G^{-1}}t_1\sigma_j\right)^p \\
&\le \sum_{j=1}^{J(\omega)+1}\left(L_{G^{-1}}\mathbb{E}[U_j] + L_{G^{-1}}t_1\sigma_j\right)^p \\
&\le \sum_{j=1}^{J(\omega)+1}\left(L_{G^{-1}}\frac{j}{N(\omega)+1} + L_{G^{-1}}t_1\frac{\sqrt{j}}{N(\omega)+1}\right)^p \\
&\le \frac{L_{G^{-1}}^p}{(N(\omega)+1)^p}\sum_{j=1}^{J(\omega)+1} j^p\left(1 + t_1\frac{1}{\sqrt{j}}\right)^p \\
&\le \frac{L_{G^{-1}}^p(1+t_1)^p}{(N(\omega)+1)^p}\sum_{j=1}^{J(\omega)+1} j^p \\
&\le \frac{L_{G^{-1}}^p(1+t_1)^p}{(N(\omega)+1)^p}\int_{x=0}^{J(\omega)+2} x^p\,dx \\
&= \frac{L_{G^{-1}}^p(1+t_1)^p}{(N(\omega)+1)^p(p+1)}(J(\omega)+2)^{p+1}.
\end{aligned}$$

Accordingly,

$$\frac{c_n}{L_{G^{-1}}^p (1+t_1)^p} \le (J(\omega)+2)^{p+1},$$

which implies that

$$J(\omega) > J_1(\omega) = L_{G^{-1}}^{-\frac{p}{p+1}} (1+t_1)^{-\frac{p}{p+1}} c_n^{\frac{1}{p+1}} - 2.$$

Next, we derive the upper bound. Without loss of generality, suppose $J(\omega) \ge 1$ since otherwise, the upper bound trivially holds. Using (EC.9) (EC.17) and (4), we obtain that

$$\begin{aligned}
n\rho_n^p &\ge \sum_{j=0}^{J(\omega)} G^{-1}\left(U_j(\omega)\right)^p \\
&\ge \sum_{j=1}^{J(\omega)} \left(G^{-1}(\mathbb{E}(U_j)) - t_1 L_{G^{-1}} \sigma_j\right)_+^p \\
&\ge \sum_{j=1}^{J(\omega)} \left(M\mathbb{E}(U_j) - t_1 L_{G^{-1}} \sigma_j\right)_+^p \\
&\ge \frac{M^p}{(N(\omega)+1)^p} \sum_{j=1}^{J(\omega)} j^p \left(1 - \frac{t_1 L_{G^{-1}}}{M} \frac{1}{\sqrt{j}}\right)_+^p \\
&\ge \frac{M^p}{(N(\omega)+1)^p} \int_0^{J(\omega)-1} x^p \left(1 - \frac{t_1 L_{G^{-1}}}{M} \frac{1}{\sqrt{x}}\right)_+^p dx.
\end{aligned} \tag{EC.18}$$

It then follows the inequality $(1-x)_+^p \ge 1 - px$ for $p \ge 1$ and $x \ge 0$ that

$$\begin{aligned}
n\rho_n^p &\ge \frac{M^p}{(N(\omega)+1)^p} \int_0^{J(\omega)-1} x^p \left(1 - \frac{t_1 L_{G^{-1}}}{M} \frac{1}{\sqrt{x}}\right)_+^p dx \\
&\ge \frac{M^p}{(N(\omega)+1)^p} \int_0^{J(\omega)-1} \left(x^p - p \frac{t_1 L_{G^{-1}}}{M} x^{p-\frac{1}{2}}\right) dx \\
&\ge \frac{M^p}{(N(\omega)+1)^p} \left(\frac{(J(\omega)-1)^{p+1}}{p+1} - p \frac{t_1 L_{G^{-1}}}{M} \frac{(J(\omega)-1)^{p-\frac{1}{2}}}{p+\frac{1}{2}}\right) \\
&\ge \frac{M^p}{(N(\omega)+1)^p} \frac{(J(\omega)-1)^{p+1}}{p+1} \left(1 - p \frac{t_1 L_{G^{-1}}}{M} \frac{p+1}{p+\frac{1}{2}} \frac{1}{\sqrt{J(\omega)-1}}\right) \\
&\ge \frac{M^p}{(N(\omega)+1)^p} \frac{(J(\omega)-1)^{p+1}}{p+1} \left(1 - (p+1) \frac{t_1 L_{G^{-1}}}{M} \frac{1}{\sqrt{J_1(\omega)-1}}\right) \\
&\ge \frac{M^p}{2(N(\omega)+1)^p} \frac{(J(\omega)-1)^{p+1}}{p+1},
\end{aligned}$$

where the last inequality holds because $J_1(\omega) \ge 2$ and $1 - (p+1)\frac{t_1 L_{G^{-1}}}{M} \frac{1}{\sqrt{J_1(\omega)-1}} \ge \frac{1}{2}$ when $n \ge n_{11}$. As a result,

$$J(\omega) \le J_2(\omega) = 2^{\frac{1}{p+1}} M^{-\frac{p}{p+1}} c_n^{\frac{1}{p+1}} + 1,$$

which completes the proof. □

Proof of Lemma EC.12. For any $x, y \in [0, G(\delta_1)]$,

$$\begin{aligned}
\left|(G^{-1})'(x) - (G^{-1})'(y)\right| &= \left|\frac{1}{g(G^{-1}(x))} - \frac{1}{g(G^{-1}(y))}\right| \\
&= \frac{\left|g(G^{-1}(y)) - g(G^{-1}(x))\right|}{\left|g(G^{-1}(y))g(G^{-1}(x))\right|} \\
&\le \frac{1}{C_1^2}\left|g(G^{-1}(y)) - g(G^{-1}(x))\right| \\
&\le \frac{L_g}{C_1^2}\left|G^{-1}(y) - G^{-1}(x)\right| \\
&\le \frac{L_g}{C_1^3}\left|x - y\right|,
\end{aligned}$$

where the first inequality follows that $0 < C_1 \le g(x) \le C_2 < \infty$ for any $x \in [0, \delta_1]$; the second inequality is due to Assumption 1 and the last inequality comes from Lemma EC.7. □

Appendix EC.5: Proofs for Section 5.1

Below we show the calculation of $\mathfrak{g}_{w,r,y}(0)$ for $y = 1$, and the case $y = -1$ can be done similarly. We have that

$$\mathfrak{g}_{w,r,1}(0) = \lim_{s \downarrow 0} \frac{\mathbb{P}_{\mathsf{true}, X|Y=1}\left(-r/\|w\|_2 < w^\top X/\|w\|_2 \le -r/\|w\|_2 + s\right)}{s}.$$

Define $f_{\varsigma|Y=1}$ to be the conditional density function of a random variable $\varsigma = w^\top x/\|w\|_2$ given $Y = 1$. Then the right-hand side above equals $f_{\varsigma|Y=1}(-r/\|w\|_2)$. In the following we derive an expression for $f_{\varsigma|Y=1}$. Using properties of the conditional density, for any Borel function $\phi : \mathbb{R} \to \mathbb{R}$,

$$\mathbb{E}[\phi(\varsigma)] = \int_{\mathbb{R}} \phi(t) f_{\varsigma|Y=1}(t)dt = \int_{\mathbb{R}} \phi(w^\top x/\|w\|_2) f_1(x)dx.$$

Using the change of variable $x \mapsto (t, v_1, \cdots, v_{d-1})$, the associated Jacobian matrix is the $d \times d$ matrix $(w/\|w\|_2, e_1, \cdots, e_{d-1})$ and its determinant is 1 or -1. It follows that

$$\begin{aligned}
\mathbb{E}[\phi(w^\top x/\|w\|_2)] &= \int_{\mathbb{R}^d} \phi(w^\top x/\|w\|_2) f_1(x)dx \\
&= \int_{\mathbb{R}^d} \phi\left(\frac{w^\top}{\|w\|_2}\left(t\frac{w}{\|w\|_2} + \sum_{k=1}^{d-1} t_k e_k\right)\right) f_1\left(t\frac{w}{\|w\|_2} + \sum_{k=1}^{d-1} t_k e_k\right) dtdv \\
&= \int_{\mathbb{R}} \phi(t)\left(\int_{\mathbb{R}^{d-1}} f_1\left(t\frac{w}{\|w\|_2} + \sum_{k=1}^{d-1} t_k e_k\right) dv\right) dt \\
&= \int_{\mathbb{R}} \phi(t)\left(\int_{\mathbb{R}^{d-1}} f_1(tw/\|w\|_2 + \sum_{k=1}^{d-1} v_k e_k)dv\right) dt.
\end{aligned}$$

Therefore,

$$f_{\varsigma|Y=1}(t) = \int_{\mathbb{R}^{d-1}} f_1(tw/\|w\|_2 + \sum_{k=1}^{d-1} v_k e_k)dv,$$

and thus

$$\mathfrak{g}_{w,r,y}(0) = \int_{\mathbb{R}^{d-1}} f_y\Big(-rw/\|w\|_2^2 + \sum_{k=1}^{d-1} v_k e_k\Big)dv,$$

for $w \neq 0$. □

Appendix EC.6: Proof of Appendix A

Proof of Lemma 3. For $p=\infty$, by Lemma 2, it holds that

$$\sup_{\mathbb{P}\in\mathcal{P}(\mathcal{Z})}\{\mathbb{E}_{\mathbb{P}}[f(Z)]:\mathcal{W}_\infty(\mathbb{P},\mathbb{Q})\le\rho\}=\mathbb{E}_{\mathbb{Q}}\left[\sup_{\tilde z\in\mathcal{Z}}\{f(\tilde z):\mathsf{d}(\tilde z,Z)\le\rho\}\right].$$

Since the set $\mathcal{Z}_{\max}(z)=\arg\max_{\tilde z\in\mathcal{Z}}\{f(\tilde z):\mathsf{d}(\tilde z,z)\le\rho\}$ is bounded, it is totally bounded due to our assumption on $\mathcal{Z}$. Since f is upper semi-continuous function, the set $\mathcal{Z}_{\max}(z)$ is measurable and non-empty. It follows from Aumann's measurable selection theorem [1, Theorem 18.26] that there exists a $(\mathcal{B}_{\mathbb{Q}}(\mathcal{Z}),\mathcal{B}(\mathcal{Z}))$-measurable map $T_*:\mathcal{Z}\to\mathcal{Z}$ such that $T_*(z)\in\mathcal{Z}_{\max}(z)$, $\mathbb{Q}$-a.e. Notice that

$$\begin{aligned}
\mathcal{W}_\infty(\mathbb{P}_*,\mathbb{Q}) &= \inf_{\pi\in\Gamma(\mathbb{P}^*,\mathbb{Q})}\pi\text{-}\operatorname*{ess\,sup}_{\mathcal{Z}\times\mathcal{Z}}\mathsf{d}(\tilde z,z)\\
&= \inf_{\pi\in\Gamma(\mathbb{Q},\mathbb{Q})}\pi\text{-}\operatorname*{ess\,sup}_{\mathcal{Z}\times\mathcal{Z}}\mathsf{d}(T_*(\tilde z),z)\\
&\le \inf_{\pi\in\Gamma(\mathbb{Q},\mathbb{Q})}\pi\text{-}\operatorname*{ess\,sup}_{\mathcal{Z}\times\mathcal{Z}}\left(\mathsf{d}(T_*(\tilde z),\tilde z)+\mathsf{d}(\tilde z,z)\right)\\
&\le \rho+\inf_{\pi\in\Gamma(\mathbb{Q},\mathbb{Q})}\pi\text{-}\operatorname*{ess\,sup}_{\mathcal{Z}\times\mathcal{Z}}\mathsf{d}(\tilde z,z)\\
&= \rho.
\end{aligned}$$

Moreover,

$$\mathbb{E}_{\mathbb{P}_*}[f(Z)]=\mathbb{E}_{\mathbb{Q}}[f(T_*(Z))]=\mathbb{E}_{\mathbb{Q}}\left[\sup_{\tilde z\in\mathcal{Z}}\{f(\tilde z):\mathsf{d}(\tilde z,Z)\le\rho\}\right].$$

This shows that $\mathbb{P}_*$ is a worst-case distribution.

The proof for $p\in[1,\infty)$ is given in [19, Corollary 1]. □

Appendix EC.7: Additional Numerical Results

Consider the linear classification problem in Section 5.1. Particularly, let $\mathcal{Z}=(\mathbb{R}^{100},\|\cdot\|_2)$ and $E=\{z\in\mathbb{R}^{100}:w^\top z>0\}$ and $w=(1,-1,2,-2,\ldots,50,-50)^\top$. Let $p=1$. Suppose $\mathbb{P}_{\mathsf{true}}$ is a Gaussian distribution $\mathcal{N}((0,\ldots,0)^\top,\Sigma)$ where $\Sigma=(\sigma_{ij})\in\mathbb{R}^{100\times100}$ and $\sigma_{ij}=0.9^{|i-j|}$ for $1\le i,j\le100$.

Using a similar numerical setup for Example 1, we compare the radius selection rule $\rho_n=\rho_0/n$ as suggested by Corollary 2 and the $1/\sqrt{n}$-rule, where we vary $\rho_0\in\{0.5,5,50\}$ for $\rho_n=\rho_0/n$ and $\rho_0\in\{0.01,0.1,1\}$ for $\rho_n=\rho_0/\sqrt{n}$.

The results in Fig. EC.1 are consistent with Fig. 2 — the robust formulation (P) with both scaling schemes yield high-confidence bounds for the true loss, yet the bounds resulting from the $1/n$-radius selection rule suggested by Corollary 2 is tighter.

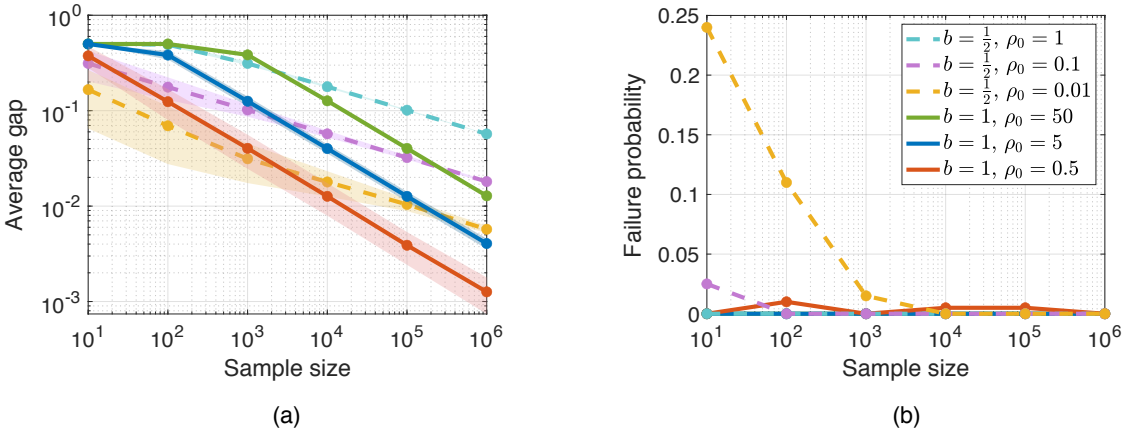

Figure EC.1 Comparisons between two scaling schemes $\rho_n \sim 1/n$ and $\rho_n \sim 1/\sqrt{n}$, represented by solid lines and dashed lines, respectively.