# DISTRIBUTIONALLY ROBUST LINEAR REGRESSION THROUGH THE LENS OF ADVERSARIAL TRAINING

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Distributionally robust optimization (DRO) studies parameter estimation under uncertainty in the underlying probability distribution and has emerged as a principled framework for analyzing robustness and generalization. In particular, Wasserstein DRO, with distributional uncertainty induced by the Wasserstein distance, generalizes several popular regularizers. This paper studies Wasserstein DRO linear regression, unifying square-root Lasso and adversarial linear regression as important special cases. We prove that many properties of these two special cases carry over to this general method. In particular, we show (i) deterministic and non-asymptotic in-sample error bounds $O(n^{-1/2})$ in general and $O(n^{-1})$ under design matrix and sparsity conditions; (ii) insensitivity to the noise level, also known as the pivotal property; and (iii) solution equivalences for small and large ambiguity sets. The key proof step is to recast the method into a quadratic form, mimicking adversarial linear regression. We also show that the method can be solved efficiently, and we validate our findings through numerical simulations.

## 1 INTRODUCTION

Distributionally robust optimization (DRO) studies parameter estimation under uncertainty in the probability distribution. Consider the traditional linear regression setting, where we have $n$ input-output samples $(\boldsymbol{x}_i, y_i)_{i=1}^n$ and we want to find a linear mapping between inputs $\boldsymbol{x}$ and outputs $y$, with squared error loss. The DRO formulation solves this problem by minimizing the expected loss with respect to a decision parameter $\boldsymbol{\beta}$ under the worst-case distribution:

$$\inf_{\boldsymbol{\beta}} \sup_{\mathbb{Q}\in\mathbb{B}_\delta(\widehat{\mathbb{P}}_n)} \mathbb{E}_{(\boldsymbol{x},y)\sim\mathbb{Q}}[(\boldsymbol{\beta}^\top \boldsymbol{x} - y)^2]. \tag{1}$$

Here, $\mathbb{B}_\delta(\widehat{\mathbb{P}}_n)$ is a ball of radius $\delta \geq 0$ around the empirical distribution $\widehat{\mathbb{P}}_n$ capturing the set of distributions, also known as the ambiguity set. DRO serves as a principled framework for capturing *distributional* uncertainty, being a natural tool for studying both generalization and out-of-distribution performance (Shafiee et al., 2025; Kuhn et al., 2025).

In this paper, we build on the broader relationship between robustness and regularization. Following the introduction of the regularization method Lasso (Tibshirani, 1996), several variants have been suggested to improve its statistical and practical properties. In particular, square-root Lasso was proposed as an alternative to Lasso that allows the regularization strength to be set without knowing the variance of the noise (Belloni et al., 2011), a property sometimes called *pivotal* estimation. The procedure has also been shown to be an instance of robust optimization (Xu et al., 2008a). More recently, with the growing interest in adversarial robustness, researchers found that adversarial linear regression has the same pivotal property (Ribeiro et al., 2023; Xie & Huo, 2024).

One interesting observation is that both square-root Lasso and adversarial linear regression can be unified as instances of DRO. Indeed, letting the ambiguity ball $\mathbb{B}_\delta(\widehat{\mathbb{P}}_n)$ be induced by the $p$-Wasserstein distance (an optimal transport distance), there are close connections to several regularizers (Kuhn et al., 2019). In particular, linear regression using the 2-Wasserstein distance is equivalent to square-root Lasso (Blanchet et al., 2019), while $\infty$-Wasserstein DRO corresponds to adversarially trained linear regression (see, e.g., Zhang et al. 2024, Theorem 2). This raise the natural question whether other methods in the $p$-Wasserstein DRO family share the same properties, such as being pivotal estimators.



Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

**Contributions.** In this paper, we study linear regression under square loss, where the ambiguity set is given by the $p$-Wasserstein distance with $2 \leq p \leq \infty$. For short, we call this method *Wasserstein DRO linear regression*. We show that many properties of the special cases square-root Lasso ($p = 2$) and adversarially trained linear regression ($p = \infty$) carry over to this more general method, including the pivotal property, slow and fast error rates, and equivalences in the small and large $\delta$ regimes. More precisely, our contributions are:

- We derive an equivalent form of the robust risk (the inner supremum of (1)) for Wasserstein DRO linear regression (Section 4).
- We prove deterministic and non-asymptotic in-sample error bounds $O(n^{-1/2})$ in general and $O(n^{-1})$ under design matrix and sparsity conditions, generalizing results that were previously only known for square-root Lasso and adversarially trained linear regression (Section 5).
- We characterize the solution for small and large radii $\delta$ in (1), respectively, where small values correspond to minimum norm interpolation in the overparametrized setting, and large values to the zero solution being optimal (Section 6).
- We provide tailored numerical solvers for Wasserstein DRO linear regression (Section 7) and conduct numerical experiments validating our findings (Section 8).

The paper is structured as follows. Section 2 covers related work. Section 3 presents preliminaries. Sections 4, 5 and 6 provide analytic properties we described in the contributions above, followed by numerical solvers (Section 7), simulations (Section 8) and conclusion (Section 9). Our implementation and experiments are available at `anonymous.4open.science/r/dro-BBDF/`.

## 2 Related work

DRO has a long history dating back to the pioneering work of Scarf et al. (1957), and has recently received substantial interest in the machine learning community as a principled way to account for distributional uncertainty in the data; see, e.g., the recent survey (Kuhn et al., 2025).

**Wasserstein DRO.** DRO with $p$-Wasserstein distance has been studied in both the classification and the regression settings, see the survey (Kuhn et al., 2019) and the references therein. In particular, Wasserstein DRO linear regression (under various loss functions) has mainly been studied for the special cases $p = 1$ and $p = 2$, see (Blanchet et al., 2019; Gao et al., 2022; Blanchet et al., 2021; Shafieezadeh-Abadeh et al., 2019; Kuhn et al., 2019; Aolaritei et al., 2026). Notably, Aolaritei et al. (2026) seek asymptotic statistical guarantees for a class of linear prediction problems in the high-dimensional regime with $p = 1$ and $p = 2$. We instead provide non-asymptotic statistical guarantees for Wasserstein DRO linear regression (with square loss) uniformly over the range $2 \leq p \leq \infty$.[1]

Wu et al. (2026) consider generalization bounds for a general learning objective, which could be applied to linear regression. However, their high-probability bound scales as $O(n^{-1/(2p)})$ with an improved ($p$-independent) slow rate of $O(n^{-1/2})$ that is not applicable to square loss (Wu et al., 2026, Assumption 1). There are also other generic high-probability generalization bounds that, e.g., assume Lipschitz loss (Shafieezadeh-Abadeh et al., 2019; An & Gao, 2021; Gao, 2023) or a compact domain (Le & Malick, 2025), neither of which holds for square loss. Additionally, these generic bounds (derived from concentration-based methods) typically have no explicit dependence on the noise. We instead derive *deterministic* error bounds with *explicit* dependence on the noise, inspired by Belloni et al. (2011); Ribeiro et al. (2023); Xie & Huo (2024) and Wainwright (2019, Chapter 7), that with high probability enjoy both slow rates $O(n^{-1/2})$ and fast rates $O(n^{-1})$ *uniformly* in $p$.

**Connection between DRO and regularization.** Several well-known regularization techniques in machine learning have found connections with DRO, such as square-root Lasso (Xu et al., 2008a; Blanchet et al., 2019) with higher-order generalizations (Montiel Olea et al., 2022), regularized logistic regression and support vector machines (Blanchet et al., 2019), ridge regression (Shafieezadeh-Abadeh et al., 2019; Li et al., 2022), adversarial training (Gao et al., 2022; Pydi & Jog, 2021), as well as close correspondences to total variation, Lipschitz-variation and gradient

[1] Here, $p < 2$ is excluded since the robust risk in (1) then becomes infinite for square loss. Thus, the natural range to consider is $2 \leq p \leq \infty$.

variation (Gao et al., 2022) as well as higher-order variation (Shafiee et al., 2026). In particular, square-root Lasso is equivalent to Wasserstein DRO linear regression with square loss and $p = 2$ (Blanchet et al., 2019), while the $p = \infty$ case coincides with adversarial linear regression, since adversarial training and $\infty$-Wasserstein DRO are equivalent (under mild conditions); see Zhang et al. (2024) for a general result. Our work builds upon this research direction by unifying these two instances as special cases of the more general Wasserstein DRO linear regression with square loss ($2 \leq p \leq \infty$), where *we extend fundamental properties such as error rates and the pivotal property to this unifying method.*

**Tractable reformulations.** There has been substantial work to make DRO tractable by reformulating the DRO objective as a finite-dimensional optimization problem, commonly via a dual representation of the inner supremum (Shafieezadeh-Abadeh et al., 2019; Shafiee et al., 2026), see Zhang et al. (2024) for a general result. In particular, using this dual representation, Shafiee et al. (2026, Proposition 2 and 3) and Kuhn et al. (2025, Theorem 7.20) both show that the DRO objective is equivalent to a finite convex program, under assumption that the loss is a (finite maximum of) concave function(s). However, such assumptions do not hold in our case due to the square loss.

The work Shafiee et al. (2026) also considers linear prediction models and shows that the dual representation then admits a lower-dimensional form, where an inner supremum (cf. Proposition 1) is recast into an expectation over a (possibly nonconvex) univariate maximization (see Shafiee et al. 2026, Theorem 4). In this paper, we note that for Wasserstein DRO linear regression (a special case of linear prediction models), this one-dimensional representation becomes unimodal, which we use in the numerical solvers. Furthermore, Kuhn et al. (2019) present convex forms for $p$-Wasserstein linear regression with square loss and $p = 2$ (and $p = 1$ for other loss functions), whereas we consider the whole range $2 \leq p \leq \infty$. Finally, we note that our dual form (Theorem 1) naturally generalize adversarial linear regression (Ribeiro et al., 2023, Proposition 1) by coupling the radius constraints, and also identifies a case in which robust optimization forms coincides with DRO (see, e.g., Gao & Kleywegt 2023, Corollary 2(iii) and An & Gao 2021, Proposition 2 for approximate results).

## 3 Background

In this section, we describe the setup and relevant results in DRO and adversarial linear regression.

**Wasserstein distance.** Let $\mathcal{P}(\mathcal{Z})$ denote the set of probability distributions with support $\mathcal{Z}$. One common choice of distance between such probability distributions is the *$p$-Wasserstein distance* $W_p$, where $1 \leq p \leq \infty$ is the order of this distance. More precisely, the $p$-Wasserstein distance $W_p$ with $1 \leq p < \infty$ is given by

$$W_p(\mathbb{P}, \widehat{\mathbb{P}}_n) := \left( \inf_{\pi \in \Pi(\mathbb{P}, \widehat{\mathbb{P}}_n)} \mathbb{E}_{(\boldsymbol{z}, \boldsymbol{z}') \sim \pi} [\|\boldsymbol{z} - \boldsymbol{z}'\|^p] \right)^{1/p}, \tag{2}$$

where $\Pi(\mathbb{P}, \widehat{\mathbb{P}}_n)$ is the set of all distributions in $\mathcal{P}(\mathcal{Z}) \times \mathcal{P}(\mathcal{Z})$ with marginals $\mathbb{P}$ and $\widehat{\mathbb{P}}_n$, respectively, and $\|\cdot\|$ is a norm. Intuitively, $\|\boldsymbol{z} - \boldsymbol{z}'\|^p$ in (2) captures the cost of moving probability mass from $\boldsymbol{z}$ to $\boldsymbol{z}'$, where the infimum picks the strategy $\pi$ with the least total cost for moving $\mathbb{P}$ to $\widehat{\mathbb{P}}_n$. The normalization with the $p$-th root is to ensure that $W_p$ is a metric (Kuhn et al., 2019). Moreover, for $p = \infty$, we have $W_\infty(\mathbb{P}, \widehat{\mathbb{P}}_n) := \inf_{\pi \in \Pi(\mathbb{P}, \widehat{\mathbb{P}}_n)} \operatorname{ess\,sup}_{(\boldsymbol{z}, \boldsymbol{z}') \sim \pi} [\|\boldsymbol{z} - \boldsymbol{z}'\|]$.

**Strong duality.** Consider the Wasserstein DRO problem

$$\inf_{\boldsymbol{\beta}} \sup_{\mathbb{Q} \in \mathbb{B}_\delta(\widehat{\mathbb{P}}_n)} \mathbb{E}_{\boldsymbol{z} \sim \mathbb{Q}}[h_{\boldsymbol{\beta}}(\boldsymbol{z})]$$

with objective $h_{\boldsymbol{\beta}} : \mathcal{Z} \to \mathbb{R}$ and ambiguity set $\mathbb{B}_\delta(\widehat{\mathbb{P}}_n) = \{\mathbb{Q} \in \mathcal{P}(\mathcal{Z}) : W_p(\mathbb{Q}, \widehat{\mathbb{P}}_n) \leq \delta\}$ induced by the $p$-Wasserstein distance centered around the empirical distribution $\widehat{\mathbb{P}}_n$, where the empirical distribution is $\widehat{\mathbb{P}}_n := \frac{1}{n} \sum_{i=1}^n \delta_{\boldsymbol{z}_i}$ for $\delta_{\boldsymbol{z}_i}$ the Dirac point mass at data sample $\boldsymbol{z}_i \in \mathcal{Z}$.

Here, we will denote by $V_\delta$ the inner supremum objective function. In particular, with the given definition of $\mathbb{B}_\delta(\widehat{\mathbb{P}}_n)$, we can write:

$$V_\delta(\boldsymbol{\beta}) := \sup_{\mathbb{Q}} \{\mathbb{E}_{\boldsymbol{z} \sim \mathbb{Q}}[h_{\boldsymbol{\beta}}(\boldsymbol{z})] : W_p(\mathbb{Q}, \widehat{\mathbb{P}}_n) \leq \delta\}, \tag{3}$$

which we will call the *robust risk*. Using strong duality, we have the following result:

**Proposition 1** (Strong duality)**.** Let $h_{\boldsymbol{\beta}} : \mathcal{Z} \to \mathbb{R}$ be a measurable function with $\mathbb{E}_{\widehat{\mathbb{P}}_n}[h_{\boldsymbol{\beta}}] > -\infty$. Then the robust risk in (3) equals:

$$V_\delta(\boldsymbol{\beta}) = \min_{\lambda \geq 0} \left( \lambda \delta^p + \mathbb{E}_{\boldsymbol{z} \sim \widehat{\mathbb{P}}_n} \Big[ \sup_{\boldsymbol{z}' \in \mathcal{Z}} \left( h_{\boldsymbol{\beta}}(\boldsymbol{z}') - \lambda \|\boldsymbol{z}' - \boldsymbol{z}\|^p \right) \Big] \right). \tag{4}$$

Proposition 1 is a straightforward application of the general result Zhang et al. (2024, Theorem 1) and the first step to our equivalent form for linear regression (Theorem 1) in Section 4.

**Wasserstein DRO linear regression.** We focus on $p$-Wasserstein DRO linear regression. That is, $\boldsymbol{z}_i = (\boldsymbol{x}_i, y_i) \in \mathcal{Z} = \mathbb{R}^{d+1}$ and $h_{\boldsymbol{\beta}}(\boldsymbol{z}) = (\boldsymbol{\beta}^\top \boldsymbol{x} - y)^2$. We restrict our attention to the norm $\|\boldsymbol{z}_i' - \boldsymbol{z}_i\| = \|\boldsymbol{x}_i' - \boldsymbol{x}_i\| + \infty |y_i' - y_i|$. Here, $\infty |y_i' - y_i|$ is shorthand for $+\infty$ if $y_i' \neq y_i$ and zero otherwise. These restrictions result in a close correspondence to square-root Lasso and adversarial linear regression.

*Remark* 1. We exclude $1 \leq p < 2$ since in this case the robust risk $V_\delta(\boldsymbol{\beta})$ becomes infinite. The reason is that the squared loss increases faster than the penalty for moving probability mass, resulting in an unbounded robust risk. See the appendix for a detailed derivation.

**Adversarial linear regression.** Our work also considers connections with adversarial linear regression. Adversarial linear regression considers linear regression where each input $\boldsymbol{x}_i \in \mathbb{R}^d$ is adversarially perturbed by $\Delta \boldsymbol{x}_i$ within a *fixed* budget $\|\Delta \boldsymbol{x}_i\| \leq \delta$ (uniform over all inputs) to form the worst-case risk:[2]

$$V_\delta^{\text{adv}}(\boldsymbol{\beta}) := \frac{1}{n} \sum_{i=1}^n \sup_{\|\Delta \boldsymbol{x}_i\| \leq \delta} ((\boldsymbol{x}_i + \Delta \boldsymbol{x}_i)^\top \boldsymbol{\beta} - y_i)^2 = \frac{1}{n} \sum_{i=1}^n (|\boldsymbol{x}_i^\top \boldsymbol{\beta} - y_i| + \delta \|\boldsymbol{\beta}\|_*)^2, \tag{5}$$

where the last equality follows by rewriting the supremum inside the sum using a simple dual norm argument; see, e.g., Proposition 1 in Ribeiro et al. (2023). In Section 4, we generalize this form to Wasserstein DRO linear regression (Theorem 1), a key step in establishing the analytical properties.

## 4 Equivalent form

In this section, we present an equivalent form of the robust risk $V_\delta(\boldsymbol{\beta})$, called the *robust quadratic form*. All analytic properties of Wasserstein DRO linear regression shown in subsequent sections are based on this form.

**Theorem 1** (Robust quadratic form)**.** The robust risk of Wasserstein DRO linear regression ($2 \leq p \leq \infty$) equals

$$V_\delta(\boldsymbol{\beta}) = \sup_{t_i \geq 0,\ \|\boldsymbol{t}\|_p \leq n^{1/p} \delta} \frac{1}{n} \sum_{i=1}^n (|\boldsymbol{x}_i^\top \boldsymbol{\beta} - y_i| + t_i \|\boldsymbol{\beta}\|_*)^2. \tag{6}$$

Crucially, the right-hand side of (6) can naturally be interpreted as an adversarial linear regression with a *variable* perturbation budget for each sample. Indeed, for $p = \infty$, we have that (6) coincides with (5) since the optimal outer constraint becomes $\|\boldsymbol{t}\|_\infty = \delta$ (optimum is attained on the boundary) yielding $\|\Delta \boldsymbol{x}_i\| \leq \delta$, i.e., a *fixed* perturbation budget $\delta$ for each input sample $\boldsymbol{x}_i$. In the general $2 \leq p \leq \infty$ case, these perturbation budgets $t_i$ can *vary* between samples, as long as the overall budget constraint $\|\boldsymbol{t}\|_p \leq n^{1/p} \delta$ is fulfilled. Thus, (6) *naturally generalizes the objective of adversarial linear regression*. This is the key step to show analytical properties, since it allows us to extend proofs from the adversarial setting.

The proof of Theorem 1 starts by recasting the supremum in (4) as a one-dimensional concave supremum from which (6) is shown using a change of variables and strong duality; see the appendix for details. Finally, Theorem 1 yields a simple proof not only that $p = \infty$ coincides with adversarial linear regression, but also that $p = 2$ corresponds to square-root Lasso (Xu et al., 2008a; Blanchet et al., 2019), by simply computing the robust risk. We summarize these findings with the following corollary (using $V_{p,\delta}(\boldsymbol{\beta})$ to temporarily stress dependence on $p$):

[2]Note that with no budget $\delta = 0$, this becomes traditional linear regression with square loss.

**Corollary 1** (Special cases)**.** Wasserstein DRO linear regression with $p = 2$ has robust risk $V_{2,\delta}(\boldsymbol{\beta}) = \left(\sqrt{\frac{1}{n}\sum_{i=1}^n (\boldsymbol{x}_i^T\boldsymbol{\beta} - y_i)^2} + \delta\|\boldsymbol{\beta}\|_*\right)^2$, and thus corresponds to square-root Lasso by taking $\|\cdot\| = \|\cdot\|_\infty$, while $p = \infty$ corresponds to adversarial linear regression with the continuous limit $V_{\infty,\delta}(\boldsymbol{\beta}) = \lim_{p\to\infty} V_{p,\delta}(\boldsymbol{\beta}) = \sum_{i=1}^n \sup_{\|\boldsymbol{v}_i\|\le\delta}((\boldsymbol{x}_i + \boldsymbol{v}_i)^T\boldsymbol{\beta} - y_i)^2$.

*Remark* 2. Weaker versions of Theorem 1 are described in other papers. The right-hand side of (6) is shown to lower-bound the robust risk (Kuhn et al., 2019, Theorem 6), but that work does not explicitly show when equality holds. For $p = \infty$, Zhang et al. (2024, Theorem 2) proves this result in general, while Kuhn et al. (2025, Proposition 6.16) shows that equality holds and a saddle-point exists, but only for compact sets. We show that equality in fact holds for all $2 \le p \le \infty$ for linear regression. See also Gao & Kleywegt (2023, Corollary 2(iii)) and An & Gao (2021, Proposition 2) for results that only hold approximately.

## 5 Error bounds

In this section, we derive error bounds for Wasserstein DRO linear regression (with $2 \le p \le \infty$). We consider the following data-generating model $y_i = \boldsymbol{x}_i^\top\boldsymbol{\beta}^* + \varepsilon_i$ with true parameter $\boldsymbol{\beta}^* \in \mathbb{R}^d$ and noise $\varepsilon_i \in \mathbb{R}$. More precisely, the quantity of interest is the average in-sample error $\frac{1}{n}\|X\widehat{\Delta}\|_2^2$, where $X$ is the $n \times d$ data matrix with row $i$ given by $\boldsymbol{x}_i$, and $\widehat{\Delta} := \widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*$ is the difference between the estimated parameter $\widehat{\boldsymbol{\beta}}$ and the true parameter $\boldsymbol{\beta}^*$, where $\widehat{\boldsymbol{\beta}}$ is the minimizer of $V_\delta(\boldsymbol{\beta})$. The aim is to upper bound $\frac{1}{n}\|X\widehat{\Delta}\|_2^2$ as a function of $\delta$, which we can then use to find good decay rates by tuning $\delta = \delta(n)$ appropriately.[3]

We derive a slow error decay rate $O(1/\sqrt{n})$ (Theorem 2) without any additional assumptions on the data. In Section 5.2, we then improve this to a fast rate $O(1/n)$ (Theorem 3) by assuming a restricted eigenvalue condition and focusing on the infinity norm $\|\cdot\| = \|\cdot\|_\infty$. Moreover, for both the slow rate and the fast rate, the pivotal property (that is, tuning $\delta = \delta(n)$ to achieve the rate is independent of the noise level) comes out for free.

### 5.1 Slow rate of $O(1/\sqrt{n})$

To show the desired error rate, we first show the following bound on $\frac{1}{n}\|X\widehat{\Delta}\|_2^2$:

**Lemma 1.** The average in-sample error is bounded by

$$\frac{1}{n}\|X\widehat{\Delta}\|_2^2 \le \frac{2}{n}\boldsymbol{\varepsilon}^\top X\widehat{\Delta} + 2\delta\frac{\|\boldsymbol{\varepsilon}\|_q}{n^{1/q}}\|\boldsymbol{\beta}^*\|_* + \delta^2\|\boldsymbol{\beta}^*\|_*^2. \tag{7}$$

Lemma 1 is a first step toward obtaining sharp in-sample error upper bounds, where the important quantity to bound is $\frac{2}{n}\boldsymbol{\varepsilon}^\top X\widehat{\Delta}$. Crucially, $\frac{2}{n}\boldsymbol{\varepsilon}^\top X\widehat{\Delta} \le \frac{2}{n}\|X^\top\boldsymbol{\varepsilon}\|\|\widehat{\Delta}\|_*$. Thus, provided $\|\widehat{\Delta}\|_*$ is bounded (true if $\|\widehat{\boldsymbol{\beta}}\|_*$ is bounded) and $\frac{2}{n}\|X^\top\boldsymbol{\varepsilon}\|$ decreases with $n$ (typical for, e.g., Gaussian noise $\boldsymbol{\varepsilon}$), we may obtain decreasing upper bounds as long as $\delta$ decreases with $n$. The next result shows that $\|\widehat{\boldsymbol{\beta}}\|_*$ can be bounded by $\|\boldsymbol{\beta}^*\|_*$:

**Lemma 2.** If $\delta > 2\frac{\|X^\top\boldsymbol{\varepsilon}\|}{n^{1/p}\|\boldsymbol{\varepsilon}\|_q}$, then

$$\|\widehat{\boldsymbol{\beta}}\|_* \le \left(3 + \frac{n^{1/q}\delta}{\|\boldsymbol{\varepsilon}\|_q}\|\boldsymbol{\beta}^*\|_*\right)\|\boldsymbol{\beta}^*\|_* \tag{8}$$

Importantly, Lemma 2 combined with Lemma 1 enables us to prove a slow rate $O(n^{-1/2})$, with the slow rate of square-root Lasso ($p = 2$) and adversarial training ($p = \infty$) as special cases:

[3]The quantity $\frac{1}{n}\|X\widehat{\Delta}\|_2^2$ is also called the excess risk because it equals the difference between the expected risk $\mathbb{E}_\varepsilon[\frac{1}{n}\sum_{i=1}^n (y_i - \boldsymbol{x}_i^\top\widehat{\boldsymbol{\beta}})^2]$ and the Bayes optimal risk $\mathbb{E}_\varepsilon[\frac{1}{n}\sum_{i=1}^n (y_i - \boldsymbol{x}_i^\top\boldsymbol{\beta}^*)^2]$ (Bach (2024, Prop. 3.3)).

**Theorem 2.** Assume $\delta > \bar{\delta} := 2\frac{\|X^\top \boldsymbol{\varepsilon}\|}{n^{1/p}\|\boldsymbol{\varepsilon}\|_q}$. Then

$$\frac{1}{n}\|X\widehat{\Delta}\|_2^2 \leq 6\frac{\|\boldsymbol{\varepsilon}\|_q}{n^{1/q}}\delta\|\boldsymbol{\beta}^*\|_* + 2\delta^2\|\boldsymbol{\beta}^*\|_*^2. \tag{9}$$

In particular, let $\boldsymbol{\varepsilon} \sim N(0, \sigma^2 I_n)$ and assume each entry of $X$ is bounded by $M > 0$. Then, with high probability (independent of the Wasserstein parameter $p$):

$$\frac{1}{n}\|X\widehat{\Delta}\|_2^2 \in O(\delta).$$

That is, the error rate coincides with $\delta$ (as long as $\delta > \bar{\delta}$). In particular, let $c > 0$ be the norm-equivalence constant between $\|\cdot\|_\infty$ and $\|\cdot\|$, that is, $\|\cdot\|_\infty \leq c\|\cdot\|$. Then, by setting $\delta = KM\sqrt{\frac{\log(d/\gamma)}{n}}$ with $K > 2c\sqrt{\pi}$, we get with probability greater than $1 - 4\gamma$:

$$\frac{1}{n}\|X\widehat{\Delta}\|_2^2 \in O(n^{-1/2}). \tag{10}$$

Theorem 2 shows that $p$-Wasserstein linear regression can achieve the rate $O(1/\sqrt{n})$, where we recover the previously obtained rates for square-root Lasso (Belloni et al., 2011) ($p = 2$) and adversarial linear regression (Ribeiro et al., 2023) ($p = \infty$).

**Pivotal property.** The tuning $\delta = KM\sqrt{\frac{\log(d/\gamma)}{n}}$ in Theorem 2 is independent of the noise level $\sigma$ of $\boldsymbol{\varepsilon} \sim N(0, \sigma^2 I)$; hence, *we do not need to estimate* $\sigma$. This independence is known as the *pivotal property*. Previously, both square-root Lasso ($p = 2$) and adversarial linear regression ($p = \infty$) were known to be pivotal (Belloni et al., 2011; Ribeiro et al., 2023). Here, we show that the same holds for the more general $p$-Wasserstein linear regression with $2 \leq p \leq \infty$, suggesting that the pivotal property is tightly linked to Wasserstein robustification.

*Sketch of proof of Theorem 2.* We bound $\frac{2}{n}\boldsymbol{\varepsilon}^\top X\widehat{\Delta}$ as $\frac{2}{n}\boldsymbol{\varepsilon}^\top X\widehat{\Delta} \leq \frac{2}{n}\|X^\top\boldsymbol{\varepsilon}\|\|\widehat{\Delta}\|_* \leq \frac{2}{n}\|X^\top\boldsymbol{\varepsilon}\|(\|\widehat{\boldsymbol{\beta}}\|_* + \|\boldsymbol{\beta}^*\|_*)$ and then use Lemma 2 with $\delta n^{1/p}\|\varepsilon\|_q > 2\|X^\top\boldsymbol{\varepsilon}\|$ to get $\frac{2}{n}\boldsymbol{\varepsilon}^\top X\widehat{\Delta} \leq 4\frac{\|\boldsymbol{\varepsilon}\|_q}{n^{1/q}}\delta\|\boldsymbol{\beta}^*\|_* + \delta^2\|\boldsymbol{\beta}^*\|_*^2$, which, when inserted into Lemma 1, yields (9). Moreover, the rates are obtained by standard high-probability bounds for Gaussians. For a full proof, see the appendix. □

### 5.2 Fast rate of $O(1/n)$ under the restricted eigenvalue condition

In this section, we show the fast in-sample error rate $O(1/n)$ by assuming a restricted eigenvalue condition and focusing our attention to the infinity norm $\|\cdot\| = \|\cdot\|_\infty$. The restricted eigenvalue condition reads (Hastie et al., 2015):

**Definition 1** (Restricted eigenvalue condition (Hastie et al., 2015))**.** The matrix $X \in \mathbb{R}^{n\times d}$ satisfies the restricted eigenvalue (RE) condition if there exists a positive constant $\kappa = \kappa(s, l) > 0$ such that

$$\kappa \leq \min\left\{\frac{\|X\boldsymbol{v}\|_2}{\sqrt{n}\|\boldsymbol{v}\|_2} : |S| \leq s, \boldsymbol{v} \in \mathbb{R}^d\backslash\{0\}, \|\boldsymbol{v}_{S^c}\| \leq l\|\boldsymbol{v}_S\|\right\}$$

where $S$ is some subset of $\{1, 2, \ldots, d\}$.

Let $\mathrm{RE}(s, l)$ denote the set of matrices $X$ that satisfy the RE condition. We have the fast rate result:

**Theorem 3.** Let $\boldsymbol{\beta}^*$ have $s$ nonzero entries indexed by $S \subseteq \{1, \ldots, d\}$ with $s = |S|$, assume $X$ is in $\mathrm{RE}(s, 3)$ with positive constant $\kappa = \kappa(s, 3)$ (same for all $n$), and consider $\|\cdot\| = \|\cdot\|_\infty$. Then with $\delta > \bar{\delta} := 2\frac{\|X^\top\boldsymbol{\varepsilon}\|}{n^{1/p}\|\boldsymbol{\varepsilon}\|_q}$, we have

$$\frac{1}{n}\|X\widehat{\Delta}\|_2^2 \leq \delta^2 \cdot \max\left\{4C^2\|\boldsymbol{\beta}^*\|_1^2, \frac{16B^2s^2}{\kappa^2}\right\} \tag{11}$$

where $B = 3\frac{\|\boldsymbol{\epsilon}\|_q}{n^{1/q}} + (2MC + \delta C + \delta)\|\boldsymbol{\beta}^*\|_*$ and $C = (3 + \frac{n^{1/q}\delta}{\|\boldsymbol{\epsilon}\|_q}\|\boldsymbol{\beta}^*\|_*)$ and each entry of $X$ is bounded by $M > 0$. In particular, for $\boldsymbol{\varepsilon} \sim N(0, \sigma^2 I_n)$, we have with high probability that $B$ and $C$ are bounded by constants and therefore (independent of the Wasserstein parameter $p$)

$$\frac{1}{n}\|X\widehat{\Delta}\|_2^2 \in O\left(\delta^2\right). \tag{12}$$

Thus, setting $\delta$ as in Theorem 2, we get the fast rate with probability greater than $1 - 5\gamma$:

$$\frac{1}{n}\|X\widehat{\Delta}\|_2^2 \in O\left(1/n\right). \tag{13}$$

Theorem 3 shows that the general $p$-Wasserstein linear regression ($2 \leq p \leq \infty$) can achieve the rate $O(1/n)$, similar to previous results obtained for square-root Lasso (Belloni et al., 2011) ($p = 2$) and for adversarial linear regression (Xie & Huo, 2024) ($p = \infty$). Moreover, just as in the slow rate, the tuning scheme $\delta = KM\sqrt{\frac{\log(d/\gamma)}{n}}$ makes the method pivotal.

## 6 Small and large ambiguity balls

Here, we characterize the minimizer $\widehat{\boldsymbol{\beta}}$ of $V_\delta(\boldsymbol{\beta})$ when the radius $\delta$ of the ambiguity set is small and large, respectively. These results generalize conditions from adversarial linear regression (Ribeiro et al., 2023, Theorem 1 and Proposition 3) to the general $p$-Wasserstein linear regression ($2 \leq p \leq \infty$). Importantly, these results also apply to square-root Lasso ($p = 2$), which is novel to the best of our knowledge. We start with the large regime.

**Large $\delta$ regime.** The first result shows that the zero solution $\boldsymbol{\beta} = 0$ minimizes $V_\delta(\boldsymbol{\beta})$ for a sufficiently large radius $\delta \geq \delta_L$. In other words, for large ambiguity sets, the optimal solution is to be maximally conservative.

**Theorem 4** (Large $\delta$ regime)**.** The zero solution $\boldsymbol{\beta} = 0$ minimizes $V_\delta(\boldsymbol{\beta})$ if and only if

$$\delta \geq \delta_L := \frac{\|X^\top \boldsymbol{y}\|}{n^{1/p}\|\boldsymbol{y}\|_q}.$$

**Small $\delta$ regime.** For the small $\delta$ regime, we follow Ribeiro et al. (2023) and consider the over-parametrized setting where we assume $X \in \mathbb{R}^{n\times d}$ to have full row rank and $n < d$. In this setting, there are multiple solutions to the original least-squares problem $X\boldsymbol{\beta} = y$, where a natural solution is to take the minimal $\|\cdot\|_N$-norm interpolator $\bar{\boldsymbol{\beta}}$ that solves $\min_{X\boldsymbol{\beta}=y}\|\boldsymbol{\beta}\|_N$ for a given norm $\|\cdot\|_N$. The next result shows that for a small enough radius $\delta \leq \delta_S$, the minimal $\|\cdot\|_*$-norm interpolator minimizes $V_\delta(\boldsymbol{\beta})$:

**Theorem 5** (Small $\delta$ regime)**.** Let $X$ be of full row rank and $Q = \arg\max_{\|X^T\boldsymbol{\alpha}\|\leq 1}\boldsymbol{\alpha}^\top \boldsymbol{y}$. Then the minimum $\|\cdot\|_*$-norm interpolator $\bar{\boldsymbol{\beta}}$ minimizes $V_\delta(\boldsymbol{\beta})$ if and only if

$$\delta \leq \delta_S := \begin{cases} \frac{1}{n\min_{\boldsymbol{\alpha}\in Q}\|\boldsymbol{\alpha}\|_\infty} & \text{for } 2 < p \leq \infty \\ \frac{1}{\sqrt{n}\min_{\boldsymbol{\alpha}\in Q}\|\boldsymbol{\alpha}\|_2} & \text{for } p = 2 \end{cases}$$

Note first that the set $\{\boldsymbol{\alpha} : \|X^\top\boldsymbol{\alpha}\| \leq 1\}$ is compact, and hence both $\min_{\boldsymbol{\alpha}\in Q}\|\boldsymbol{\alpha}\|_\infty$ and $\min_{\boldsymbol{\alpha}\in Q}\|\boldsymbol{\alpha}\|_2$ are finite. Thus, *interpolation always occurs* for small enough $\delta$. Moreover, the interpolation threshold $\delta_S$ *is the same for all* $2 < p \leq \infty$ while it is *higher* for $p = 2$:

$$\frac{1}{\sqrt{n}\min_{\boldsymbol{\alpha}\in Q}\|\boldsymbol{\alpha}\|_2} \geq \frac{1}{n\min_{\boldsymbol{\alpha}\in Q}\|\boldsymbol{\alpha}\|_\infty},$$

by the norm inequality $\|\boldsymbol{\alpha}\|_2 \leq \sqrt{n}\|\boldsymbol{\alpha}\|_\infty$. In fact, this inequality is typically strict since the norm inequality is typically strict (see Remark 3 in the appendix: for $\|\cdot\| = \|\cdot\|_2$, the set $Q$ is a singleton $\bar{\boldsymbol{\alpha}}$, and varying $X$ and $\boldsymbol{y}$ generally gives different $\infty$-norms and 2-norms). That is, *interpolation occurs sooner* for square-root Lasso ($p = 2$) than for $2 < p \leq \infty$ (e.g., adversarial linear regression). We also stress that $\delta_S$ is easy to compute via a convex program, see Remark 3 in the appendix.

## 7 NUMERICAL SOLVERS

We provide two solvers tailored specific for Wasserstein DRO linear regression. The first solver, called the *saddle-point solver*, considers a novel saddle-point formulation of the robust risk $\min_{\boldsymbol{\beta}} nV_\delta(\boldsymbol{\beta}) = \min_{\boldsymbol{\beta}} \sup_{\boldsymbol{\gamma}\geq 0} H(\boldsymbol{\beta}, \boldsymbol{\gamma})$ with convex-concave function $H(\boldsymbol{\beta}, \boldsymbol{\gamma})$ on the form:

$$H(\boldsymbol{\beta}, \boldsymbol{\gamma}) := n^{1/p}\delta\|\boldsymbol{\beta}\|_*\|\boldsymbol{\gamma}\|_1^{1/q} + \sum_{i=1}^{n} \gamma_i^{1/q}|\boldsymbol{x}_i^T\boldsymbol{\beta} - y_i| - \gamma_i^{2/q}/4.$$

See appendix for a derivation. We solve this saddle problem using disciplined saddle programming (Schiele et al., 2024), an extension of disciplined convex programming (Diamond & Boyd, 2016).

The second solver, called the *$\eta$-trick solver*, uses what is sometimes referred to as the $\eta$-trick (Bach, 2019) to iteratively alternative between solving a weighted ridge regression problem and updating the weights in closed-form. More precisely, we fix $\|\cdot\| = \|\cdot\|_\infty$ and rewrite $\|\boldsymbol{\beta}\|_* = \|\boldsymbol{\beta}\|_1$ in (6) as a minimum of a weighted quadratic form, with weights $\boldsymbol{\eta}$. This becomes $V_\delta(\boldsymbol{\beta}) = \inf_{\boldsymbol{\eta}} \Phi(\boldsymbol{\beta}, \boldsymbol{t}^\star, \boldsymbol{\eta})$ for a function $\Phi$ that is jointly convex in $(\boldsymbol{\beta}, \boldsymbol{\eta})$, and $\boldsymbol{t}^\star = \boldsymbol{t}^\star(\boldsymbol{\beta})$ attains the supremum in (6). We sequentially update these three variables repeatedly until convergence:

1. Given $\boldsymbol{\beta}$, we update $\boldsymbol{t}^\star$;
2. Given $\boldsymbol{t}^\star$ and $\boldsymbol{\beta}$, we compute the next $\boldsymbol{\eta}$;
3. Given new $\boldsymbol{t}^\star, \boldsymbol{\eta}$, we compute the next $\boldsymbol{\beta}$, and so on.

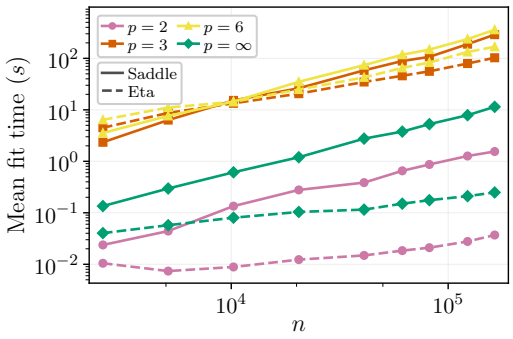



Figure 1: Mean computation times over 10 repetitions for the saddle-point solver (solid) and the $\eta$-trick solver (dashed) in the fast-rate experiment.

Crucially, all these subproblems can be solved fast: the $t^\star$-update can be solved by a scalar formulation, the $\boldsymbol{\eta}$-update has a closed-form solution and the $\boldsymbol{\beta}$-update is weighted ridge regression. The idea follows closely (Ribeiro et al., 2025b) proposed for adversarial linear training. We provide details about implementation and convergence in the appendix. The result is the improved speed we show in the numerical experiments.

## 8 NUMERICAL EXPERIMENTS

This section showcases our findings through simulations. We start with the fast rate $O(n^{-1})$ and the slow rate $O(n^{-1/2})$. Then, we illustrate the small and large $\delta$ regimes. We fix $\|\cdot\| = \|\cdot\|_\infty$, and let $p$ vary as $p \in \{2, 3, 6, \infty\}$. For the rates, we use 10 trials for each $(n, p)$ with $\pm 1$ standard deviation bands (although, so tight they are barely visible). All simulations were run on a MacBook Air M3 in the matter of hours. For Figure 2 and Figure 3, we used the saddle-point solver (see Section 7). In Figure 1 we compare the performance of the solvers. See the appendix for additional simulations.

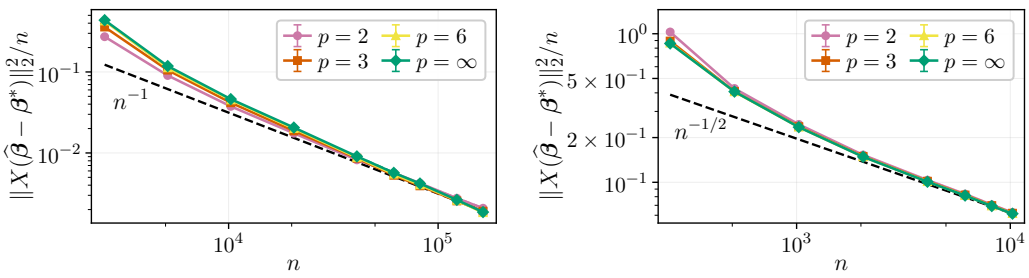



Figure 2: Fast rates (left) and slow rates (right) for $p \in \{2, 3, 6, \infty\}$. Both show a clear convergence toward the predicted rates $O(n^{-1})$ and $O(n^{-1/2})$, respectively. The curves are mean over 10 runs with $\pm 1$ standard deviation bands.

**Fast rate.** For the fast rate, we consider a simple setup $y_i = \boldsymbol{x}_i^\top \boldsymbol{\beta}^* + \varepsilon_i$ with $d = 10$, true parameter with $\beta_i^* = 3$ for the first 5 entries and otherwise zero (i.e., $s = 5$), Gaussian noise $\boldsymbol{\varepsilon} \sim N(0, \sigma^2 I)$ with $\sigma = 0.5$, and entries $x_{ij} \in X$ independently and uniformly sampled from $[-1, 1]$ (so $M = 1$ in Theorem 3). Here, the *independent* sampling of $X$ ensures that $X \in \mathrm{RE}(s, 3)$ in Theorem 3; see the appendix for a standard derivation. We compute $\widehat{\boldsymbol{\beta}} \in \arg\min_\beta V_\delta(\boldsymbol{\beta})$ and plot the in-sample error $\|X(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)\|_2^2/n$ as a function of $n$, shown in Figure 2 (left). We clearly see the rate $O(n^{-1})$ as predicted by Theorem 3. This rate hinges on the fact that $X \in \mathrm{RE}(s, 3)$ due to the independent sampling. In contrast, if the entries in $X$ are *correlated*, they may violate the RE condition and instead yield an $O(n^{-1/2})$ rate, as the next example illustrates.

**Slow rate.** For the slow rate, we consider $d = 2$, $\boldsymbol{\beta}^* = [2, -1]^\top$, noise $\boldsymbol{\varepsilon} \sim N(0, \sigma^2 I)$ with $\sigma = 0.5$, and entries of $X$ still uniformly bounded but now *correlated*. This correlation breaks the RE condition (see the appendix for details on $X$). Thus, we can only guarantee an $O(n^{-1/2})$ rate by Theorem 2, also seen in Figure 2 (right).

**Small and large $\boldsymbol{\delta}$.** We conclude with the extreme regimes of $\delta$. We see that for large $\delta \geq \delta_L$ in Figure 3, $\widehat{\boldsymbol{\beta}}$ becomes the zero solution as in Theorem 4, while small $\delta \leq \delta_S$ in the overparametrized setting results in interpolation as in Theorem 5. See the appendix for setup details.

**Saddle-point and $\eta$-trick solver.** In the fast-rate experiment, both solvers achieve nearly identical prediction errors and robust-risk values; see the appendix for the full comparison. In Figure 1, we see that the $\eta$-trick solver is faster at larger sample sizes, with the largest gains for $p = 2$ and $p = \infty$.

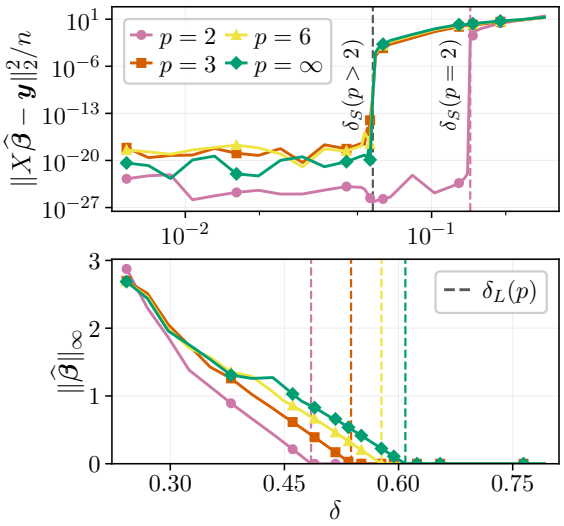



Figure 3: Large $\delta$ (bottom) and small $\delta$ (top) in the overparametrized regime, with zero solution and interpolation for $\delta \geq \delta_L$ and $\delta \leq \delta_S$, respectively.

## 9 Conclusion

In this paper, we have studied Wasserstein DRO linear regression with $2 \leq p \leq \infty$. We first proved an equivalent robust quadratic form of the robust risk. Using this form, we generalized several important properties from the two special cases square-root Lasso ($p = 2$) and adversarial linear regression ($p = \infty$) to the general $2 \leq p \leq \infty$ setting, including a slow error rate $O(n^{-1/2})$ in general, a fast rate $O(n^{-1})$ using a restricted eigenvalue condition, the pivotal property, and equivalent solutions when the ambiguity radius $\delta$ is small and large, respectively. We also proposed efficient numerical solvers and validated our theoretical findings via simulations.

Here, we only study Wasserstein DRO linear regression. An interesting future direction is to investigate conditions under which the analysis in this paper applies to other loss functions or to nonparametric regression problems. For instance, some of the adversarial techniques we used have already been adapted to kernel methods (Ribeiro et al., 2025a). Another future direction is to consider overparametrized regimes where one lets both $d$ and $n$ tend to infinity, under some fixed finite ratio $\rho := d/n$. The analytic properties in this paper do not explicitly depend on $d$ (except for Theorem 3, where $s = d$ if no sparse solution exists), and may therefore naturally serve as a start. Finally, the deterministic error bounds in this paper depend only very mildly on $p$, with error rates independent of $p$. This is also reflected in the numerical simulations. It would be interesting if one could find cases studies in which different $p$ results in significantly different error (which could still be possible since we only provide upper bounds on the error).

**AI use statement.** In this work, we used generative AI tools to implement methods. We have also partially used generative AI tools to provide critical ingredients for proving mathematical claims. We have not used generative AI tools to help develop theoretical models or conceptual frameworks, formulate mathematical claims, assist in the writing of proofs and interpret results. We have also used AI tools to identify relevant literature (beyond the manual literature survey), search for information and improve readability. We have reviewed all AI-assisted work. In particular, the paper including all proofs are fully written and checked by the authors and the LLM-generated code is verified and tested for correctness.

**Reproducibility statement.** We provide the source code, model implementations and evaluation procedures in the anonymous repository linked (see the link on page 2).

## REFERENCES


Yang An and Rui Gao. Generalization bounds for (wasserstein) robust optimization. *Advances in Neural Information Processing Systems*, 34:10382–10392, 2021.

Liviu Aolaritei, Soroosh Shafiee, and Florian Dörfler. Wasserstein distributionally robust estimation in high dimensions: performance analysis and optimal hyperparameter tuning *Mathematical Programming*, pp. 1–85, 2026.

Francis Bach. The "$\eta$-trick" or the effectiveness of reweighted least-squares – Machine Learning Research Blog, 2019.

Francis Bach. *Learning Theory from First Principles* MIT Press, 2024.

Alexandre Belloni, Victor Chernozhukov, and Lie Wang. Square-Root Lasso: Pivotal Recovery of Sparse Signals via Conic Programming. *Biometrika*, 98(4):791–806, 2011.

Dimitri P Bertsekas. *Control of uncertain systems with a set-membership description of the uncertainty.* PhD thesis, Massachusetts Institute of Technology, 1971.

Dimitri P Bertsekas. *Convex optimization theory*, volume 1. Athena Scientific Belmont, 2009.

Dimitri P Bertsekas, Angelia Nedi, and Asuman E Ozdaglar. *Convex Analysis and Optimization*. 2003.

Jose Blanchet, Yang Kang, and Karthyek Murthy. Robust wasserstein profile inference and applications to machine learning. *Journal of Applied Probability*, 56(3):830–857, 2019.

Jose Blanchet, Karthyek Murthy, and Viet Anh Nguyen. Statistical analysis of wasserstein distributionally robust estimators. In *Tutorials in Operations Research: Emerging optimization methods and modeling techniques with applications*, pp. 227–254. INFORMS, 2021.

Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford University Press, 2013.

Stephen Boyd, John Duchi, Mert Pilanci, and Lieven Vandenberghe. Subgradients (Lecture Notes). Technical Report Notes for EE364b, Stanford University, 2022.

Stephen Boyd and Lieven Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.

Frank H. Clarke. *Optimization and Nonsmooth Analysis*. Society for Industrial and Applied Mathematics, 1990.

John M Danskin. *The Theory of Max-min and its Application to Weapons Allocation Problems*. Springer Berlin, Heidelberg, 1967.

Steven Diamond and Stephen Boyd. CVXPY: A Python-embedded modeling language for convex optimization. *Journal of Machine Learning Research*, 17(83):1–5, 2016.

Rui Gao. Finite-sample guarantees for Wasserstein distributionally robust optimization: Breaking the curse of dimensionality. *Operations Research*, 71(6):2291–2306, 2023.

Rui Gao and Anton J. Kleywegt. Distributionally robust stochastic optimization with wasserstein distance. *Mathematics of Operations Research*, 48(2):603–655, 2023.

Rui Gao, Xi Chen, and Anton J. Kleywegt. Wasserstein distributionally robust optimization and variation regularization. *Operations Research*, 72(3):1177–1191, 2022.

Trevor Hastie, Robert Tibshirani, and Martin J. Wainwright. *Statistical Learning with Sparsity: The Lasso and Generalizations*. Chapman and Hall/CRC, 2015.

Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. *An introduction to statistical learning: with applications in R*, volume 2. Springer, 2021.

Hidetoshi Komiya. Elementary proof for Sion's minimax theorem. *Kodai Mathematical Journal*, 11(1):5–7, 1988.

Daniel Kuhn, Peyman Mohajerin Esfahani, Viet Anh Nguyen, and Soroosh Shafieezadeh-Abadeh. Wasserstein Distributionally Robust Optimization: Theory and Applications in Machine Learning. *Tutorials in Operations Research*, pp. 130–166, October 2019.

Daniel Kuhn, Soroosh Shafiee, and Wolfram Wiesemann. Distributionally robust optimization. *Acta Numerica*, 34:579–804, 2025.

Tor Lattimore and Csaba Szepesvári. *Bandit Algorithms*. Cambridge University Press, 2020.

Tam Le and Jérôme Malick. Universal generalization guarantees for wasserstein distributionally robust models. In *International Conference on Learning Representations*, volume 2025, pp. 20887–20921, 2025.

Jiajin Li, Sirui Lin, Jose Blanchet, and Viet Anh Nguyen. Tikhonov regularization is optimal transport robust under martingale constraints. *Advances in Neural Information Processing Systems*, 35:17677–17689, 2022.

José Luis Montiel Olea, Cynthia Rush, Amilcar Velez, and Johannes Wiesel. The out-of-sample prediction error of the square-root-lasso and related estimators. *arXiv preprint arXiv:2211.07608*, 2022.

Muni Sreenivas Pydi and Varun Jog. The many faces of adversarial risk. *Advances in Neural Information Processing Systems*, 34:10000–10012, 2021.

Antônio H. Ribeiro, David Vävinggren, Dave Zachariah, Thomas B. Schön, and Francis Bach. Kernel Learning with Adversarial Features: Numerical Efficiency and Adaptive Regularization. *Advances in Neural Information Processing Systems (NeurIPS)*, December 2025.

Antônio H. Ribeiro, Dave Zachariah, Francis Bach, and Thomas B. Schön. Regularization properties of adversarially-trained linear regression. *Advances in Neural Information Processing Systems*, 36:23658–23670, 2023.

Antônio H. Ribeiro, Thomas B. Schön, Dave Zachariah, and Francis Bach. Efficient optimization algorithms for linear adversarial training. In *International Conference on Artificial Intelligence and Statistics*, 2025b.

Herbert E Scarf, KJ Arrow, and S Karlin. A min-max solution of an inventory problem. Technical report, Rand Corporation Santa Monica, 1957.

Philipp Schiele, Eric Sager Luxenberg, and Stephen Boyd. Disciplined saddle programming. *Transactions on Machine Learning Research*, 2024.

Soroosh Shafiee, Liviu Aolaritei, Florian Dörfler, and Daniel Kuhn. Nash Equilibria, Regularization and Computation in Optimal Transport-Based Distributionally Robust Optimization, June 2025. arXiv:2303.03900.

Soroosh Shafiee, Liviu Aolaritei, Florian Dörfler, and Daniel Kuhn. Nash equilibria, regularization, and computation in optimal transport-based distributionally robust optimization. *Operations Research*, 74(3):1689–1709, 2026.

Soroosh Shafieezadeh-Abadeh, Daniel Kuhn, and Peyman Mohajerin Esfahani. Regularization via mass transportation. *Journal of Machine Learning Research*, 20(103):1–68, 2019.

Maurice Sion. On general minimax theorems. *Pacific Journal of Mathematics*, 8(1):171–176, 1958.

Robert Tibshirani. Regression shrinkage and selection via the LASSO. *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 267–288, 1996.

Sara van de Geer. Some exercises with the lasso and its compatibility constant. *arXiv preprint arXiv:1701.03326*, 2017.

Pauli Virtanen, Ralf Gommers, et al. SciPy 1.0–Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 17(3):261–272, 2020.

Martin J. Wainwright. *High-dimensional statistics: a non-asymptotic viewpoint*. Cambridge University Press, 2019.

Qinyu Wu, Jonathan Yu-Meng Li, and Tiantian Mao. On generalization and regularization via wasserstein distributionally robust optimization. *Management Science*, 72(7):6104–6119, 2026.

Yiling Xie and Xiaoming Huo. High-dimensional (group) adversarial training in linear regression. *Advances in Neural Information Processing Systems*, 37:31708–31735, 2024.

Huan Xu, Constantine Caramanis, and Shie Mannor. Robust regression and lasso. *Advances in neural information processing systems*, 21, 2008.

Luhao Zhang, Jincheng Yang, and Rui Gao. A Short and General Duality Proof for Wasserstein Distributionally Robust Optimization. *Operations Research*, 2024.

# Distributionally robust linear regression through the lens of adversarial training

# Appendix

*In this appendix, we provide additional material supporting the main text, in particular, all the proofs. The structure follows the order of the sections in the main text.*

---

CONTENTS


**A Background: Additional details** **ii**

**B Equivalent forms: Proofs and additional details** **ii**
B.1 Proof of Theorem 1 . . . . . ii
B.2 Proof of Corollary 1 . . . . . iv

**C Error bounds** **v**
C.1 Slow rate: Proofs . . . . . v
C.2 Fast rate: Proofs . . . . . x

**D Small and large ambiguity balls: Proofs and additional details** **xiii**
D.1 Proof of Theorem 4 . . . . . xiii
D.2 Proof of Theorem 5 . . . . . xiv
D.3 Computing $\delta_S$ via a convex program . . . . . xvi

**E Numerical solvers: Additional details** **xvii**
E.1 Saddle-point formulation and solver . . . . . xvii
E.2 $\eta$-trick solver . . . . . xviii
E.3 Fast robust risk evaluations . . . . . xxi

**F Numerical experiments: Additional details** **xxii**
F.1 Fast rate . . . . . xxii
F.2 Slow rate . . . . . xxii
F.3 Small and large $\delta$ . . . . . xxiv
F.4 Comparisons between the saddle-point solver and the $\eta$-trick solver . . . xxiv


---

## A Background: Additional details

In this section, we provide additional details of Section 3.

**Proof of Remark 1.** We show that $V_\delta(\boldsymbol{\beta}) = \infty$ for $\boldsymbol{\beta} \neq 0$ if $1 \leq p < 2$ as stated in Remark 1. This follows from Lemma 3 by noting that the supremum inside the sum is infinite, since the quadratic term with $t_i$ ($\|\boldsymbol{\beta}\|_* \neq 0$) increases faster than $t_i^p$ (since $p < 2$). Thus, $V_\delta(\boldsymbol{\beta}) = \infty$.

## B Equivalent forms: Proofs and additional details

In this section, we provide the proofs and additional details of Section 4.

### B.1 Proof of Theorem 1

To prove Theorem 1, we will use the following lemma that reduces the robust risk to several scalar problems:

**Lemma 3.** The robust risk of Wasserstein DRO linear regression for $1 \leq p < \infty$ equals

$$nV_\delta(\boldsymbol{\beta}) = \min_{\lambda \geq 0} \left[ n\delta^p \lambda + \sum_{i=1}^n \sup_{t_i \geq 0} \left\{ \left( |r_i| + t_i \|\boldsymbol{\beta}\|_* \right)^2 - \lambda t_i^p \right\} \right]. \tag{A.1}$$

with $r_i := \boldsymbol{x}_i^\top \boldsymbol{\beta} - y_i$.

Lemma 3 is a direct application of Theorem 4 (ii) in Shafiee et al. (2026). For completeness, we provide a short proof here for the linear regression case:

*Proof of Lemma 3.* We start from Proposition 1, where we note that the expectation in (4) turns into a sum due to the empirical distribution $\widehat{\mathbb{P}}_n := \frac{1}{n} \sum_{i=1}^n \delta_{\boldsymbol{z}_i}$, yielding:

$$nV_\delta(\boldsymbol{\beta}) = \min_{\lambda \geq 0} \left[ n\delta^p \lambda + \sum_{i=1}^n \sup_{\boldsymbol{z}' \in \mathcal{Z}} \left\{ h_{\boldsymbol{\beta}}(\boldsymbol{z}') - \lambda \|\boldsymbol{z}' - \boldsymbol{z}_i\| \right\} \right]. \tag{A.2}$$

In our case, $\mathcal{Z} = \mathbb{R}^{d+1}$ and $h_{\boldsymbol{\beta}} : \mathbb{R}^{d+1} \to \mathbb{R}$ is the square loss combined with the linear predictor model, i.e., for $\boldsymbol{z} = (\boldsymbol{x}, y)$, we have $h_{\boldsymbol{\beta}}(\boldsymbol{z}) = \left( \boldsymbol{\beta}^\top \boldsymbol{x} - y \right)^2$, and $\|\boldsymbol{z}' - \boldsymbol{z}\|$ with $z = (\boldsymbol{x}, y)$ and $z' = (\boldsymbol{x}', y')$ equals

$$\|\boldsymbol{z}' - \boldsymbol{z}\| = \begin{cases} \|\boldsymbol{x}' - \boldsymbol{x}\|^p & \text{if } y = y', \\ +\infty & \text{otherwise.} \end{cases}$$

Since $\|\boldsymbol{z}' - \boldsymbol{z}\| = \infty$ whenever $y \neq y'$, the supremum in (A.2) never moves $\boldsymbol{z}'$ away from $\boldsymbol{z}_i$ in the $y$-coordinate.[4] In other words, we have that

$$nV_\delta(\boldsymbol{\beta}) = \min_{\lambda \geq 0} \left[ n\delta^p \lambda + \sum_{i=1}^n \sup_{\boldsymbol{x}' \in \mathbb{R}^d} \left\{ \left( \boldsymbol{\beta}^\top \boldsymbol{x}' - y_i \right)^2 - \lambda \|\boldsymbol{x}' - \boldsymbol{x}_i\|^p \right\} \right].$$

Now center the problem by letting $\boldsymbol{u}_i = \boldsymbol{x}' - \boldsymbol{x}_i$. Then

$$\boldsymbol{\beta}^\top \boldsymbol{x}' - y_i = \boldsymbol{\beta}^\top (\boldsymbol{u}_i + \boldsymbol{x}_i) - y_i = \boldsymbol{\beta}^\top \boldsymbol{x}_i - y_i + \boldsymbol{\beta}^\top \boldsymbol{u}_i = r_i + \boldsymbol{\beta}^\top \boldsymbol{u}_i,$$

and therefore

$$\left( \boldsymbol{\beta}^\top \boldsymbol{x}' - y_i \right)^2 - \lambda \|\boldsymbol{x}' - \boldsymbol{x}_i\|^p = \left( r_i + \boldsymbol{\beta}^\top \boldsymbol{u}_i \right)^2 - \lambda \|\boldsymbol{u}_i\|^p.$$

Thus, we have

$$nV_\delta(\boldsymbol{\beta}) = \min_{\lambda \geq 0} \left[ n\delta^p \lambda + \sum_{i=1}^n \sup_{\boldsymbol{u} \in \mathbb{R}^d} \left\{ \left( r_i + \boldsymbol{\beta}^\top \boldsymbol{u} \right)^2 - \lambda \|\boldsymbol{u}\|^p \right\} \right],$$

[4] Indeed, such a deviation would let the outer minimization in (A.2) pick any $\lambda > 0$, resulting in an $-\infty$ cost due to $\|\boldsymbol{z}' - \boldsymbol{z}\| = \infty$, which is strictly suboptimal for the inner supremum (since the zero deviation $\boldsymbol{z}' = \boldsymbol{z}$ results in a bounded cost).

where we drop the subscript on $\boldsymbol{u}$ for convenience. We note that the penalty $\lambda\|\boldsymbol{u}\|^p$ depends only on the magnitude of $\boldsymbol{u}$, and thus its direction is irrelevant to the cost. To decompose direction and magnitude, let $\boldsymbol{u} = t\boldsymbol{v}$ with $t \geq 0$ and $\|\boldsymbol{v}\| = 1$. Then, the supremum can be split into two parts:

$$\sup_{t\geq 0} \sup_{\|\boldsymbol{v}\|=1} \left\{ \left(r_i + t\boldsymbol{\beta}^\top \boldsymbol{v}\right)^2 - \lambda t^p \right\} = \sup_{t\geq 0} \left[ \sup_{\|\boldsymbol{v}\|=1} \left\{ \left(r_i + t\boldsymbol{\beta}^\top \boldsymbol{v}\right)^2 \right\} - \lambda t^p \right].$$

The remaining inner problem $\sup_{\|\boldsymbol{v}\|=1}(r_i + t\boldsymbol{\beta}^\top \boldsymbol{v})^2$ can be solved in closed form:

$$\sup_{\|\boldsymbol{v}\|=1} \left\{ \left(r_i + t\boldsymbol{\beta}^\top \boldsymbol{v}\right)^2 \right\} = \sup_{\|\boldsymbol{v}\|=1} \left\{ \left(|r_i| + t\boldsymbol{\beta}^\top \boldsymbol{v}\right)^2 \right\} = \left(|r_i| + t \sup_{\|\boldsymbol{v}\|=1} \{\boldsymbol{\beta}^\top \boldsymbol{v}\}\right)^2 = \left(|r_i| + t\|\boldsymbol{\beta}\|_*\right)^2,$$

where the first equality is due to symmetry, the second holds since $w \mapsto (a + tw)^2$ is increasing for $a, t \geq 0$, and the third holds since $\|\boldsymbol{\beta}\|_* = \sup_{\|\boldsymbol{v}\|=1}\{\boldsymbol{\beta}^\top \boldsymbol{v}\}$. This leaves us with the scalar problem

$$nV_\delta(\boldsymbol{\beta}) = \min_{\lambda\geq 0} \left[ n\delta^p\lambda + \sum_{i=1}^n \sup_{t\geq 0} \left\{ \left(|r_i| + t\|\boldsymbol{\beta}\|_*\right)^2 - \lambda t^p \right\} \right],$$

completing the proof. □

We now prove Theorem 1:

*Proof of Theorem 1.* We first consider the case $p < \infty$. We want to show that

$$F(\boldsymbol{\beta}) := \sup_{t_i\geq 0, \sum_{i=1}^n t_i^p \leq n\delta^p} \sum_{i=1}^n \left(|r_i| + t_i\|\boldsymbol{\beta}\|_*\right)^2. \tag{A.3}$$

is equal to $nV_\delta(\boldsymbol{\beta})$. For $\delta = 0$, we trivially have $F(\boldsymbol{\beta}) = nV_\delta(\boldsymbol{\beta}) = \sum_{i=1}^n r_i^2$ using (A.1). For the case $\delta > 0$, we show equality using standard strong duality theory (Boyd & Vandenberghe, 2004, Chapter 5). More precisely, consider (A.3), and apply the change of variable $\nu_i = t_i^p$, leading to

$$F(\boldsymbol{\beta}) = \sup_{\nu_i\geq 0, \sum_{i=1}^n \nu_i \leq n\delta^p} \sum_{i=1}^n \left(|r_i| + \nu_i^{1/p}\|\boldsymbol{\beta}\|_*\right)^2 =$$

$$\sup_{\nu_i\geq 0, \sum_{i=1}^n \nu_i \leq n\delta^p} \sum_{i=1}^n \underbrace{r_i^2 + 2|r_i|\|\boldsymbol{\beta}\|_*\nu_i^{1/p} + \|\boldsymbol{\beta}\|_*^2\nu_i^{2/p}}_{\phi_i(\nu_i):=} = \sup_{\nu_i\geq 0, \sum_{i=1}^n \nu_i \leq n\delta^p} \sum_{i=1}^n \phi_i(\nu_i).$$

Here, $\phi_i(\nu_i) = r_i^2 + 2|r_i|\|\boldsymbol{\beta}\|_*\nu_i^{1/p} + \|\boldsymbol{\beta}\|_*^2\nu_i^{2/p}$ is concave in $\nu_i \geq 0$ since $p \geq 2$ implies $1/p \in [0, 1/2]$ and $2/p \in [0, 1]$. In other words, $F(\boldsymbol{\beta})$ is a concave maximization over an affine domain. Equivalently, to work with the more standard convex setting (Boyd & Vandenberghe, 2004, Chapter 5), we have that

$$\tilde{F}(\boldsymbol{\beta}) := -F(\boldsymbol{\beta}) = \inf_{\nu_i\geq 0, \sum_{i=1}^n \nu_i \leq n\delta^p} \sum_{i=1}^n -\phi_i(\nu_i) \tag{A.4}$$

is a convex minimization over an affine domain. The dual problem of (A.4) with respect to the constraint $\sum_{i=1}^n \nu_i \leq n\delta^p$ (leaving $\nu_i \geq 0$ as part of the domain) is[5]

$$\sup_{\lambda\geq 0} \inf_{\boldsymbol{\nu}\geq 0} L(\nu, \lambda) := \sup_{\lambda\geq 0} \inf_{\boldsymbol{\nu}\geq 0} \left[ \sum_{i=1}^n (-\phi_i(\nu_i)) + \lambda \left( \sum_{i=1}^n (\nu_i) - n\delta^p \right) \right] \tag{A.5}$$

Moreover, $\nu_i = 0$ yields $\sum_{i=1}^n \nu_i = 0 < n\delta^p$ for $\delta > 0$, so Slater's condition holds. Therefore, strong duality holds (Boyd & Vandenberghe, 2004, Chapter 5) and hence (A.4) is equal to (A.5).

[5] We compactly write $\boldsymbol{\nu} \geq 0$ meaning $\nu_i \geq 0$ for each $i$.

Equivalently,

$$F(\boldsymbol{\beta}) = -\sup_{\lambda\geq 0}\inf_{\boldsymbol{\nu}\geq 0}\left[\sum_{i=1}^{n}(-\phi_i(\nu_i)) + \lambda\left(\sum_{i=1}^{n}(\nu_i) - n\delta^p\right)\right] =$$

$$\inf_{\lambda\geq 0}\sup_{\boldsymbol{\nu}\geq 0}\left[\sum_{i=1}^{n}(\phi_i(\nu_i)) + \lambda\left(n\delta^p - \sum_{i=1}^{n}\nu_i\right)\right] = \inf_{\lambda\geq 0}\left[n\delta^p\lambda + \sum_{i=1}^{n}\sup_{\nu_i\geq 0}(\phi_i(\nu_i) - \lambda\nu_i)\right] =$$

$$\inf_{\lambda\geq 0}\left[n\delta^p\lambda + \sum_{i=1}^{n}\sup_{t_i\geq 0}(\phi_i(t_i^p) - \lambda t_i^p)\right] = \inf_{\lambda\geq 0}\left[n\delta^p\lambda + \sum_{i=1}^{n}\sup_{t_i\geq 0}\left\{\left(|r_i| + t_i\|\boldsymbol{\beta}\|_*\right)^2 - \lambda t_i^p\right\}\right] =$$

$$nV_\delta(\boldsymbol{\beta})$$

where the last equality is due to (A.1). This completes the proof for $p < \infty$.

For $p = \infty$, we have by Theorem 2 in Zhang et al. (2024) that

$$nV_\delta(\boldsymbol{\beta}) = \sum_{i=1}^{n}\sup_{\|\boldsymbol{v}_i\|\leq\delta}((\boldsymbol{x}_i + \boldsymbol{v}_i)^T\boldsymbol{\beta} - y_i)^2$$

which can be rewritten to the desired form using a dual norm argument, see, e.g., Proposition 1 in Ribeiro et al. (2023). For completeness, we provide this argument here too:

$$\sup_{\|v_i\|\leq\delta}((\boldsymbol{x}_i + \boldsymbol{v}_i)^T\boldsymbol{\beta} - y_i)^2 = \sup_{\|v_i\|\leq\delta}(\boldsymbol{v}_i^T\boldsymbol{\beta} + \boldsymbol{x}_i^T\boldsymbol{\beta} - y_i)^2 = \sup_{\|v_i\|\leq\delta}(\boldsymbol{v}_i^T\boldsymbol{\beta} + |\boldsymbol{x}_i^T\boldsymbol{\beta} - y_i|)^2 =$$

$$\sup_{\|v_i\|\leq 1}(\delta\boldsymbol{v}_i^T\boldsymbol{\beta} + |\boldsymbol{x}_i^T\boldsymbol{\beta} - y_i|)^2 = (\delta\sup_{\|v_i\|\leq 1}\boldsymbol{v}_i^T\boldsymbol{\beta} + |\boldsymbol{x}_i^T\boldsymbol{\beta} - y_i|)^2 = (\delta\|\boldsymbol{\beta}\|_* + |\boldsymbol{x}_i^T\boldsymbol{\beta} - y_i|)^2$$

and thus

$$V_\delta(\boldsymbol{\beta}) = \sum_{i=1}^{n}(|\boldsymbol{x}_i^T\boldsymbol{\beta} - y_i| + \delta\|\boldsymbol{\beta}\|_*)^2 = \sup_{t_i\geq 0,\|\boldsymbol{t}\|\leq\delta}\sum_{i=1}^{n}(|\boldsymbol{x}_i^T\boldsymbol{\beta} - y_i| + t_i\|\boldsymbol{\beta}\|_*)^2$$

completing the proof for $p = \infty$ and hence of Theorem 1. □

### B.2 Proof of Corollary 1

We continue by proving Corollary 1, using the saddle-point formulation of the robust risk.

*Proof of Corollary 1.* We will use Theorem 7 (that we prove latter in the appendix). This result state that

$$nV_\delta(\boldsymbol{\beta}) = \sup_{\boldsymbol{\alpha}\geq 0}K(\boldsymbol{\beta},\boldsymbol{\alpha}) = \sup_{\boldsymbol{\alpha}\geq 0}\left[n^{1/2}\delta\|\boldsymbol{\beta}\|_*\|\boldsymbol{\alpha}\|_2 + \sum_{i=1}^{n}\alpha_i|\boldsymbol{x}_i^\top\boldsymbol{\beta} - y_i| - \alpha_i^2/4\right]$$

We start with $p = 2$ and compute (A.18) noting that $q = 2$. For brevity, let $D$ be the vector with entries $D_i = |\boldsymbol{x}_i^T\boldsymbol{\beta} - y_i|$. Then:

$$\begin{aligned}
nV_\delta(\boldsymbol{\beta}) &= \sup_{\boldsymbol{\alpha}\geq 0}\left[n^{1/2}\delta\|\boldsymbol{\beta}\|_*\|\boldsymbol{\alpha}\|_2 + \langle\boldsymbol{\alpha},\boldsymbol{D}\rangle - \|\alpha\|_2^2/4\right] \\
&= \sup_{\boldsymbol{\alpha}\geq 0}\left[n^{1/2}\delta\|\boldsymbol{\beta}\|_*\|\boldsymbol{\alpha}\|_2 + \|\boldsymbol{\alpha}\|_2\|\boldsymbol{D}\|_2 - \|\boldsymbol{\alpha}\|_2^2/4\right] \\
&= \sup_{t\geq 0}\left[n^{1/2}\delta\|\boldsymbol{\beta}\|_*t + t\|\boldsymbol{D}\|_2 - t^2/4\right],
\end{aligned}$$

where we used the Cauchy–Schwarz inequality $\langle\boldsymbol{\alpha},\boldsymbol{D}\rangle \leq \|\boldsymbol{\alpha}\|_2\|\boldsymbol{D}\|_2$ with equality if $\boldsymbol{\alpha} = l\cdot\boldsymbol{D}$ for some scalar $l \geq 0$ (note that such an $\boldsymbol{\alpha}$ is indeed feasible since $\boldsymbol{D} \geq 0$), and then transformed the supremum to a one-dimensional optimization by setting $t = \|\boldsymbol{\alpha}\|_2$. This one-dimensional optimization is quadratic, with solution

$$(\|\boldsymbol{D}\|_2 + n^{1/2}\delta\|\boldsymbol{\beta}\|_*)^2.$$

Therefore

$$V_\delta(\boldsymbol{\beta}) = \left(\frac{1}{n^{1/2}}\|\boldsymbol{D}\|_2 + \delta\|\boldsymbol{\beta}\|_*\right)^2 = \left(\sqrt{\frac{1}{n}\sum_{i=1}^n(\boldsymbol{x}_i^\top\boldsymbol{\beta} - y_i)^2} + \delta\|\boldsymbol{\beta}\|_*\right)^2,$$

which coincides with square-root Lasso (Belloni et al., 2011) for $\|\cdot\|_* = \|\cdot\|_1$, i.e., $\|\cdot\| = \|\cdot\|_\infty$.

To show the adversarial linear regression case, we have by Theorem 1 that

$$V_{\infty,\delta}(\boldsymbol{\beta}) = \sup_{t_i\geq 0, \|t_i\|\leq\delta}\sum_{i=1}^n(|\boldsymbol{x}_i^\top\boldsymbol{\beta} - y_i| + t_i\|\boldsymbol{\beta}\|_*)^2 = \sum_{i=1}^n(|\boldsymbol{x}_i^\top\boldsymbol{\beta} - y_i| + \delta\|\boldsymbol{\beta}\|_*)^2 = \sum_{i=1}^n\sup_{\|\boldsymbol{v}_i\|\leq\delta}((\boldsymbol{x}_i + \boldsymbol{v}_i)^\top\boldsymbol{\beta} - y_i)^2$$

where the last step follows by the dual norm argument in the proof of Theorem 1. Here, we identify $v_i$ as the adversarial noise on the input $\boldsymbol{x}_i$. In other words, $p$-Wasserstein linear regression with $p = \infty$ coincides with adversarial linear regression. It remains to show that $\lim_{p\to\infty} V_{p,\delta}(\boldsymbol{\beta}) = V_{\infty,\delta}(\boldsymbol{\beta})$. Again, using Theorem 7, we have that:

$$\lim_{p\to\infty} V_{p,\delta}(\boldsymbol{\beta}) = \lim_{p\to\infty}\sup_{\boldsymbol{\alpha}\geq 0} K(\boldsymbol{\beta},\boldsymbol{\alpha}) = \lim_{p\to\infty}\sup_{\boldsymbol{\alpha}\geq 0}\left[n^{1/p}\delta\|\boldsymbol{\beta}\|_*\|\boldsymbol{\alpha}\|_q + \sum_{i=1}^n\alpha_i|\boldsymbol{x}_i^\top\boldsymbol{\beta} - y_i| - \alpha_i^2/4\right].$$

We want to swap the limit and the supremum. To this end, fix $\boldsymbol{\beta}$ and let $f_p(\boldsymbol{\alpha}) := K(\boldsymbol{\beta},\boldsymbol{\alpha})$, where we stress the dependence on $p$. We have that $\|\boldsymbol{\alpha}\|_q \leq n^{1/q}\|\boldsymbol{\alpha}\|_2$ by a norm inequality and $\sum_{i=1}^n\alpha_i|\boldsymbol{x}_i^\top\boldsymbol{\beta} - y_i| \leq \|\boldsymbol{\alpha}\|_2\|\boldsymbol{D}\|_2$ by Cauchy–Schwarz inequality. Thus,

$$f_p(\boldsymbol{\alpha}) \leq n^{1/p}\delta\|\boldsymbol{\beta}\|_* n^{1/q}\|\boldsymbol{\alpha}\|_2 + \|\boldsymbol{\alpha}\|_2\|\boldsymbol{D}\|_2 - \|\boldsymbol{\alpha}\|_2^2/4 = (n\delta\|\boldsymbol{\beta}\|_* + \|\boldsymbol{D}\|_2)\|\boldsymbol{\alpha}\|_2 - \|\boldsymbol{\alpha}\|_2^2/4 \to -\infty$$

when $\|\boldsymbol{\alpha}\|_2 \to \infty$ independent of $p$. Thus, there exists a radius $R > 0$ such that the maximum of $f_p(\boldsymbol{\alpha})$ for $\boldsymbol{\alpha} \geq 0$ is attained in $\Lambda := \{\boldsymbol{\alpha} \geq 0 : \|\boldsymbol{\alpha}\|_2 \leq R\}$. Let

$$f_\infty(\boldsymbol{\alpha}) := \lim_{p\to\infty} f_p(\boldsymbol{\alpha}) = \delta\|\boldsymbol{\beta}\|_*\|\boldsymbol{\alpha}\|_1 + \sum_{i=1}^n\alpha_i|\boldsymbol{x}_i^T\boldsymbol{\beta} - y_i| - \alpha_i^2/4$$

and note that $f_\infty(\boldsymbol{\alpha})$ also attains its maximum on $\Lambda$. Moreover, since $\Lambda$ is compact, $f_p(\boldsymbol{\alpha})$ converges uniformly to $f_\infty(\boldsymbol{\alpha})$ on $\Lambda$. Due to this uniform convergence, we can swap the limit and the supremum:

$$\lim_{p\to\infty}\sup_{\boldsymbol{\alpha}\geq 0} f_p(\boldsymbol{\alpha}) = \lim_{p\to\infty}\sup_{\boldsymbol{\alpha}\in\Lambda} f_p(\boldsymbol{\alpha}) = \sup_{\boldsymbol{\alpha}\in\Lambda}\lim_{p\to\infty} f_p(\boldsymbol{\alpha}) = \sup_{\boldsymbol{\alpha}\in\Lambda} f_\infty(\boldsymbol{\alpha}) = \sup_{\boldsymbol{\alpha}\geq 0} f_\infty(\boldsymbol{\alpha})$$

Therefore,

$$\lim_{p\to\infty} V_{p,\delta}(\boldsymbol{\beta}) = \lim_{p\to\infty}\sup_{\boldsymbol{\alpha}\geq 0} f_p(\boldsymbol{\alpha}) = \sup_{\boldsymbol{\alpha}\geq 0} f_\infty(\boldsymbol{\alpha}) =$$

$$\sup_{\boldsymbol{\alpha}\geq 0}\left[\delta\|\boldsymbol{\beta}\|_*\|\boldsymbol{\alpha}\|_1 + \sum_{i=1}^n\alpha_i|\boldsymbol{x}_i^\top\boldsymbol{\beta} - y_i| - \alpha_i^2/4\right] = \sup_{\boldsymbol{\alpha}\geq 0}\left[\sum_{i=1}^n(\delta\|\boldsymbol{\beta}\|_* + |\boldsymbol{x}_i^\top\boldsymbol{\beta} - y_i|)\alpha_i - \alpha_i^2/4\right] =$$

$$\sum_{i=1}^n\sup_{\alpha_i\geq 0}\left[(\delta\|\boldsymbol{\beta}\|_* + |\boldsymbol{x}_i^\top\boldsymbol{\beta} - y_i|)\alpha_i - \alpha_i^2/4\right] = \sum_{i=1}^n(\delta\|\boldsymbol{\beta}\|_* + |\boldsymbol{x}_i^\top\boldsymbol{\beta} - y_i|)^2,$$

The last expression coincides with $V_{\infty,\delta}(\boldsymbol{\beta})$, completing the proof. □

# C Error bounds

## C.1 Slow rate: Proofs

In this section, we provide the proofs of Section 5.1.

To prove Lemma 1, we first show the following lemma:

**Lemma 4.** We have the following bounds:

$$nV_\delta(\boldsymbol{\beta}^*) \le n\delta^2\|\boldsymbol{\beta}^*\|_*^2 + 2n^{1/p}\delta\|\boldsymbol{\beta}^*\|_*\|\boldsymbol{\varepsilon}\|_q + \|\boldsymbol{\varepsilon}\|_2^2 \tag{A.6}$$

$$nV_\delta(\widehat{\boldsymbol{\beta}}) \ge n\delta^2\|\widehat{\boldsymbol{\beta}}\|_*^2 + 2\delta\|\widehat{\boldsymbol{\beta}}\|_*\|r(\widehat{\boldsymbol{\beta}})\|_1 + \|r(\widehat{\boldsymbol{\beta}})\|_2^2, \tag{A.7}$$

where $r_i(\boldsymbol{\beta}) = |\boldsymbol{x}_i^T\boldsymbol{\beta} - y_i|$.

*Proof.* We obtain these bounds using the representation from Theorem 1:

$$nV_\delta(\boldsymbol{\beta}) = \max_{\boldsymbol{t}\ge 0, \|\boldsymbol{t}\|_p \le 1} \sum_{i=1}^n \Big(n^{1/p}\delta\|\boldsymbol{\beta}\|_* t_i + |\boldsymbol{x}_i^\top\boldsymbol{\beta} - y_i|\Big)^2 =$$

$$\max_{\boldsymbol{t}\ge 0, \|\boldsymbol{t}\|_p \le 1} \Big(n^{2/p}\delta^2\|\boldsymbol{\beta}\|_*^2\|\boldsymbol{t}\|_2^2 + 2n^{1/p}\delta\|\boldsymbol{\beta}\|_* \sum_{i=1}^n t_i r_i(\boldsymbol{\beta}) + \|r(\boldsymbol{\beta})\|_2^2\Big).$$

To upper bound $V_\delta(\boldsymbol{\beta}^*)$, we note that $r(\boldsymbol{\beta}^*) = |\boldsymbol{\varepsilon}|$, resulting in

$$nV_\delta(\boldsymbol{\beta}^*) = \max_{\boldsymbol{t}\ge 0, \|\boldsymbol{t}\|_p \le 1} \Big(n^{2/p}\delta^2\|\boldsymbol{\beta}^*\|_*^2\|\boldsymbol{t}\|_2^2 + 2n^{1/p}\delta\|\boldsymbol{\beta}^*\|_* \sum_{i=1}^n t_i|\varepsilon_i| + \|\boldsymbol{\varepsilon}\|_2^2\Big).$$

Then we use Hölder's inequality $\sum_{i=1}^n t_i|\varepsilon_i| \le \|\boldsymbol{t}\|_p\|\boldsymbol{\varepsilon}\|_q \le \|\boldsymbol{\varepsilon}\|_q$ and the norm inequality $\|\boldsymbol{t}\|_2^2 \le n^{1-2/p}\|\boldsymbol{t}\|_p^2 \le n^{1-2/p}$ to get

$$nV_\delta(\boldsymbol{\beta}^*) \le n\delta^2\|\boldsymbol{\beta}^*\|_*^2 + 2n^{1/p}\delta\|\boldsymbol{\beta}^*\|_*\|\boldsymbol{\varepsilon}\|_q + \|\boldsymbol{\varepsilon}\|_2^2.$$

We next lower bound $nV_\delta(\widehat{\boldsymbol{\beta}})$. For this, we simply pick the point $s_i = n^{-1/p}$ (which satisfies $\|s\|_p = 1$ and $s \ge 0$), yielding

$$nV_\delta(\widehat{\boldsymbol{\beta}}) \ge n\delta^2\|\widehat{\boldsymbol{\beta}}\|_*^2 + 2\delta\|\widehat{\boldsymbol{\beta}}\|_*\|r(\widehat{\boldsymbol{\beta}})\|_1 + \|r(\widehat{\boldsymbol{\beta}})\|_2^2.$$

This completes the proof. □

**Proof of Lemma 1.** We now prove Lemma 1:

*Proof.* By optimality, $nV_\delta(\widehat{\boldsymbol{\beta}}) \le nV_\delta(\boldsymbol{\beta}^*)$. Combined with Lemma 4 yields

$$n\delta^2\|\widehat{\boldsymbol{\beta}}\|_*^2 + 2\delta\|\widehat{\boldsymbol{\beta}}\|_*\|r(\widehat{\boldsymbol{\beta}})\|_1 + \|r(\widehat{\boldsymbol{\beta}})\|_2^2 \le n\delta^2\|\boldsymbol{\beta}^*\|_*^2 + 2n^{1/p}\delta\|\boldsymbol{\beta}^*\|_*\|\boldsymbol{\varepsilon}\|_q + \|\boldsymbol{\varepsilon}\|_2^2$$

Noting that $\|r(\widehat{\boldsymbol{\beta}})\|_2^2 = \|X\widehat{\Delta}\|_2^2 - 2\varepsilon^T X\widehat{\Delta} + \|\boldsymbol{\varepsilon}\|_2^2$, we get

$$n\delta^2\|\widehat{\boldsymbol{\beta}}\|_*^2 + 2\delta\|\widehat{\boldsymbol{\beta}}\|_*\|r(\widehat{\boldsymbol{\beta}})\|_1 + \|X\widehat{\Delta}\|_2^2 \le 2\boldsymbol{\varepsilon}^\top X\widehat{\Delta} + n\delta^2\|\boldsymbol{\beta}^*\|_*^2 + 2n^{1/p}\delta\|\boldsymbol{\beta}^*\|_*\|\boldsymbol{\varepsilon}\|_q.$$

The lemma now follows by dropping the nonnegative terms $n\delta^2\|\widehat{\boldsymbol{\beta}}\|_*^2 + 2\delta\|\widehat{\boldsymbol{\beta}}\|_*\|r(\widehat{\boldsymbol{\beta}})\|_1$ and dividing by $n$. □

We continue with the proof of Lemma 2, where we use the following two lemmas:

**Lemma 5.** Let $\boldsymbol{t}^*$ be a maximizer of (6) for $\boldsymbol{\beta} = \boldsymbol{\beta}^*$. Then

$$\|X\widehat{\Delta}\|_2^2 \le 2\boldsymbol{\varepsilon}^\top X\widehat{\Delta} + 2(\|\boldsymbol{\beta}^*\|_* - \|\widehat{\boldsymbol{\beta}}\|_*)\langle|\boldsymbol{\varepsilon}|, \boldsymbol{t}^*\rangle + 2\|\widehat{\boldsymbol{\beta}}\|_*\langle|X\widehat{\Delta}|, \boldsymbol{t}^*\rangle + \|\boldsymbol{t}^*\|_2^2(\|\boldsymbol{\beta}^*\|_*^2 - \|\widehat{\boldsymbol{\beta}}\|_*^2). \tag{A.8}$$

*Proof.* Using (6) and the definition of $\boldsymbol{t}^*$, we get that

$$nV_\delta(\boldsymbol{\beta}^*) = \sum_{i=1}^n (|x_i^T\boldsymbol{\beta}^* - y_i| + t_i^*\|\boldsymbol{\beta}^*\|_*)^2 = \sum_{i=1}^n (|\varepsilon_i| + t_i^*\|\boldsymbol{\beta}^*\|_*)^2$$

and

$$nV_\delta(\widehat{\boldsymbol{\beta}}) = \sup_{t_i \geq 0,\ \sum_{i=1}^n t_i^p \leq n\delta^p} \sum_{i=1}^n (|x_i^\top \widehat{\boldsymbol{\beta}} - y_i| + t_i \|\widehat{\boldsymbol{\beta}}\|_*)^2 \geq$$
$$\sum_{i=1}^n (|x_i^\top \widehat{\boldsymbol{\beta}} - y_i| + t_i^* \|\widehat{\boldsymbol{\beta}}\|_*)^2 = \sum_{i=1}^n (|x_i^\top \widehat{\Delta} - \varepsilon_i| + t_i^* \|\widehat{\boldsymbol{\beta}}\|_*)^2.$$

By optimality, $V_\delta(\widehat{\boldsymbol{\beta}}) \leq V_\delta(\boldsymbol{\beta}^*)$. Therefore,

$$\sum_{i=1}^n (|x_i^\top \widehat{\Delta} - \varepsilon_i| + t_i^* \|\widehat{\boldsymbol{\beta}}\|_*)^2 \leq nV_\delta(\widehat{\boldsymbol{\beta}}) \leq nV_\delta(\boldsymbol{\beta}^*) = \sum_{i=1}^n (|\varepsilon_i| + t_i^* \|\boldsymbol{\beta}^*\|_*)^2.$$

Expanding both sides yields

$$\|X\widehat{\Delta} - \boldsymbol{\varepsilon}\|_2^2 + 2\|\widehat{\boldsymbol{\beta}}\|_* \langle |X\widehat{\Delta} - \boldsymbol{\varepsilon}|, \boldsymbol{t}^* \rangle + \|\boldsymbol{t}^*\|_2^2 \|\widehat{\boldsymbol{\beta}}\|_*^2 \leq \|\boldsymbol{\varepsilon}\|_2^2 + 2\|\boldsymbol{\beta}^*\|_* \langle |\boldsymbol{\varepsilon}|, \boldsymbol{t}^* \rangle + \|\boldsymbol{t}^*\|_2^2 \|\boldsymbol{\beta}^*\|_*^2$$

which with $\|X\widehat{\Delta} - \boldsymbol{\varepsilon}\|_2^2 = \|X\widehat{\Delta}\|_2^2 - 2\boldsymbol{\varepsilon}^\top X\widehat{\Delta} + \|\boldsymbol{\varepsilon}\|_2^2$ results in

$$\|X\widehat{\Delta}\|_2^2 + 2\|\widehat{\boldsymbol{\beta}}\|_* \langle |X\widehat{\Delta} - \varepsilon|, \boldsymbol{t}^* \rangle + \|\boldsymbol{t}^*\|_2^2 \|\widehat{\boldsymbol{\beta}}\|_*^2 \leq 2\boldsymbol{\varepsilon}^\top X\widehat{\Delta} + 2\|\boldsymbol{\beta}^*\|_* \langle |\boldsymbol{\varepsilon}|, \boldsymbol{t}^* \rangle + \|\boldsymbol{t}^*\|_2^2 \|\boldsymbol{\beta}^*\|_*^2.$$

To get the final expression, we use the reverse triangle inequality

$$\langle |X\widehat{\Delta} - \boldsymbol{\varepsilon}|, \boldsymbol{t}^* \rangle \geq \langle |\boldsymbol{\varepsilon}| - |X\widehat{\Delta}|, \boldsymbol{t}^* \rangle = \langle |\boldsymbol{\varepsilon}|, \boldsymbol{t}^* \rangle - \langle |X\widehat{\Delta}|, \boldsymbol{t}^* \rangle$$

to obtain

$$\|X\widehat{\Delta}\|_2^2 \leq 2\boldsymbol{\varepsilon}^\top X\widehat{\Delta} + 2(\|\boldsymbol{\beta}^*\|_* - \|\widehat{\boldsymbol{\beta}}\|_*) \langle |\boldsymbol{\varepsilon}|, \boldsymbol{t}^* \rangle + 2\|\widehat{\boldsymbol{\beta}}\|_* \langle |X\widehat{\Delta}|, \boldsymbol{t}^* \rangle + \|\boldsymbol{t}^*\|_2^2 (\|\boldsymbol{\beta}^*\|_*^2 - \|\widehat{\boldsymbol{\beta}}\|_*^2).$$

This completes the proof. □

**Lemma 6.** We have

$$0 \leq 2\|X^T\boldsymbol{\varepsilon}\| \|\widehat{\Delta}\|_* + 2(\|\boldsymbol{\beta}^*\|_* - \|\widehat{\boldsymbol{\beta}}\|_*) \|\boldsymbol{\varepsilon}\|_q n^{1/p} \delta + n\delta^2 \|\boldsymbol{\beta}^*\|_*^2.$$

*Proof.* By Lemma 5, we have that

$$0 \leq$$
$$2\varepsilon^T X\widehat{\Delta} + 2(\|\boldsymbol{\beta}^*\|_* - \|\widehat{\boldsymbol{\beta}}\|_*) \langle |\boldsymbol{\varepsilon}|, \boldsymbol{t}^* \rangle + 2\|\widehat{\boldsymbol{\beta}}\|_* \langle |X\widehat{\Delta}|, \boldsymbol{t}^* \rangle - \|X\widehat{\Delta}\|_2^2 + \|\boldsymbol{t}^*\|_2^2 (\|\boldsymbol{\beta}^*\|_*^2 - \|\widehat{\boldsymbol{\beta}}\|_*^2) \leq$$
$$2\boldsymbol{\varepsilon}^\top X\widehat{\Delta} + 2(\|\boldsymbol{\beta}^*\|_* - \|\widehat{\boldsymbol{\beta}}\|_*) \langle |\varepsilon|, \boldsymbol{t}^* \rangle + 2\|\widehat{\boldsymbol{\beta}}\|_* \|X\widehat{\Delta}\|_2 \|\boldsymbol{t}^*\|_2 - \|X\widehat{\Delta}\|_2^2 + \|\boldsymbol{t}^*\|_2^2 (\|\boldsymbol{\beta}^*\|_*^2 - \|\widehat{\boldsymbol{\beta}}\|_*^2)$$

where we used the Cauchy–Schwarz inequality for the last inequality. Note now that

$$2\|\widehat{\boldsymbol{\beta}}\|_* \|X\widehat{\Delta}\|_2 \|\boldsymbol{t}^*\|_2 - \|X\widehat{\Delta}\|_2^2 \leq \|\widehat{\boldsymbol{\beta}}\|_*^2 \|\boldsymbol{t}^*\|_2^2,$$

obtained by maximizing the expression with respect to $\|X\widehat{\Delta}\|_2$. Thus,

$$0 \leq 2\boldsymbol{\varepsilon}^\top X\widehat{\Delta} + 2(\|\boldsymbol{\beta}^*\|_* - \|\widehat{\boldsymbol{\beta}}\|_*) \langle |\boldsymbol{\varepsilon}|, \boldsymbol{t}^* \rangle + \|\boldsymbol{t}^*\|_2^2 \|\boldsymbol{\beta}^*\|_*^2.$$

Finally, by the definition of the dual norm,

$$2\boldsymbol{\varepsilon}^\top X\widehat{\Delta} \leq 2\|X^\top \boldsymbol{\varepsilon}\| \|\widehat{\Delta}\|_*,$$

Hölder's inequality

$$\langle |\boldsymbol{\varepsilon}|, \boldsymbol{t}^* \rangle \leq \|\boldsymbol{\varepsilon}\|_q \|\boldsymbol{t}^*\|_p \leq \|\boldsymbol{\varepsilon}\|_q n^{1/p} \delta,$$

and the norm inequality

$$\|\boldsymbol{t}^*\|_2^2 \leq n^{1-2/p} \|\boldsymbol{t}^*\|_p^2 \leq n\delta^2,$$

we get

$$0 \leq 2\|X^\top \boldsymbol{\varepsilon}\| \|\widehat{\Delta}\|_* + 2(\|\boldsymbol{\beta}^*\|_* - \|\widehat{\boldsymbol{\beta}}\|_*) \|\boldsymbol{\varepsilon}\|_q n^{1/p} \delta + n\delta^2 \|\boldsymbol{\beta}^*\|_*^2.$$

This completes the proof. □

**Proof of Lemma 2.** We now prove Lemma 2:

*Proof.* We have by Lemma 6 and the triangle inequality $\|\widehat{\Delta}\|_* \leq \|\boldsymbol{\beta}^*\|_* + \|\widehat{\boldsymbol{\beta}}\|_*$ that

$$0 \leq 2\|X^\top \boldsymbol{\varepsilon}\|(\|\boldsymbol{\beta}^*\|_* + \|\widehat{\boldsymbol{\beta}}\|_*) + 2(\|\boldsymbol{\beta}^*\|_* - \|\widehat{\boldsymbol{\beta}}\|_*)\|\boldsymbol{\varepsilon}\|_q n^{1/p}\delta + n\delta^2\|\boldsymbol{\beta}^*\|_*^2.$$

Moreover, $2\|X^T\boldsymbol{\varepsilon}\| < \delta n^{1/p}\|\boldsymbol{\varepsilon}\|_q$ by assumption. Hence,

$$0 \leq \delta n^{1/p}\|\boldsymbol{\varepsilon}\|_q(\|\boldsymbol{\beta}^*\|_* + \|\widehat{\boldsymbol{\beta}}\|_*) + 2(\|\boldsymbol{\beta}^*\|_* - \|\widehat{\boldsymbol{\beta}}\|_*)\|\boldsymbol{\varepsilon}\|_q n^{1/p}\delta + n\delta^2\|\boldsymbol{\beta}^*\|_*^2 = 3\delta n^{1/p}\|\boldsymbol{\varepsilon}\|_q\|\boldsymbol{\beta}^*\|_* - \delta n^{1/p}\|\boldsymbol{\varepsilon}\|_q\|\widehat{\boldsymbol{\beta}}\|_* + n\delta^2\|\boldsymbol{\beta}^*\|_*^2$$

which we can solve for $\|\widehat{\boldsymbol{\beta}}\|_*$ to get

$$\|\widehat{\boldsymbol{\beta}}\|_* \leq 3\|\boldsymbol{\beta}^*\|_* + \frac{n^{1/q}\delta}{\|\boldsymbol{\varepsilon}\|_q}\|\boldsymbol{\beta}^*\|_*^2$$

from which the lemma follows. □

**Proof of Theorem 2.** We continue with the proof of Theorem 2:

*Proof.* By Lemma 1, we have that

$$\frac{1}{n}\|X\widehat{\Delta}\|_2^2 \leq \frac{2}{n}\boldsymbol{\varepsilon}^\top X\widehat{\Delta} + 2\delta\frac{\|\boldsymbol{\varepsilon}\|_q}{n^{1/q}}\|\boldsymbol{\beta}^*\|_* + \delta^2\|\boldsymbol{\beta}^*\|_*^2.$$

We bound the term $\frac{2}{n}\boldsymbol{\varepsilon}^\top X\widehat{\Delta}$ using the definition of a dual norm, the triangle inequality, the assumption $\delta > 2\frac{\|X^\top\boldsymbol{\varepsilon}\|}{n^{1/p}\|\boldsymbol{\varepsilon}\|_q}$, and Lemma 2:

$$\frac{2}{n}\boldsymbol{\varepsilon}^\top X\widehat{\Delta} \leq \frac{2}{n}\|X^\top\boldsymbol{\varepsilon}\|\|\widehat{\Delta}\|_* \leq \frac{2}{n}\|X^\top\boldsymbol{\varepsilon}\|(\|\widehat{\boldsymbol{\beta}}\|_* + \|\boldsymbol{\beta}^*\|_*) \leq \frac{2}{n}\|X^\top\boldsymbol{\varepsilon}\|\left(4\|\boldsymbol{\beta}^*\|_* + \frac{n^{1/q}}{\|\boldsymbol{\varepsilon}\|_q}\|\boldsymbol{\beta}^*\|_*^2\right) \leq \frac{1}{n}\delta n^{1/p}\|\boldsymbol{\varepsilon}\|_q\left(4\|\boldsymbol{\beta}^*\|_* + \frac{n^{1/q}}{\|\boldsymbol{\varepsilon}\|_q}\|\boldsymbol{\beta}^*\|_*^2\right) = 4\delta\frac{\|\boldsymbol{\varepsilon}\|_q}{n^{1/q}}\|\boldsymbol{\beta}^*\|_* + \delta^2\|\boldsymbol{\beta}^*\|_*^2,$$

from which we get

$$\frac{1}{n}\|X\widehat{\Delta}\|_2^2 \leq 6\delta\frac{\|\boldsymbol{\varepsilon}\|_q}{n^{1/q}}\|\boldsymbol{\beta}^*\|_* + 2\delta^2\|\boldsymbol{\beta}^*\|_*^2.$$

We have thus shown (9).

To show the high-probability rate in (10), note first that by a norm inequality (since $1 \leq q \leq 2$):

$$6\frac{\|\boldsymbol{\varepsilon}\|_q}{n^{1/q}} \leq 6\frac{n^{1/q-1/2}\|\boldsymbol{\varepsilon}\|_2}{n^{1/q}} = 6\frac{\|\boldsymbol{\varepsilon}\|_2}{\sqrt{n}}.$$

We bound $f(\boldsymbol{\varepsilon}) = 6\frac{\|\boldsymbol{\varepsilon}\|_2}{\sqrt{n}}$ with high probability. Since $f$ is Lipschitz with respect to the 2-norm with constant $L = \frac{6}{\sqrt{n}}$, a concentration bound for Lipschitz functions with Gaussian input (Boucheron et al., 2013, Theorem 5.6) yields:

$$\mathbb{P}(f(\boldsymbol{\varepsilon}) - \mathbb{E}[f(\boldsymbol{\varepsilon})] \geq t) \leq \exp\left(-\frac{t^2}{2\sigma^2L^2}\right).$$

By setting $\gamma := \exp\left(-\frac{t^2}{2\sigma^2L^2}\right)$ we can solve for $t$ to get $t = \sigma L\sqrt{2\log(1/\gamma)} = 6\sigma\sqrt{\frac{2\log(1/\gamma)}{n}}$. Thus with probability $\geq 1-\gamma$, we have that $f(\boldsymbol{\varepsilon}) < \mathbb{E}[f(\boldsymbol{\varepsilon})] + 6\sigma\sqrt{\frac{2\log(1/\gamma)}{n}}$, and since $\mathbb{E}[f(\boldsymbol{\varepsilon})] = 6\sigma$, we get that the bound

$$6\frac{\|\boldsymbol{\varepsilon}\|_q}{n^{1/q}} \leq 6\sigma + 6\sigma\sqrt{\frac{2\log(1/\gamma)}{n}}$$

holds with probability greater than $1-\gamma$. Therefore, due to (9), the rate of $\frac{1}{n}\|X\widehat{\Delta}\|_2^2$ is with high probability determined by the rate of $\delta$. That is, $\frac{1}{n}\|X\widehat{\Delta}\|_2^2 \in O(\delta)$ with high probability ($\geq 1-\gamma$).

We now obtain the rate for the particular choice $\delta = KM\sqrt{\frac{\log(d/\gamma)}{n}}$. To this end, we first upper bound $\bar{\delta}$. By a norm inequality

$$\bar{\delta} = \frac{2}{\sqrt{n}} \frac{\frac{\|X^\top \boldsymbol{\varepsilon}\|}{\sqrt{n}}}{\frac{\|\boldsymbol{\varepsilon}\|_q}{n^{1/q}}} \leq \frac{2}{\sqrt{n}} \frac{\frac{\|X^\top \boldsymbol{\varepsilon}\|}{\sqrt{n}}}{\frac{\|\boldsymbol{\varepsilon}\|_1}{n}}.$$

We upper bound $\|X^\top \boldsymbol{\varepsilon}\|$ and lower bound $\frac{\|\boldsymbol{\varepsilon}\|_1}{n}$. To lower bound $\frac{\|\boldsymbol{\varepsilon}\|_1}{n}$, note that $f(\boldsymbol{\varepsilon}) = \frac{\|\boldsymbol{\varepsilon}\|_1}{n} \leq \frac{\|\boldsymbol{\varepsilon}\|_2}{\sqrt{n}}$ is 2-norm Lipschitz with constant $L = 1/\sqrt{n}$. Hence, a concentration bound for Gaussians (Boucheron et al., 2013, Theorem 5.6) yield

$$\mathbb{P}(\mathbb{E}[f(\boldsymbol{\varepsilon})] - f(\boldsymbol{\varepsilon}) \geq t) \leq \exp\left(-\frac{t^2}{2\sigma^2 L^2}\right).$$

Setting $\gamma = \exp\left(-\frac{t^2}{2\sigma^2 L^2}\right)$ so that $t = \sigma L\sqrt{2\log(1/\gamma)} = \sigma\sqrt{\frac{2\log(1/\gamma)}{n}}$, and since $\mathbb{E}[f(\boldsymbol{\varepsilon})] = \sigma\sqrt{\frac{2}{\pi}}$, we get with probability $\geq 1-\gamma$:

$$\frac{\|\boldsymbol{\varepsilon}\|_1}{n} \geq \sigma\sqrt{\frac{2}{\pi}} - \sigma\sqrt{\frac{2\log(1/\gamma)}{n}}.$$

Moreover, we have with probability $\geq 1-2\gamma$ that $\frac{\|X^\top \boldsymbol{\varepsilon}\|_\infty}{\sqrt{n}} \leq M\sigma\sqrt{2\log(d/\gamma)}$ (see, e.g., (Ribeiro et al., 2023, Section C.3)), which implies that with probability $\geq 1-2\gamma$:

$$\frac{\|X^\top \boldsymbol{\varepsilon}\|}{\sqrt{n}} \leq \frac{c\|X^\top \boldsymbol{\varepsilon}\|_\infty}{\sqrt{n}} \leq cM\sigma\sqrt{2\log(d/\gamma)},$$

where we used that $\|\cdot\|_\infty \leq c\|\cdot\|$. Finally, we combine these two bounds to get a bound for $\bar{\delta}$. Concretely, with probability greater than $1-3\gamma$ (due to a union bound):

$$\bar{\delta} = \frac{2}{\sqrt{n}} \frac{\frac{\|X^\top \boldsymbol{\varepsilon}\|}{\sqrt{n}}}{\frac{\|\boldsymbol{\varepsilon}\|_1}{n}} \leq \frac{2}{\sqrt{n}} \frac{cM\sigma\sqrt{2\log(d/\gamma)}}{\sigma\sqrt{\frac{2}{\pi}} - \sigma\sqrt{\frac{2\log(1/\gamma)}{n}}} = \frac{2}{\sqrt{n}} \frac{cM\sqrt{\log(d/\gamma)}}{\sqrt{\frac{1}{\pi}} - \sqrt{\frac{\log(1/\gamma)}{n}}} =: \frac{C_1(n)}{\sqrt{n}}, \tag{A.9}$$

valid for $n > \pi\log(1/\gamma)$ since the denominator is then positive. Note that the function $C_1(n)$ is decreasing with $n$ (given that $n > \pi\log(1/\gamma)$). In other words, $\bar{\delta} \in O(1/\sqrt{n})$ with probability $\geq 1-3\gamma$.[6]

Finally, we tune $K$ in $\delta = KM\sqrt{\frac{\log(d/\gamma)}{n}}$ to satisfy $\delta > \bar{\delta}$ for large enough $n$, which is true if

$$K > \frac{2c}{\sqrt{\frac{1}{\pi}} - \sqrt{\frac{\log(1/\gamma)}{n}}} \tag{A.10}$$

for large enough $n$. Since $\frac{2c}{\sqrt{\frac{1}{\pi}} - \sqrt{\frac{\log(1/\gamma)}{n}}} \to 2c\sqrt{\pi}$ as $n \to \infty$, a sufficient condition for satisfying $\delta > \bar{\delta}$ for large enough $n$ (with probability $\geq 1-3\gamma$) is that $K > 2c\sqrt{\pi}$, where a larger $K$ results in $\delta > \bar{\delta}$ being satisfied sooner (smaller $n$). Thus, for $K > 2c\sqrt{\pi}$, we have $\frac{1}{n}\|X\widehat{\Delta}\|_2^2 \in O(\delta) = O(1/\sqrt{n})$ (with probability $\geq 1-4\gamma$ due to a union bound), achieving the slow rate. This completes the proof. □

[6] That we do not have an upper bound for $\bar{\delta}$ over the transition period $n \leq \pi\log(1/\gamma)$ is immaterial since $\bar{\delta}$ is almost surely finite over these finitely many value of $n$.

### C.2 Fast rate: Proofs

In this section, we provide the proofs of Section 5.2.

To prove Theorem 3, we first need an intermediate result. This result holds for any norm $\|\cdot\|$ (not only $\|\cdot\| = \|\cdot\|_\infty$):

**Lemma 7.** Let $\widehat{\boldsymbol{\beta}}$ be a minimizer of $V_\delta(\boldsymbol{\beta})$ and consider $\widehat{\Delta} = \widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*$. Assume each entry of $X$ is bounded by $M > 0$ and let $r := \sup_{\boldsymbol{x}\neq 0} \frac{\|\boldsymbol{x}\|}{\|\boldsymbol{x}\|_\infty}$.[a] Then for $\delta > \bar{\delta} = 2\frac{\|X^\top \boldsymbol{\varepsilon}\|}{n^{1/p}\|\boldsymbol{\varepsilon}\|_q}$, we have

$$\frac{1}{n}\|X\widehat{\Delta}\|_2^2 \leq \delta\left(3\frac{\|\boldsymbol{\varepsilon}\|_q}{n^{1/q}} + (2MCr + \delta C + \delta)\|\boldsymbol{\beta}^*\|_*\right)\|\widehat{\Delta}\|_* \tag{A.11}$$

where $C = (3 + \frac{n^{1/q}\delta}{\|\boldsymbol{\varepsilon}\|_q}\|\boldsymbol{\beta}^*\|_*)$.

[a] We introduce $r$ here to keep it general. Note that $r = 1$ for $\|\cdot\| = \|\cdot\|_\infty$.

*Proof.* By Lemma 5, we have that

$$\|X\widehat{\Delta}\|_2^2 \leq 2\boldsymbol{\varepsilon}^\top X\widehat{\Delta} + 2(\|\boldsymbol{\beta}^*\|_* - \|\widehat{\boldsymbol{\beta}}\|_*)\langle|\boldsymbol{\varepsilon}|, \boldsymbol{t}^*\rangle + 2\|\widehat{\boldsymbol{\beta}}\|_*\langle|X\widehat{\Delta}|, \boldsymbol{t}^*\rangle + \|\boldsymbol{t}^*\|_2^2(\|\boldsymbol{\beta}^*\|_*^2 - \|\widehat{\boldsymbol{\beta}}\|_*^2).$$

To prove Lemma 7, we simply bound each term. Concretely, by definition of the dual norm and since $\delta > \bar{\delta} = 2\frac{\|X^\top \boldsymbol{\varepsilon}\|}{n^{1/p}\|\boldsymbol{\varepsilon}\|_q}$:

$$2\boldsymbol{\varepsilon}^\top X\widehat{\Delta} \leq 2\|X^\top\boldsymbol{\varepsilon}\|\|\widehat{\Delta}\|_* \leq \delta n^{1/p}\|\boldsymbol{\varepsilon}\|_q\|\widehat{\Delta}\|_*.$$

Furthermore, the reverse triangle inequality yields

$$\|\boldsymbol{\beta}^*\|_* - \|\widehat{\boldsymbol{\beta}}\|_* \leq \|\boldsymbol{\beta}^* - \widehat{\boldsymbol{\beta}}\|_* = \|\widehat{\Delta}\|_*,$$

and by Hölder's inequality

$$\langle|\boldsymbol{\varepsilon}|, \boldsymbol{t}^*\rangle \leq \|\boldsymbol{\varepsilon}\|_q\|\boldsymbol{t}^*\|_p \leq \|\boldsymbol{\varepsilon}\|_q n^{1/p}\delta.$$

Therefore,

$$2(\|\boldsymbol{\beta}^*\|_* - \|\widehat{\boldsymbol{\beta}}\|_*)\langle|\boldsymbol{\varepsilon}|, \boldsymbol{t}^*\rangle \leq 2\|\boldsymbol{\varepsilon}\|_q n^{1/p}\delta\|\widehat{\Delta}\|_*$$

To bound $\langle|X\widehat{\Delta}|, \boldsymbol{t}^*\rangle$, note first that

$$\langle|X\widehat{\Delta}|, \boldsymbol{t}^*\rangle = \sum_{i=1}^n \operatorname{sign}(\boldsymbol{x}_i^\top\widehat{\Delta})\boldsymbol{x}_i^\top\widehat{\Delta}t_i^* \leq \left\|\sum_{i=1}^n \operatorname{sign}(\boldsymbol{x}_i^\top\widehat{\Delta})\boldsymbol{x}_i t_i^*\right\|\|\widehat{\Delta}\|_* \leq$$
$$\sum_{i=1}^n \|\boldsymbol{x}_i\|t_i^*\|\widehat{\Delta}\|_* \leq \|\boldsymbol{\tau}\|_q\|\boldsymbol{t}^*\|_p\|\widehat{\Delta}\|_*$$

where the last inequality holds due to Hölder's inequality with $\tau_i := \|\boldsymbol{x}_i\|$. Since $X$ has entries bounded by $M > 0$, we get

$$\|\boldsymbol{\tau}\|_q \leq \left(\sum_{i=1}^n (Mr)^q\right)^{1/q} = n^{1/q}Mr$$

and by definition of $\boldsymbol{t}^*$,

$$\|\boldsymbol{t}^*\|_p \leq n^{1/p}\delta,$$

and therefore

$$\langle|X\widehat{\Delta}|, \boldsymbol{t}^*\rangle \leq n^{1/q}Mrn^{1/p}\delta\|\widehat{\Delta}\|_* = nMr\delta\|\widehat{\Delta}\|_*.$$

Thus, we have the bound

$$2\|\widehat{\boldsymbol{\beta}}\|_*\langle|X\widehat{\Delta}|, \boldsymbol{t}^*\rangle \leq 2\|\widehat{\boldsymbol{\beta}}\|_* nMr\delta\|\widehat{\Delta}\|_*.$$

Finally, we have that

$$\|\boldsymbol{t}^*\|_2^2(\|\boldsymbol{\beta}^*\|_*^2 - \|\widehat{\boldsymbol{\beta}}\|_*^2) = \|\boldsymbol{t}^*\|_2^2(\|\boldsymbol{\beta}^*\|_* + \|\widehat{\boldsymbol{\beta}}\|_*)(\|\boldsymbol{\beta}^*\|_* - \|\widehat{\boldsymbol{\beta}}\|_*) \leq \|\boldsymbol{t}^*\|_2^2(\|\boldsymbol{\beta}^*\|_* + \|\widehat{\boldsymbol{\beta}}\|_*)\|\widehat{\Delta}\|_*$$

by the reverse triangle inequality, and since

$$\|\boldsymbol{t}^*\|_2 \leq n^{1/2-1/p}\|\boldsymbol{t}^*\|_p \leq n^{1/2-1/p}n^{1/p}\delta = n^{1/2}\delta,$$

by a norm inequality, we get that $\|\boldsymbol{t}^*\|_2^2 \leq n\delta^2$ and so

$$\|\boldsymbol{t}^*\|_2^2(\|\boldsymbol{\beta}^*\|_*^2 - \|\widehat{\boldsymbol{\beta}}\|_*^2) \leq n\delta^2(\|\boldsymbol{\beta}^*\|_* + \|\widehat{\boldsymbol{\beta}}\|_*)\|\widehat{\Delta}\|_*.$$

Combining all these bounds, we get

$$\begin{aligned}\|X\widehat{\Delta}\|_2^2 \leq \delta n^{1/p}\|\boldsymbol{\varepsilon}\|_q\|\widehat{\Delta}\|_* + 2\|\boldsymbol{\varepsilon}\|_q n^{1/p}\delta\|\widehat{\Delta}\|_* + \\ 2\|\widehat{\boldsymbol{\beta}}\|_* nMr\delta\|\widehat{\Delta}\|_* + n\delta^2(\|\boldsymbol{\beta}^*\|_* + \|\widehat{\boldsymbol{\beta}}\|_*)\|\widehat{\Delta}\|_* = \\ \left(3\delta n^{1/p}\|\boldsymbol{\varepsilon}\|_q + 2\|\widehat{\boldsymbol{\beta}}\|_* nMr\delta + n\delta^2(\|\boldsymbol{\beta}^*\|_* + \|\widehat{\boldsymbol{\beta}}\|_*)\right)\|\widehat{\Delta}\|_*\end{aligned}$$

and therefore

$$\frac{1}{n}\|X\widehat{\Delta}\|_2^2 \leq \delta\left(3\frac{\|\boldsymbol{\varepsilon}\|_q}{n^{1/q}} + 2\|\widehat{\boldsymbol{\beta}}\|_* Mr + \delta(\|\boldsymbol{\beta}^*\|_* + \|\widehat{\boldsymbol{\beta}}\|_*)\right)\|\widehat{\Delta}\|_*.$$

Finally, we bound $\|\widehat{\boldsymbol{\beta}}\|_*$ using Lemma 2 to obtain

$$\frac{1}{n}\|X\widehat{\Delta}\|_2^2 \leq \delta\left(3\frac{\|\boldsymbol{\varepsilon}\|_q}{n^{1/q}} + (2MCr + \delta C + \delta)\|\boldsymbol{\beta}^*\|_*\right)\|\widehat{\Delta}\|_*$$

completing the proof. □

**Proof of Theorem 3.** We now prove Theorem 3:

*Proof.* We divide the proof into two cases, taking inspiration from the two cases in the proof of Theorem 2.3 in Xie & Huo (2024).

**Case 1: Assume $\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_1 + 2(\|\boldsymbol{\beta}^*\|_1 - \|\widehat{\boldsymbol{\beta}}\|_1) \leq 0$.** In this case, we will not need the RE condition. Indeed, note first that

$$\|\boldsymbol{\beta}^*\|_1 - \|\widehat{\boldsymbol{\beta}}\|_1 = \|\widehat{\boldsymbol{\beta}}\|_1 - \|\boldsymbol{\beta}^*\|_1 + 2(\|\boldsymbol{\beta}^*\|_1 - \|\widehat{\boldsymbol{\beta}}\|_1) \leq \|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_1 + 2(\|\boldsymbol{\beta}^*\|_1 - \|\widehat{\boldsymbol{\beta}}\|_1) \leq 0$$

Next, using Lemma 5 with $\|\cdot\|_* = \|\cdot\|_1$, we have:

$$\|X\widehat{\Delta}\|_2^2 \leq 2\boldsymbol{\varepsilon}^\top X\widehat{\Delta} + 2(\|\boldsymbol{\beta}^*\|_1 - \|\widehat{\boldsymbol{\beta}}\|_1)\langle|\boldsymbol{\varepsilon}|, \boldsymbol{t}^*\rangle + 2\|\widehat{\boldsymbol{\beta}}\|_1\langle|X\widehat{\Delta}|, \boldsymbol{t}^*\rangle + \|\boldsymbol{t}^*\|_2^2(\|\boldsymbol{\beta}^*\|_1^2 - \|\widehat{\boldsymbol{\beta}}\|_1^2).$$

We bound the first term as

$$2\boldsymbol{\varepsilon}^\top X\widehat{\Delta} \leq 2\|X^\top\boldsymbol{\varepsilon}\|_\infty\|\widehat{\Delta}\|_1 \leq \delta n^{1/p}\|\boldsymbol{\varepsilon}\|_q\|\widehat{\Delta}\|_1$$

using a dual norm inequality and the assumption $\delta > \bar{\delta} = 2\frac{\|X^\top\boldsymbol{\varepsilon}\|}{n^{1/p}\|\boldsymbol{\varepsilon}\|_q}$. Moreover, we bound the second term as

$$\langle|\boldsymbol{\varepsilon}|, \boldsymbol{t}^*\rangle \leq \|\boldsymbol{\varepsilon}\|_q\|\boldsymbol{t}^*\|_p \leq \|\boldsymbol{\varepsilon}\|_q n^{1/p}\delta,$$

using Hölder's inequality and the $p$-norm constraint on $\boldsymbol{t}^*$. Combining these bounds, we get

$$\begin{aligned}\|X\widehat{\Delta}\|_2^2 \leq \\ \delta n^{1/p}\|\boldsymbol{\varepsilon}\|_q\|\widehat{\Delta}\|_1 + 2(\|\boldsymbol{\beta}^*\|_1 - \|\widehat{\boldsymbol{\beta}}\|_1)\|\boldsymbol{\varepsilon}\|_q n^{1/p}\delta + 2\|\widehat{\boldsymbol{\beta}}\|_1\langle|X\widehat{\Delta}|, \boldsymbol{t}^*\rangle + \|\boldsymbol{t}^*\|_2^2(\|\boldsymbol{\beta}^*\|_1^2 - \|\widehat{\boldsymbol{\beta}}\|_1^2) = \\ \delta n^{1/p}\|\boldsymbol{\varepsilon}\|_q\big[\|\widehat{\Delta}\|_1 + 2(\|\boldsymbol{\beta}^*\|_1 - \|\widehat{\boldsymbol{\beta}}\|_1)\big] + 2\|\widehat{\boldsymbol{\beta}}\|_1\langle|X\widehat{\Delta}|, \boldsymbol{t}^*\rangle + \|\boldsymbol{t}^*\|_2^2(\|\boldsymbol{\beta}^*\|_1^2 - \|\widehat{\boldsymbol{\beta}}\|_1^2) \leq \\ 2\|\widehat{\boldsymbol{\beta}}\|_1\langle|X\widehat{\Delta}|, \boldsymbol{t}^*\rangle.\end{aligned}$$

where we used the assumption $\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_1 + 2(\|\boldsymbol{\beta}^*\|_1 - \|\widehat{\boldsymbol{\beta}}\|_1) \leq 0$ and that $\|\boldsymbol{\beta}^*\|_1 - \|\widehat{\boldsymbol{\beta}}\|_1 \leq 0$ in the last step. Moreover, by a norm inequality,

$$\|\boldsymbol{t}^*\|_2 \leq n^{1/2-1/p}\|\boldsymbol{t}^*\|_p \leq n^{1/2-1/p}n^{1/p}\delta = n^{1/2}\delta,$$

we get

$$\langle|X\widehat{\Delta}|, \boldsymbol{t}^*\rangle \leq \|X\widehat{\Delta}\|_2\|\boldsymbol{t}^*\|_2 \leq \|X\widehat{\Delta}\|_2 n^{1/2}\delta,$$

using the Cauchy–Schwarz inequality. Hence, $\|X\widehat{\Delta}\|_2^2 \leq 2\|\widehat{\boldsymbol{\beta}}\|_1\|X\widehat{\Delta}\|_2 n^{1/2}\delta$ which implies $\frac{1}{\sqrt{n}}\|X\widehat{\Delta}\|_2 \leq 2\delta\|\widehat{\boldsymbol{\beta}}\|_1$ and so

$$\frac{1}{n}\|X\widehat{\Delta}\|_2^2 \leq 4\delta^2\|\widehat{\boldsymbol{\beta}}\|_1^2.$$

We finally bound $\|\widehat{\boldsymbol{\beta}}\|_1^2$ using Lemma 2 yielding $\|\widehat{\boldsymbol{\beta}}\|_1^2 \leq C^2\|\boldsymbol{\beta}^*\|_1^2$ and

$$\frac{1}{n}\|X\widehat{\Delta}\|_2^2 \leq 4\delta^2 C^2\|\boldsymbol{\beta}^*\|_1^2.$$

**Case 2: Assume $\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_1 + 2(\|\boldsymbol{\beta}^*\|_1 - \|\widehat{\boldsymbol{\beta}}\|_1) \geq 0$.** In this case, we will use the RE condition. More precisely, we have

$$\|\widehat{\Delta}\|_1 \leq \|\widehat{\Delta}_S\|_1 + \|\widehat{\Delta}_{S^c}\|_1$$

and

$$\|\boldsymbol{\beta}^*\|_1 - \|\widehat{\boldsymbol{\beta}}\|_1 = \|\boldsymbol{\beta}^*_S\|_1 - \|\boldsymbol{\beta}^*_S + \widehat{\Delta}_S + \widehat{\Delta}_{S^c}\|_1 =$$
$$\|\boldsymbol{\beta}^*_S\|_1 - \|\boldsymbol{\beta}^*_S + \widehat{\Delta}_S\|_1 - \|\widehat{\Delta}_{S^c}\|_1 \leq \|\widehat{\Delta}_S\|_1 - \|\widehat{\Delta}_{S^c}\|_1.$$

Hence,

$$0 \leq \|\widehat{\Delta}\|_1 + 2(\|\boldsymbol{\beta}^*\|_1 - \|\widehat{\boldsymbol{\beta}}\|_1) \leq \|\widehat{\Delta}_S\|_1 + \|\widehat{\Delta}_{S^c}\|_1 + 2(\|\widehat{\Delta}_S\|_1 - \|\widehat{\Delta}_{S^c}\|_1)$$

which implies that $\|\widehat{\Delta}_{S^c}\|_1 \leq 3\|\widehat{\Delta}_S\|_1$. In other words, we can use $\mathrm{RE}(s, 3)$ to get

$$\|\widehat{\Delta}\|_1 = \|\widehat{\Delta}_S\|_1 + \|\widehat{\Delta}_{S^c}\|_1 \leq 4\|\widehat{\Delta}_S\|_1 \leq 4s\|\widehat{\Delta}_S\|_2 \leq 4s\frac{\|X\widehat{\Delta}\|_2}{\kappa\sqrt{n}},$$

where we also used the norm inequality $\|\widehat{\Delta}_S\|_1 \leq s\|\widehat{\Delta}_S\|_2$. This bound combined with Lemma 7 yields

$$\frac{1}{n}\|X\widehat{\Delta}\|_2^2 \leq \delta B\|\widehat{\Delta}\|_1 \leq 4\delta Bs\frac{\|X\widehat{\Delta}\|_2}{\kappa\sqrt{n}} \quad \Rightarrow \quad \frac{1}{\sqrt{n}}\|X\widehat{\Delta}\|_2 \leq \frac{4\delta Bs}{\kappa}.$$

Therefore, $\frac{1}{n}\|X\widehat{\Delta}\|_2^2 \leq \frac{16\delta^2 B^2 s^2}{\kappa^2}$, completing the second case.

Combining Cases 1 and 2, we conclude that

$$\frac{1}{n}\|X\widehat{\Delta}\|_2^2 \leq \delta^2 \cdot \max\left\{4C^2\|\boldsymbol{\beta}^*\|_1^2, \frac{16B^2s^2}{\kappa^2}\right\}.$$

That is, we have proved (11). It remains to show the rates. To this end, assume $\boldsymbol{\varepsilon} \sim N(0, \sigma I)$. We show that $B$ and $C$ are bounded by constants with high probability. We start with $C$. By a norm inequality and the proof of Theorem 2, we have with probability $\geq 1 - \gamma$:

$$\frac{\|\boldsymbol{\varepsilon}\|_q}{n^{1/q}} \geq \frac{\|\boldsymbol{\varepsilon}\|_1}{n} \geq \sigma\sqrt{\frac{2}{\pi}} - \sigma\sqrt{\frac{2\log(1/\gamma)}{n}}.$$

Therefore, with probability $\geq 1 - \gamma$:

$$C = \left(3 + \frac{n^{1/q}\delta}{\|\boldsymbol{\varepsilon}\|_q}\|\boldsymbol{\beta}^*\|_*\right) \leq \left(3 + \frac{\delta}{\sigma\sqrt{\frac{2}{\pi}} - \sigma\sqrt{\frac{2\log(1/\gamma)}{n}}}\|\boldsymbol{\beta}^*\|_*\right),$$

valid for $n > \pi\log(1/\gamma)$ (to ensure the denominator is positive). In other words, $C$ is bounded by a constant with high probability.[7] Moreover, by the proof of Theorem 2, we have with probability $\geq 1 - \gamma$:

$$\frac{\|\boldsymbol{\varepsilon}\|_q}{n^{1/q}} \leq \sigma + \sigma\sqrt{\frac{2\log(1/\gamma)}{n}}.$$

[7] For the finitely many $n$ such that $n \leq \pi\log(1/\gamma)$, we have that $C$ is almost surely bounded.

Hence, applying a union bound, we have with probability $\geq 1-2\gamma$:

$$B = 3\frac{\|\boldsymbol{\varepsilon}\|_q}{n^{1/q}} + (2M+\delta)\|\boldsymbol{\beta}^*\|_* C + \delta\|\boldsymbol{\beta}^*\|_* \leq$$

$$3\left(\sigma + \sigma\sqrt{\frac{2\log(1/\gamma)}{n}}\right) + (2M+\delta)\|\boldsymbol{\beta}^*\|_* \left(3 + \frac{\delta}{\sigma\sqrt{\frac{2}{\pi}} - \sigma\sqrt{\frac{2\log(1/\gamma)}{n}}}\|\boldsymbol{\beta}^*\|_*\right) + \delta\|\boldsymbol{\beta}^*\|_*$$

valid for $n > \pi\log(1/\gamma)$. We conclude that $B$ is also bounded by a constant with high probability. As a result, we have that $\frac{1}{n}\|X\widehat{\Delta}\|_2^2 \in O(\delta^2)$ with high probability ($\geq 1-2\gamma$), independent of $p$ since the upper bounds on $B$ and $C$ are independent of $p$. Finally, by setting $\delta$ as in Theorem 2 and following its proof (where $\delta > \bar{\delta}$ with probability $\geq 1-3\gamma$), we analogously obtain that $\frac{1}{n}\|X\widehat{\Delta}\|_2^2 \in O(\delta^2) = O(1/n)$ for $K > 2c\sqrt{\pi}$ with probability $\geq 1-5\gamma$ due to a union bound. This completes the proof. □

# D Small and large ambiguity balls: Proofs and additional details

In this section, we provide the proofs and additional details of Section 6.

## D.1 Proof of Theorem 4

To prove Theorem 4, we will use a generalization of Danskin's theorem (Danskin, 1967) proved in the PhD thesis of Dimitri P. Bertsekas (Bertsekas, 1971, Proposition A.22). Indeed, Theorem 6 is a straightforward special case of (Bertsekas, 1971, Proposition A.22) adapted to our setting:[8]

**Theorem 6** (Bertsekas, 1971, Proposition A.22, Simplified)**.** Let $\phi : \mathbb{R}^d \times \mathbb{R}^n \to \mathbb{R}$ be a continuous function such that $\phi(\cdot, \boldsymbol{t})$ is convex for each $\boldsymbol{t} \in T$ in a compact set $T$, and consider $f(\boldsymbol{\beta}) := \sup_{\boldsymbol{t}\in T} \phi(\boldsymbol{\beta}, \boldsymbol{t})$. Then, for each $\boldsymbol{\beta} \in R^d$, the subdifferential of $f$ equals

$$\partial f(\boldsymbol{\beta}) = \operatorname{conv}\{\partial_{\boldsymbol{\beta}}\phi(\boldsymbol{\beta}, \boldsymbol{t}) : \boldsymbol{t} \in T(\boldsymbol{\beta})\}, \quad T(\boldsymbol{\beta}) := \{\boldsymbol{t} \in T : \phi(\boldsymbol{\beta}, \boldsymbol{t}) = \max_{\boldsymbol{t}\in T} \phi(\boldsymbol{\beta}, \boldsymbol{t})\}, \tag{A.12}$$

where $\partial_{\boldsymbol{\beta}}\phi(\boldsymbol{\beta}, \boldsymbol{t})$ is the subdifferential of $\phi(\cdot, \boldsymbol{t})$ at $\boldsymbol{\beta}$.

*Proof of Theorem 4.* Note that $\widehat{\boldsymbol{\beta}} = \mathbf{0}$ minimizes the robust risk $V_\delta(\boldsymbol{\beta})$ if and only if $\mathbf{0} \in \partial V_\delta(\mathbf{0})$, since $V_\delta(\boldsymbol{\beta})$ is convex in $\boldsymbol{\beta}$. We compute $\partial V_\delta(\mathbf{0})$ using Theorem 6. More precisely, note that $nV_\delta(\boldsymbol{\beta}) = \sup_{\boldsymbol{t}\in T} \phi(\boldsymbol{\beta}, \boldsymbol{t})$ with

$$\phi(\boldsymbol{\beta}, \boldsymbol{t}) = \sum_{i=1}^n \left(|\boldsymbol{x}_i^\top\boldsymbol{\beta} - y_i| + t_i\|\boldsymbol{\beta}\|_*\right)^2$$
$$T = \left\{\boldsymbol{t} \in \mathbb{R}^n : t_i \geq 0, \|\boldsymbol{t}\|_p \leq n^{1/p}\delta\right\}$$

It is easy to verify that the conditions of Theorem 6 hold, yielding:

$$\partial_{\boldsymbol{\beta}}(nV_\delta(\mathbf{0})) = \operatorname{conv}\{\partial_{\boldsymbol{\beta}}\phi(\mathbf{0}, \boldsymbol{t}) : \boldsymbol{t} \in T(\mathbf{0})\}. \tag{A.13}$$

We compute $\partial_{\boldsymbol{\beta}}\phi(\mathbf{0}, \boldsymbol{t})$ and $T(\mathbf{0})$. In general,

$$\partial_{\boldsymbol{\beta}}\phi(\boldsymbol{\beta}, \boldsymbol{t}) = 2\sum_{i=1}^n \left(|\boldsymbol{x}_i^\top\boldsymbol{\beta} - y_i| + t_i\|\boldsymbol{\beta}\|_*\right)\left(\partial_{\boldsymbol{\beta}}|\boldsymbol{x}_i^\top\boldsymbol{\beta} - y_i| + t_i\partial_{\boldsymbol{\beta}}\|\boldsymbol{\beta}\|_*\right), \tag{A.14}$$

[8] The subdifferential of a function $\varphi : \mathbb{R}^d \to \mathbb{R}$ at a point $\boldsymbol{z} \in \mathbb{R}^d$ is the set

$$\{\boldsymbol{v} \in \mathbb{R}^d : \varphi(\boldsymbol{z}') - \varphi(\boldsymbol{z}) \geq \langle \boldsymbol{v}, \boldsymbol{z}' - \boldsymbol{z}\rangle, \forall \boldsymbol{z}' \in \mathbb{R}^d\}.$$

For standard properties of subdifferentials, see, e.g., (Bertsekas et al., 2003; Clarke, 1990; Boyd et al., 2022).

where

$$\partial_{\boldsymbol{\beta}}|\boldsymbol{x}_i^\top\boldsymbol{\beta}-y_i| = \boldsymbol{x}_i\text{sign}(\boldsymbol{x}_i^\top\boldsymbol{\beta}-y_i) \quad \text{with} \quad \text{sign}(a) := \begin{cases}\{1\} & \text{if } a>0\\ \{-1\} & \text{if } a<0\\ [-1,1] & \text{if } a=0\end{cases}$$

$$\partial_{\boldsymbol{\beta}}\|\boldsymbol{\beta}\|_* = \{\boldsymbol{z}\in\mathbb{R}^d : \|\boldsymbol{z}\|\leq 1, \langle \boldsymbol{z},\boldsymbol{\beta}\rangle = \|\boldsymbol{\beta}\|_*\}$$

by standard derivations.[9] In particular, for $\boldsymbol{\beta}=\mathbf{0}$, we have $\partial_{\boldsymbol{\beta}}\|\boldsymbol{\beta}\|_* = \mathbb{B} := \{\boldsymbol{z}\in\mathbb{R}^d : \|\boldsymbol{z}\|\leq 1\}$ and $\partial_{\boldsymbol{\beta}}|\boldsymbol{x}_i^\top\boldsymbol{\beta}-y_i| = -\boldsymbol{x}_i\text{sign}(y_i)$. Hence,

$$\partial_{\boldsymbol{\beta}}\phi(\mathbf{0},\boldsymbol{t}) = 2\sum_{i=1}^n |y_i|\left(-\boldsymbol{x}_i\text{sign}(y_i)+t_i\mathbb{B}\right) =$$

$$2\sum_{i=1}^n(-\boldsymbol{x}_i y_i) + 2\sum_{i=1}^n(t_i|y_i|\mathbb{B}) = 2\big(-X^T\boldsymbol{y}+\langle \boldsymbol{t},|\boldsymbol{y}|\rangle\mathbb{B}\big)$$

where we used that $|y_i|\text{sign}(y_i)=y_i$, and that the Minkowski sum of balls with the same center (the origin) equals the ball with their radii added. Next, we note that $\phi(\mathbf{0},\boldsymbol{t}) = \sum_{i=1}^n y_i^2$ for each $\boldsymbol{t}\in T$, so $T(\mathbf{0})=T$. Therefore, (A.13) yields

$$\partial_{\boldsymbol{\beta}}(nV_\delta(\mathbf{0})) = \text{conv}\left[\bigcup_{\boldsymbol{t}\in T} 2\big(-X^\top\boldsymbol{y}+\langle\boldsymbol{t},|\boldsymbol{y}|\rangle\mathbb{B}\big)\right] =$$

$$-X^\top\boldsymbol{y} + \big(\max_{\boldsymbol{t}\in T}\langle\boldsymbol{t},|\boldsymbol{y}|\rangle\big)\mathbb{B} = -X^\top\boldsymbol{y} + n^{1/p}\delta\|\boldsymbol{y}\|_q\mathbb{B}$$

since the convex hull of a union of balls with the same center equals the ball with largest radius (i.e., $\max_{\boldsymbol{t}\in T}\langle\boldsymbol{t},|\boldsymbol{y}|\rangle$ in our case) and

$$\max_{\boldsymbol{t}\in T}\langle\boldsymbol{t},|\boldsymbol{y}|\rangle = n^{1/p}\delta\max_{\|\boldsymbol{t}\|_p\leq 1}\langle\boldsymbol{t},|\boldsymbol{y}|\rangle = n^{1/p}\delta\|\boldsymbol{y}\|_q.$$

Therefore, $\mathbf{0}\in\partial_{\boldsymbol{\beta}}V_\delta(\mathbf{0})$ if and only if

$$\mathbf{0}\in -X^\top\boldsymbol{y}+n^{1/p}\delta\|\boldsymbol{y}\|_q\mathbb{B} \quad\Leftrightarrow\quad \|X^\top\boldsymbol{y}\|\leq n^{1/p}\delta\|\boldsymbol{y}\|_q \quad\Leftrightarrow\quad \frac{\|X^\top\boldsymbol{y}\|}{n^{1/p}\|\boldsymbol{y}\|_q}\leq\delta$$

completing the proof. □

### D.2 Proof of Theorem 5

We continue with the proof of Theorem 5. We need the following lemma:

**Lemma 8.** Let $\bar{\boldsymbol{\beta}}$ be the minimum $\|\cdot\|_*$-norm interpolator. Then $X^T\widehat{\boldsymbol{\alpha}}\in\partial\|\bar{\boldsymbol{\beta}}\|_*$ if and only if $\widehat{\boldsymbol{\alpha}}$ solves $\max_{\|X^\top\boldsymbol{\alpha}\|\leq 1}\boldsymbol{\alpha}^\top\boldsymbol{y}$.

*Proof.* Analogously to the proof of Lemma 1 in Ribeiro et al. (2023), we have by strong duality that

$$\|\bar{\boldsymbol{\beta}}\|_* = \min_{X\boldsymbol{\beta}=\boldsymbol{y}}\|\boldsymbol{\beta}\|_* = \max_{\boldsymbol{\alpha}}\min_{\boldsymbol{\beta}}\left(\|\boldsymbol{\beta}\|_*+\boldsymbol{\alpha}^\top(\boldsymbol{y}-X\boldsymbol{\beta})\right) =$$

$$\max_{\boldsymbol{\alpha}}\left(\boldsymbol{\alpha}^\top\boldsymbol{y}+\min_{\boldsymbol{\beta}}(\|\boldsymbol{\beta}\|_*-\boldsymbol{\alpha}^\top X\boldsymbol{\beta})\right) = \max_{\|X^\top\boldsymbol{\alpha}\|\leq 1}\boldsymbol{\alpha}^\top\boldsymbol{y},$$

where the last step used that the Fenchel conjugate of $\|\boldsymbol{\beta}\|_*$ is the indicator function on the unit ball with norm $\|\cdot\|$. Moreover,

$$\partial\|\bar{\boldsymbol{\beta}}\|_* = \{\boldsymbol{z}\in\mathbb{R}^d : \|\boldsymbol{z}\|\leq 1, \boldsymbol{z}^\top\bar{\boldsymbol{\beta}} = \|\bar{\boldsymbol{\beta}}\|_*\}$$

so $X^\top\widehat{\boldsymbol{\alpha}}\in\partial\|\bar{\boldsymbol{\beta}}\|_*$ implies that $\|X^\top\widehat{\boldsymbol{\alpha}}\|\leq 1$ and

$$\widehat{\boldsymbol{\alpha}}^\top\boldsymbol{y} = \widehat{\boldsymbol{\alpha}}^\top X\bar{\boldsymbol{\beta}} = (X^\top\widehat{\boldsymbol{\alpha}})^\top\bar{\boldsymbol{\beta}} = \|\bar{\boldsymbol{\beta}}\|_* = \max_{\|X^\top\boldsymbol{\alpha}\|\leq 1}\boldsymbol{\alpha}^\top\boldsymbol{y}.$$

[9] Note that we are working with sets where addition is the Minkowski addition, standard for subdifferentials.

That is, $\widehat{\boldsymbol{\alpha}}$ solves $\max_{\|X^\top \boldsymbol{\alpha}\| \le 1} \boldsymbol{\alpha}^\top \boldsymbol{y}$. Conversely, assume $\widehat{\boldsymbol{\alpha}}$ solves $\max_{\|X^\top \boldsymbol{\alpha}\| \le 1} \boldsymbol{\alpha}^\top \boldsymbol{y}$. Then $\|X^\top \widehat{\boldsymbol{\alpha}}\| \le 1$ and

$$(X^\top \widehat{\boldsymbol{\alpha}})^\top \bar{\boldsymbol{\beta}} = \widehat{\boldsymbol{\alpha}}^\top \boldsymbol{y} = \max_{\|X^\top \boldsymbol{\alpha}\| \le 1} \boldsymbol{\alpha}^\top \boldsymbol{y} = \|\bar{\boldsymbol{\beta}}\|_*$$

so $X^\top \widehat{\boldsymbol{\alpha}} \in \partial \|\bar{\boldsymbol{\beta}}\|_*$, completing the proof. □

*Proof of Theorem 5.* Note that $\bar{\boldsymbol{\beta}}$ minimizes $V_\delta(\boldsymbol{\beta})$ if and only if $\mathbf{0} \in \partial_{\boldsymbol{\beta}} V_\delta(\bar{\boldsymbol{\beta}})$. We divide the proof into two subcases: $2 < p \le \infty$ and $p = 2$.

Assume first that $2 < p \le \infty$. We compute $\partial_{\boldsymbol{\beta}} V_\delta(\bar{\boldsymbol{\beta}})$ using Theorem 6. To this end, let $\phi$ and $T$ be as in the proof of Theorem 4, so that $nV_\delta(\boldsymbol{\beta}) = \sup_{\boldsymbol{t} \in T} \phi(\boldsymbol{\beta}, \boldsymbol{t})$. Theorem 6 implies that

$$\partial(nV_\delta(\bar{\boldsymbol{\beta}})) = \operatorname{conv}\{\partial_{\boldsymbol{\beta}} \phi(\bar{\boldsymbol{\beta}}, \boldsymbol{t}) : \boldsymbol{t} \in T(\bar{\boldsymbol{\beta}})\}. \tag{A.15}$$

We compute $\partial_{\boldsymbol{\beta}} \phi(\bar{\boldsymbol{\beta}}, \boldsymbol{t})$ and $T(\bar{\boldsymbol{\beta}})$. By (A.14), we have

$$\partial_{\boldsymbol{\beta}} \phi(\bar{\boldsymbol{\beta}}, \boldsymbol{t}) = 2 \sum_{i=1}^n t_i \|\bar{\boldsymbol{\beta}}\|_* (\boldsymbol{x}_i \mathbb{I} + t_i \partial_{\boldsymbol{\beta}} \|\bar{\boldsymbol{\beta}}\|_*)$$

since $X\bar{\boldsymbol{\beta}} = \boldsymbol{y}$, where $\mathbb{I} = [-1, 1]$. Moreover,

$$\phi(\bar{\boldsymbol{\beta}}, \boldsymbol{t}) = \sum_{i=1}^n t_i^2 \|\bar{\boldsymbol{\beta}}\|_*^2 = \|\bar{\boldsymbol{\beta}}\|_*^2 \|\boldsymbol{t}\|_2^2$$

and therefore

$$\sup_{\boldsymbol{t} \in T} \phi(\bar{\boldsymbol{\beta}}, \boldsymbol{t}) = \|\bar{\boldsymbol{\beta}}\|_*^2 \sup_{\boldsymbol{t} \in T} \|\boldsymbol{t}\|_2^2 = \|\bar{\boldsymbol{\beta}}\|_*^2 n\delta^2$$

with the supremum attained only at $t_i = \delta$ due to a norm inequality (recall that $t_i \ge 0$ in $T$). That is, $T(\bar{\boldsymbol{\beta}})$ equals the singleton $\bar{\boldsymbol{t}}$ with entries $\bar{t}_i = \delta$.[10] Hence, (A.15) yields

$$\partial_{\boldsymbol{\beta}} nV_\delta(\bar{\boldsymbol{\beta}}) = \partial_{\boldsymbol{\beta}} \phi(\bar{\boldsymbol{\beta}}, \bar{\boldsymbol{t}}) = 2 \sum_{i=1}^n \delta \|\bar{\boldsymbol{\beta}}\|_* (\boldsymbol{x}_i \mathbb{I} + \delta \partial_{\boldsymbol{\beta}} \|\bar{\boldsymbol{\beta}}\|_*).$$

Pick an arbitrary point in this set:

$$2 \sum_{i=1}^n \delta \|\bar{\boldsymbol{\beta}}\|_* (\boldsymbol{x}_i \rho_i + \delta \boldsymbol{g}_{\boldsymbol{i}}), \tag{A.16}$$

where $\rho_i \in \mathbb{I} = [-1, 1]$ and $\boldsymbol{g}_{\boldsymbol{i}} \in \partial_{\boldsymbol{\beta}} \|\bar{\boldsymbol{\beta}}\|_*$. Since $\partial_{\boldsymbol{\beta}} \|\bar{\boldsymbol{\beta}}\|_*$ is a convex set, we have $\sum_{i=1}^n \boldsymbol{g}_{\boldsymbol{i}} = n\boldsymbol{g}$ for some $\boldsymbol{g} \in \partial_{\boldsymbol{\beta}} \|\bar{\boldsymbol{\beta}}\|_*$, which simplifies (A.16) to

$$2\delta \|\bar{\boldsymbol{\beta}}\|_* \left[ \sum_{i=1}^n (\boldsymbol{x}_i \rho_i) + \delta n \boldsymbol{g} \right] = 2\delta \|\bar{\boldsymbol{\beta}}\|_* \left[ X^\top \boldsymbol{\rho} + \delta n \boldsymbol{g} \right].$$

Thus, $\mathbf{0} \in \partial_{\boldsymbol{\beta}} nV_\delta(\bar{\boldsymbol{\beta}})$ if and only if

$$\boldsymbol{g} = -\frac{X^\top \boldsymbol{\rho}}{n\delta}$$

for some $\boldsymbol{g} \in \partial_{\boldsymbol{\beta}} \|\bar{\boldsymbol{\beta}}\|_*$ and $\|\boldsymbol{\rho}\|_\infty \le 1$. By Lemma 8, this is if and only if there exists some $\|\boldsymbol{\rho}\|_\infty \le 1$ such that

$$\widehat{\boldsymbol{\alpha}} := -\frac{\boldsymbol{\rho}}{n\delta}$$

solves $\max_{\|X^\top \boldsymbol{\alpha}\| \le 1} \boldsymbol{\alpha}^\top \boldsymbol{y}$. Equivalently, there exists $\widehat{\boldsymbol{\alpha}}$ solving $\max_{\|X^\top \boldsymbol{\alpha}\| \le 1} \boldsymbol{\alpha}^\top \boldsymbol{y}$ such that

$$\|\widehat{\boldsymbol{\alpha}}\|_\infty \le \frac{1}{n\delta}.$$

[10] This argument holds provided $p > 2$. For $p = 2$, the set $T(\bar{\boldsymbol{\beta}})$ has more than one element since the supremum $\sup_{t \in T} \|t\|_2^2$ has multiple solutions, which is why we consider the case $p = 2$ separately.

Crucially, this condition holds if and only if

$$\|\bar{\boldsymbol{\alpha}}\|_\infty \leq \frac{1}{n\delta}$$

where $\bar{\boldsymbol{\alpha}}$ solves $\max_{\|X^\top \boldsymbol{\alpha}\| \leq 1} \boldsymbol{\alpha}^\top \boldsymbol{y}$ and has the *minimum* $\infty$-norm among all such solutions. In other words, $\bar{\boldsymbol{\alpha}}$ solves $\min_{\boldsymbol{\alpha} \in Q} \|\boldsymbol{\alpha}\|_\infty$ with $Q = \arg\max_{\|X^\top \boldsymbol{\alpha}\| \leq 1} \boldsymbol{\alpha}^\top \boldsymbol{y}$, completing the proof for $2 < p \leq \infty$.

We now continue with the case $p = 2$. By Corollary 1, we have that

$$nV_\delta(\boldsymbol{\beta}) = \left( \sqrt{\frac{1}{n} \sum_{i=1}^n (\boldsymbol{x}_i^\top \boldsymbol{\beta} - y_i)^2} + \delta \|\boldsymbol{\beta}\|_* \right)^2 = \left( \frac{1}{\sqrt{n}} \|X\boldsymbol{\beta} - \boldsymbol{y}\|_2 + \delta \|\boldsymbol{\beta}\|_* \right)^2$$

To ease the computation of subgradients, we consider instead the equivalent objective

$$\sqrt{nV_\delta(\boldsymbol{\beta})} = \frac{1}{\sqrt{n}} \|X\boldsymbol{\beta} - \boldsymbol{y}\|_2 + \delta \|\boldsymbol{\beta}\|_*.$$

Indeed, $\bar{\boldsymbol{\beta}}$ minimizes $nV_\delta(\boldsymbol{\beta})$ if and only if $\bar{\boldsymbol{\beta}}$ minimizes $\sqrt{nV_\delta(\boldsymbol{\beta})}$, which in turn is if and only if $\boldsymbol{0} \in \partial_{\boldsymbol{\beta}} \sqrt{nV_\delta(\bar{\boldsymbol{\beta}})}$. We compute $\partial_{\boldsymbol{\beta}} \sqrt{nV_\delta(\bar{\boldsymbol{\beta}})}$:

$$\partial_{\boldsymbol{\beta}} \sqrt{nV_\delta(\bar{\boldsymbol{\beta}})} = \frac{1}{\sqrt{n}} \partial_{\boldsymbol{\beta}} \|X\bar{\boldsymbol{\beta}} - \boldsymbol{y}\|_2 + \delta \partial_{\boldsymbol{\beta}} \|\bar{\boldsymbol{\beta}}\|_*. \tag{A.17}$$

By the chain rule for affine transformations (see, e.g., (Bertsekas, 2009, Proposition 5.4.5)), we have that

$$\partial_{\boldsymbol{\beta}} \|X\bar{\boldsymbol{\beta}} - \boldsymbol{y}\|_2 = X^T \partial_{\boldsymbol{r}} \|\boldsymbol{r}\|_2,$$

where $\partial_{\boldsymbol{r}} \|\boldsymbol{r}\|_2$ is the subgradient of the norm $\|\cdot\|_2$ at $\boldsymbol{r} = X\bar{\boldsymbol{\beta}} - y$. Since $\boldsymbol{r} = X\bar{\boldsymbol{\beta}} - \boldsymbol{y} = \boldsymbol{0}$, we get $\partial_{\boldsymbol{r}} \|\boldsymbol{r}\|_2 = \{\boldsymbol{u} : \|\boldsymbol{u}\|_2 \leq 1\}$ and therefore

$$\partial_{\boldsymbol{\beta}} \|X\bar{\boldsymbol{\beta}} - \boldsymbol{y}\|_2 = \{X^\top \boldsymbol{u} : \|\boldsymbol{u}\|_2 \leq 1\}.$$

Thus, the subgradient in (A.17) contains zero if and only if

$$\frac{1}{\sqrt{n}} X^\top \boldsymbol{u} + \delta \boldsymbol{g} = \boldsymbol{0} \quad \Leftrightarrow \quad \boldsymbol{g} = \frac{1}{\delta\sqrt{n}} X^\top \boldsymbol{u}$$

for some $\|\boldsymbol{u}\|_2 \leq 1$ and $\boldsymbol{g} \in \partial_{\boldsymbol{\beta}} \|\bar{\boldsymbol{\beta}}\|_*$. By Lemma 8, this is if and only if there exists $\boldsymbol{u} \in \mathbb{R}^n$ such that $\|\boldsymbol{u}\|_2 \leq 1$ and

$$\widehat{\boldsymbol{\alpha}} := -\frac{\boldsymbol{u}}{\delta\sqrt{n}}$$

solves $\max_{\|X^\top \boldsymbol{\alpha}\| \leq 1} \boldsymbol{\alpha}^\top \boldsymbol{y}$. Equivalently, there exists $\widehat{\boldsymbol{\alpha}}$ solving $\max_{\|X^\top \boldsymbol{\alpha}\| \leq 1} \boldsymbol{\alpha}^\top \boldsymbol{y}$ such that

$$\|\widehat{\boldsymbol{\alpha}}\|_2 \leq \frac{1}{\delta\sqrt{n}}.$$

Similarly to the case $2 < p \leq \infty$, this holds if and only if

$$\|\bar{\boldsymbol{\alpha}}\|_2 \leq \frac{1}{\delta\sqrt{n}},$$

where $\bar{\boldsymbol{\alpha}}$ solves $\max_{\|X^\top \boldsymbol{\alpha}\| \leq 1} \boldsymbol{\alpha}^\top \boldsymbol{y}$ and has the *minimum* 2-norm among all such solutions. In other words, $\bar{\boldsymbol{\alpha}}$ solves $\min_{\boldsymbol{\alpha} \in Q} \|\boldsymbol{\alpha}\|_2$ with $Q = \arg\max_{\|X^\top \boldsymbol{\alpha}\| \leq 1} \boldsymbol{\alpha}^\top \boldsymbol{y}$, completing the proof for $p = 2$. Combining both cases completes the proof in general. □

### D.3 Computing $\delta_S$ via a convex program

We conclude by showing that $\delta_S$ in Theorem 5 is easy to compute via a convex program:

*Remark* 3 (Computing $\delta_S$). For brevity, let $\bar{\boldsymbol{\alpha}}$ denote any solution of $\min_{\boldsymbol{\alpha}\in Q}\|\boldsymbol{\alpha}\|_\infty$ for $2<p\leq\infty$ and $\min_{\boldsymbol{\alpha}\in Q}\|\boldsymbol{\alpha}\|_2$ for $p=2$. To compute the bound $\delta_S$ in Theorem 5, we need to compute $\bar{\boldsymbol{\alpha}}$ (any will do). Fortunately, $\bar{\boldsymbol{\alpha}}$ can be easily obtained from a simple convex program. Indeed, solve the inner maximization $v^\star := \max_{\|X^\top\boldsymbol{\alpha}\|\leq 1}\boldsymbol{\alpha}^\top\boldsymbol{y}$ first, which is a convex program with linear objective and a compact convex constraint $\|X^\top\boldsymbol{\alpha}\|\leq 1$ since $X^\top$ is injective ($X$ has full rank). Then solve the outer minimization $\min_{\boldsymbol{\alpha}\in Q}\|\boldsymbol{\alpha}\|_\infty$ or $\min_{\boldsymbol{\alpha}\in Q}\|\boldsymbol{\alpha}\|_2$, with convex and affine constraints $Q=\{\boldsymbol{\alpha}:\|X^\top\boldsymbol{\alpha}\|\leq 1,\boldsymbol{\alpha}^\top\boldsymbol{y}=v^\star\}$ to obtain $\bar{\boldsymbol{\alpha}}$. Furthermore, for $\boldsymbol{y}\neq 0$, we have a unique solution for the inner maximization provided the set $\{\boldsymbol{\alpha}:\|X^\top\boldsymbol{\alpha}\|\leq 1\}$ is strictly convex.[11] In this case, $\bar{\boldsymbol{\alpha}}$ equals this unique solution (for any $p$). An example is the Euclidean norm $\|\cdot\|=\|\cdot\|_2$ where an easy derivation yields $\bar{\boldsymbol{\alpha}}=(XX^T)^{-1}\boldsymbol{y}/\sqrt{\boldsymbol{y}^\top(XX^T)^{-1}\boldsymbol{y}}$ (with $XX^T$ invertible since $X$ has full rank).[12]

# E Numerical solvers: Additional details

In this section, we provide additional details for the numerical solvers described in Section 7. More precisely, we present two solvers that minimize the robust risk, the saddle-point solver (Appendix E.1) and $\eta$-trick solver (Appendix E.2), and one solver that instead evaluates the robust risk quickly, by exploiting the one-dimensional subproblems (Appendix E.3).

## E.1 Saddle-point formulation and solver

We use the following saddle-point formulation of the robust risk:

**Theorem 7** (Saddle-point formulation of robust risk)**.** The robust risk of Wasserstein DRO linear regression ($2\leq p<\infty$) with square loss equals

$$nV_\delta(\boldsymbol{\beta})=\sup_{\boldsymbol{\alpha}\geq 0}K(\boldsymbol{\beta},\boldsymbol{\alpha}):=\sup_{\boldsymbol{\alpha}\geq 0}\left[n^{1/p}\delta\|\boldsymbol{\beta}\|_*\|\boldsymbol{\alpha}\|_q+\sum_{i=1}^n\alpha_i|\boldsymbol{x}_i^T\boldsymbol{\beta}-y_i|-\alpha_i^2/4\right]. \quad \text{(A.18)}$$

Moreover, by applying the bijection $\gamma_i=\alpha_i^q$, where $1/p+1/q=1$, we get the following form:

$$nV_\delta(\boldsymbol{\beta})=\sup_{\boldsymbol{\gamma}\geq 0}K(\boldsymbol{\beta},\boldsymbol{\gamma}^{1/q})=\sup_{\boldsymbol{\gamma}\geq 0}\left[n^{1/p}\delta\|\boldsymbol{\beta}\|_*\|\boldsymbol{\gamma}\|_1^{1/q}+\sum_{i=1}^n\gamma_i^{1/q}|\boldsymbol{x}_i^T\boldsymbol{\beta}-y_i|-\gamma_i^{2/q}/4\right] \quad \text{(A.19)}$$

Crucially, $K(\boldsymbol{\beta},\boldsymbol{\gamma}^{1/q})$ is a convex-concave function with respect to $(\boldsymbol{\beta},\gamma)$. Thus, minimizing the robust risk is equivalent to a convex-concave optimization problem $\min_{\boldsymbol{\beta}}nV_\delta(\boldsymbol{\beta})=\min_{\boldsymbol{\beta}}\sup_{\boldsymbol{\gamma}\geq 0}H(\boldsymbol{\beta},\boldsymbol{\gamma})$, where $H(\boldsymbol{\beta},\boldsymbol{\gamma}):=K(\boldsymbol{\beta},\boldsymbol{\gamma}^{1/q})$.

The concave-convex formulation allows us to write

$$\min_{\boldsymbol{\beta}}nV_\delta(\boldsymbol{\beta})=\min_{\boldsymbol{\beta}}\sup_{\boldsymbol{\gamma}\geq 0}H(\boldsymbol{\beta},\boldsymbol{\gamma}),$$

and we solve the problem using disciplined saddle-point programming (Schiele et al., 2024). We also use this result in the proof of Corollary 1 in Appendix B.2.

*Proof.* We first note that we do not need to prove (A.19). Indeed, (A.19) follows from (A.18) by using the bijection $\gamma_i=\alpha_i^q$. Moreover, $\|\boldsymbol{\gamma}\|_1^{1/q}$, $\gamma_i^{1/q}$ and $-\gamma_i^{2/q}$ are all concave in $\boldsymbol{\gamma}$ since $1<q\leq 2$, so $K(\boldsymbol{\beta},\boldsymbol{\gamma}^{1/q})$ is concave in $\boldsymbol{\gamma}\geq 0$. Finally, $K(\boldsymbol{\beta},\boldsymbol{\alpha})$ is convex in $\boldsymbol{\beta}$ (since $\|\boldsymbol{\beta}\|_*$ and $|\boldsymbol{x}_i^T\boldsymbol{\beta}-y_i|$ are convex), so $K(\boldsymbol{\beta},\boldsymbol{\gamma}^{1/q})$ is also convex in $\boldsymbol{\beta}$. Thus, $K(\boldsymbol{\beta},\boldsymbol{\gamma}^{1/q})$ is a convex-concave

[11] Indeed, the set $\{\boldsymbol{\alpha}:\|X^\top\boldsymbol{\alpha}\|\leq 1\}$ contains a ball around the origin and thus spans the directions where $\boldsymbol{\alpha}^\top\boldsymbol{y}$ increases. Thus, strict convexity of $\{\boldsymbol{\alpha}:\|X^\top\boldsymbol{\alpha}\|\leq 1\}$ guarantees uniqueness of $\max_{\|X^\top\boldsymbol{\alpha}\|\leq 1}\boldsymbol{\alpha}^\top\boldsymbol{y}$.

[12] To see this, let $M:=XX^T$. By Cauchy–Schwarz $\boldsymbol{\alpha}^\top\boldsymbol{y}=(X^\top\boldsymbol{\alpha})^\top(X^\top M^{-1}\boldsymbol{y})\leq\|X^\top\boldsymbol{\alpha}\|_2\|X^\top M^{-1}\boldsymbol{y}\|_2$ with equality iff $X^\top\boldsymbol{\alpha}=cX^\top M^{-1}\boldsymbol{y}$ for some $c>0$. The constraint $\|X^\top\boldsymbol{\alpha}\|_2=1$ (attained at optimality) yields $c=1/\|X^\top M^{-1}\boldsymbol{y}\|_2$ and thus $X^\top\boldsymbol{\alpha}=cX^\top M^{-1}\boldsymbol{y}$ iff $\boldsymbol{\alpha}=cM^{-1}\boldsymbol{y}=M^{-1}\boldsymbol{y}/\|X^\top M^{-1}\boldsymbol{y}\|_2$ (since $X^\top$ is injective), completing the derivation.

function with respect to $(\boldsymbol{\beta}, \boldsymbol{\gamma})$. Therefore, what remains to prove is (A.18). Theorem 1 yields

$$nV_\delta(\boldsymbol{\beta}) = \max_{\boldsymbol{t}\geq 0, \|t\|_p^p \leq n\delta^p} \sum_{i=1}^n \Big(\|\boldsymbol{\beta}\|_* t_i + |\boldsymbol{x}_i^\top \boldsymbol{\beta} - y_i|\Big)^2 =$$

$$\max_{\boldsymbol{t}\geq 0, \|t\|_p \leq 1} \sum_{i=1}^n \Big(n^{1/p}\delta\|\boldsymbol{\beta}\|_* t_i + |\boldsymbol{x}_i^\top \boldsymbol{\beta} - y_i|\Big)^2 =$$

$$\max_{\boldsymbol{t}\geq 0, \|t\|_p \leq 1} \left[\sum_{i=1}^n \sup_{\alpha_i \geq 0} \Big(n^{1/p}\delta\|\boldsymbol{\beta}\|_* t_i + |\boldsymbol{x}_i^\top \boldsymbol{\beta} - y_i|\Big)\alpha_i - \alpha_i^2/4\right] =$$

$$\max_{\boldsymbol{t}\geq 0, \|\boldsymbol{t}\|_p \leq 1} \sup_{\alpha \geq 0} \left[n^{1/p}\delta\|\boldsymbol{\beta}\|_* \boldsymbol{t}^\top \alpha + \sum_{i=1}^n \alpha_i |\boldsymbol{x}_i^\top \boldsymbol{\beta} - y_i| - \alpha_i^2/4\right] =$$

$$\sup_{\alpha \geq 0} \left[n^{1/p}\delta\|\boldsymbol{\beta}\|_* \max_{t\geq 0, \|\boldsymbol{t}\|_p \leq 1} \{\boldsymbol{t}^\top \alpha\} + \sum_{i=1}^n \alpha_i |\boldsymbol{x}_i^\top \boldsymbol{\beta} - y_i| - \alpha_i^2/4\right] =$$

$$\sup_{\boldsymbol{\alpha}\geq 0} \left[n^{1/p}\delta\|\boldsymbol{\beta}\|_* \|\boldsymbol{\alpha}\|_q + \sum_{i=1}^n \alpha_i |\boldsymbol{x}_i^\top \boldsymbol{\beta} - y_i| - \alpha_i^2/4\right] = \sup_{\boldsymbol{\alpha}\geq 0} K(\boldsymbol{\beta}, \boldsymbol{\alpha})$$

where we used that $\|\boldsymbol{\alpha}\|_q = \max_{\|\boldsymbol{t}\|_p \leq 1} \boldsymbol{t}^\top \boldsymbol{\alpha} = \max_{\boldsymbol{t}\geq 0, \|\boldsymbol{t}\|_p \leq 1} \boldsymbol{t}^\top \boldsymbol{\alpha}$ since $\boldsymbol{\alpha} \geq 0$. This proves (A.18) and completes the proof. □

### E.2 $\eta$-TRICK SOLVER

In this section, we describe in detail the $\eta$-trick solver that iteratively alternative between solving a weighted ridge regression problem and updating the weights in closed-form. We take inspiration from previous works on adversarial linear regression (Ribeiro et al., 2025b;a) and use what is sometimes referred to as the "$\eta$-trick" (see, e.g., Bach 2019) to minimize the risk $V_\delta(\boldsymbol{\beta})$ with respect to $\boldsymbol{\beta}$. We only consider the case where we have the $\infty$-norm in the Wasserstein cost, i.e., $\|\cdot\| = \|\cdot\|_\infty$ and $\|\cdot\|_* = \|\cdot\|_1$. We leave the extension to other norms as future work.

The $\eta$-trick refers to a family of variational identities that are useful for optimization (Bach, 2019). For the squared 1-norm:

$$\|\boldsymbol{w}\|_1^2 = \inf_{\boldsymbol{\eta}\in\Delta_d} \sum_{j=1}^d \frac{w_j^2}{\eta_j},$$

where $\Delta_d = \{\boldsymbol{\eta} \in \mathbb{R}^d_{>0} : \sum_{j=1}^d \eta_j = 1\}$ is the $d$-dimensional simplex and $\mathbb{R}_{>0} = (0, \infty)$. Now, consider the formulation of the risk in Theorem 1

$$V_\delta(\boldsymbol{\beta}) = \sup_{\boldsymbol{t}\in T} \frac{1}{n} \sum_{i=1}^n (r_i + t_i\|\boldsymbol{\beta}\|_1)^2 \quad \text{with} \quad T = \{t \in \mathbb{R}^n_+ : \|\boldsymbol{t}\|_p \leq n^{1/p}\delta\}, \tag{A.20}$$

where $\mathbb{R}_+ = [0, \infty)$, $r_i = |\boldsymbol{x}_i^\top \boldsymbol{\beta} - y_i|$ and we used that $\|\cdot\|_* = \|\cdot\|_1$. Letting $\boldsymbol{w}^{(i)} = [r_i, t_i\beta_1, \ldots, t_i\beta_d]^\top$ and applying the trick then gives

$$(r_i + t_i\|\boldsymbol{\beta}\|_1)^2 = \|\boldsymbol{w}^{(i)}\|_1^2 = \inf_{\boldsymbol{\eta}^{(i)}\in\Delta_{d+1}} \left\{\frac{r_i^2}{\eta_0^{(i)}} + t_i^2 \sum_{j=1}^d \frac{\beta_j^2}{\eta_j^{(i)}}\right\}.$$

Plugging this into (A.20) we get

$$V_\delta(\boldsymbol{\beta}) = \sup_{\boldsymbol{t}\in T} \frac{1}{n} \sum_{i=1}^n \inf_{\boldsymbol{\eta}^{(i)}\in\Delta_{d+1}} \left\{\frac{r_i^2}{\eta_0^{(i)}} + t_i^2 \sum_{j=1}^d \frac{\beta_j^2}{\eta_j^{(i)}}\right\},$$

where we can swap the summation and the infimum since the problem is separable over $i$:

$$V_\delta(\boldsymbol{\beta}) = \sup_{\boldsymbol{t}\in T} \inf_{\boldsymbol{\eta}\in\Delta_{d+1}^n} \underbrace{\frac{1}{n} \sum_{i=1}^n \left\{\frac{r_i^2}{\eta_0^{(i)}} + t_i^2 \sum_{j=1}^d \frac{\beta_j^2}{\eta_j^{(i)}}\right\}}_{\Phi(\boldsymbol{\beta}, \boldsymbol{t}, \boldsymbol{\eta})}. \tag{A.21}$$

Here, we denote $\boldsymbol{\eta} := [\boldsymbol{\eta}^{(1)}, \ldots, \boldsymbol{\eta}^{(n)}]^\top \in \Delta_{d+1}^n$ and $\Phi$ the current objective function. If we make the substitution $z_i = t_i^2$, we get

$$\tilde{\Phi}(\boldsymbol{\beta}, \boldsymbol{z}, \boldsymbol{\eta}) = \frac{1}{n} \sum_{i=1}^{n} \left\{ \frac{r_i^2}{\eta_0^{(i)}} + z_i \sum_{j=1}^{d} \frac{\beta_j^2}{\eta_j^{(i)}} \right\} \quad \text{and} \quad Z = \{z \in \mathbb{R}_+^n : \|\boldsymbol{z}\|_{p/2} \leq n^{2/p}\delta^2\}.$$

Then,

- $\tilde{\Phi}(\boldsymbol{\beta}, \cdot, \boldsymbol{\eta})$ is continuous and affine over $Z$ for every fixed $(\boldsymbol{\beta}, \boldsymbol{\eta}) \in \mathbb{R}^d \times \Delta_{d+1}^n$.
- $\tilde{\Phi}(\boldsymbol{\beta}, \boldsymbol{z}, \cdot)$ is continuous and convex over $\Delta_{d+1}^n$ for every fixed $(\boldsymbol{\beta}, \boldsymbol{z}) \in \mathbb{R}^d \times Z$.
- $Z$ is compact convex since $p \geq 2$.
- $\Delta_{d+1}^n$ is convex.

Thus, Sion's minimax theorem (Sion, 1958; Komiya, 1988) applies[13] and we can swap supremum and infimum in (A.21):

$$V_\delta(\boldsymbol{\beta}) = \inf_{\boldsymbol{\eta} \in \Delta_{d+1}^n} \sup_{\boldsymbol{z} \in Z} \frac{1}{n} \sum_{i=1}^{n} \left\{ \frac{r_i^2}{\eta_0^{(i)}} + z_i \sum_{j=1}^{d} \frac{\beta_j^2}{\eta_j^{(i)}} \right\}. \tag{A.22}$$

The supremum can now be solved in closed form. Let $u_i := \sum_{j=1}^{d} \beta_j^2 / \eta_j^{(i)}$ so that

$$\sup_{\boldsymbol{z} \in Z} \frac{1}{n} \sum_{i=1}^{n} \left\{ \frac{r_i^2}{\eta_0^{(i)}} + z_i \sum_{j=1}^{d} \frac{\beta_j^2}{\eta_j^{(i)}} \right\} = \frac{1}{n} \sum_{i=1}^{n} \frac{r_i^2}{\eta_0^{(i)}} + \frac{1}{n} \sup_{\boldsymbol{z} \in Z} \boldsymbol{z}^\top \boldsymbol{u}.$$

Then, by letting $k(p)$ be the conjugate exponent of $p/2$, i.e.,

$$k(p) = \begin{cases} \infty & p = 2 \\ p/(p-2) & 2 < p < \infty \\ 1 & p = \infty \end{cases} \qquad \text{so that} \qquad \frac{1}{p/2} + \frac{1}{k(p)} = 1, \ \forall p \in [2, \infty],$$

we have that

$$\sup_{\boldsymbol{z} \in Z} \boldsymbol{z}^\top \boldsymbol{u} = n^{2/p} \delta^2 \|\boldsymbol{u}\|_{k(p)}$$

by the dual-norm identity since $u_i \geq 0$. Thus, we can define

$$G_\delta(\boldsymbol{\beta}, \boldsymbol{\eta}) := \frac{1}{n} \sum_{i=1}^{n} \frac{r_i^2}{\eta_0^{(i)}} + n^{2/p-1} \delta^2 \|\boldsymbol{u}\|_{k(p)},$$

and therefore,

$$V_\delta(\boldsymbol{\beta}) = \inf_{\boldsymbol{\eta} \in \Delta_{d+1}^n} G_\delta(\boldsymbol{\beta}, \boldsymbol{\eta}). \tag{A.23}$$

We note here that $G_\delta$ is jointly convex in $(\boldsymbol{\beta}, \boldsymbol{\eta})$ since the quadratic-over-linear function $(a, b) \mapsto a^2/b$ is jointly convex for $b > 0$ (Boyd & Vandenberghe, 2004, Chapter 3.1.5) and since the norm preserves the convexity. Hence, we can alternate between minimization over $\boldsymbol{\beta}$ and minimization over $\boldsymbol{\eta}$, particularly the optimal $\boldsymbol{\eta}$ has a structure that turns the subsequent $\boldsymbol{\beta}$-update into a weighted ridge regression problem. The $\boldsymbol{\eta}$-subproblem itself can be solved numerically (for example, by using CVXPY), but we can also recover its solution explicitly from the maximizer of the supremum in Theorem 1.

To derive this update for $\delta > 0$ and nonzero residuals $r_i \neq 0$, fix $\boldsymbol{\beta} \neq 0$ and from (A.20) let $\boldsymbol{t}^\star$ maximize

$$\frac{1}{n} \sum_{i=1}^{n} (r_i + t_i \|\boldsymbol{\beta}\|_1)^2 \qquad \text{over } \boldsymbol{t} \in T.$$

[13] We use the version for example found in Lattimore & Szepesvári (2020, Theorem 28.12) where either $Z$ or $\Delta_{d+1}^n$ needs to be compact, see also the original Corollary 3.3 in Sion (1958).

Finding $\boldsymbol{t}^\star$ is the same problem as evaluating $V_\delta(\boldsymbol{\beta})$ and can be solved using the scalar reduction derived in Appendix E.3. We now show how $\boldsymbol{t}^\star$ allows us to recover a minimizer of $G_\delta(\boldsymbol{\beta}, \cdot)$. From (A.21), recall that $\boldsymbol{t}^\star$ maximizes $\Phi(\boldsymbol{\beta}, \boldsymbol{t}, \boldsymbol{\eta})$ after minimization over $\boldsymbol{\eta}$:

$$\boldsymbol{t}^\star \in \arg\max_{\boldsymbol{t}\in T} \inf_{\boldsymbol{\eta}\in\Delta_{d+1}^n} \Phi(\boldsymbol{\beta}, \boldsymbol{t}, \boldsymbol{\eta}).$$

That is, (A.21) states

$$\inf_{\boldsymbol{\eta}\in\Delta_{d+1}^n} \Phi(\boldsymbol{\beta}, \boldsymbol{t}^\star, \boldsymbol{\eta}) = V_\delta(\boldsymbol{\beta}). \tag{A.24}$$

On the other hand, $G_\delta$ was obtained by solving the supremum in (A.22):

$$G_\delta(\boldsymbol{\beta}, \boldsymbol{\eta}) = \sup_{\boldsymbol{z}\in Z} \tilde{\Phi}(\boldsymbol{\beta}, \boldsymbol{z}, \boldsymbol{\eta}) = \sup_{\boldsymbol{t}\in T} \Phi(\boldsymbol{\beta}, \boldsymbol{t}, \boldsymbol{\eta})$$

for an *arbitrary* $\boldsymbol{\eta}$. Thus, $\boldsymbol{t}^\star$ is not guaranteed to attain this supremum, but its feasibility gives

$$G_\delta(\boldsymbol{\beta}, \boldsymbol{\eta}) \geq \Phi(\boldsymbol{\beta}, \boldsymbol{t}^\star, \boldsymbol{\eta}).$$

Minimizing over $\boldsymbol{\eta}$ on both sides give

$$V_\delta(\boldsymbol{\beta}) = \inf_{\boldsymbol{\eta}\in\Delta_{d+1}^n} G_\delta(\boldsymbol{\beta}, \boldsymbol{\eta}) \geq \inf_{\boldsymbol{\eta}\in\Delta_{d+1}^n} \Phi(\boldsymbol{\beta}, \boldsymbol{t}^\star, \boldsymbol{\eta}) = V_\delta(\boldsymbol{\beta}),$$

where we have used (A.23) and (A.24), respectively. We use this to execute the $\boldsymbol{\eta}$-update. More precisely, for a fixed current parameter $\widetilde{\boldsymbol{\beta}}$ (in our algorithm), the $\boldsymbol{\eta}$-update is thus the minimizer $\hat{\boldsymbol{\eta}}$ of

$$\Phi(\widetilde{\boldsymbol{\beta}}, \boldsymbol{t}^\star, \boldsymbol{\eta}) = \frac{1}{n}\sum_{i=1}^n \left\{ \frac{\tilde{r}_i^2}{\eta_0^{(i)}} + (t_i^\star)^2 \sum_{j=1}^d \frac{\widehat{\beta}_j^2}{\eta_j^{(i)}} \right\}, \tag{A.25}$$

where $\tilde{r}_i = |\boldsymbol{x}_i^\top \widetilde{\boldsymbol{\beta}} - y_i|$. From Ribeiro et al. (2025b, Proposition 4), this minimizer $\hat{\boldsymbol{\eta}}$ is uniquely determined and given by

$$\hat{\eta}_0^{(i)} := \frac{\tilde{r}_i}{\tilde{r}_i + t_i^\star \|\widehat{\boldsymbol{\beta}}\|_1}, \qquad \hat{\eta}_j^{(i)} := \frac{t_i^\star |\widehat{\beta}_j|}{\tilde{r}_i + t_i^\star \|\widehat{\boldsymbol{\beta}}\|_1}, \quad j = 1, \ldots, d. \tag{A.26}$$

This completes the $\boldsymbol{\eta}$-update. Next, we want to execute the $\boldsymbol{\beta}$-update (since we alternate), with updated parameter denoted by $\boldsymbol{\beta}^+$. To obtain $\boldsymbol{\beta}^+$ from the $\boldsymbol{\beta}$-update, we plug $\hat{\boldsymbol{\eta}}$ back into (A.25), which gives

$$\Phi(\boldsymbol{\beta}, \boldsymbol{t}^\star, \hat{\boldsymbol{\eta}}) = \frac{1}{n}\sum_{i=1}^n \left\{ r_i^2 \frac{\tilde{r}_i + t_i^\star \|\widetilde{\boldsymbol{\beta}}\|_1}{\tilde{r}_i} + (t_i^\star)^2 \sum_{j=1}^d \beta_j^2 \frac{\tilde{r}_i + t_i^\star \|\widetilde{\boldsymbol{\beta}}\|_1}{t_i^\star |\widetilde{\beta}_j|} \right\} \tag{A.27}$$

$$= \frac{1}{n}\sum_{i=1}^n \hat{w}_i r_i^2 + \sum_{j=1}^d \hat{\gamma}_j \beta_j^2, \tag{A.28}$$

where

$$\hat{w}_i = \frac{\tilde{r}_i + t_i^\star \|\widetilde{\boldsymbol{\beta}}\|_1}{\tilde{r}_i} \quad \text{and} \quad \hat{\gamma}_j = \left[ \frac{1}{n}\sum_{i=1}^n t_i^\star (\tilde{r}_i + t_i^\star \|\widetilde{\boldsymbol{\beta}}\|_1) \right] \frac{1}{|\widetilde{\beta}_j|}, \quad \forall (i,j) \in [n] \times [d], \tag{A.29}$$

using the shorthand $[n] = \{1, 2, \ldots, n\}$, and same for $[d]$. We here note that minimizing $G_\delta(\cdot, \hat{\boldsymbol{\eta}})$ (and therefore obtain $\boldsymbol{\beta}^+$) is now a weighted ridge regression problem with $\widehat{W} = \mathrm{diag}(\hat{w}_1, \ldots, \hat{w}_n)$ and $\widehat{\Gamma} = \mathrm{diag}(\hat{\gamma}_1, \ldots, \hat{\gamma}_d)$. The $\boldsymbol{\beta}$-update is therefore obtained by solving

$$(X^\top \widehat{W} X + n\widehat{\Gamma})\boldsymbol{\beta}^+ = X^\top \widehat{W} y$$

and assigning the solution $\boldsymbol{\beta}^+$ as our new $\widetilde{\boldsymbol{\beta}}$, from which we again perform an $\boldsymbol{\eta}$-update, and so on. This alternating scheme is repeated until a stopping criterion is satisfied. We initialize with uniform weights $\eta_j^{(i)} = 1/(d+1)$. Then $u_i = (d+1)\|\boldsymbol{\beta}\|_2^2$, so $\|\boldsymbol{u}\|_{k(p)} = n^{1/k(p)}(d+1)\|\boldsymbol{\beta}\|_2^2$. Since $2/p + 1/k(p) = 1$, substitution gives

$$G_\delta(\boldsymbol{\beta}, \boldsymbol{\eta}_{\mathrm{unif}}) = (d+1)\left[ \frac{1}{n}\|X\boldsymbol{\beta} - \boldsymbol{y}\|_2^2 + \delta^2 \|\boldsymbol{\beta}\|_2^2 \right].$$

**Algorithm 1:** $\eta$-trick solver

**input:** $X, \boldsymbol{y}$; $\delta \geq 0$; $p \in [2, \infty]$; $\epsilon > 0$; tolerance $\tau > 0$; iterations $M$
**init:** *solve* $(X^\top X + n\delta^2 I)\boldsymbol{\beta} = X^\top \boldsymbol{y}$ (uniform $\boldsymbol{\eta}$ init)
**if** $\delta = 0$ **then**
  **return** $\boldsymbol{\beta}$
**for** $m = 1, \ldots, M$ **do**
  ▷ *smoothed magnitudes and worst-case transport*
  $r_i \leftarrow \sqrt{(\boldsymbol{x}_i^\top \boldsymbol{\beta} - y_i)^2 + \epsilon^2}, \quad b_j \leftarrow \sqrt{\beta_j^2 + \epsilon^2}$
  $B \leftarrow \sum_j b_j$
  $\boldsymbol{t}^\star \leftarrow \texttt{ScalarReduction}(\boldsymbol{r}, B, \delta, p)$

  ▷ *compute weights and solve weighted ridge*
  $w_i \leftarrow (r_i + t_i^\star B)/r_i, \quad \gamma_j \leftarrow [\frac{1}{n}\sum_i t_i^\star (r_i + t_i^\star B)]/b_j$
  $W \leftarrow \mathrm{diag}(\boldsymbol{w}), \quad \Gamma \leftarrow \mathrm{diag}(\boldsymbol{\gamma})$
  *solve* $(X^\top W X + n\Gamma)\boldsymbol{\beta}^+ = X^\top W \boldsymbol{y}$

  ▷ *update and check stopping criterion*
  change $\leftarrow \|\boldsymbol{\beta}^+ - \boldsymbol{\beta}\|_2 / (1 + \|\boldsymbol{\beta}\|_2)$
  $\boldsymbol{\beta} \leftarrow \boldsymbol{\beta}^+$
  **if** change $\leq \tau$ **then**
    **break**
**return** $\boldsymbol{\beta}$

Thus, the initial $\boldsymbol{\beta}$-update solves $(X^\top X + n\delta^2 I)\boldsymbol{\beta} = X^\top \boldsymbol{y}$. We summarize this algorithm by Algorithm 1, where we also add smoothing for numerical stability. The smoothing adds only terms that are constant in $\boldsymbol{\beta}$ and therefore leaves this initialization unchanged. Moreover, `ScalarReduction` in Algorithm 1 corresponds to solving the scalar form in Lemma 3, described in Appendix E.3. We use this form since it is easier to solve it than the robust quadratic form given by Theorem 1. Note that $\boldsymbol{t}^\star$ is the same for both these forms due to strong duality, see, e.g., the proof of Theorem 1.

### E.3 Fast robust risk evaluations

To evaluate only the robust risk $V_\delta(\boldsymbol{\beta})$ (rather than minimize it), we can compute it faster by using Lemma 3. First, the special cases $\delta = 0$, $p = 2$, and $p = \infty$ have closed-form expressions (described above for minimizing the robust risk) and can be readily computed. The same holds for $\boldsymbol{\beta} = 0$. We may therefore assume $\delta > 0$ and $2 < p < \infty$ and $\boldsymbol{\beta} \neq 0$. We compute the robust risk for this general case using Lemma 3. Note first that $\lambda = 0$ is suboptimal in this case since the inner supremum of (A.1) becomes infinite. Hence, we may assume $\lambda > 0$. Given $\lambda > 0$, we can easily compute the inner supremum $\sup_{t_i \geq 0} \left\{ \left(|r_i| + t_i \|\boldsymbol{\beta}\|_*\right)^2 - \lambda t_i^p \right\}$ of (A.1). Indeed, let $f_i$ denote the objective function of this supremum:

$$f_i(t_i) := \left(|r_i| + t_i \|\boldsymbol{\beta}\|_*\right)^2 - \lambda t_i^p = t_i^2 \|\boldsymbol{\beta}\|_*^2 + 2|r_i| t_i \|\boldsymbol{\beta}\|_* + r_i^2 - \lambda t_i^p.$$

Then,

$$f_i'(t_i) = 2t_i \|\boldsymbol{\beta}\|_*^2 + 2|r_i| \|\boldsymbol{\beta}\|_* - \lambda p t_i^{p-1}, \qquad f_i''(t_i) = 2\|\boldsymbol{\beta}\|_*^2 - \lambda p(p-1) t_i^{p-2}.$$

Given $2 < p < \infty$, we have that $f_i''$ is strictly decreasing with $f_i''(0) > 0$, so $f_i'$ first increases and then decreases. Since $f_i'(0) = 2|r_i| \|\boldsymbol{\beta}\|_* \geq 0$ and $f_i'(t_i) \to -\infty$, it follows that $f_i'$ changes sign at most once. Hence $f_i$ is unimodal on $[0, \infty)$, with maximizer either at a stationary point $t_i^\star$ (unique if it exists) or at the boundary $t_i = 0$. The inner supremum $\sup_{t_i \geq 0} f_i(t)$ in (A.1) is therefore straightforward to solve. Concretely, we minimize the negated objective using SciPy (Virtanen et al., 2020), where we first apply the method `scipy.optimize.elementwise.bracket_minimum` to locate a small interval on $[0, \infty)$ where the minimum is attained, and then use `scipy.optimize.elementwise.find_minimum` to compute the minimum exactly. For further details, see the code implementation.

For the outer problem in (A.1), note that $f_i$ is affine in $\lambda$, and hence $\sup_{t_i \geq 0} f_i(t)$ is convex in $\lambda$. Adding $n\delta^p\lambda$ preserves this convexity, so the outer minimization over $\lambda$ is a minimization of a convex function. We solve this outer problem using `scipy.optimize.minimize_scalar`, recomputing the inner maximizers for each candidate $\lambda$, to obtain the robust risk $V_\delta(\boldsymbol{\beta})$ to numerical accuracy. For further details, see the code implementation.

# F Numerical experiments: Additional details

In this section, we provide additional details of the numerical simulations in Section 8, starting with the setup details for the fast rate and slow rate, followed by details of the small and large $\delta > 0$ simulations.

## F.1 Fast rate

In this section, we present additional details for the fast rate simulations, showing that $X$ satisfies the RE condition and providing some additional simulations.

***X* satisfies the RE condition.** We here show that $X$ with entries sampled independently and uniformly in $[-1, 1]$ is in $\mathrm{RE}(s, \ell)$ for any $s$ and $\ell$ with $\kappa = 1/\sqrt{6}$, which holds with arbitrarily high probability for sufficiently large $n$. The argument is standard and based on Rayleigh–Ritz variational characterization (see, e.g., (Wainwright, 2019, Equation (6.3))) to obtain a sufficient criterion for satisfying the RE condition. To this end, pick any $s$ and $\ell$, consider an index set $S \subseteq \{1, \ldots, d\}$, and define the cone

$$\mathcal{C}_S = \left\{ \boldsymbol{v} \in \mathbb{R}^d : \|\boldsymbol{v}_{S^c}\|_1 \leq \ell \|\boldsymbol{v}_S\|_1 \right\}.$$

Moreover, let $G = X^\top X/n$ and $\lambda_{\min}(G)$ be the smallest eigenvalue of $G$. Using standard results from random matrix theory, we have that $\lambda_{\min}(G)$ is lower bounded by $1/3 - \epsilon$ with arbitrarily high probability for any given $\epsilon > 0$ and sufficiently large $n$.[14] For convenience, pick $\epsilon = 1/6$ so that

$$\kappa = \frac{1}{\sqrt{6}} \leq \sqrt{\max\{\lambda_{\min}(G), 0\}}$$

holds with arbitrarily high probability for large enough $n$. Then, the Rayleigh–Ritz variational characterization (Wainwright, 2019, Equation (6.3)) yields for every $\boldsymbol{v} \in \mathbb{R}^d$:

$$\frac{\|X\boldsymbol{v}\|_2^2}{n} = \boldsymbol{v}^\top \frac{1}{n} X^\top X \boldsymbol{v} = \boldsymbol{v}^\top G \boldsymbol{v} \geq \lambda_{\min}(G) \|\boldsymbol{v}\|_2^2 \geq \kappa^2 \|\boldsymbol{v}\|_2^2,$$

which holds with arbitrarily high probability for large enough $n$. Since this is valid for every $\boldsymbol{v} \in \mathbb{R}^d$, it is also valid for every $\boldsymbol{v}$ in $\mathcal{C}_S$. In other words, we have that $X$ is in $\mathrm{RE}(s, \ell)$ for $\kappa = 1/\sqrt{6}$ with arbitrarily high probability (for large enough $n$), completing the derivation.

**Additional plots.** We also provide plots that support the claim that the conditions in Theorem 3 hold for the fast rate simulations in Section 8, seen in Figure A.1. Indeed, we see that both the RE condition and that $\delta \geq \bar{\delta}$ are fulfilled. Moreover, we also plot $e_q := \frac{\|\boldsymbol{\epsilon}\|_q}{n^{1/q}}$ in Figure A.2. We see that as $n$ increases, $e_q$ converges to a constant as predicted. Moreover, since both $B$ and $C$ in Theorem 3 are functions of $e_q$ that converge to constants if $e_q$ converges to a constant, we have that the maximum in (11) becomes a constant. Thus, we indeed get the rate $O(\delta^2)$ which equals $O(n^{-1})$ since $\delta$ decreases as $1/\sqrt{n}$.

## F.2 Slow rate

We continue with the slow rate. More precisely, we here provide the details on how $X$ is set and why the RE condition fails for it. First of all, $\boldsymbol{\beta}^* = [2, -1]^\top$ is 2-sparse, so we need an RE condition

[14] This is seen as follows. Let $\Sigma := \mathbb{E}[\boldsymbol{v}\boldsymbol{v}^T]$ where $\boldsymbol{v} \in \mathbb{R}^d$ has independent entries uniformly sampled in $[-1, 1]$. Then, $\Sigma = I/3$ due to independence and variance $\mathbb{E}[v_i^2] = 1/3$. Moreover, by Equation (6.7) in Wainwright (2019) (Weyl's theorem), we have that $\lambda_{\min}(\Sigma) - \lambda_{\min}(G) \leq \|G - \Sigma\|_2$ and so $\lambda_{\min}(G) \geq \lambda_{\min}(\Sigma) - \|G - \Sigma\|_2 = 1/3 - \|G - \Sigma\|_2$. By Theorem 6.5 in Wainwright (2019) for some universal constants $(c_i)_{i=1}^3$, we have that $\|G - \Sigma\|_2 < c_1[d/n + \sqrt{d/n}] + \epsilon$ holds for any $\epsilon \geq 0$ with probability $\geq 1 - c_2 e^{-c_3 n \min(\epsilon, \epsilon^2)}$. That is, $\lambda_{\min}(G) \geq 1/3 - (c_1[d/n + \sqrt{d/n}] + \epsilon)$ with probability $\geq 1 - c_2 e^{-c_3 n \min(\epsilon, \epsilon^2)}$, which yields the result.

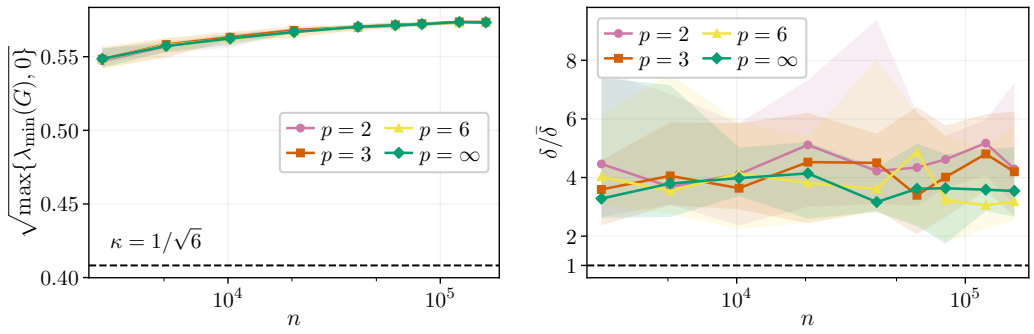



Figure A.1: Additional plots for the fast rate simulations in Section 8 showing that the conditions in Theorem 3 hold. Concretely, the left plot shows that the RE condition is satisfied for $\kappa = 1/\sqrt{6}$ (for all $n$) since it indeed lower-bounds the minimum eigenvalue $\lambda_{\min}(G)$ of $G = X^T X/n$ (see (A.30) and the corresponding derivation for details). Moreover, the right plot shows that $\delta \geq \bar{\delta}$ (which it should be with high probability). Therefore, the conditions in Theorem 3 hold. Both plots show median over 10 runs with shaded bands showing minimum and maximum across the repetitions.

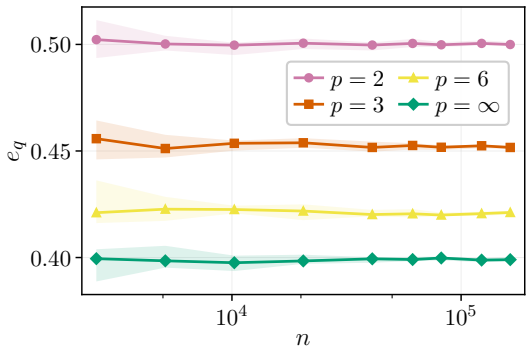



Figure A.2: Convergence for $e_q := \frac{\|\boldsymbol{\epsilon}\|_q}{n^{1/q}}$ in the fast rate simulations from Section 8. We see that all $e_q$ quickly stabilizes toward a constant as $n$ increases. Since $e_q$ converges to a constant, so do $B$ and $C$ in Theorem 3. The plot shows the median over 10 repeats with shaded bands being the 10th-90th percentile range.

of the form $X \in \mathrm{RE}(2, \ell)$ for some fixed $\ell$ and $\kappa > 0$. We construct $X$ so that there is no such $\ell$ and $\kappa > 0$ such that $X \in \mathrm{RE}(2, \ell)$ holds for all $n$. This construction closely resembles van de Geer (2017); see also references therein. Importantly, we want to construct $X$ such that the entries are *correlated* in a way that breaks the RE condition. We achieve this by setting the first column $\boldsymbol{X}_1$ and the second column $\boldsymbol{X}_2$ of $X$ to the following:

$$\boldsymbol{X}_1 = \sqrt{\frac{1+\rho_n}{2}}\boldsymbol{u} + \sqrt{\frac{1-\rho_n}{2}}\boldsymbol{w}, \quad \boldsymbol{X}_2 = \sqrt{\frac{1+\rho_n}{2}}\boldsymbol{u} - \sqrt{\frac{1-\rho_n}{2}}\boldsymbol{w},$$

where $\boldsymbol{u}$ and $\boldsymbol{w}$ are any vectors in $\{-1, 1\}^n$ such that $\boldsymbol{u}$ is orthogonal to $\boldsymbol{w}$ for even $n$ (and for odd $n$ set them so that they are "almost" orthogonal $\boldsymbol{w}^\top \boldsymbol{u} = \pm 1$), and $\rho_n = 1 - c_0/\sqrt{n}$ for some constant $c_0 > 0$. It is easy to see that the entries of $X$ are uniformly bounded with $M = \sqrt{2}$. Crucially, this construction is made so that for every even $n$,

$$G = \frac{1}{n} X^T X = \begin{pmatrix} 1 & \rho_n \\ \rho_n & 1 \end{pmatrix}$$

becomes close to a singular matrix. Indeed, note that $\lambda_{\min}(G) = 1 - \rho_n = c_0/\sqrt{n}$ is the smallest eigenvalue. Thus, the Rayleigh–Ritz variational characterization (Wainwright, 2019, Equation (6.3)) implies that for any even $n$, there exists a vector $\boldsymbol{v} \in \mathbb{R}^2$ such that:

$$\frac{\|X\boldsymbol{v}\|_2^2}{n} = \boldsymbol{v}^\top \frac{1}{n} X^\top X \boldsymbol{v} = \boldsymbol{v}^\top G \boldsymbol{v} = \lambda_{\min}(G)\|\boldsymbol{v}\|_2^2 = \frac{c_0}{\sqrt{n}}\|\boldsymbol{v}\|_2^2. \tag{A.30}$$

Thus, since $\boldsymbol{v}$ is trivially in the cone $\mathcal{C}_S = \{\boldsymbol{v} \in \mathbb{R}^2 : \|\boldsymbol{v}_{S^c}\|_1 \leq \ell \|v_S\|_1\} = \mathbb{R}^2$ for $S = \{1, 2\}$ and $\lambda_{\min}(G) = \frac{c_0}{\sqrt{n}} \to 0$ as $n \to \infty$, we conclude that no $\kappa > 0$ exists such that $X \in \mathrm{RE}(2, \ell)$ for large enough $n$, proving the claim.

Concretely, in the experiments, we set $c_0 = 3$ and set $\boldsymbol{u}$ and $\boldsymbol{w}$ to $n/4$ stacked copies of the vectors $[1, 1, -1, -1]^\top$ and $[1, -1, 1, -1]^T$, respectively (with $n$ a multiple of 4) and plot only such $n$, namely, $n = 4096, 6144, 8192, 10240$. The reason for this is twofold. First, it is easy to implement stacked copies. Second, there is a convenient analytic interpretation. Indeed, since $\boldsymbol{u}$ and $\boldsymbol{w}$ always have zero mean (their entries sum to zero), the parameter $\rho_n$ equals the standard empirical correlation between the columns $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$. To see this, consider the sample correlation coefficient (see, e.g., James et al. (2021, Equation (3.18))) between $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ given by

$$r = \frac{\sum_{i=1}^{n} (X_1[i] - \mathrm{mean}(\boldsymbol{X}_1))(X_2[i] - \mathrm{mean}(\boldsymbol{X}_2))}{\|\boldsymbol{X}_1\|_2 \|\boldsymbol{X}_2\|_2}$$

where $X_1[i]$ is the $i$-th entry of $\boldsymbol{X}_1$ (and similarly for $\boldsymbol{X}_2$), and $\mathrm{mean}(\boldsymbol{X}_i)$ is the mean over all entries of $\boldsymbol{X}_i$. In our case, $\mathrm{mean}(\boldsymbol{X}_i) = 0$ by construction, and a simple computation yields $r = \rho_n$ as desired. Therefore, another interpretation of what happens is as follows. As $n \to \infty$, we have that $r = \rho_n \to 1$; in other words, $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ become maximally correlated. This correlation breaks down the RE condition and hinders the fast rate $O(n^{-1})$. As a result, we get only the slow rate $O(n^{-1/2})$.

### F.3 Small and large $\delta$

Finally, we describe the setup for the small and large $\delta > 0$ simulations. In both cases, we let $y_i = \boldsymbol{x}_i^\top \boldsymbol{\beta}^* + \varepsilon_i$ with $n = 30$ and $d = 60$, true parameter iid sampled as $\boldsymbol{\beta}_i^* \sim N(0, 1)$, noise $\boldsymbol{\varepsilon} \sim N(0, \sigma^2 I)$ with $\sigma = 0.5$, and entries $x_{ij} \in X$ iid sampled as $x_{ij} \sim N(0, 1)$. Note that $d > n$, that is, we are in the overparametrized regime. Moreover, since we sample $x_{ij} \sim N(0, 1)$, we have that $X$ is of full row rank almost surely. However, as an extra precaution, we also check that $X$ is indeed of full rank to be sure that the conditions in Theorem 5 hold; For all simulations, $X$ was indeed of full row rank.

For the small $\delta$ case, we plot the average error $\|X\widehat{\boldsymbol{\beta}} - \boldsymbol{y}\|_2^2/n$ as a function of $\delta$ as seen in Figure 3 (upper) for $p \in \{2, 3, 6, \infty\}$. Moreover, we also compute the threshold $\delta_S$ for $p = 2$ (pink dashed) and $\delta_S$ for $2 < p \leq \infty$ (gray dashed, identical for all $2 < p \leq \infty$), as given by Theorem 5, where $\min_{\boldsymbol{\alpha} \in Q} \|\boldsymbol{\alpha}\|_\infty$ and $\min_{\boldsymbol{\alpha} \in Q} \|\boldsymbol{\alpha}\|_2$ in the expression for $\delta_S$ are computed using the convex program as detailed by Remark 3. For each $p \in \{2, 3, 6, \infty\}$, we observe that the average error $\|X\widehat{\boldsymbol{\beta}} - \boldsymbol{y}\|_2^2/n$ decreases rapidly for $\delta \leq \delta_S$, reaching values in the numerical precision regime, indicating that interpolation occurs. Furthermore, this interpolation transition happens simultaneously for all $2 < p \leq \infty$ at their $\delta_S$, while it occurs sooner for the $p = 2$ case since its threshold $\delta_S$ is larger. All these observations are in line with Theorem 5.

For the large $\delta$ case, we plot $\|\widehat{\boldsymbol{\beta}}\|_\infty$ as a function of $\delta$, as seen in Figure 3 (lower) for $p \in \{2, 3, 5, \infty\}$ with threshold $\delta_L$ for each $p$ (dashed lines with matching color). We see that $\|\widehat{\boldsymbol{\beta}}\|_\infty$ decreases in the region $\delta \leq \delta_L$ and becomes zero for $\delta \geq \delta_L$ as predicted by Theorem 4. Note also that the threshold $\delta_L$ is different for each $p$ (as opposed to $\delta_S$ in the small $\delta$ case).

### F.4 Comparisons between the saddle-point solver and the $\eta$-trick solver

Finally, we compare the saddle-point solver and the $\eta$-trick solver as introduced in Section 7 with details in Appendix E.

**Correctness.** First, we run both solvers on the fast rate setup as detailed in Section 8, with results shown in Figure A.3. We see that both solvers achieve near-identical results (left plot). Indeed, the difference in robust risk error is $10^{-6}$ to $10^{-8}$ (right plot). This provide numerical support that both solvers are correctly solving the given optimization problem.

**Speed.** Next we compare the speed of the two solvers and how their computation times scale with $n$, using the fast rate setup. The computation time is shown in Figure A.4 (left). Here, we see that the $\eta$-trick solver is faster in general, being slightly faster for $p \in \{3, 6\}$ (especially for large $n$) and several orders of magnitude faster for $p \in \{2, \infty\}$. We can also see this more clearly by looking at

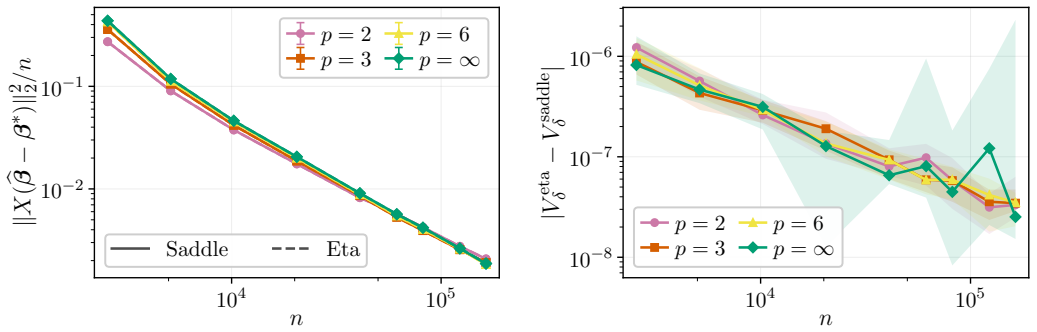



Figure A.3: We plot the error of the saddle-point solver and the $\eta$-trick solver for the fast rate setup (left). We run 10 repeats and plot the mean with $\pm 1$ standard deviation bands. We also plot the robust risk for the obtained estimate for both solvers (right). Here we plot the median over 10 repeats with 10-90 percentiles. We note that both output near-identical results.

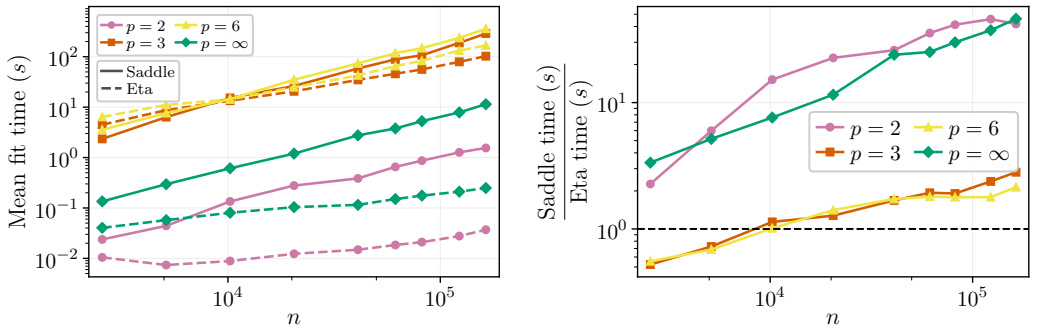



Figure A.4: Computation time (left) and time ratio (right) of the saddle-point solver and the $\eta$-trick solver for the fast rate setup. We see that $\eta$-trick solver is, in general, faster, particularly for large $n$ and $p \in \{2, \infty\}$. Both methods are slower for $p \in \{3, 6\}$, since the robust risk then lacks a closed-form expression. Both plots show mean over 10 runs.

their time ratio, illustrated in Figure A.4 (right), where the region above the dashed line indicates that the $\eta$-trick solver is faster. The reason for the improved speed could be that the saddle-point formulation leads to slower convergence in general or that the DSP implementation is slower in particular; future work aims to find the underlying reason for this difference. Finally, we also note in Figure A.4 that it takes considerably more time to solve for $p \in \{3, 6\}$ since the robust risk lacks a closed-form expression. Future work should look into other variations to improve the speed. In particular, using the scalar optimization form in Lemma 3 could yield faster computations.