Correcting the Coverage Bias of Quantile Regression

Isaac Gibbs[*†] John J. Cherian[*‡] Emmanuel J. Candès[§]

October 4, 2025

Abstract

We develop a collection of methods for adjusting the predictions of quantile regression to ensure coverage. Our methods are model agnostic and can be used to correct for high-dimensional overfitting bias with only minimal assumptions. Theoretical results show that the estimates we develop are consistent and facilitate accurate calibration in the proportional asymptotic regime where the ratio of the dimension of the data and the sample size converges to a constant. This is further confirmed by experiments on both simulated and real data. One of the key components of our work is a new connection between the leave-one-out coverage and the fitted values of variables appearing in the dual formualtion of the quantile regression. This facilitates the use of cross-validation in a variety of settings at significantly reduced computational costs.

Keywords: prediction set, cross-validation, high-dimensional statistics.

[*]The first two authors contributed equally to this work.
[†]Department of Statistics, University of California, Berkeley. Corresponding author: igibbs@berkeley.edu.
[‡]Department of Statistics, Stanford University.
[§]Departments of Mathematics and Statistics, Stanford University.

1 Introduction

Quantile regression is a popular tool for bounding the tail of a target outcome. This method has a long history dating back to the foundational work of Koenker & Bassett (1978) and has found widespread applications across a variety of areas (Koenker & Hallock 2001, Koenker 2017)). Classical results demonstrate that as the sample size increases quantile regression estimates are consistent and normally distributed around their population analogs (Koenker & Bassett 1978, Angrist et al. 2006) and, perhaps most critically, achieve their target coverage level (Jung et al. 2023, Duchi 2025).

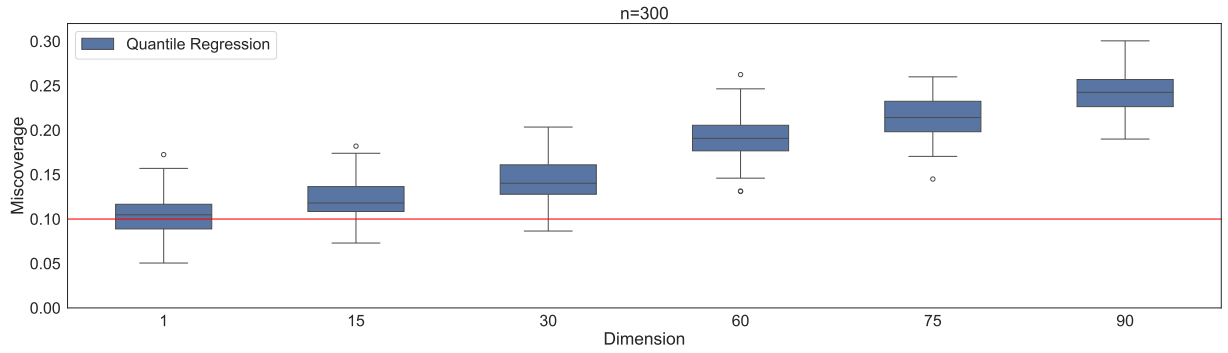

Figure 1: Miscoverage of (unregularized) quantile regression with model $Y_i \sim \beta_0 + X_i^\top \beta$ on i.i.d. data $\{(X_i, Y_i)\}_{i=1}^n$ sampled from the Gaussian linear model $Y_i = X_i^\top \tilde{\beta} + \epsilon_i$ for $X_i \sim \mathcal{N}(0, I_d)$ and $\epsilon \sim \mathcal{N}(0, 1)$ with $\epsilon_i \perp\!\!\!\perp X_i$. Boxplots in the figure show the empirical distribution of the training-conditional coverage, $\mathbb{P}(Y_{n+1} \leq \hat{\beta}_0 + X_{n+1}^\top \hat{\beta} \mid \{(X_i, Y_i)\}_{i=1}^n)$ where $(\hat{\beta}_0, \hat{\beta})$ denote the estimated coefficients at quantile level $\tau = 0.9$ and (X_{n+1}, Y_{n+1}) is an independent sample from the same model. The results come from 100 trials where in each trial the coverage is evaluated over a test set of size 2000 and the population coefficients are sampled as $\tilde{\beta} \sim \mathcal{N}(0, I_d/d)$. The red line shows the target miscoverage level of $1 - \tau = 0.1$.

Although the classical theory can be accurate for large sample sizes, it is often insufficient to fully capture the realities of finite samples. Figure 1 shows the realized miscoverage of quantile estimates fit at target level $\tau = 0.9$ in a well specified linear model $Y_i = X_i^\top \tilde{\beta} + \epsilon_i$ with

$\epsilon_i \perp\!\!\!\perp X_i$. In agreement with the classical theory, we see that when X_i is very low-dimensional (e.g., $d = 1$) quantile regression reliably obtains the target miscoverage rate of $1 - \tau = 0.1$. However, the scope of this theory is limited and the coverage shows visible bias in what might be typically considered to be small or moderate dimensions (e.g. $d \in \{15, 30\}$ compared to a sample size of $n = 300$). Perhaps unsurprisingly, this issue only worsens as the dimension increases and quantile regression exhibits over 2 times the target error rate at dimension $d = 90$.

A formal characterization of the coverage bias of quantile regression was first given in Bai et al. (2021). They eschew classical theory and instead work under a proportional asymptotic framework in which the ratio of the dimension of the data and the sample size converges to a constant. Under a stylized linear model, they show that in this regime the coverage of quantile regression converges to value different from the target level and provide an exact formula for quantifying this bias. Interestingly, while both under- and overcoverage are possible, they demonstrate that in most settings quantile regression will tend to undercover.* This is consistent with the results in Figure 1 as well as additional empirical results that we will present in Section 4.

Two proposals have been made in the literature for correcting quantile regression's bias. Under the same linear model assumptions, Bai et al. (2021) derive a simple method for adjusting the nominal level to account for overfitting. While quite effective, this procedure is

*As a matter of terminology, if $\hat{q}_\tau$ is an estimate of the $\tau \in [1/2, 1]$ quantile of Y we say that $\hat{q}_\tau$ undercovers if $\mathbb{P}(Y \leq \hat{q}_\tau) < \tau$ and overcovers if $\mathbb{P}(Y \leq \hat{q}_\tau) > \tau$. For $\tau < 1/2$ this terminology is reversed and we say that $\hat{q}_\tau$ undercovers if $\mathbb{P}(Y \leq \hat{q}_\tau) > \tau$ and overcovers otherwise. This is motivated by the fact that for $\tau > 1/2$ (resp. $\tau < 1/2$) the τ-quantile is designed to be a high probability upper (resp. lower) bound on Y and we use the terms undercoverage and overcoverage to reflect these goals.

limited in scope to small aspect ratios and a restrictive model for the data. A more generic procedure that does not require any such modeling assumptions was given in Gibbs et al. (2025). They employ a technique known as full conformal inference, which augments the regression fit with a guess of the unseen test point. This mimics the effect of overfitting the training data to the test point, thereby eliminating the resulting bias. In general, this approach has two main drawbacks. First, it requires randomization in order to obtain the desired coverage level. As we will show shortly in Section 1.1, this randomization can be significant and may cause the quantile estimate to vary by over 200% in magnitude. Second, additional computation is required for every test point in order to accurately incorporate it into the fit. This contrasts sharply with standard quantile regression, which once fitted can issue new predictions at the cost of computing just a single inner product. Depending on the application, significant additional test-time computational complexity of this form may not be permissible.

In this article, we develop a number of alternative procedures for adjusting the quantile regression fit. All of these methods are deterministic and many of them require per test point computation that is identical to standard quantile regression. After giving a brief overview of the work of Gibbs et al. (2025), we formally introduce our main methods in Section 2. Theoretical results showing the consistency of our proposals in the asymptotic regime are presented in Section 3, while Sections 2.3 and 4 give empirical results demonstrating the accuracy of our proposals in finite samples. Overall, our results show that all of our proposed methods are robust and provide reliable coverage irrespective of the dimension of the data. A more fine-grained discussion on the relative merits of our various procedures is given in Section **??**.

A central component of two of our proposed methods is a new procedure for efficiently computing the leave-one-out coverage of quantile regression. By exploiting a connection between the leave-one-out coverage indicators and a set of dual variables to the regression we show that the entire leave-one-coverage can be computed in time identical to that of running a single regression fit. This allows us to derive two methods which utilize cross-validation to efficiently find hyperparameter tunings that guarantee coverage.

The theoretical results in this paper contribute to a growing literature on characterizing and correcting for overfitting bias in high dimensions (e.g., Karoui et al. (2013), Donoho & Montanari (2013), Zhang & Zhang (2013), Javanmard & Montanari (2014), van de Geer et al. (2014), Thrampoulidis et al. (2018), Hastie et al. (2022)). Of particular relevance to our work are the Gaussian comparison inequalities of Gordon (1985, 1988) and their development for high dimensional M-estimation problems in Thrampoulidis et al. (2018). These tools will allow us to characterize the asymptotic behaviour of the quantile regression dual variables and, through their connection to leave-one-out coverage, to prove the consistency of our cross-validation estimates. There is a large body of literature investigating the consistency of cross-validation in high dimensional parameter tuning (e.g., Steinberger & Leeb (2016), Rad et al. (2020), Bayle et al. (2020), Austern & Zhou (2020), Xu et al. (2021), Patil et al. (2021, 2022), Steinberger & Leeb (2023), Zou et al. (2025)). On a technical level, these articles often require smoothness and/or strong convexity assumptions on the loss in order to derive exact formula for the leave-one-out coefficients. In contrast, we will be interested in the behaviour of the leave-one-out coverage of quantile regression, which is a discontinuous objective taken over parameter estimates coming from a non-differentiable loss. Here, our connection to the dual program will be critical in allowing us to avoid technical problems present in prior work and utilize technical tools which are typically unavailable in studies of cross-validation.

Notation: In the remainder of this article we let $\{(X_i, Y_i)\}_{i=1}^{n+1} \in \mathbb{R}^d \times \mathbb{R}$ denote a set of covariate-response pairs, where the first n points denote the training set and last entry is the test point for which Y_{n+1} is unobserved. Given a target level $\tau \in (0, 1)$, we will be interested in quantile regression estimates of the form

$$(\hat{\beta}_0, \hat{\beta}) = \underset{(\beta_0,\beta)\in\mathbb{R}^{d+1}}{\operatorname{argmin}} \sum_{i=1}^{n} \ell_\tau(Y_i - \beta_0 - X_i^\top \beta) + \mathcal{R}(\beta),$$

where $\ell_\tau(r) = \tau r - \min\{r, 0\}$ is the usual pinball loss and $\mathcal{R} : \mathbb{R}^d \to \mathbb{R}$ is an optional regularization function. For d fixed and n tending to infinity, the quantile regression estimates satisfy the target coverage guarantee $\mathbb{P}(Y_{n+1} \leq \hat{\beta}_0 + X_{n+1}^\top \hat{\beta}) \to \tau$. Our goal in this article is to adjust the quantile regression procedure to recover this guarantee even in cases where $d/n \to \gamma \in (0, \infty)$ converges to a constant.

1.1 Overview of the methods of Gibbs et al. (2025)

As discussed above, the task of removing the coverage bias of quantile regression has been previously considered by Gibbs et al. (2025). They propose to adjust the quantile regression by adding an imputed guess for the test point into the fit. Concretely, they consider unpenalized regressions of the form

$$(\hat{\beta}_0^{\text{adj.},y}, \hat{\beta}^{\text{adj.},y}) = \underset{(\beta_0,\beta)\in\mathbb{R}^{d+1}}{\operatorname{argmin}} \sum_{i=1}^{n} \ell_\tau(Y_i - \beta_0 - X_i^\top \beta) + \ell_\tau(y - \beta_0 - X_{n+1}^\top \beta), \tag{1.1}$$

and define the adjusted quantile estimate

$$\hat{q}_{\text{GCC}}(X_{n+1}) = \sup\{y : y \leq \hat{\beta}_0^{\text{adj.},y} + X_{n+1}^T \hat{\beta}^{\text{adj.},y}\},$$

as the maximum value of y that is covered by the regression fit with y in place of Y_{n+1}. Under no assumptions on the data beyond that they are i.i.d., this adjustment has the conservative coverage guarantee $\mathbb{P}(Y_{n+1} \leq \hat{q}_{\text{GCC}}(X_{n+1})) \geq \tau$.

Unfortunately, the coverage of this method is not typically tight and the authors find that $\hat{q}_{\text{GCC}}(X_{n+1})$ can exhibit significant overcoverage bias in high dimensions. To further correct this and obtain exact coverage, they additionally introduce a smaller, randomized threshold that is constructed using the quantile regression dual. More formally, let $r_{n+1} = y - \beta_0 - X_{n+1}^\top \beta$ and $r_i = Y_i - \beta_0 - X_i^\top \beta$ for $i \in \{1, \ldots, n\}$ denote a set of primal variables that are constrained to be equal to the residuals. Let $\eta \in \mathbb{R}^{n+1}$ denote the corresponding dual variables for these constraints. Then, the adjusted quantile regression (1.1) can be equivlanely written in its primal form as

$$\begin{aligned}(\hat{\beta}_0^{\text{adj.,y}}, \hat{\beta}^{\text{adj.,y}}, \hat{r}^{\text{adj.,y}}) = \underset{(\beta_0,\beta)\in\mathbb{R}^{d+1}, r\in\mathbb{R}^{n+1}}{\text{argmin}} \quad & \sum_{i=1}^{n+1} \ell_\tau(r_i) \\ \text{subject to} \quad & r_{n+1} = y - \beta_0 - X_{n+1}^\top \beta, \\ & r_i = Y_i - \beta_0 - X_i^\top \beta, \ \forall i \in \{1, \ldots, n\},\end{aligned}$$

with associated Lagriangian,

$$L(\beta_0, \beta, r, \eta) = \sum_{i=1}^{n+1} \ell_\tau(r_i) + \sum_{i=1}^{n} \eta_i (Y_i - \beta_0 - X_i^\top \beta - r_i) + \eta_{n+1}(y - \beta_0 - X_{n+1}^\top \beta - r_{n+1}),$$

and dual program,

$$\begin{aligned}\hat{\eta}^{\text{adj.,}y} = \underset{\eta\in\mathbb{R}^{n+1}}{\text{argmax}} & \sum_{i=1}^{n} \eta_i Y_i + \eta_{n+1} y \\ \text{subject to} \ & \sum_{i=1}^{n+1} \eta_i = 0, \ \sum_{i=1}^{n+1} \eta_i X_i = 0, \ -(1-\tau) \preceq \eta \preceq \tau.\end{aligned}$$

To connect the dual variables to coverage, note that differentiating the Lagriangian with respect to r_{n+1} gives the first-order condition

$$\hat{\eta}_{n+1}^{\text{adj.,}y} \in \begin{cases} \{\tau\}, \ y > \hat{\beta}_0^{\text{adj.,}y} + X_{n+1}^\top \hat{\beta}^{\text{adj.,}y}, \\ \{-(1-\tau)\}, \ y < \hat{\beta}_0^{\text{adj.,}y} + X_{n+1}^\top \hat{\beta}^{\text{adj.,}y}, \\ [-(1-\tau), \tau], \ y = \hat{\beta}_0^{\text{adj.,}y} + X_{n+1}^\top \hat{\beta}^{\text{adj.,}y}. \end{cases}$$

This connection, along with some additional calculations, motivates the randomized quantile adjustment $\hat{q}_{\text{GCC, rand.}}(X_{n+1}) = \sup\{y : \hat{\eta}^y_{n+1} \leq U\}$, where $U \sim \text{Unif}(-(1-\tau), \tau)$. Crucially, this method has the desired exact coverage guarantee, $\mathbb{P}(Y_{n+1} \leq \hat{q}_{\text{GCC, rand.}}(X_{n+1})) = \tau$.

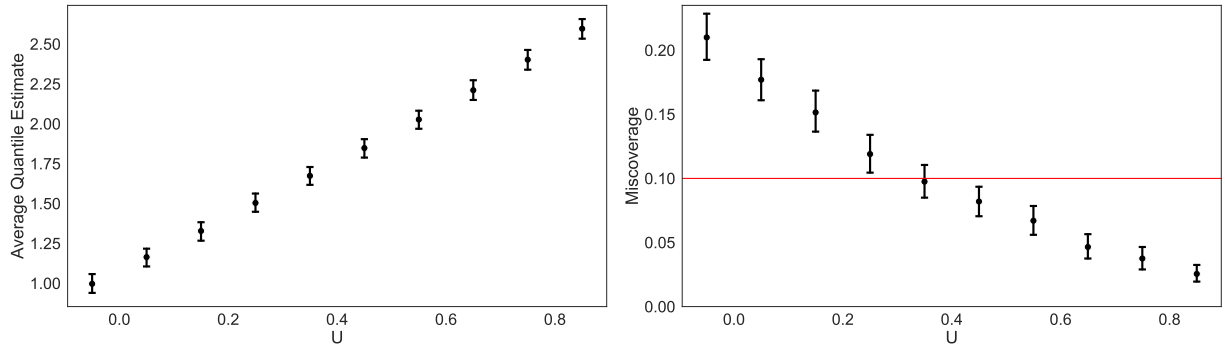

Figure 2: Empirical estimates of the average adjusted quantile (left panel) and miscoverage (right panel) of the randomized method of Gibbs et al. (2025) conditional on the cutoff, U. Data for these experiment are sampled from the Gausian linear model $Y_i = X_i^\top \beta + \epsilon_i$ where $X_i \sim \mathcal{N}(0, I_d)$ and $\epsilon_i \sim \mathcal{N}(0, 1)$ with $X_i \perp\!\!\!\perp \epsilon_i$. Dots and error bars show means and 95% confidence intervals obtained over 2000 samples of the combined training and test dataset $\{(X_i, Y_i)\}_{i=1}^{n+1}$. Throughout we set $d = 40$ and $n = 200$. The red line in the right panel indicates the target miscoverage level of $1 - \tau = 0.1$.

As discussed above, this method has two shortcomings. The first is that to compute the cutoff we need to evaluate the solution path of $\hat{\eta}^y_{n+1}$ as y varies. Although Gibbs et al. (2025) give some strategies for accomplishing this in an efficient manner, their methods still typically require additional computational time of at least $O(d^3)$[†] per test point. Adapting their methods to penalized regression is more challenging and requires even higher computational complexity. This contrasts sharply with standard quantile regression which can issue predictions quickly at the low cost of computing the inner product $X_{n+1}^\top \hat{\beta}$. The second major shortcoming of

[†]This comes from the cost of inverting a $d \times d$ matrix, which we shorthand as requiring $O(d^3)$ time although some algorithms with faster scaling are known.

$\hat{q}_{\text{GCC, rand.}}(X_{n+1})$ is that its value depends heavily on the randomized choice of U. Figure 2 displays estimates of the average conditional cutoff, $\mathbb{E}[\hat{q}_{\text{GCC, rand.}}(X_{n+1}) \mid U]$ and miscoverage, $\mathbb{P}(Y_{n+1} > \hat{q}_{\text{GCC, rand.}}(X_{n+1}) \mid U)$ as U varies on data sampled from the Gaussian linear model with $d/n = 0.2$. We see that the average cutoff can change by a factor of almost 2.5 and the miscoverage can vary by over $0.5 - 2$ times the target level depending on the sampled value of U. As an aside, we note that the exact magnitude of these values depends directly on the aspect ratio. In the classical case where $d/n \to 0$ the randomization disappears and the method (asymptotically) produces a fixed cutoff, while larger aspect ratios produce greater variability.

2 Methods

2.1 Debiasing quantile regression

We now introduce three alternative methods for debiasing quantile regression. As shown theoretically in Section 3 and empirically in Sections 2.3 and 4, all of these methods provide (asymptotically) exact coverage. Notably, this does not mean that their performance is identical. In Section 4 we compare the three approaches across a number of additional metrics (e.g. prediction set length, conditional coverage properties) and observe considerable variability. After reading the introduction to each method below, readers who are primarily interested in practical recommendations may choose to skip ahead to these results.

Fixed dual thresholding: Our first procedure makes a simple adjustment to $\hat{q}_{\text{GCC, rand.}}(X_{n+1})$ by replacing the randomized cutoff with a fixed threshold. This gives us

the adjusted quantile estimate

$$\hat{q}_{\text{dual thresh.}}(X_{n+1}; t) = \sup\left\{y : \hat{\eta}_{n+1}^{\text{adj.},y} \leq t\right\}.$$

At threshold t, the coverage of this method is given by $\mathbb{P}(Y_{n+1} \leq \hat{q}_{\text{dual thresh.}}(X_{n+1}; t)) = \mathbb{P}(\hat{\eta}^{\text{adj.},Y_{n+1}} \leq t)$. So, to obtain the target coverage level of τ we see that we should set t as the τ quantile of $\hat{\eta}^{\text{adj.},Y_{n+1}}$. Since this quantity is unknown, we replace it with the empirical estimate

$$\hat{t} = \text{Quantile}\left(\tau, \frac{1}{n}\sum_{i=1}^{n} \delta_{\hat{\eta}_i}\right),$$

where $\hat{\eta}$ denotes the dual variables fit using just the training data $\{(X_i, Y_i)\}_{i=1}^n$ and $\text{Quantile}(\tau, P)$ denotes the τ quantile of the distribution P. Corollary 3.1 below verifies that this quantile estimate is consistent in high dimensions and thus that this method provides the desired asymptotic coverage.

While this approach is derandomized, it still retains the same test-time computational complexity as the method of Gibbs et al. (2025). Our next two proposals will address this shortcoming.

Level adjustment: The second method we will consider is to modify the nominal level used in the quantile regression loss. In particular, we consider replacing the loss ℓ_τ with the adjustment $\ell_{\hat{\tau}^{\text{adj.}}}$, where $\hat{\tau}^{\text{adj.}}$ is tuned to ensure a final coverage level of τ. A similar approach has been previously proposed by Bai et al. (2021). They showed that when the aspect ratio is small and the data come from a stylized linear model the adjustment $\hat{\tau}^{\text{adj.}} = (\tau - \frac{1}{2}\frac{d}{n})/(1 - \frac{1}{2}\frac{d}{n})$ will asymptoically provide (approximately) the desired coverage. Here, we extend this procedure to be model agnostic by instead tuning the adjustment using leave-one-out cross-validation.

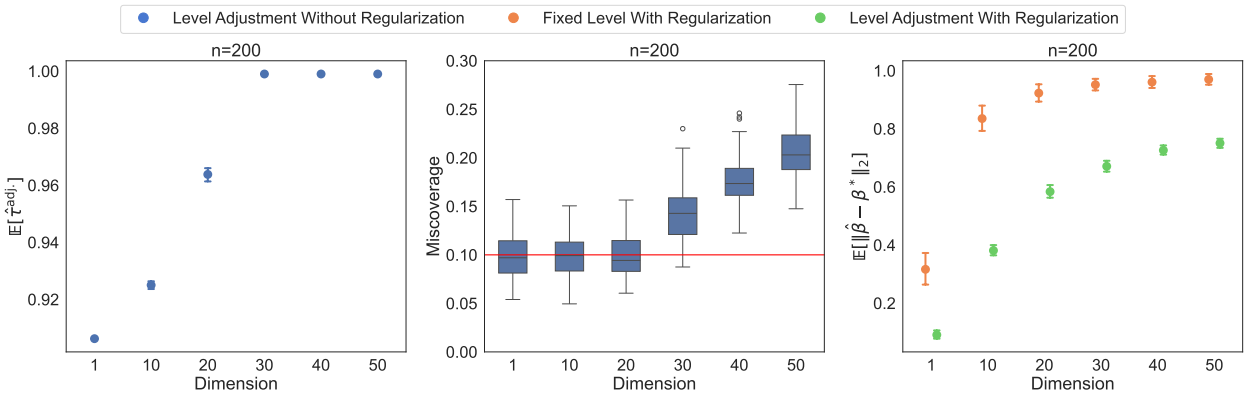

Figure 3: Average value of $\hat{\tau}^{\text{adj.}}$ (left panel), empirical miscoverage (center panel), and mean coefficient estimation error (right panel) of quantile regression fit with an adjusted level (blue), adjusted regularization (orange) and a joint level and regularization adjustment (green) as the dimension of the data varies. Data for these experiments are sampled from the Gaussian linear model $Y_i = X_i^\top \tilde{\beta} + \epsilon$ with $X_i \sim \mathcal{N}(0, I_d)$, $\epsilon_i \sim \mathcal{N}(0,1)$ and $\epsilon_i \perp\!\!\!\perp X_i$. Dots and error bars in the left and right panel show estimated means and 95% confidence intervals from 100 trials. Boxplots in the center panel show the empirical distribution of the training-conditional miscoverage evaluated over the same 100 trials where in each trial the miscoverage is estimated on a test set of size 2000. The red line shows the target miscoverage of $1 - \tau = 0.1$.

Unfortunately, tuning the level alone is not sufficient to regain coverage at higher aspect ratios. The center panel of Figure 3 shows the realized miscoverage of this method for increasing values of d/n on data generated from the Gaussian linear model. We see that for $d/n \leq 0.1$ the method successfully finds an adjusted level that restores coverage. On the other hand, for larger aspect ratios all values of $\hat{\tau}^{\text{adj.}}$ undercover. Results in the figure show the best possible choice of $\hat{\tau}^{\text{adj.}} \approx 1^{\ddagger}$ which still realizes a coverage bias of almost 5% at $d/n = 0.15$.

To obtain uniform coverage across higher aspect ratios, we will add regularization to the regression. For simplicity, we choose to focus our experiments on ridge regularization, though

[‡]We set $\hat{\tau}^{\text{adj.}}$ to be slightly less than 1 to have a well-defined quantile regression fit.

we anticipate that other choices would also be effective. This gives us the quantile regression,

$$(\hat{\beta}_0(\lambda, \tau^{\text{adj.}}), \hat{\beta}(\lambda, \tau^{\text{adj.}})) = \underset{(\beta_0,\beta)\in\mathbb{R}^{d+1}}{\text{argmin}} \sum_{i=1}^{n} \ell_{\tau^{\text{adj.}}}(Y_i - \beta_0 - X_i^\top \beta) + \lambda\|\beta\|_2^2.$$

Let $(\hat{\beta}_0^{-i}(\lambda, \tau^{\text{adj.}}), \hat{\beta}^{-i}(\lambda, \tau^{\text{adj.}}))$ denote coefficients fit by this regression when the i_{th} sample is left out. To obtain the desired coverage, we set the hyperparameter as any values,

$$(\hat{\lambda}, \hat{\tau}^{\text{adj.}}) \in \left\{(\lambda, \tau^{\text{adj.}}) : \left|\frac{1}{n}\sum_{i=1}^{n} \mathbb{1}\left\{Y_i \le \hat{\beta}_0^{-i}(\lambda, \tau^{\text{adj.}}) + X_i^\top \hat{\beta}^{-i}(\lambda, \tau^{\text{adj.}})\right\} - \tau\right| \le \epsilon\right\}, \tag{2.1}$$

that give a leave-one-out coverage of approximately τ. This gives us the final adjusted quantile estimate,

$$\hat{q}_{\text{level-reg.}}(X_{n+1}) = \hat{\beta}_0(\hat{\lambda}, \hat{\tau}^{\text{adj.}}) + X_{n+1}^\top \hat{\beta}(\hat{\lambda}, \hat{\tau}^{\text{adj.}}).$$

As an aside, we note that the numerical error tolerance parameter ϵ introduced here is not critical and is used simply to account for the fact that the leave-one-out coverage can only take values in $\{0, \frac{1}{n}, \frac{2}{n}, \ldots, 1\}$. In all of our experiments we simply set $\epsilon = 1/n$. Additionally, we remark that, in general, there will be more than one set of hyperparameters that satisfy (2.1). To choose a specific setting, we will use an auxiliary multiaccuracy condition. We defer a precise definition of this procedure to Section 4.1 where we discuss other goals for quantile regression beyond marginal coverage.

Before moving on, it is worthwhile to ask if level-tuning is necessary or if coverage could more easily be obtained by simply holding $\tau^{\text{adj.}} = \tau$ fixed and adjusting the regularization alone. Empirically, we find that while such a strategy is feasible, it typically leads to over regularization. To illustrate this, the right panel of Figure 3 compares the estimation error of tuning just the regularization alone against that of tuning the regularization and nominal level jointly on data generated from a well-specified Gaussian linear model. In both cases, we set the hyperparameters in order to guarantee a fixed coverage level of $\tau = 0.9$. More

concretely, joint regularization and level tuning is done using the procedure outlined in (4.3) below, while when tuning the regularization level alone we set λ to be the smallest value that gives a leave-one-out coverage of at least $\tau - 1/n$. We find that joint regularization and level tuning gives a smaller estimation error uniformly across all aspect ratios. As a result, we will prefer this method in the sections that follow and omit further investigation of sole regularization tuning.

Additive adjustment: The final method we will consider is applying an additive adjustment to the quantile estimate. One way to implement such an adjustment would be to fit the parameters $(\hat{\beta}_0, \hat{\beta})$ using a standard quantile regression and then, at prediction time, output the corrected estimate $c + \hat{\beta}_0 + X_{n+1}^\top \hat{\beta}$ for some constant $c \in \mathbb{R}$. This approach has been previously considered by Romano et al. (2019) under the name conformalized quantile regression. They propose to fit the parameter c using a held out subset of the training data that is not used in the quantile regression. In high-dimensional problems where data is scarce withholding data from the initial regression may lead to a considerable drop in efficiency. In the following section, we will develop a computationally efficient leave-one-out cross-validation procedure that facilitates accurate parameter tuning without data splitting. To leverage that theory here, we now introduce an alternative method for computing an additive adjustment.

For any fixed $c \in \mathbb{R}$, let $\hat{\beta}^c$ denote the coefficients fit in the intercept-less quantile regression,

$$\hat{\beta}^c = \operatorname*{argmin}_{\beta \in \mathbb{R}^d} \sum_{i=1}^{n} \ell_\tau(Y_i - c - X_i^\top \beta). \tag{2.2}$$

Let $\hat{\beta}^{c,-i}$ denote the coefficients fit by same regression with datapoint i excluded. Our goal is to find a parameter setting

$$\hat{c} \in \left\{ c : \left| \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}\{Y_i \leq c + X_i^\top \hat{\beta}^{c,-i}\} - \tau \right| \leq \epsilon \right\}, \tag{2.3}$$

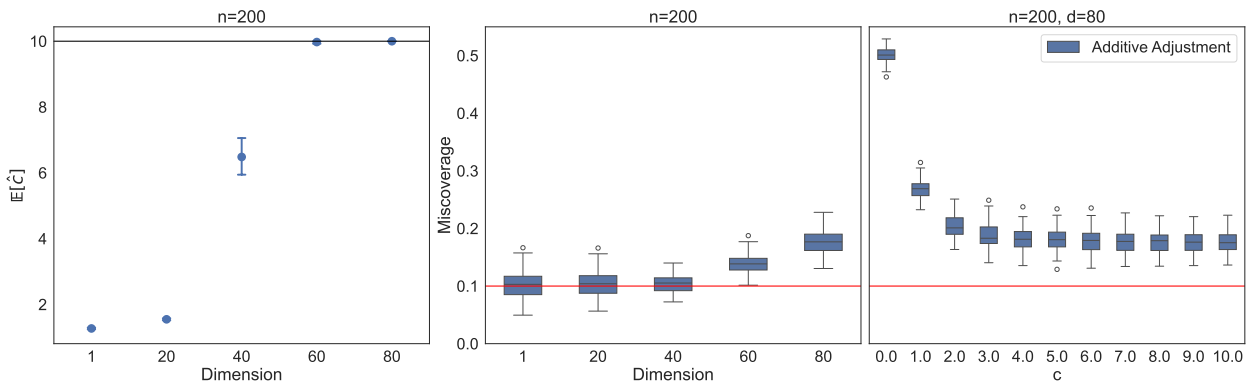

Figure 4: Empirical estimate of the mean selected value of $\hat{c}$ (left panel), realized miscoverage for varying dimension (center panel), and realized miscoverage as c varies (right panel) of the unregularized additive adjustment (equations (2.2) and (2.3)) Data for this experiment are sampled from the Gaussian linear model $Y_i = X_i^\top \tilde{\beta} + \epsilon$ with $X_i \sim \mathcal{N}(0, I_d)$, $\epsilon_i \sim \mathcal{N}(0, 1)$ and $\epsilon_i \perp\!\!\!\perp X_i$. Dots and error bars in the left panel show estimated means and 95% confidence intervals taken over 100 trials. Boxplots in the center and right panel show the empirical distribution of the training-conditional miscoverage evaluated over the same 100 trials where in each trial the miscoverage is estimated on a test set of size 2000. The black line in the left panel shows the maximum allowable value for $\hat{c}$, while the red lines in the center and right panel shows the target miscoverage of $1 - \tau = 0.1$.

that gives a leave-one-out coverage of approximately τ. However, similar to level-tuning, we find empirically that for large aspect ratios there is often no value of c that satisfies this criteria. Figure 4 demonstrates this on simulated data from a Gaussian linear model. In this experiment the maximum allowable value for $\hat{c}$ is capped at 10. We see that for $d/n \geq 0.3$ the method almost always selects this value and, despite using the largest possible choice of $\hat{c}$, still undercovers. This issue cannot be alleviated by increasing the cap on $\hat{c}$ as larger values do not change the coverage (right panel).

As was the case for tuning of the nominal level, coverage at higher aspect ratios can be

restored by adding regularization. Let

$$\hat{\beta}^{c,\lambda} = \underset{\beta \in \mathbb{R}^d}{\operatorname{argmin}} \sum_{i=1}^{n} \ell_\tau(Y_i - c - X_i^\top \beta) + \lambda \|\beta\|_2^2,$$

denote the coefficients fit with added adjustment c and regularization level λ. Let $\hat{\beta}^{c,\lambda,-i}$ denote the same coefficients when point i is excluded from the fit and

$$(\hat{c}, \hat{\lambda}) \in \left\{ c : \left| \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}\{Y_i \leq c + X_i^\top \hat{\beta}^{c,\lambda,-i}\} - \tau \right| \leq \epsilon \right\} \tag{2.4}$$

denote any parameter setting that gives a leave-one-out coverage of approximately ϵ. Then, we consider the adjusted quantile estimate,

$$\hat{q}_{\text{add.-reg.}}(X_{n+1}; \hat{c}) = \hat{c} + X_{n+1}^\top \hat{\beta}^{\hat{c},\hat{\lambda}}.$$

Similar to level-tuning, in general there will be multiple parameter settings that satisfy (2.4). As above, we will choose a specific setting using an axillary multiaccuracy whose formal definition is deferred so Section 4.1.

2.2 Efficient leave-one-out cross-validation

Two of the methods developed in the previous section use leave-one-out cross-validation to select their hyperparameters. The typical implementation of these procedures requires fitting n quantile regressions across a large range of parameter settings. In this section, we derive a connection between the leave-one-out coverage and the dual variables that allows us to obtain all n leave-one-out coverage indicators with just a single fit. Hyperparameter tuning can then be performed at the cost of just a few regression fits across a range of parameter values.

To introduce this method, we define a few pieces of additional notation. Throughout this section we consider quantile regressions of the form

$$\hat{w} = \underset{w \in \mathbb{R}^p}{\operatorname{argmin}} \sum_{i=1}^{n} \ell_\tau(\tilde{Y}_i - \tilde{X}_i^\top \hat{w}) + \mathcal{R}(w). \tag{2.5}$$

Note that unlike the previous sections, here we have chosen to omit an explicit intercept parameter. This allows us to unify the notation to encompass both our level-based adjustment, in which $\tilde{Y}_i = Y_i$, $\tilde{X}_i = (1, X_i)$, and $p = d+1$, and our additive adjustment, in which $\tilde{Y}_i = Y_i + c$, $\tilde{X}_i = X_i$, and $p = d$. Following the same steps as in Section 1.1, a useful dual for this regression can be obtained by defining the additional primal variables $r_i = \tilde{Y}_i - \tilde{X}_i^\top w$ for $i \in \{1, \dots, n\}$ and corresponding dual variables $\eta \in \mathbb{R}^n$ for these constraints. This gives the dual program

$$\hat{\eta} = \operatorname*{argmax}_{\eta \in \mathbb{R}^n} \sum_{i=1}^n \eta_i \tilde{Y}_i - \mathcal{R}^* \left(\sum_{i=1}^n \eta_i \tilde{X}_i \right)$$
$$\text{subject to } -(1-\tau) \preceq \eta_i \preceq \tau,$$

where $\mathcal{R}^*(\cdot)$ denotes the convex conjugate of $\mathcal{R}(\cdot)$. Finally, in what follows we let $\hat{w}^{-i}$ denote the primal solution when the i_{th} sample is omitted from the fit.

Our first result derives a general connection between the leave-one-out coverage and the sign of the dual variables.

Proposition 2.1. *Assume that $\mathcal{R}(\cdot)$ is convex. Then, all dual solutions $\hat{\eta}$ and leave-one-out primal solutions $\hat{w}^{-i}$ satisfy the conditions*

$$\tilde{Y}_i < \tilde{X}_i^\top \hat{w}^{-i} \implies \hat{\eta}_i \leq 0,$$

and

$$\tilde{Y}_i > \tilde{X}_i^\top \hat{w}^{-i} \implies \hat{\eta}_i \geq 0.$$

Now, recall that our goal is to compute the leave-one-out coverage, $\frac{1}{n} \sum_{i=1}^n \mathbb{1}\{\tilde{Y}_i \leq \tilde{X}_i^\top \hat{w}^{-i}\}$. The above proposition suggests that this quantity should be comparable to $\frac{1}{n} \sum_{i=1}^n \mathbb{1}\{\hat{\eta}_i \leq 0\}$. Unfortunately however, deriving an exact equivalence between these two quantities is not possible due to the ambiguity around the edge cases $\tilde{Y}_i = \tilde{X}_i^\top \hat{w}^{-i}$ and $\hat{\eta}_i = 0$. We are not

aware of any simple method for resolving these cases in full generality. One of the key difficulties is that without additional assumptions both the primal and dual solutions are not unique and at these edge cases the coverage can vary depending on which solution we select. The following example illustrates one such instance where this occurs.

Example 2.1. *Consider fitting an intercept only quantile regression with $\tau = 1/2$ to find the median of the three data points (Y_1, Y_2, Y_3). For simplicity, assume that $Y_1 < Y_2 < Y_3$. The primal solution is $\hat{w} = Y_2$ with corresponding dual variables $\hat{\eta} = (-1/2, 0, 1/2)$. Critically, we have that $\hat{\eta}_2 = 0$. Now, consider the leave-one-out problem when Y_2 is omitted. Then, the median is any point $\hat{w}^{-2} \in [Y_1, Y_3]$ and it is ambiguous whether Y_2 is covered.*

We will now introduce two different techniques for modifying the regression to avoid the above ambiguity. For simplicity, we focus specifically on cases where $\mathcal{R}(\cdot)$ is a quadratic regularizer, although we expect similar results to hold for other choices. Our first method is to perturb the covariates by adding a small amount of independent noise to each of their values. The magnitude of this noise is not critical and can be made arbitrarily small such that it has a vanishing impact on the quantile regression objective. Our insight is that even a small amount of noise is sufficient to push the dual solutions away from zero and, correspondingly, to enforce a unique value for the leave-one-out coverage. To illustrate this, the following demonstrates how added noise removes the ambiguity observed in Example 2.1.

Example 2.2. *We consider adding noise to the intercept parameter in Example 2.1. More concretely, consider fitting an intercept-less quantile regression on the three data points $\{(1 + \xi_1, Y_1), (1 + \xi_2, Y_2), (1 + \xi_3, Y_3)\}$ where ξ_1, ξ_2, and ξ_3 are i.i.d. continuously distributed random variables independent of (Y_1, Y_2, Y_3). As before, assume for simplicity that $Y_1 < Y_2 < Y_3$. For sufficiently small values of (ξ_1, ξ_2, ξ_3), the dual solution is uniquely specified as*

$\hat{\eta} = (-1/2, \frac{\xi_1 - \xi_3}{2(1+\xi_2)}, 1/2)$. *Moreover the leave-one-out primal solution with point* $(1 + \xi_2, Y_2)$ *omitted is (with probability one) unique and given by*

$$\hat{w}^{-2} = \begin{cases} \frac{1+\xi_1}{2(1+\xi_3)}, & \text{if } |1 + \xi_1| < |1 + \xi_3|, \\ -\frac{1+\xi_3}{2(1+\xi_1)}, & \text{if } |1 + \xi_1| > |1 + \xi_3|. \end{cases}$$

By the independence of (ξ_1, ξ_2, ξ_3) *we see that* $\mathbb{P}(Y_2 = (1 + \xi_2)\hat{w}^{-2}) = 0$ *and thus there is no ambiguity in the coverage of the leave-one-out solution.*

The second method we will consider is to add non-zero L_2 regularization to all of the primal variables. Similar to the added noise, the magnitude of this regularization can be arbitrary and, in particular, can be taken to be vanishingly small such that it has almost no impact on the regression. The only important consideration is that the regularization makes the fitted leave-one-out solutions unique and thus removes ambiguity in the coverage.

Assumptions 1 and 2 give more formal statements of our two approaches for ensuring leave-one-out uniqueness. We note that both of these assumptions require that the distribution of $\tilde{Y}_i \mid \tilde{X}_i$ is continuous. This can always be guaranteed by adding a small amount of noise to $\tilde{Y}_i$. The main result of this section is stated in Theorem 2.1 which shows that these assumptions are sufficient to ensure a one-to-one equivalence between the leave-one-out coverage and the signs of the dual variables.

Assumption 1. *The distribution of* $\tilde{Y}_i \mid \tilde{X}_i$ *is continuous. Moreover, the regularization can be written as* $\mathcal{R}(w) = \sum_{j=1}^{p} \lambda_j w_j^2$ *for some non-negative constants* $\lambda_1, \ldots, \lambda_p \in \mathbb{R}_{\geq 0}$ *and the covariates can be written as* $\tilde{X}_i = Z_i + \xi_i$ *where* $\xi_i \perp\!\!\!\perp (Z_i, Y_i)$ *has independent, continuously distributed entries. Finally, we have that* $p < n$.

Assumption 2. *The distribution of* $\tilde{Y}_i \mid \tilde{X}_i$ *is continuous. Moreover, the regularization can*

be written as $\mathcal{R}(w) = \sum_{j=1}^{p} \lambda_j w_j^2$ *for some positive constants* $\lambda_1, \ldots, \lambda_p > 0$.

Theorem 2.1. *Assume that* $\{(\tilde{X}_i, \tilde{Y}_i)\}_{i=1}^n$ *are i.i.d. and that the conditions of either Assumption 1 or 2 are satisfied. Then, with probability one we have that for all* $i \in \{1, \ldots, n\}$ *either all dual solutions satisfy* $\hat{\eta}_i < 0$ *or all dual solutions satisfy* $\hat{\eta}_i > 0$. *Similarly, with probability one either all leave-one-out primal solutions satisfy* $\tilde{Y}_i < \tilde{X}_i^\top \hat{w}^{-i}$ *or all leave-one-out primal solutions satisfy* $\tilde{Y}_i > \tilde{X}_i^\top \hat{w}^{-i}$. *Finally, letting* $\hat{\eta}$ *and* $\{\hat{w}^{-i}\}_{i=1}^n$ *denote any such solutions we have that*

$$\frac{1}{n}\sum_{i=1}^{n} \mathbb{1}\{\hat{\eta}_i \leq 0\} \overset{a.s.}{=} \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}\{\tilde{Y}_i \leq \tilde{X}_i^\top \hat{w}^{-i}\}.$$

In general, on real data we find that the conditions outlined in Assumptions 1 and 2 tend to be redundant. In our experiments in Sections 4 and 2.3 we will ignore these assumptions and use the dual variables to estimate the leave-one-out coverage and perform hyperparameter selection across a variety of different datasets and regularization settings that do not satisfy these conditions. In all cases, we find that the dual estimate is accurate and facilitates the selection of hyperparameter values that yield reliable coverage. As a result, outside of rare edge cases we find that $\frac{1}{n}\sum_{i=1}^{n} \mathbb{1}\{\hat{\eta}_i \leq 0\}$ can typically be used to estimate the leave-one-out coverage without the need to modify the data or estimation procedure.

2.3 Simulated example

To conclude this section, we give a brief simulated example demonstrating that all of the methods proposed above give accurate coverage in high dimensions. More extensive comparisons that evaluate the methods across a number of additional metrics are given in Section 4. Similar to Figure 1 from the introduction, we generate data from the Gaussian linear model $Y_i = X_i^\top \tilde{\beta} + \epsilon$ with $X_i \sim \mathcal{N}(0, I_d)$, $\epsilon_i \sim \mathcal{N}(0, 1)$, and $\epsilon_i \perp\!\!\!\perp X_i$. Figure 5 shows

the realized coverage of the three methods proposed in Section 2.1. As a baseline, we also show the coverage of standard quantile regression. We see that all three methods offer robust coverage irrespective of the aspect ratio which becomes more tightly concentrated on the target level as n and d increase.

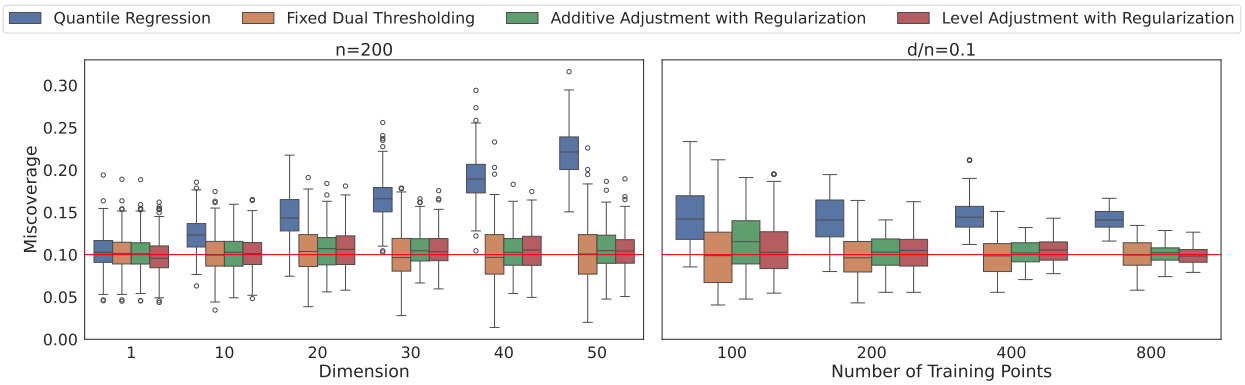

Figure 5: Empirical distribution of the training-conditional miscoverage of quantile regression (blue) and our fixed dual thresholding (orange), additive adjustment (green), and level adjustment (red) methods. The left panel shows results obtained with varying dimension and a fixed sample size of $n = 200$, while the right panel varies n and d together at a fixed aspect ratio of $d/n = 0.1$. Boxplots show results from 200 trials where in each trial the miscoverage is evaluated empirically over a test set of size 2000.

3 High-dimensional consistency

We now develop our main theoretical results establishing the high-dimensional consistency of the estimates proposed in the previous section. Throughout, we will work in a stylized linear model with Gaussian covariates that is commonly used in work in this area (e.g., Donoho & Montanari (2013), Thrampoulidis et al. (2018), Hastie et al. (2022)). While we will not pursue this in detail, universality results derived for related problems suggest that one should expect our results to also hold under more relaxed assumptions (e.g. X_i having i.i.d. entries) (Han & Shen 2023). This is validated by the empirical results presented in the following

section which demonstrate the robustness of our methods on real datasets.

Assumption 3. *The data* $\{(X_i, Y_i)\}_{i=1}^n$ *are i.i.d. and distributed as* $Y_i = X_i^\top \tilde{\beta} + \epsilon_i$ *with* $X_i \sim \mathcal{N}(0, I_d)$, $\epsilon_i \sim P_\epsilon$, *and* $\epsilon_i \perp\!\!\!\perp X_i$. *Moreover, the error distribution* P_ϵ *is continuous, mean zero, has positive density on* $\mathbb{R}$, *and at least two bounded moment. The true coefficients are themselves random and sampled as* $(\sqrt{d}\tilde{\beta}_j)_{j=1}^d \overset{i.i.d.}{\sim} P_\beta$ *independent of* $\{(X_i, \epsilon_i)\}_{i=1}^n$.

We will focus on quantile regressions of the form

$$\min_{(\beta_0,\beta)\in\mathbb{R}^{d+1}} \sum_{i=1}^n \ell_\tau(Y_i - \beta_0 - X_i^\top \beta) + \lambda_d \sum_{j=1}^d \nu(\beta_j). \tag{3.1}$$

We make three remarks about this set-up. First, for simplicity, we have chosen to focus on regressions containing an intercept. To obtain results for our additive adjustment method we will also need to consider cases where β_0 is replaced by a fixed constant. This extension is stated at the end of this section in Theorem 3.3. Second, here we have allowed the regularization level λ_d to depend explicitly on the dimension. This is done to account for the fact that the regularization level may need to be rescaled as n and d increase. Third, although we have chosen to focus on separable penalties, we expect our results can be readily extended to other choices. Our formal assumptions on ν are stated in Assumption **??** the appendix. At a high-level, we require that $\nu(\cdot)$ is convex with $\nu(0) = 0$ and that the data have enough bounded moments to ensure that various functions of ν satisfy the law of large numbers. Lemma **??** verifies that are assumptions are met if P_β has four bounded moments and $\lambda_d \sum_{j=1}^d \nu(\beta_j) = d^{-1/2}\lambda\|\beta\|_1$ or $\lambda_d \sum_{j=1}^d \nu(\beta_j) = \lambda\|\beta\|_2^2$ is L_1 or L_2 regularization.

We now state the main result of this section, which establishes that the coordinate-wise empirical distribution of the dual variables converges to an asymptotic limit. Although we only state this result for aspect ratios $d/n \to \gamma \in (0, \sqrt{2/\pi})$, we expect a similar result to hold for $\gamma \geq \sqrt{2/\pi}$ under appropriate assumptions on the regularization.

Theorem 3.1. *Fix any $\tau \in (0,1)$ and let $\hat{\eta}$ denote the dual variables to the quantile regression (3.1) where the data satisfy Assumption 3 and the regularizer satisfies Assumption* **??**. *Suppose that $d, n \to \infty$ with $d/n \to \gamma \in (0,1)$. Then, there exists a limiting distribution P_η such that for any bounded, Lipschitz function ψ,*

$$\frac{1}{n}\sum_{i=1}^{n} \psi(\hat{\eta}_i) \xrightarrow{\mathbb{P}} \mathbb{E}_{Z\sim P_\eta}[\psi(Z)].$$

The distribution P_η is supported on $[-(1-\tau), \tau]$ with discrete masses at $-(1-\tau)$ and τ and a continuous distribution with positive density on $(-(1-\tau), \tau)$.

Theorem 3.1 has two critical corollaries for our debiasing methods. The first shows that the quantile estimated used by our fixed dual thresholding method is consistent.

Corollary 3.1. *Under the setting of Theorem 3.1,*

$$\textit{Quantile}\left(\tau, \frac{1}{n}\sum_{i=1}^{n} \delta_{\hat{\eta}_i}\right) \xrightarrow{\mathbb{P}} \textit{Quantile}\left(\tau, P_\eta\right).$$

Our second corollary establishes the consistency of our leave-one-out coverage estimates. As above, we focus on the case of L_2 regularization though we expect similar results to hold for other choices.

Corollary 3.2. *Let (X_{n+1}, Y_{n+1}) denote an independent sample from the same distribution as $\{(X_i, Y_i)\}_{i=1}^n$ and $(\hat{\beta}_0, \hat{\beta})$ denote any minimizer of (3.1). Suppose that $\mathcal{R}_d(\beta) = d\lambda\|\beta\|_2^2$ for some fixed $\lambda \geq 0$. Then, under the assumptions of Theorem 3.1,*

$$\frac{1}{n}\sum_{i=1}^{n} \mathbb{1}\{\hat{\eta}_i \leq 0\} - \mathbb{P}\left(Y_{n+1} \leq \hat{\beta}_0 + X_{n+1}^\top \hat{\beta}\right) \xrightarrow{\mathbb{P}} 0.$$

Proofs of Theorem 3.1 as well as Corollaries 3.1 and 3.2 are given in the appendix. Our arguments build heavily on Gordon's comparison inequalities (Gordon 1985, 1988) and their

application to high-dimensional regression developed in Thrampoulidis et al. (2018). At a high level, these results allow us to derive a correspondence between the dual quantile regression problem and a simplified auxiliary optimization that replaces the covariate matrix with vector-valued random variables. The main technical difficulty is then to characterize the solutions of this auxiliary optimization. A key difference between our result and that of the original work of Thrampoulidis et al. (2018) is that we consider the behaviour of the solutions under arbitrary bounded, Lipschitz functions. That said, we are not the first to derive an extension of this type. However, to the best of our knowledge previous extensions typically rely on strong convexity of the auxillary optimization (see e.g., Abbasi et al. (2016), Miolane & Montanari (2021), Celentano et al. (2023)). Here, we derive a similar result under weaker conditions.

It is worthwhile to constrast Theorem 2.1 with the results of Bai et al. (2021). In that paper, the authors derive a number of asymptotic consistency results for the primal quantile regression estimates $(\hat{\beta}_0, \hat{\beta})$. Here, we provide a set of complimentary asymptotics for the dual. In addition, we also treat a more general setting that removes the restrictions to small aspect ratios and unregularized regressions present in their work.

Theorem 3.2. *Under the assumptions of Theorem 3.1, for all C_u and C_{β_0} sufficiently large the asymptotic program (B.2) admits unique solutions (M_u^*, β_0^*) for (M_u, β_0) with $M_u^* < C_u$ and $|\beta_0^*| < C_\beta^0$. Moreover, we have that for all $\delta > 0$*

$$\mathbb{P}\left(\textit{For all primal solutions to (B.1)}, \; |\|\hat{\beta} - \tilde{\beta}\|_2 - M_u^*| < \delta \textit{ and } |\hat{\beta}_0 - \beta_0^*| < \delta\right) \to 1.$$

Finally, as a last remark, we note that all of the conclusions stated above also hold for the intercept-less regression used in the additive adjustment procedure. The proof of this result is identical to that of Theorem 3.1 and thus is ommitted.

Theorem 3.3. *Under identical assumptions, the conclusions of Theorem 3.1 and Corollaries 3.1 and 3.2 remain true when the intercept β_0 is replaced by a fixed, real-valued constant.*

4 Real data experiments

4.1 Methods and Metrics

We now undertake a series of empirical comparisons of our proposed methods. As baselines, we also evaluate the performance of standard quantile regression, the randomized method of Gibbs et al. (2025), and the conformalized quantile regression (CQR) method of Romano et al. (2019). In all experiments we implement conformalized quantile regression so 75% of the data is used to fit the quantile regression and 25% is used to calibrate its coverage.

To evaluate these methods, we compare the quality of prediction sets constructed using their estimated quantiles. More precisely, for a given miscoverage level $\alpha \in (1/2, 1)$ (taken to be 0.1 in our experiments) we compute the (adjusted) quantile estimates $\hat{q}^{\alpha/2}(X_{n+1})$ and $\hat{q}^{1-\alpha/2}(X_{n+1})$ using each of the methods. We then evaluate the resulting prediction interval $[\hat{q}^{\alpha/2}(X_{n+1}), \hat{q}^{1-\alpha/2}(X_{n+1})]$ in terms of three criteria: 1) marginal coverage, $\mathbb{P}(\hat{q}^{\alpha/2}(X_{n+1}) \le Y_{n+1} \le \hat{q}^{1-\alpha/2}(X_{n+1}))$, 2) interval length, $\max\{\hat{q}^{1-\alpha/2}(X_{n+1}) - \hat{q}^{\alpha/2}(X_{n+1}), 0\}$, and 3) maximum multiaccuracy error.

Multiaccuracy as introduced in Hébert-Johnson et al. (2018), Kim et al. (2019) is a general criteria for measuring the bias of a predictor over reweightings of the covariate space. In the context of quantile regression, Duchi (2025) showed that (under appropriate tail bounds on the data) in the classical regime where $d \log(n)/n \to 0$ the vanilla quantile regression

estimates $(\hat{q}_{\mathrm{QR}}^{\alpha/2}, \hat{q}_{\mathrm{QR}}^{1-\alpha/2})$ satisfy the multiaccuracy condition,

$$\sup_{v\in\mathbb{R}^d} \mathbb{E}\left[X_{n+1}^\top v(\mathbb{1}\{\hat{q}_{\mathrm{QR}}^{\alpha/2}(X_{n+1}) \le Y_{n+1} \le \hat{q}_{\mathrm{QR}}^{1-\alpha/2}(X_{n+1})\} - (1-\alpha)) \mid \{(X_i,Y_i)\}_{i=1}^n\right] \xrightarrow{\mathbb{P}} 0. \quad (4.1)$$

As a concrete example to motivate the utility of this condition, consider fitting quantile regression with a feature $X_{i,j} = \mathbb{1}\{X_i \in G\}$ that indicates whether sample i falls into group G. Then, applying (4.1) with $v = e_i$ gives the conditional coverage statement,

$$\mathbb{P}(\hat{q}_{\mathrm{QR}}^{\alpha/2}(X_{n+1}) \le Y_{n+1} \le \hat{q}_{\mathrm{QR}}^{1-\alpha/2}(X_{n+1}) \mid X_{n+1} \in G, \{(X_i,Y_i)\}_{i=1}^n) \xrightarrow{\mathbb{P}} 1-\alpha.$$

More generally, by designing the features appropriately multiaccuracy conditions of this form can be used to ensure that the prediction set provides accurate performance across sensitive attributes of the population.

Motivated by this, Gibbs et al. (2025) extend (4.1) to the high-dimensional setting and show that their randomized adjustment satisfies

$$\mathbb{E}\left[X_{n+1}^\top v(\mathbb{1}\{\hat{q}_{\mathrm{GCC,\ rand.}}^{\alpha/2}(X_{n+1}) \le Y_{n+1} \le \hat{q}_{\mathrm{GCC,\ rand.}}^{1-\alpha/2}(X_{n+1})\} - (1-\alpha))\right], \ \forall v \in \mathbb{R}^d.$$

Notably, this statement is not directly comparable to (4.1) since here the expectation is taken marginally over the random draw of the training set. In general, one cannot expect to obtain training-conditional convergence uniformly over v in the high-dimensional regime. Nevertheless, as we will see shortly, empirically $\hat{q}_{\mathrm{GCC,\ rand.}}(X_{n+1})$ can provide approximate validity when v is restricted to a smaller set (e.g. to the coordinate axes).

The methods developed in the previous section do not explicitly target multiaccuracy. Regardless, since they are built on top of quantile regression one may hope that they still approximately satisfy these conditions. To evaluate this, we will examine the coordinatewise multiaccuracy error of each method defined as

$$\max_{j\in\{1,\ldots,d\}} \frac{\mathbb{E}[X_{n+1,j}(\mathbb{1}\{\hat{q}^{\alpha/2}(X_{n+1}) \le Y_{n+1} \le \hat{q}^{1-\alpha/2}(X_{n+1})\} - (1-\alpha)) \mid \{(X_i,Y_i)\}_{i=1}^n]}{\mathbb{E}[|X_{n+1,j}|]}. \quad (4.2)$$

In order to improve the performance on this metric, we will choose the parameters for our level and additive adjustment procedures in order to minimize a leave-one-out estimate of (4.2). For instance, at each $\tau \in \{\alpha/2, 1-\alpha/2\}$ we will choose the parameters $(\hat{\lambda}(\tau), \hat{\tau}^{\text{adj.}}(\tau))$ as a solution to the program

$$\begin{aligned} &\underset{\lambda, \tau^{\text{adj.}}}{\text{minimize}} \max_{j \in \{1,\ldots,d\}} \frac{\frac{1}{n}\sum_{i=1}^{n} X_{i,j}(\mathbb{1}\{\hat{\eta}_i(\lambda, \tau^{\text{adj.}}) \leq 0\} - \tau)}{\frac{1}{n}\sum_{i=1}^{n} |X_{i,j}|} \\ &\text{subject to } \left|\frac{1}{n}\sum_{i=1}^{n}(\mathbb{1}\{\hat{\eta}_i(\lambda, \tau^{\text{adj.}}) \leq 0\} - \tau\right| \leq \frac{1}{n}, \end{aligned} \tag{4.3}$$

where $\hat{\eta}(\lambda, \tau^{\text{adj.}})$ denotes the dual variables fit with parameters $(\lambda, \tau^{\text{adj.}})$ and the minimization is over a grid of values for $(\lambda, \tau^{\text{adj.}})$. The parameters for the additive adjustment method are chosen similarly.

4.2 Results

Our first experiment considers the Communities and Crime dataset (Dua & Graff (2017), Redmond & Baveja (2002)). Here, the goal is to predict the per capita violent crime rate of various communities. After filtering out features with missing values, the dataset has 1994 samples and 99 covariates. We normalize all the features to have mean zero and variance one and then compare the methods discussed above in terms of their miscoverage, median length, and multiaccuracy error.

Figure 6 shows the outcome of this experiment. Results in the figure summarize 20 trials where in each trial the data are randomly split into a training set of size 400 and a test set of size 1594 and a random subset $d \in \{1, 20, 40, 60, 80, 99\}$ of the features are selected for use. As shown in the left panel, all methods provide the desired marginal coverage except for standard quantile regression which realizes significant bias as the dimension increases. Among the methods with the desired coverage, our level adjustment procedure yields the

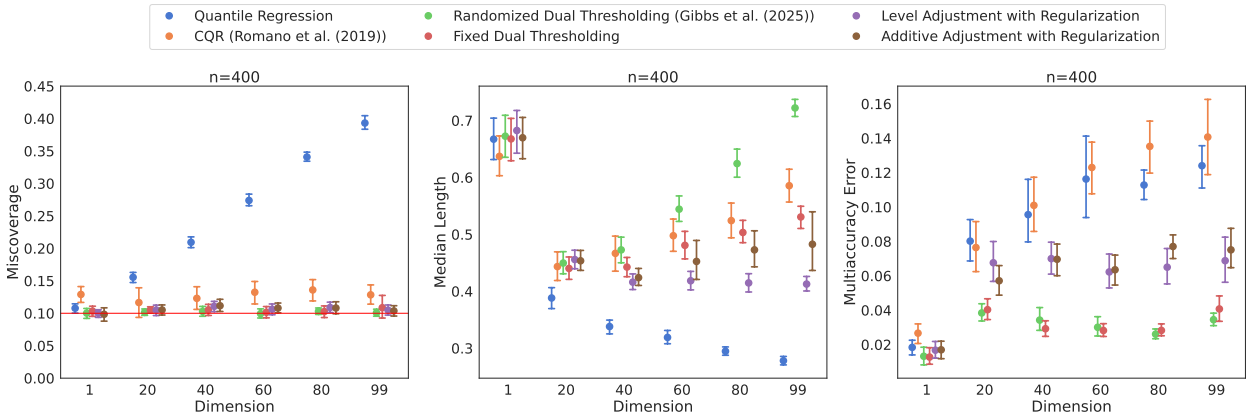

Figure 6: Empirical miscoverage (left panel), multiaccuracy error (center panel), and median length (right panel) of quantile regression (blue), the baseline methods of Romano et al. (2019) (orange) and Gibbs et al. (2025) (blue) and our fixed dual thresholding (red), level adjustment (purple), and additive adjustment (brown) methods on the communities and crime dataset. Dots and error bars show means and 95% confidence intervals obtained over 20 trials where in each trial the methods are evaluated on a test set of size 1594. The red line in the left panel shows the target level of $\alpha = 0.1$.

smallest intervals. The largest intervals are produced by the randomized method of Gibbs et al. (2025), which obtains a median interval length of almost two times that of the level adjustment procedure in higher dimensions.

In terms of multiaccuracy, the lowest error is obtained by the dual thresholding methods. Interestingly, while randomization is necessary to obtain a theoretical multiaccuracy bias of zero, we find that the fixed thresholding method offers nearly identical performance in practice. On the other hand, our level adjustment and additive adjustment procedures realize a higher multiaccuracy error. This is to be expected since by adding regularization to these methods we have introduced bias. To see this, note that letting $(\hat{\beta}_0(\lambda), \hat{\beta}(\lambda))$ denote the fitted coefficients at quantile level τ with L_2 regularization λ and $\tilde{\beta}$ denote the population

quantile regression coefficients, we have that in the classical regime where $d\log(n)/n \to 0$,

$$\mathbb{E}\left[X_{n+1}^\top v(\mathbb{1}\{Y_{n+1} \le \hat{\beta}_0(\lambda) + X_{n+1}^\top \hat{\beta}(\lambda)\} - \tau) \mid \{(X_i, Y_i)\}_{i=1}^n\right] \xrightarrow{\mathbb{P}} -2\lambda v^\top \tilde{\beta}.$$

This follows directly from the first order conditions of quantile regression and the arguments of Duchi (2025). Notably, while non-neglgible, we find that this bias is small relative to the effects of overfitting bias and our level and additive adjustment procedures still produce much lower mutliaccuracy error than the baselines approaches of quantile regression and conformalized quantile regression (right panel of Figure 6).

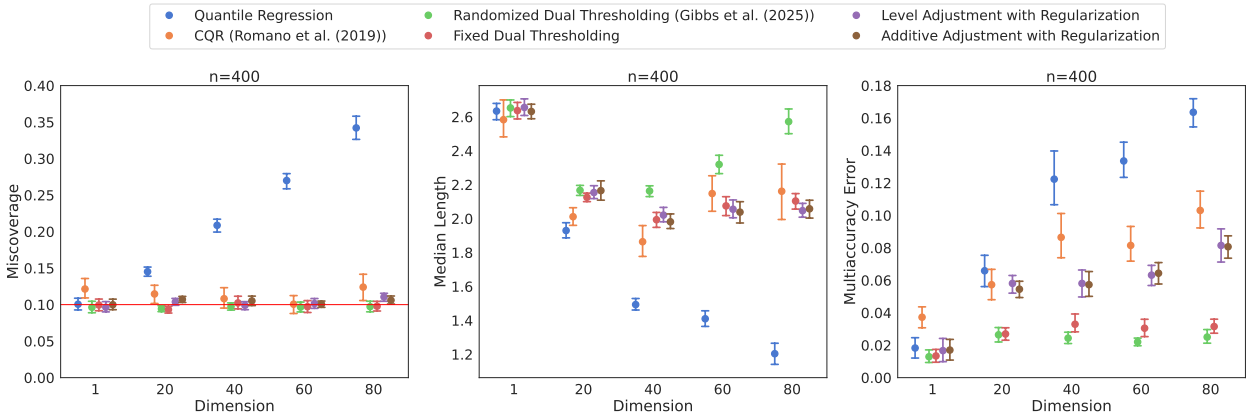

Figure 7: Empirical miscoverage (left panel), multiaccuracy error (center panel), and median length (right panel) of quantile regression (blue), the baseline methods of Romano et al. (2019) (orange) and Gibbs et al. (2025) (blue) and our fixed dual thresholding (red), level adjustment (purple), and additive adjustment (brown) methods for predicting the critical temperature of superconductors. Dots and error bars show means and 95% confidence intervals obtained over 20 trials where in each trial the methods are evaluated on a test set of size 2000. The red line in the left panel shows the target level of $\alpha = 0.1$.

To investigate the robustness of these conclusions we run similar experiments on two additional datasets where the goals are to predict the critical temperature of superconductors (Hamidieh 2018) and the number of times an article was shared online (Fernandes et al. 2015). After removing non-linearly independent features. these datasets have 81 and and 55 covariates,

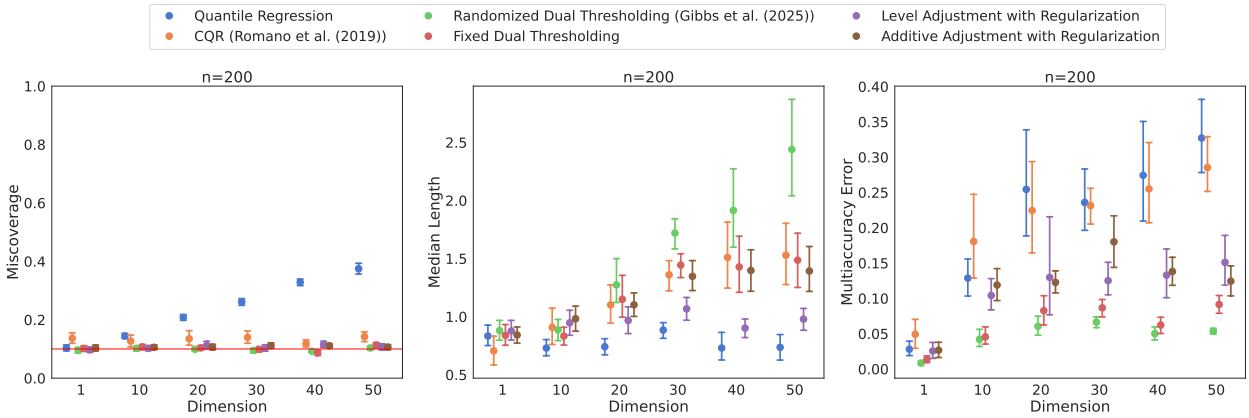

Figure 8: Empirical miscoverage (left panel), multiaccuracy error (center panel), and median length (right panel) of quantile regression (blue), the baseline methods of Romano et al. (2019) (orange) and Gibbs et al. (2025) (blue) and our fixed dual thresholding (red), level adjustment (purple), and additive adjustment (brown) methods for predicting the number of times an article was shared online. Dots and error bars show means and 95% confidence intervals obtained over 20 trials where in each trial the methods are evaluated on a test set of size 2000. The red line in the left panel shows the target level of $\alpha = 0.1$.

respectively. As above, we normalize the covariates to have mean zero and variance one. We additionally normalize the response in the same way. This latter normalization is done simply to simplify our hyperparameter search and does not otherwise impact the results.

Figures 7 and 8 show the results of these experiments. We find that all methods behave similarly as in the experiment on the communities and crimes dataset. Namely, all three of our methods offer exact marginal coverage and outperform the quantile regression and conformalized quantile regression baselines in terms of interval length and/or multiaccuracy error. The best performing methods are fixed dual thresholding and the level adjustment procedure with the former providing lower multiaccuracy error, while the latter gives smaller interval lengths.

References

Abbasi, E., Thrampoulidis, C. & Hassibi, B. (2016), General performance metrics for the lasso, *in* '2016 IEEE Information Theory Workshop (ITW)', pp. 181–185.

Angrist, J., Chernozhukov, V. & Fernández-Val, I. (2006), 'Quantile regression under misspecification, with an application to the u.s. wage structure', *Econometrica* **74**(2), 539–563.

Austern, M. & Zhou, W. (2020), 'Asymptotics of cross-validation', *arXiv preprint* . arxiv:2001.11111.

Bai, Y., Mei, S., Wang, H. & Xiong, C. (2021), Understanding the under-coverage bias in uncertainty estimation, *in* 'Advances in Neural Information Processing Systems'.

Bauschke, H. H. & Combettes, P. L. (2017), *Convex Analysis and Monotone Operator Theory in Hilbert Spaces*, 2 edn, Springer, Cham.

Bayle, P., Bayle, A., Janson, L. & Mackey, L. (2020), Cross-validation confidence intervals for test error, *in* 'Advances in Neural Information Processing Systems', Vol. 33, Curran Associates, Inc., pp. 16339–16350.

Bertsekas, D. P. (2009), *Convex Optimization Theory*, Athena Scientific.

Celentano, M., Montanari, A. & Wei, Y. (2023), 'The Lasso with general Gaussian designs with applications to hypothesis testing', *The Annals of Statistics* **51**(5), 2194 – 2220.

Donoho, D. L. & Montanari, A. (2013), 'High dimensional robust m-estimation: asymptotic variance via approximate message passing', *Probability Theory and Related Fields* **166**, 935 – 969.

Dua, D. & Graff, C. (2017), 'UCI machine learning repository'.
URL: *http://archive.ics.uci.edu/ml*

Duchi, J. C. (2025), 'A few observations on sample-conditional coverage in conformal prediction', *arXiv preprint* . arXiv:2503.00220.

Fernandes, K., Vinagre, P. & Cortez, P. (2015), A proactive intelligent decision support system for predicting the popularity of online news, *in* 'Progress in Artificial Intelligence', Springer International Publishing, Cham, pp. 535–546.

Gibbs, I., Cherian, J. J. & Candès, E. J. (2025), 'Conformal prediction with conditional guarantees', *Journal of the Royal Statistical Society Series B: Statistical Methodology* p. qkaf008.

Gordon, Y. (1985), 'Some inequalities for gaussian processes and applications', *Israel Journal of Mathematics* **50**, 265–289.

Gordon, Y. (1988), On milman's inequality and random subspaces which escape through a mesh in $\mathbb{R}^n$, *in* J. Lindenstrauss & V. D. Milman, eds, 'Geometric Aspects of Functional Analysis', Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 84–106.

Hamidieh, K. (2018), 'A data-driven statistical model for predicting the critical temperature of a superconductor', *Computational Materials Science* **154**, 346–354.

Han, Q. & Shen, Y. (2023), 'Universality of regularized regression estimators in high dimensions', *The Annals of Statistics* **51**(4), 1799 – 1823.
URL: *https://doi.org/10.1214/23-AOS2309*

Hastie, T., Montanari, A., Rosset, S. & Tibshirani, R. J. (2022), 'Surprises in high-dimensional ridgeless least squares interpolation', *The Annals of Statistics* **50**(2), 949 – 986.

Hébert-Johnson, U., Kim, M., Reingold, O. & Rothblum, G. (2018), Multicalibration: Calibration for the (computationally-identifiable) masses, *in* 'International Conference on Machine Learning', PMLR, pp. 1939–1948.

Javanmard, A. & Montanari, A. (2014), 'Confidence intervals and hypothesis testing for high-dimensional regression', *Journal of Machine Learning Research* **15**(82), 2869–2909.

Jung, C., Noarov, G., Ramalingam, R. & Roth, A. (2023), Batch multivalid conformal prediction, *in* 'International Conference on Learning Representations'.

Karoui, N. E., Bean, D., Bickel, P. J., Lim, C. & Yu, B. (2013), 'On robust regression with high-dimensional predictors', *Proceedings of the National Academy of Sciences* **110**(36), 14557–14562.

Kim, M. P., Ghorbani, A. & Zou, J. (2019), Multiaccuracy: Black-box post-processing for fairness in classification, *in* 'Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society', pp. 247–254.

Koenker, R. (2017), 'Quantile regression: 40 years on', *Annual Review of Economics* **9**(Volume 9, 2017), 155–176.

Koenker, R. & Bassett, G. (1978), 'Regression quantiles', *Econometrica* **46**(1), 33–50.

Koenker, R. & Hallock, K. F. (2001), 'Quantile regression', *The Journal of Economic Perspectives* **15**(4), 143–156.

Miescke, K.-J. & Liese, F. (2008), *Statistical Decision Theory: Estimation, Testing, and Selection*, Springer, New York.

Miolane, L. & Montanari, A. (2021), 'The distribution of the Lasso: Uniform control over sparse balls and adaptive parameter tuning', *The Annals of Statistics* **49**(4), 2313 – 2335.

Patil, P., Rinaldo, A. & Tibshirani, R. (2022), Estimating functionals of the out-of-sample error distribution in high-dimensional ridge regression, *in* 'Proceedings of The 25th International Conference on Artificial Intelligence and Statistics', Vol. 151 of *Proceedings of Machine Learning Research*, PMLR, pp. 6087–6120.

Patil, P., Wei, Y., Rinaldo, A. & Tibshirani, R. (2021), Uniform consistency of cross-validation estimators for high-dimensional ridge regression, *in* 'Proceedings of The 24th International Conference on Artificial Intelligence and Statistics', Vol. 130 of *Proceedings of Machine Learning Research*, PMLR, pp. 3178–3186.

Rad, K. R., Zhou, W. & Maleki, A. (2020), Error bounds in estimating the out-of-sample prediction error using leave-one-out cross validation in high-dimensions, *in* 'Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics', Vol. 108 of *Proceedings of Machine Learning Research*, PMLR, pp. 4067–4077.

Redmond, M. & Baveja, A. (2002), 'A data-driven software tool for enabling cooperative information sharing among police departments', *European Journal of Operational Research* **141**(3), 660–678.

Rockafellar, R. T. & Wets, R. J. B. (1997), *Variational Analysis*, Springer, Berlin.

Romano, Y., Patterson, E. & Candes, E. (2019), Conformalized quantile regression, *in* 'Advances in Neural Information Processing Systems', Vol. 32, Curran Associates, Inc.

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

Steinberger, L. & Leeb, H. (2016), 'Leave-one-out prediction intervals in linear regression models with many variables', *arXiv preprint* . arXiv:1602.05801.

Steinberger, L. & Leeb, H. (2023), 'Conditional predictive inference for stable algorithms', *The Annals of Statistics* **51**(1), 290 – 311.

Thrampoulidis, C., Abbasi, E. & Hassibi, B. (2018), 'Precise error analysis of regularized m -estimators in high dimensions', *IEEE Transactions on Information Theory* **64**(8), 5592–5628.

Thrampoulidis, C., Oymak, S. & Hassibi, B. (2015), Regularized linear regression: A precise analysis of the estimation error, *in* P. Grünwald, E. Hazan & S. Kale, eds, 'Proceedings of The 28th Conference on Learning Theory', Vol. 40 of *Proceedings of Machine Learning Research*, PMLR, Paris, France, pp. 1683–1709.

van de Geer, S., Bühlmann, P., Ritov, Y. & Dezeure, R. (2014), 'On asymptotically optimal confidence regions and tests for high-dimensional models', *The Annals of Statistics* **42**(3), 1166 – 1202.

Xu, J., Maleki, A., Rad, K. R. & Hsu, D. (2021), 'Consistent risk estimation in moderately high-dimensional linear regression', *IEEE Transactions on Information Theory* **67**(9), 5997–6030.

Yin, Y., Bai, Z. & Krishnaiah, P. (1988), 'On the limit of the largest eigenvalue of the large dimensional sample covariance matrix', *Probability Theory and Related Fields* **78**, 509–521.

Zhang, C.-H. & Zhang, S. S. (2013), 'Confidence intervals for low dimensional parameters in high dimensional linear models', *Journal of the Royal Statistical Society Series B: Statistical Methodology* **76**(1), 217–242.

Zou, H., Auddy, A., Rad, K. R. & Maleki, A. (2025), Theoretical analysis of leave-one-out cross validation for non-differentiable penalties under high-dimensional settings, *in* 'Proceedings of The 28th International Conference on Artificial Intelligence and Statistics', Vol. 258 of *Proceedings of Machine Learning Research*, PMLR, pp. 4033–4041.

A Proofs for Section 2.2

We will now give formal proofs for the results stated in Section 2.2. To begin, we prove a useful technical lemma that relates a dual value of zero to interpolation of the leave-one-out prediction.

Lemma A.1. *Suppose there exists a leave-one-out primal solution with $\tilde{Y}_i = \tilde{X}_i^\top \hat{w}^{-i}$. Then, there exists a dual solution to the full program with $\hat{\eta}_i = 0$. Symmetrically, if there exists a dual solution to the full program with $\hat{\eta}_i = 0$, then there exists a leave-one-out primal solution with $\tilde{Y}_i = \tilde{X}_i^\top \hat{w}^{-i}$.*

Proof. For notational simplicity we will focus on the case $i = n$. Suppose there exists a leave-one-out primal solution with $\tilde{Y}_n = \tilde{X}_n^\top \hat{w}^{-n}$. For $i \in \{1, \dots, n-1\}$, let $\hat{r}_i^{-n} = \tilde{Y}_i = \tilde{X}_i^\top \hat{w}^{-n}$ denote the additional primal variables. Let $\hat{\eta}^{-n} \in \mathbb{R}^{n-1}$ denote a corresponding dual solution to the leave-one-out program. The Lagrangian for the leave-one-out program is

$$L_{-n}(w^{-n}, r^{-n}, \eta^{-n}) = \sum_{j=1}^{n-1} \ell_\tau(r_j^{n-1}) + \sum_{j=1}^{n-1} \eta_j^{-n}(\tilde{Y}_j - \tilde{X}_j^\top w^{-n} - r_j^{-n}) + \mathcal{R}(w),$$

and the Lagriangian for the full program is

$$L(w, r, \eta) = \sum_{j=1}^{n} \ell_\tau(r_j) + \sum_{j=1}^{n} \eta_j(\tilde{Y}_j - \tilde{X}_j^\top w^{-n} - r_j) + \mathcal{R}(w). \tag{A.1}$$

By assumption, $(\hat{w}^{-n}, \hat{r}^{-n}, \hat{\eta}^{-n})$ is a saddle point of $L_{-n}(\cdot)$. Using this fact and taking first-order derivatives, it is straightforward to verify that $(\hat{w}^{-n}, (\hat{r}^{-n}, 0), (\hat{\eta}^{-n}, 0))$ is a saddle point of $L(\cdot)$. Thus, $(\hat{\eta}^{-n}, 0)$ is a solution to the full dual program, as desired.

For the reverse direction, suppose there exists a solution to the full dual program with $\hat{\eta}_n = 0$. Let $(\hat{w}, \hat{r}, \hat{\eta})$ denote the corresponding saddle point of $L(\cdot)$. By taking the first-order derivative of $L(\cdot)$ in r_n we see that we must have $\hat{r}_n = 0$ and so, recalling the constraint on $\hat{r}$, that

$\tilde{Y}_n - \tilde{X}_n^\top \hat{w} = \hat{r}_n = 0$. Then, using the notation $v_{1:(n-1)}$ to denote the first $n-1$ entries of a vector $v \in \mathbb{R}^n$ and taking first-order derivatives of L_{-n}, it is straightforward to verify that $(\hat{w}, \hat{r}_{1:(n-1)}, \hat{\eta}_{1:(n-1)})$ is a solution to the leave-one-out program with point n excluded. Since $\tilde{Y}_n = \tilde{X}_n^\top \hat{w}$, this proves the desired result. □

To prove Proposition 2.1 we will need one additional technical result demonstrating that the dual solution, $\hat{\eta}_i$, behaves monotonically in $\tilde{Y}_i$. This result was originally proven in Gibbs et al. (2025) where it was used to derive efficient algorithms for computing $\hat{q}_{\text{GCC, rand.}}(\cdot)$. Here, we will leverage this result to derive a relationship between the dual and leave-one-out solutions. To state this result formally, let us focus for simplicity on the case $i = n$ and define

$$\hat{\eta}^{\tilde{Y}_n \to y} = \underset{\eta \in [-(1-\tau), \tau]^n}{\operatorname{argmax}} \sum_{j=1}^{n-1} \eta_j \tilde{Y}_j + \eta_n y - \mathcal{R}^* \left(\sum_{j=1}^{n} \eta_j \tilde{X}_j \right), \tag{A.2}$$

to be the dual solution obtained when $\tilde{Y}_n$ is replaced with $y \in \mathbb{R}$. We have the following lemma.

Lemma A.2. *[Theorem 4 of Gibbs et al. (2025)] Let $\{\hat{\eta}^{\tilde{Y}_n \to y}\}_{y \in \mathbb{R}}$ denote any collection of solutions to (A.2). Then, $y \mapsto \hat{\eta}^{\tilde{Y}_n \to y}$ is non-decreasing.*

We are now ready to prove Proposition 2.1.

Proof of Proposition 2.1. For notational simplicity, we focus on the case $i = n$. Suppose there exists a leave-one-out primal solution with $\tilde{Y}_n < \tilde{X}_n^\top \hat{w}^{-n}$. By Lemma A.1, when $y = \tilde{X}_n^\top \hat{w}^{-n}$ there exists a solution to (A.2) with $\hat{\eta}_n^{\tilde{Y}_n \to \tilde{X}_n^\top \hat{w}^{-n}} = 0$. By the monotonicity of the dual solutions (Lemma A.2), this immediately implies that any dual solution to the full program must satisfy $\hat{\eta}_n \leq \hat{\eta}_n^{\tilde{Y}_n \to \tilde{X}_n^\top \hat{w}^{-n}} = 0$, as desired.

The case where $\tilde{Y}_n > \tilde{X}_n^\top \hat{w}^{-n}$ follows by an identical argument. □

We conclude this section with a proof of Theorem 2.1. To aid in this proof, we introduce a number of additional pieces of notation. We let $\tilde{X} \in \mathbb{R}^{n \times p}$ denote the matrix with rows $\tilde{X}_1, \ldots, \tilde{X}_n$ and $\tilde{Y} \in \mathbb{R}^n$ denote the vector with entries $\tilde{Y}_1, \ldots, \tilde{Y}_n$. For any vector $v \in \mathbb{R}^k$ and set $I \subseteq \{1, \ldots, k\}$ we let $(v_I) = (v_i)_{i \in I}$ denote the subvector consistency of the entries in I. Similarly, for any $I \subseteq \{1, \ldots, n\}$ and $J \in \{1, \ldots, p\}$ we let $\tilde{X}_{I,J} = (\tilde{X}_{ij})_{i \in I, j \in J}$ denote the submatrix containing the rows in I and columns in J. Finally, for any $k \in \mathbb{N}$ we let $[k] = \{1, \ldots, k\}$.

We now present a preliminary lemma bounding the number of points interpolated by any quantile regression solution.

Lemma A.3. *Assume that the distribution of $\tilde{Y}_i \mid \tilde{X}_i$ is continuous. Then, with probability one all primal solutions $\hat{w}$ satisfy,*

$$\operatorname{rank}\left(\tilde{X}_{\{i:\tilde{Y}_i = \tilde{X}_i^\top \hat{w}\},[p]}\right) = |\{i : \tilde{Y}_i = \tilde{X}_i^\top \hat{w}\}|.$$

Proof. Fix any primal solution $\hat{w}$. Let $I_=(\hat{w}) = \{i \in \{1, \ldots, n\} : \tilde{Y}_i = \tilde{X}_i^\top \hat{w}\}$ denote the set of interpolated points. We have that

$$\tilde{Y}_{I_=(\hat{w})} = \tilde{X}_{I_=(\hat{w}),[p]}.$$

For the sake of deriving a contradiction, suppose that $\operatorname{rank}\left(\tilde{X}_{\{i:\tilde{Y}_i = \tilde{X}_i^\top \hat{w}\},[p]}\right) < |I_=(\hat{w})|$. Let $I(\hat{w}) \subset |I_=(\hat{w})|$ be such that $\tilde{X}_{I(\hat{w}),[p]}$ is of maximal rank. Then, conditional on $\tilde{X}$, $\tilde{Y}_{I_=(\hat{w}) \setminus I(\hat{w})} = \tilde{X}_{I_=(\hat{w}) \setminus I(\hat{w}),[p]} \hat{w}$ is a deterministic function of $(\tilde{X}_{I(\hat{w}),[p]}, \tilde{Y}_{I(\hat{w})})$. However, for any fixed sets $I' \subseteq I'_= \subseteq [n]$, the distribution of $\tilde{Y}_{I'_= \setminus I'} \mid \tilde{X}$ is continuous. So, taking a union bound over all choices of $I_=(\hat{w})$ and $I(\hat{w})$ we find that this occurs with probability zero, as claimed.

□

We now prove Theorem 2.1.

Proof of Theorem 2.1. We split into two cases corresponding to the two sets of assumptions.

Case 1, Assumption 2 holds: In this case the primal program is strongly convex. Thus, the leave-one-out solutions $\hat{w}^{-i}$ is unique and by the continuity of the distribution of $\tilde{Y}_i \mid \tilde{X}_i$ we must have us that $\mathbb{P}(\tilde{Y}_i = \tilde{X}_i^\top \hat{w}^{-i}) = 0$. By Lemma A.1, this implies that $\mathbb{P}(\text{any dual solution satisfies } \hat{\eta}_i = 0) = 0$ as well. The desired result then follows from the convexity of the space of primal and dual solutions.

Case 2, Assumption 1 holds: This case is considerably more involved. First, note that by the convexity of the space of primal and dual solutions and the results of Proposition 2.1 and Lemma A.1 it is sufficient to show that $\mathbb{P}(\text{there exists a dual solution with } \hat{\eta}_n = 0) = 0$. Now, given a fixed dual solution $\hat{\eta}$ we group the sample indices into three sets depending on the value of the duals. In particular, we let

$$I_{-(1-\tau)}(\hat{\eta}) = \{i \in \{1, \dots, n\} : \hat{\eta}_i = -(1-\tau)\},\ I_{\tau}(\hat{\eta}) = \{i \in \{1, \dots, n\} : \hat{\eta}_i = \tau\},$$

$$\text{and } I_{\text{int.}}(\hat{\eta}) = \{i \in \{1, \dots, n\} : -(1-\tau) < \hat{\eta}_i < \tau\}.$$

Let $(\hat{w}, \hat{r})$ denote any corresponding primal solution. The Lagrangian for this primal-dual pair is

$$L(w, r, \eta) = \sum_{i=1}^{n} \ell_\tau(r_i) + \sum_{i=1}^{n} \eta_i(\tilde{Y}_i - \tilde{X}_i^\top w - r_i) + \sum_{j=1}^{p} \lambda_j w_j^2.$$

Taking a derivative in w gives the first order condition

$$\eta^\top \tilde{X} = (2\lambda_j w_j)_{j=1}^{p}.$$

Let $J_+ = \{j \in \{1, \dots, d\} : \lambda_j > 0\}$ denote the set coordinates which receive positive regularization. Let $\Lambda_{J_+} = \text{diag}((\lambda_j)_{j \in J_+})$. Minimizing over w and r in the Lagrangian gives

us that the dual program can be written as

$$\begin{aligned} &\underset{\eta \in [-(1-\tau),\tau]^n}{\text{minimize}} \ \eta^\top \tilde{Y}_i - \frac{1}{4} \tilde{X}_{[n],J_+} \Lambda_{J_+}^{-1} \tilde{X}_{[n],J_+} \\ &\text{subject to } \eta^\top \tilde{X}_{[n],J_+^c} = 0. \end{aligned}$$

Moreover, examining the first order condition of this maximization problem over the indices in $I_{\text{int.}}(\hat{\eta})$ gives

$$\tilde{Y}_{I_{\text{int.}}(\hat{\eta})} - \frac{1}{2} \tilde{X}_{I_{\text{int.}}(\hat{\eta}),J_+} \Lambda_{J_+}^{-1} \tilde{X}^\top_{I_{\text{int.}}(\hat{\eta})^c,J_+} \hat{\eta}_{I_{\text{int.}}(\hat{\eta})^c} - \frac{1}{2} \tilde{X}_{I_{\text{int.}}(\hat{\eta}),J_+} \Lambda_{J_+}^{-1} \tilde{X}^\top_{I_{\text{int.}}(\hat{\eta}),J_+} \hat{\eta}_{I_{\text{int.}}(\hat{\eta})} = 0,$$

and combining this with the constraint on the dual variables we arrive at the equation

$$\begin{bmatrix} \frac{1}{2} \tilde{X}_{I_{\text{int.}}(\hat{\eta}),J_+} \Lambda_{J_+}^{-1} \tilde{X}^\top_{I_{\text{int.}}(\hat{\eta}),J_+} \\ \tilde{X}^\top_{I_{\text{int.}}(\hat{\eta}),J_+^c} \end{bmatrix} \hat{\eta}_{I_{\text{int.}}(\hat{\eta})} = \begin{pmatrix} \tilde{Y}_{I_{\text{int.}}(\hat{\eta})} - \frac{1}{2} \tilde{X}_{I_{\text{int.}}(\hat{\eta}),J_+} \Lambda_{J_+}^{-1} \tilde{X}^\top_{I_{\text{int.}}(\hat{\eta})^c,J_+} \hat{\eta}_{I_{\text{int.}}(\hat{\eta})^c} \\ \tilde{X}^\top_{I_{\text{int.}}(\hat{\eta})^c,J_+^c} \hat{\eta}_{I_{\text{int.}}(\hat{\eta})^c} \end{pmatrix} \quad \text{(A.3)}$$

We claim that the matrix on the right-hand-side of this expression is of rank $|I_{\text{int.}}(\hat{\eta})|$. To see this, first note that for any $v \in \mathbb{R}^{|I_{\text{int.}}(\hat{\eta})|}$,

$$\tilde{X}_{I_{\text{int.}}(\hat{\eta}),J_+} \Lambda_{J_+}^{-1} \tilde{X}^\top_{I_{\text{int.}}(\hat{\eta}),J_+} v = 0 \implies v^\top \tilde{X}_{I_{\text{int.}}(\hat{\eta}),J_+} \Lambda_{J_+}^{-1} \tilde{X}^\top_{I_{\text{int.}}(\hat{\eta}),J_+} v = 0 \implies \tilde{X}^\top_{I_{\text{int.}}(\hat{\eta}),J_+} v = 0.$$

Thus,

$$\begin{bmatrix} \frac{1}{2} \tilde{X}_{I_{\text{int.}}(\hat{\eta}),J_+} \Lambda_{J_+}^{-1} \tilde{X}^\top_{I_{\text{int.}}(\hat{\eta}),J_+} \\ \tilde{X}^\top_{I_{\text{int.}}(\hat{\eta}),J_+^c} \end{bmatrix} v = 0 \implies \tilde{X}^\top_{I_{\text{int.}}(\hat{\eta}),[p]} v = 0.$$

By taking a derivative in r in the Lagriangian, we must have that $I_{\text{int.}}(\hat{\eta}) \subseteq I_=(\hat{w}) := \{i \in \{1,\dots,n\} : \tilde{Y}_i = \tilde{X}_i^\top \hat{w}\}$. So, by Lemma A.3, we have $\text{rank}(\tilde{X}_{I_{\text{int.}}(\hat{\eta}),[p]}) = |I_{\text{int.}}(\hat{\eta})|$ and thus the above is only possible if $v = 0$. This proves the claim.

Now, using this claim, we may find submatrices $I_{\text{sub.}}(\hat{\eta}) \subseteq I_{\text{int.}}(\hat{\eta})$ and $J_{\text{sub}}(\hat{\eta}) \subseteq J_+^c$ such that $|I_{\text{sub.}}(\hat{\eta})| + |J_{\text{sub}}(\hat{\eta})| = |I_{\text{int.}}(\hat{\eta})|$ and the matrix,

$$\begin{bmatrix} \frac{1}{2} \tilde{X}_{I_{\text{sub.}}(\hat{\eta}),J_+} \Lambda_{J_+}^{-1} \tilde{X}^\top_{I_{\text{int.}}(\hat{\eta}),J_+} \\ \tilde{X}^\top_{I_{\text{int.}}(\hat{\eta}),J_{\text{sub.}}(\hat{\eta})} \end{bmatrix},$$

is of full rank. Applying this to (A.3) gives us

$$\hat{\eta}_{I_{\text{int.}}(\hat{\eta})} = \begin{bmatrix} \frac{1}{2}\tilde{X}_{I_{\text{sub.}}(\hat{\eta}),J_+}\Lambda_{J_+}^{-1}\tilde{X}_{I_{\text{int.}}(\hat{\eta}),J_+}^\top \\ \tilde{X}_{I_{\text{int.}}(\hat{\eta}),J_{\text{sub.}}(\hat{\eta})}^\top \end{bmatrix}^{-1} \begin{pmatrix} \tilde{Y}_{I_{\text{sub.}}(\hat{\eta})} - \frac{1}{2}\tilde{X}_{I_{\text{sub.}}(\hat{\eta}),J_+}\Lambda_{J_+}^{-1}\tilde{X}_{I_{\text{int.}}(\hat{\eta})^c,J_+}^\top \hat{\eta}_{I_{\text{int.}}(\hat{\eta})^c} \\ \tilde{X}_{I_{\text{int.}}(\hat{\eta})^c,J_{\text{sub.}}(\hat{\eta})}^\top \hat{\eta}_{I_{\text{int.}}(\hat{\eta})^c} \end{pmatrix}.$$

Now, let us consider the behavior of this random variable when the sets appearing above and the value of $\hat{\eta}_{I_{\text{int.}}(\hat{\eta})^c}$ are fixed, i.e., fix any sets $I_{\text{sub.}} \subseteq I_{\text{int.}} \subseteq [n]$, $J_{\text{sub.}} \subseteq J_+^c$ and vector $\eta_{I_{\text{int.}}^c} \in \{-(1-\tau), \tau\}^{|I_{\text{int.}}^c|}$ that do not depend on the data. We are interested in the behaviour of the random variable

$$\eta_{I_{\text{int.}}} = \begin{bmatrix} \frac{1}{2}\tilde{X}_{I_{\text{sub.}},J_+}\Lambda_{J_+}^{-1}\tilde{X}_{I_{\text{int.}},J_+}^\top \\ \tilde{X}_{I_{\text{int.}},J_{\text{sub.}}}^\top \end{bmatrix}^{-1} \begin{pmatrix} \tilde{Y}_{I_{\text{sub.}}} - \frac{1}{2}\tilde{X}_{I_{\text{sub.}},J_+}\Lambda_{J_+}^{-1}\tilde{X}_{I_{\text{int.}}^c,J_+}^\top \eta_{I_{\text{int.}}^c} \\ \tilde{X}_{I_{\text{int.}}^c,J_{\text{sub.}}}^\top \eta_{I_{\text{int.}}^c} \end{pmatrix}. \tag{A.4}$$

Now, on the event that the matrix inverse above exists we must have that $\text{rank}(\frac{1}{2}\tilde{X}_{I_{\text{sub.}},J_+}\Lambda_{J_+}^{-1}) = |I_{\text{sub.}}|$. Moreover, by our assumptions on $\tilde{X}$, conditional conditional on $(\tilde{Y}_{I_{\text{sub.}}}, \tilde{X}_{I_{\text{int.}},[p]}, Z_{[n],[p]})$, $\tilde{X}_{I_{\text{int.}}^c,J_+}^\top \eta_{I_{\text{int.}}^c}$ and $\tilde{X}_{I_{\text{int.}}^c,J_{\text{sub.}}}^\top \eta_{I_{\text{int.}}^c}$ are independent and continuously distributed random variables with independent entries. So, on the event that the matrix inverse appearing in (A.4) exists, we find that conditional on $(\tilde{Y}_{I_{\text{sub.}}}, \tilde{X}_{I_{\text{int.}},[p]}, Z_{[n],[p]})$, $\eta_{I_{\text{int.}}}$ is the product of an invertible matrix and a continuously distributed vector. Since this invertible matrix must have all non-zero rows, we conclude that the distribution of $\eta_{I_{\text{int.}}}$ is continuous. Critically, this implies that with probability one all coordinates of $\eta_{I_{\text{int.}}}$ are non-zero. The desired result then follows by taking a union bound over all possible choices of the sets $I_{\text{sub.}}$, $I_{\text{int.}}$, and $J_{\text{sub.}}$ and vector $\eta_{I_{\text{int.}}^c}$.

□

B Proofs for Section 3

The bulk of this section is devoted to a proof of Theorem 3.1. Proofs of Corollaries 3.1 and 3.2 are then given at the end of the section. In what follows, we use $X \in \mathbb{R}^{n\times p}$ to denote the matrix with rows $X_1, \dots, X_n$ and $Y \in \mathbb{R}^n$ and $\epsilon \in \mathbb{R}^n$ to denote the vectors with entries $(Y_1, \dots, Y_n)$ and $(\epsilon_1, \dots, \epsilon_n)$, respectively. For notational convenience, we let $\mathcal{R}_d(\beta) = \lambda_d \sum_{i=1}^d \nu(\beta_i)$ denote the regularizer and $\tilde{\mathcal{R}}_d = n^{-1/2}\mathcal{R}_d$ denote a rescaling of this quantity. Additionally, for any convex function $f : \mathbb{R} \to \mathbb{R}$, $x \in \mathbb{R}$, and $\rho > 0$ we recall the definition of the Moreau envelope.

$$e_f(x;\rho) = \min_{v\in\mathbb{R}} \frac{1}{2\rho}\|x - v\|^2 + f(v),$$

For $f : \mathbb{R}^d \to \mathbb{R}$ we additionally recall the definition of the convex conjugate,

$$f^*(x) = -\inf_{v\in\mathbb{R}} f(v) - v^\top x$$

The formal assumptions of Theorem 3.1 are as follows.

Assumption 4. *The regularization function and P_β are such that*

1. *There exists a convex function $\tilde{\nu}$ with the property that for any $c \in \mathbb{R}$, $\rho > 0$, and $h \sim N(0, I_d)$,*

$$\frac{1}{d}\sum_{i=1}^d e_{\tilde{\mathcal{R}}_d\nu}\left(ch_i + \sqrt{n}\tilde{\beta}_i; \rho\right) \xrightarrow{\mathbb{P}} \mathbb{E}[e_{\tilde{\nu}}(ch_1 + \gamma\sqrt{d}\tilde{\beta}_1; \rho)],$$

   and $d/n \to \gamma \in (0, \infty)$.

2. *For any $\rho > 0$, $\partial_x e_{\tilde{\nu}^*}$ and $\partial_x^2 e_{\tilde{\nu}^*}$ exist almost everywhere and satisfy the equations*

$$\frac{d}{dc}\mathbb{E}[e_{\tilde{\nu}^*}(ch_1 + \gamma\sqrt{d}\tilde{\beta}_1; \rho)] = \mathbb{E}[h_1\partial_x e_{\tilde{\nu}^*}(ch_1 + \gamma\sqrt{d}\tilde{\beta}_1; \rho)] = c\mathbb{E}[\partial_x^2 e_{\tilde{\nu}^*}(ch_1 + \gamma\sqrt{d}\tilde{\beta}_1; \rho)].$$

3. *For any compact sets* $A \subseteq \mathbb{R}_{\geq 0}$ *and* $B \subseteq \mathbb{R}_{>0}$ *we have that*

$$\inf_{c \in A, \rho \in B} c\mathbb{E}[\partial_x^2 e_{\tilde{\nu}^*}(ch_1 + \gamma\sqrt{d}\tilde{\beta}_1; \rho)] > 0.$$

4. *For all* $\beta \in \mathbb{R}^d$, $\mathcal{R}(\beta) \geq 0$ *and* $\mathcal{R}(0) = 0$.

5. *For any* $C > 0$, *the subderivatives of* $\mathcal{R}_d$ *are bounded as*

$$\sup_{d \in \mathbb{N}} \sup_{\|\beta\|_2 \leq C} \frac{1}{d} \|\partial \mathcal{R}_d(\beta)\|_2 < \infty.$$

The first part of Assumption 4 will follow for most regularizers by the law of large numbers. The second part of the assumption will follow by using the dominated convergence theorem to swap the derivative and expectation and using Stein's lemma to compare the last two terms. This will be possible if, for example, $e_{\tilde{\nu}^*}$ is twice continuously differentiable and appropriate moment conditions are met. Finally, the last part will simply require the second derivative of $e_{\tilde{\nu}^*}$ not be identically zero. The following lemma verifies that all these conditions are satisfied by L_1 and L_2 regularization.

Lemma B.1. *Assume* P_β *as four bounded moments. Then, for any* $\lambda \geq 0$ *the conditions of Assumption 4 are met for* $\mathcal{R}_d(\beta) = \lambda\|\beta\|_2^2$ *and* $\mathcal{R}_d(\beta) = \lambda\|\beta\|_2^2$

Proof. For $\mathcal{R}_d(\beta) = \lambda\|\beta\|_2^2$ define $\tilde{\nu}(b) = \lambda b^2$. Then, by a direct calculation we have that

$$e_\nu(x; \rho) = \frac{\lambda x^2}{1 + 2\lambda\rho}, \quad \tilde{\nu}^*(b) = b^2/(4\lambda), \quad e_{\tilde{\nu}^*}(x; \rho) = \frac{x^2}{4\lambda + 2\rho}.$$

Part one of Assumption 4 follows immediately by Chernoff's bound. Part 2 follows by the dominated convergence theorem and Stein's lemma and part 3 is immediate since $\partial_x^2 e_{\tilde{\nu}^*}(x; \rho) = (2\lambda + \rho)^{-1} > 0$.

On the other hand suppose $\mathcal{R}_d(\beta) = d^{-1/2}\lambda\|\beta\|_1$. Define $\tilde{\nu}(b) = \lambda|b|$. Then,

$$e_\nu(x;\rho) = \begin{cases} \lambda x - \frac{\lambda^2\rho}{2}, & x > \lambda\rho, \\ \frac{x^2}{2\rho}, & |x| \le \lambda\rho, \\ -x\lambda - \frac{\lambda^2\rho}{2}, & x < -\lambda\rho, \end{cases} \qquad \tilde{\nu}^*(b) = \begin{cases} 0, & |x| \le \lambda, \\ \infty, & |x| > \lambda, \end{cases} \qquad e_{\tilde{\nu}^*}(x;\rho) = \begin{cases} 0, & |x| \le \lambda, \\ \frac{(|x|-\lambda)^2}{2\rho}, & |x| > \lambda. \end{cases}.$$

Moreover, one can verify that $e_{\tilde{\nu}^*}(x;\rho)$ is twice piecewise continuously differentiable with

$$\partial_x e_{\tilde{\nu}^*}(x;\rho) = \begin{cases} 0, & |x| \le \lambda, \\ \frac{x - \mathrm{sgn}(x)\lambda}{\rho}, & |x| > \lambda, \end{cases} \qquad \partial_x e_{\tilde{\nu}^*}(x;\rho) = \begin{cases} 0, & |x| < \lambda, \\ \frac{1}{\rho}, & |x| > \lambda. \end{cases}$$

The desired results once again follow by the dominated convergence theorem and Stein's lemma.

□

Our main point of study is the joint min-max formulation of the quantile regression,

$$\max_{\eta} \min_{\beta_0,\beta,r} \frac{1}{n}\sum_{i=1}^n \ell(r_i) + \frac{1}{n}\eta^\top(Y - \beta_0\mathbf{1}_n - X\beta) + \frac{1}{\sqrt{n}}\tilde{\mathcal{R}}_d(\beta).$$

Letting $u = \tilde{\beta} - \beta$ and re-writing $\tilde{\mathcal{R}}_d(\beta)$ in terms of the convex conjugate, this can be equivalently formulated as

$$\max_{\eta,s} \min_{\beta_0,u,r} \frac{1}{n}\sum_{i=1}^n \ell(r_i) + \frac{1}{n}\eta^\top(\epsilon - \beta_0\mathbf{1}_n - Xu - r) + \frac{1}{\sqrt{n}}s^\top(\tilde{\beta} + u) - \frac{1}{\sqrt{n}}\tilde{\mathcal{R}}_d^*(s). \tag{B.1}$$

We will prove Theorem 3.1 in four main steps. First, in Section B.1 we give a number of preliminary simplifications of the program. Section B.2 then begins our main study of B.1. We show that the solutions B.1 are characterized by an auxiliary optimization program in which the matrix X is replaced by vector-valued Gaussian random variables and that,

more over, the solutions to this auxillary optimization are themselves characterized by the deterministic asymptotic program

$$\min_{(|\beta_0|\le C_{\beta_0}, 0\le M_u\le C_u, 0\le\rho_1\le C_1)} \max_{(0<\rho_2<C_2, c_\eta\le M_\eta\le C_\eta)} \left(\mathbb{E}\left[e_{\ell_\alpha}\left(M_u g + \epsilon - \beta_0; \frac{\rho_1}{M_\eta}\right)\right] - \frac{M_\eta^2 M_u \gamma}{2\rho_2} + \frac{M_\eta \rho_1}{2} - \frac{M_u \rho_2}{2} - \mathbb{E}\left[e_\nu \left(\frac{M_u M_\eta}{\rho_2} h_1 + \gamma\sqrt{d}\tilde{\beta}_1; \frac{M_2}{\rho_2}\right)\right]\right), \tag{B.2}$$

where $g_1, h_1 \sim \mathcal{N}(0,1)$ independent of $\tilde{\beta}_1$ and ϵ_1, $(\beta_0, M_u, \rho_1, \rho_2, M_\eta)$, are the optimization variables, $(C_{\beta_0}, C_u, C_1, C_2, c_\eta, C_\eta)$ are constants that we will define shorty, and e_f denotes the Moreau envelope of f. The solutions to this asymptotic program are characterized in Section **??**. Using all the aforementioned result. Section **??** then gives a proof of Theorem 3.1. Finally, Section gives some final results whose proofs were deferred earlier and Section **??** gives proofs of Corollaries 3.1 and 3.2.

The overall analysis framework we used is based on the work of Thrampoulidis et al. (2018) who studied the asymptotics of solutions to general regularized regressions. In what follows we will focus on the aspects of the analysis that are new to this work and omit the proofs of some small details that are minor variations of results appearing in Thrampoulidis et al. (2018).

B.1 Preliminaries

We give two simplifying preliminaries to the results of Theorem 3.1. Our first result bounds the range of the solutions for η.

Lemma B.2. *Under the assumptions of Theorem 3.1, there $C_\eta > c_\eta > 0$ such that*

$$\mathbb{P}(\textit{For all dual solutions to (B.1)}, \sqrt{n}c_\eta \le \|\hat{\eta}\|_2 \le \sqrt{n}C_\eta) \to 1.$$

Proof. In what follows we use $\hat{\eta}$ to denote a generic dual solution. First, note that by

considering the first order conditions in r of (B.1) we have that $\hat{\eta} \in [-(1-\tau), \tau]^n$ and $Y_i \neq \hat{\beta}_0 + X_i^\top \hat{\beta} \implies \hat{\eta} \in \{-(1-\tau), \tau\}$. Taking $C_\eta = \max\{(1-\tau), \tau\}$ gives the upper bound. To get the lower bound, note that by Lemma A.3 any primal solution can interpolate at most $d+1$ of the data points. Thus, we must have $\|\hat{\eta}\|_2 \geq \sqrt{n-d-1}\min\{(1-\tau), \tau\}$ and so setting $c_\eta = (1/2)\sqrt{1-\gamma}\min\{(1-\tau), \tau\}$ gives the desired result.

□

Our next lemma gives a similar set of bound on the primal variables.

Lemma B.3. *Under the assumptions of Theorem 3.1, there exists constants $C_u, C_{\beta_0} > 0$ such that*

$$\mathbb{P}\left(\textit{For all primal solutions to (B.1), } \|\hat{\beta} - \tilde{\beta}\|_2 \leq C_u \textit{ and } |\hat{\beta}_0| \leq C_{\beta_0}\right) \to 1.$$

Proof. Let $(\hat{\beta}_0, \hat{\beta})$ denote any primal solution. By the law of large numbers and the optimality of $(\hat{\beta}_0, \hat{\beta})$ we have that

$$\begin{aligned}
\mathbb{E}[\ell_\tau(Y_1)] &\geq \frac{1}{n}\sum_{i=1}^n \ell_\tau(Y_i) - o_{\mathbb{P}}(1) \geq \frac{1}{n}\sum_{i=1}^n \ell_\tau(Y_i - X_i^\top\hat{\beta} - \hat{\beta}_0) + \mathcal{R}_d(\hat{\beta}) - o_{\mathbb{P}}(1) \\
&\geq \min\{1-\tau, \tau\}\frac{1}{n}\sum_{i=1}^n |X_i^\top(\hat{\beta} - \tilde{\beta}) + \hat{\beta}_0| - \min\{1-\tau, \tau\}\frac{1}{n}\sum_{i=1}^n |\epsilon_i| - o_{\mathbb{P}}(1) \\
&\geq \min\{1-\tau, \tau\}\max\{\|\hat{\beta} - \tilde{\beta}\|_2, |\hat{\beta}_0|\} \inf_{(\|u\|_2 \leq 1, |\beta_0| \leq 1, \max\{\|u\|_2, |\beta_0|\} = 1)} \frac{1}{n}\sum_{i=1}^n |X_i^\top u + \beta_0| \\
&\quad - \min\{1-\tau, \tau\}\mathbb{E}[\epsilon_1] - o_{\mathbb{P}}(1).
\end{aligned}$$

Lemma C.1 shows that

$$\liminf_{n,d\to\infty} \inf_{(\|u\|_2 \leq 1, |\beta_0| \leq 1, \max\{\|u\|_2, |\beta_0|\} = 1)} \frac{1}{n}\sum_{i=1}^n |X_i^\top u + \beta_0| \xrightarrow{\mathbb{P}} \sqrt{2}\pi - \gamma.$$

Applying this to the above we conclude that

$$\max\{\|\hat{\beta} - \tilde{\beta}\|_2, |\hat{\beta}_0|\} \leq \frac{\mathbb{E}[\ell_\tau(Y_1)] + \min\{1-\tau, \tau\}\mathbb{E}[\epsilon_1]}{\min\{1-\tau, \tau\}(\sqrt{2/\pi} - \gamma)} + o_{\mathbb{P}}(1),$$

where it should be understood that the $o_{\mathbb{P}}(1)$ term on the right hand side is uniform over all primal solutions. Taking $C_u = C_{\beta_0} = 2\frac{\mathbb{E}[\ell_\tau(Y_1)]+\min\{1-\tau,\tau\}\mathbb{E}[\epsilon_1]}{\min\{1-\tau,\tau\}(\sqrt{2/\pi}-\gamma)}$ gives the desired result. □

Our next final preliminary lemma bounds the size of solutions for r and s. In what follows we use the notation $\hat{r}$ and $\hat{s}$ to denote any solutions for r and s in (B.1).

Lemma B.4. *Under the conditions of Theorem 3.1, there exists constants $C_r > 0$ and $C_2 > 0$ such that*

$$\mathbb{P}(\textit{For all solutions to (B.1), } \|\hat{s}\|_2 \leq C_s\sqrt{n} \textit{ and } \|\hat{r}\|_2 \leq C_r\sqrt{n}) \to 1.$$

Proof. Fix any solution to $(\hat{\eta}, \hat{s}, \hat{\beta}_0, \hat{u}, \hat{r})$ to (B.1). By the first order conditions of (B.1) in η we must have

$$\|\hat{r}\|_2 = \|\epsilon - \hat{\beta}_0\mathbf{1}_n - X\hat{u}\|_2 \leq \|\epsilon\|_2 + \sqrt{n}|\hat{\beta}_0| + \lambda_{\max}(X)\|\hat{u}\|_2.$$

By standard results (e.g. Theorem 3.1 of Yin et al. (1988)) we have that $\lambda_{\max}(X)/\sqrt{n}$ is converging in probability to a constant. Moreover, by the law of large $\|\epsilon\|_2/\sqrt{n} \xrightarrow{\mathbb{P}} \sqrt{\mathbb{E}[\epsilon_i^2]}$. Combining these facts with the bounds on $|\hat{\beta}_0|$ and $\|\hat{u}\|_2$ given by Lemma B.3 gives the desired bound on $\|r\|_2$.

To bound $\|s\|_2$, note that by standard facts regarding the convex conjugate (e.g. Proposition 5.4.3 of Bertsekas (2009)), $\hat{s} \in \partial\tilde{\mathcal{R}}_d(\tilde{\beta} + \hat{u})$. Now, Lemma **??** and the law of large numbers there exists $C > 0$ such that with probability converging to one, $\|\tilde{\beta} + \hat{u}\|_2 \leq C$. So,

$$\frac{1}{\sqrt{n}}\|\hat{s}\|_2 \leq \sup_{\|v\|_2 \leq C} \left\| \frac{1}{\sqrt{n}}\partial\tilde{\mathcal{R}}_d(v) \right\|_2 = \sup_{\|v\|_2 \leq C} \left\| \frac{1}{n}\partial\mathcal{R}_d(v) \right\|_2.$$

This last quantity is bounded by our assumptions on $\mathcal{R}_d$. □

B.2 Reduction to the auxiliary optimization problem

We will now reduce (B.1) to a simpler asymptotic program that is easier to study. Our main tool will be the Gaussian comparison inequalities of Gordon (1985, 1988) and their application to regression problems developed in Thrampoulidis et al. (2018). In particular, we will apply the following proposition. This result is a minor extension of Theorem 3 of Thrampoulidis et al. (2018) (see also Thrampoulidis et al. (2015)) and thus its proof is ommitted.

Proposition B.1 (Extension of Theorem 3 of Thrampoulidis et al. (2018))**.** *Fix and $d, n \in \mathbb{R}$. Let $X \in \mathbb{R}^{n\times d}$ be distributed as $(X_{ij})_{i\in[n],j\in[d]} \overset{i.i.d.}{\sim} \mathcal{N}(0,1)$ and define $g \sim \mathcal{N}(0, I_n)$ and $h \sim \mathcal{N}(0, I_d)$ to be independent Gaussian vectors. Let $Q(r, \beta_0, s, u, \eta) : \mathbb{R}^n \times \mathbb{R} \times \mathbb{R}^d \times \mathbb{R}^d \times \mathbb{R}^n \to \mathbb{R}$ be continuous and jointly convex in (r, β_0, u) and concave in (s, η). Fix any compact sets $A \subseteq \mathbb{R}^n \times \mathbb{R} \times \mathbb{R}^d$ and $B \subseteq \mathbb{R}^d \times \mathbb{R}^n$. Define the values*

$$\Phi = \min_{(r,\beta_0,u)\in A} \max_{(s,\eta)\in B} \eta^\top X u + Q(r, \beta_0, su, \eta),$$

$$\phi = \min_{(r,\beta_0,u)\in A} \max_{(s,\eta)\in B} \|u\|_2 \eta^\top g + \|\eta\|_2 u^\top h + Q(r, \beta_0, su, \eta).$$

Then, for all $c \in \mathbb{R}$,

$$\mathbb{P}(\Phi < c) \le 2\mathbb{P}(\phi \le c).$$

If in addition A and B and convex, then we additionally have that for all $c \in \mathbb{R}$

$$\mathbb{P}(\Phi > c) \le 2\mathbb{P}(\phi \ge c).$$

To apply this result in our context, first note that by Lemma B.3 we may find constants $C_u, C_{\beta_0} > 0$ such that with probability approaching one, all primal solutions $(\hat{\beta}_0, \hat{\beta})$ satisfy $\|\hat{\beta} - \beta\|_2 \le C_u$ and $|\hat{\beta}_0| \le C_{\beta_0}$. Let

$$\Phi(S) = \max_{(\eta\in S, \|s\|_2 \le C_s\sqrt{n})} \min_{(|\beta_0|\le C_{\beta_0}, \|u\|_2 \le C_u, \|r\|_2 \le C_r\sqrt{n})} \frac{1}{n}\sum_{i=1}^n \ell(r_i) + \frac{1}{n}\eta^\top(\epsilon - \beta_0 \mathbf{1}_n - Xu - r)$$

$$+\frac{1}{\sqrt{n}}s^\top(\tilde{\beta}+u)-\frac{1}{\sqrt{n}}\tilde{\mathcal{R}}_d^*(s).$$

By Lemma B.3 and Lemma B.4, asymptotically $\Phi(\mathbb{R})$ agrees with (B.1). Our goal will be to compare the value of $\Phi(\mathbb{R})$ to $\phi(S)$ when S is more restricted set. The key insight of Thrampoulidis et al. (2018) is that for this purpose it is sufficient to study the value of the auxiliary optimization,

$$\phi(S) := \min_{(\|r\|_2\leq C_r\sqrt{n},|\beta_0|}$$
$$+\frac{1}{n}\eta^\top\epsilon-\frac{1}{n}\beta_0\sum_{i=1}^n\eta_i-\frac{1}{n}\eta^\top r+\frac{1}{n}\sum_{i=1}^n\ell_\tau(r_i)+\frac{1}{\sqrt{n}}s^\top(\tilde{\beta}+u)-\frac{1}{\sqrt{n}}\tilde{\mathcal{R}}_d^*(s)\Big),$$

where $h\sim\mathcal{N}(0,I_d)$ and $g\sim\mathcal{N}(0,I_n)$ are Gaussian vectors sampled such that (g,h,ϵ) is jointly independent. The following proposition formalizes this.

Proposition B.2 (Corollary of Lemma 7 in Thrampoulidis et al. (2018))**.** *Suppose that the assumptions of Theorem 3.1 hold with the additional restriction that $X_i\sim\mathcal{N}(0,I_d)$. Let S be any set such that*

1. *S is compact,*
2. *There exists constants $\rho,\delta,\xi>0$ such that $\min\{\mathbb{P}(\phi(\mathbb{R})\geq\rho+\delta),\mathbb{P}(\phi(S)\leq\rho-\delta)\}\geq 1-\xi$.*

Then,

$$\mathbb{P}(\textit{For all dual solutions, }\hat{\eta}\in S)\geq 1-4\xi.$$

Proof. This result follows immediately by applying Proposition B.1 and repeating the steps of Lemma 7 in Thrampoulidis et al. (2018). □

Our goal now is to lower bound $\phi(\mathbb{R})$ and upper bound $\phi(S)$ for some restricted set S. We will focus initially on $\psi(\mathbb{R})$. Let c_η and C_η be the constants appearing in Lemma B.2. We

have that

$$\begin{aligned}
\phi(\mathbb{R}) &\geq \phi(\{\eta : c_\eta\sqrt{n} \leq \|\eta\|_2 \leq C_\eta\sqrt{n}\}) \\
&= \min_{(\|r\|_2 \leq C_r\sqrt{n}, |\beta_0| \leq C_{\beta_0}, 0 \leq M_u \leq C_u)} \max_{(\|s\|_2 \leq C_s\sqrt{n}, c_\eta\sqrt{n} \leq \|\eta\|_2 \leq C_\eta\sqrt{n})} \Bigg(- M_u \left\| \frac{1}{n}\|\eta\|_2 h + \frac{1}{n} s \right\|_2 \\
&\qquad \frac{1}{n} M_u \eta^\top g + \frac{1}{n}\eta^\top \epsilon - \frac{1}{n}\beta_0 \sum_{i=1}^n \eta_i - \frac{1}{n}\eta^\top r + \frac{1}{n}\sum_{i=1}^n \ell_\tau(r_i) + \frac{1}{\sqrt{n}} s^\top \tilde{\beta} - \frac{1}{\sqrt{n}}\tilde{\mathcal{R}}_d^*(s) \Bigg) \\
&= \min_{(\|r\|_2 \leq C_r\sqrt{n}, |\beta_0| \leq C_{\beta_0}, 0 \leq M_u \leq C_u)} \max_{(\|s\|_2 \leq C_s\sqrt{n}, c_\eta \leq M_\eta \leq C_\eta)} \Bigg(- M_u \left\| \frac{1}{\sqrt{n}} M_\eta h + \frac{1}{\sqrt{n}} s \right\|_2 \\
&+ M_\eta \left\| \frac{1}{\sqrt{n}} M_u g + \frac{1}{\sqrt{n}}\epsilon - \frac{1}{\sqrt{n}}\beta_0 \mathbf{1}_n - \frac{1}{\sqrt{n}} r \right\|_2 + \frac{1}{n}\sum_{i=1}^n \ell_\tau(r_i) + \frac{1}{\sqrt{n}} s^\top \tilde{\beta} - \frac{1}{\sqrt{n}}\tilde{\mathcal{R}}_d^*(s) \Bigg).
\end{aligned}$$

Notably, this last optimization problem is convex-concave. Now, recall that for any vector x and $C \geq \|x\|_2$, $\|x\|_2 = \min_{0<\tau\leq C} \frac{\|x\|^2}{2\tau} + \frac{\tau}{2}$. By weak law of large numbers and our bounds on r and s, there exists constants $C_1, C_2 > 0$ such that with probability converging to one,

$$\begin{aligned}
\left\| \frac{1}{\sqrt{n}} M_u g + \frac{1}{\sqrt{n}}\epsilon - \frac{1}{\sqrt{n}}\beta_0 \mathbf{1}_n - \frac{1}{\sqrt{n}} r \right\|_2 &\leq C_1, \\
\left\| \frac{1}{\sqrt{n}} M_\eta h + \frac{1}{\sqrt{n}} s \right\|_2 &\leq C_2.
\end{aligned}$$

So, applying these facts and using Sion's minimax theorem to swap the order of minimization and maximization (Sion 1958), we have that the above can be rewritten as

$$\begin{aligned}
\min_{(|\beta_0| \leq C_{\beta_0}, 0 \leq M_u \leq C_u)} \max_{(c_\eta \leq M_\eta \leq C_\eta)} \min_{0<\rho_1\leq C_1} \max_{0<\rho_2\leq C_2} \min_{\|r\|_2 \leq C_r\sqrt{n}} \max_{\|s\|_2 \leq C_s\sqrt{n}} \Bigg(& \frac{M_\eta}{2n\rho_1} \|M_u g + \epsilon - \beta_0 \mathbf{1}_n - r\|_2^2 \\
& + \frac{M_\eta \rho_1}{2} - \frac{M_u}{2n\rho_2}\|M_\eta h + s\|_2^2 - \frac{M_u \rho_2}{2} + \frac{1}{n}\sum_{i=1}^n \ell_\tau(r_i) + \frac{1}{\sqrt{n}} s^\top \tilde{\beta} - \frac{1}{\sqrt{n}}\tilde{\mathcal{R}}_d^*(s) \Bigg).
\end{aligned}$$

Now, for any convex function f let

$$e_f(x;\rho) = \min_v \frac{1}{2\rho}\|x - v\|^2 + f(v),$$

denote its Moureau envelope. For notational convienence, define a continuous extension at $\rho = 0$ by $e_f(x;0) = f(x)$ (see Lemma C.7). We would like to simplify the above using this definition. This is done in the following lemma.

Lemma B.5. *There exists constant $\tilde{C}_s, \tilde{C}_r > 0$, such that with probability tending to 1, we have that for any $|\beta_0| \leq C$, $0 \leq M_u \leq C_u$, $c_\eta \leq M_\eta \leq C_\eta$, and $\rho_2, \rho_1 > 0$.*

$$\min_{\|r\|_2 \leq \tilde{C}_r\sqrt{n}} \frac{M_\eta}{2n\rho_1}\|M_u g + \epsilon - \beta_0 \mathbf{1}_n - r\|_2^2 + \frac{1}{n}\sum_{i=1}^n \ell_\tau(r_i) = \frac{1}{n} e_{\ell_\tau}\left(M_u g + \epsilon - \beta_0 \mathbf{1}_n; \frac{\rho_1}{M_\eta}\right),$$

and

$$\begin{aligned}&\max_{\|s\|_2 \leq \tilde{C}_s\sqrt{n}} -\frac{M_u}{2n\rho_2}\|M_\eta h + s\|_2^2 + \frac{1}{\sqrt{n}} s^\top \tilde{\beta} - \frac{1}{\sqrt{n}}\tilde{\mathcal{R}}_d^*(s) \\ &= -\frac{M_\eta^2 M_u}{2n\rho_2}\|h\|_2^2 + \frac{1}{n} e_{\mathcal{R}_d}\left(\frac{M_\eta M_u}{\rho_2} h + \sqrt{n}\tilde{\beta}; \frac{M_u}{\rho_2}\right).\end{aligned}$$

Proof. For this first part of the lemma, first note that the cast $\rho_1 = 0$ is immediate. So fix $\rho_1 > 0$. Recall the definition of the proximal function

$$\text{prox}_f(x;\tau) = \underset{v}{\text{argmin}} \frac{1}{2\tau}\|x - v\|_2^2 + f(v).$$

By definition of ℓ_τ, we have $\text{prox}_{\ell_\tau}(x;\tau) = 0$ for any $\tau \in (0,1)$. Then, since the proximal function is non-expansive (Proposition 12.28 of Bauschke & Combettes (2017)) it holds that for any $x \in \mathbb{R}$ and $\tau > 0$,

$$\|\text{prox}_{\ell_\tau}(x;\tau)\|_2 = \|\text{prox}_{\ell_\tau}(x;\tau) - \text{prox}_{\ell_\tau}(0;\tau)\|_2 \leq \|x - 0\|_2 = \|x\|_2.$$

So, in particular,

$$\text{prox}(M_u g + \epsilon - \beta_0 \mathbf{1}_n; \rho_1/M_\eta) \leq \|M_u g + \epsilon - \beta_0 \mathbf{1}_n\|_2.$$

This last quantity must be bounded by the law of large numbers and our restrictions on M_u and β_0. This proves the first part of the lemma. The second part of the lemma follows equations 86-88 of Thrampoulidis et al. (2018) and an identical argument. □

Now, without loss of generality we may assume that $C_r \geq \tilde{C}_r$ and $C_s \geq \tilde{C}_s$. So, applying Lemma B.5 and taking a continuous extension at $\rho_1 = 0$ to our previous calculations gives us the optimization problem

$$\min_{(|\beta_0| \leq C_{\beta_0}, 0 \leq M_u \leq C_u, 0 \leq \rho_1 \leq C_1)} \max_{(c_\eta \leq M_\eta \leq C_\eta, 0 < \rho_2 \leq C_2)} \left(\frac{1}{n} e_{\ell_\tau} \left(M_u g + \epsilon - \beta_0 \mathbf{1}_n; \frac{\rho_1}{M_\eta} \right) \right.$$
$$\left. - \frac{M_\eta^2 M_u}{2n\rho_2} \|h\|_2^2 + \frac{1}{n} e_{\mathcal{R}_d} \left(\frac{M_\eta M_u}{\rho_2} h + \sqrt{n}\tilde{\beta}; \frac{M_u}{\rho_2} \right) + \frac{M_\eta \rho_1}{2} - \frac{M_u \rho_2}{2} \right), \tag{B.3}$$

Our final step is to replace all the random quantities above with their asymptotic limits. This will leave us with a deterministic program. We will show that the solutions to this deterministic program are unique and thus, that the solutions to the above optimization problem converge to this unique limit. A key technical tool in this argument will be the following lemma which states that pointwise convergence can be converted to convergence of the minimum of a convex function. This result is a minor variant of Lemma 10 of Thrampoulidis et al. (2018). We include a proof of the first part of the lemma for completeness.

Lemma B.6 (Extension of Lemma 10 of Thrampoulidis et al. (2018).)**.** *Fix $b > a$ and let $f_n : [a, b] \to \mathbb{R}$ be a sequence of random convex functions converging pointwise in probability to a continuous, convex function $f : [a, b] \to \mathbb{R}$. Then,*

$$\inf_{x \in [a,b]} f_n(x) \xrightarrow{\mathbb{P}} \inf_{x \in [a,b]} f(x).$$

Similarly, if $f_n : (a, b] \to \mathbb{R}$ is a sequence of random convex functions converging pointwise in probability to a continuous, convex function $f : (a, b] \to \mathbb{R}$, then

$$\inf_{x \in (a,b]} f_n(x) \xrightarrow{\mathbb{P}} \inf_{x \in (a,b]} f(x).$$

Proof. We will prove the first part of the lemma. Proof of the second part is similar and is omitted. We have that f is convex and thus continuous on $[a, b]$. Let $x^* \in [a, b]$ be a point at

which f achieves its minimum. Then,

$$\limsup_{n\to\infty} \inf_{x\in[a,b]} f_n(x) \leq \limsup_{n\to\infty} f_n(x^*) \overset{\mathbb{P}}{=} f(x^*) = \inf_{x\in[a,b]} f(x).$$

It suffices to prove a matching lower bound. By Lemma 7.75 of Miescke & Liese (2008) we have that for any interval $a < x_1 < x_2 < b$,

$$\sup_{x\in[x_1,x_2]} |f_n(x) - f(x)| \overset{\mathbb{P}}{\to} 0,$$

and thus also,

$$\inf_{x\in[x_1,x_2]} f_n(x) \overset{\mathbb{P}}{\to} \inf_{x\in[x_1,x_2]} f(x). \tag{B.4}$$

So, we just need to check what happens on the boundary. We will focus on the lower boundary. First, suppose a is not a minimizer of f. Then, we may find $x_1 < x_2 < x^*$ such that $\inf_{x\in[x_1,x_2]} f(x) > \inf_{x\in[a,b]} f(x)$. Applying (B.4) to this interval we conclude that

$$\lim_{n\to\infty} \inf_{x\in[x_1,x_2]} f_n(x) \overset{\mathbb{P}}{=} \inf_{x\in[x_1,x_2]} f(x) > \inf_{x\in[a,b]} f(x) = \lim_{n\to\infty} f_n(x^*).$$

By the convexity of f_n, it follows immediately that $\liminf_{n\to\infty} \inf_{x\in[a,x_2]} f_n(x) \overset{\mathbb{P}}{\geq} \inf_{x\in[a,b]} f(x)$ as well. On the other hand, suppose a is a minimzer of f. Fix $\delta > 0$ and let $a < x_1 < x_2 < b$ be such that $f(x_1), f(x_2) < \inf_{x\in[a,b]} f(x) + \delta$. Then, for any $x \in [a, x_1]$,

$$\begin{aligned}
f_n(x_1) = f_n\left(\frac{x_2 - x_1}{x_2 - x}x + \left(1 - \frac{x_2 - x_1}{x_2 - x}\right)x_2\right) &\leq \frac{x_2 - x_1}{x_2 - x} f_n(x) + \left(1 - \frac{x_2 - x_1}{x_2 - x}\right) f_n(x_2) \\
\implies f_n(x) &\geq \frac{x_2 - x}{x_2 - x_1} f_n(x_1) - \frac{x_2 - x}{x_2 - x_1}\left(1 - \frac{x_2 - x_1}{x_2 - x}\right) f_n(x_2) \\
&\overset{\mathbb{P}}{\geq} \frac{x_2 - x}{x_2 - x_1} f(x_1) - \frac{x_2 - x}{x_2 - x_1}\left(1 - \frac{x_2 - x_1}{x_2 - x}\right) f(x_2) \\
&\geq \inf_{x'\in[a,b]} f(x) - \delta \sup_{x'\in[a,x_1]} \frac{x_2 - x'}{x_2 - x_1}\left(1 - \frac{x_2 - x_1}{x_2 - x'}\right) \\
&= \inf_{x'\in[a,b]} f(x') - \frac{x_1 - a}{x_2 - x_1}\delta.
\end{aligned}$$

Thus,

$$\liminf_{n\to\infty} \inf_{x\in[a,x_1]} f_n(x) \overset{\mathbb{P}}{\geq} \inf_{x\in[a,b]} f(x).$$

By a matching argument, we may also find $x_1' \in (a,b)$ such that

$$\liminf_{n\to\infty} \inf_{x\in[x_1',b]} f_n(x) \overset{\mathbb{P}}{\geq} \inf_{x\in[a,b]} f(x).$$

Combining these two facts and using (B.4) to get convergence on $[x_1, x_1']$ gives the desired result. □

Combining all of the previous results we arrive at the following.

Proposition B.3. *Let V denote the value of the asymptotic program defined in (B.2) with the constants $(C_{\beta_0}, C_u, c_\eta, C_\eta, C_1, C_2)$ defined as in (B.3). Then, under the conditions of Theorem 3.1,*

$$\liminf_{n,d\to\infty} \Phi(\mathbb{R}) \overset{\mathbb{P}}{\geq} V.$$

Proof. This lemma follows immediately by the law of large numbers and repeated applications of Lemma B.6 to (B.3). □

B.3 Analysis of the asymptotic auxillary program

In this section we prove a number of useful results regarding the asymptotic auxiliary program defined in (B.3). In what follows we use

$$\begin{aligned} A(\beta_0, M_u, \rho_1, M_\eta, \rho_2) = \mathbb{E}\left[e_{\ell_\alpha}\left(M_u g_1 + \epsilon_1 - \beta_0; \frac{\rho_1}{M_\eta}\right)\right] - \frac{M_\eta^2 M_u \gamma}{2\rho_2} \\ + \gamma\mathbb{E}\left[e_\nu\left(\frac{M_\eta M_u}{\rho_2} h_1 + \gamma\sqrt{d}\tilde{\beta}_1; \frac{M_u}{\rho_2}\right)\right] + \frac{M_\eta \rho_1}{2} - \frac{M_u \rho_2}{2}, \end{aligned}$$

to denote the objective of this optimization.

Lemma B.7. *Under the assumptions of Theorem 3.1, $A(\beta_0, M_u, \rho_1, M_\eta, \rho_2)$ is jointly continuous, convex in (β_0, M_u, ρ_1), and concave in (M_η, ρ_2) on the domain $\mathbb{R}\times\mathbb{R}_{\geq 0}\times\mathbb{R}_{\geq 0}\times\mathbb{R}_{>0}\times\mathbb{R}_{>0}$.*

Proof. The last four terms of A are jointly continuous by assumption. Moreover, joint of the first follows immediately by inspecting the form of e_{ℓ_α} (Lemma C.3) and applying the dominated convergence theorem. The fact that A is convex-concave follows follows directly from the fact that it is the pointwise limit of a sequence of convex-concave functions. □

Lemma B.8. *Fix any $C_\eta > c_\eta > 0$ and $C_2 > 0$. Under the conditions of Theorem 3.1, the function*

$$(\beta_0, M_u, \rho_1) \mapsto \max_{c_\eta \leq M_\eta \leq C_\eta, 0 \leq \rho_2 \leq C_2} A(\beta_0, M_u, \rho_1, M_\eta, \rho_2),$$

is jointly strictly convex on $\mathbb{R}\times\mathbb{R}_{\geq 0}\mathbb{R}_{>0}$. Moreover, for $\rho_1 = 0$ this function is jointly strictly convex in (β_0, M_u)

Proof. We first consider the case where $\rho_1 > 0$. Fix any $M_\eta \in [c_\eta, C_\eta]$ and pair of distict points $(\beta_0, M_u, \rho_1), (\beta_0', M_u', \rho_1') \in \mathbb{R}\times\mathbb{R}_{\geq 0}\times\mathbb{R}_{>0}$. For $\theta \in [0,1]$ define the function

$$w(\theta) = \mathbb{E}\left[e_{\ell_\tau}\left(((1-\theta)M_u + \theta M_u')g + \epsilon_1 - (1-\theta)\beta_0 - \theta\beta_0'; \frac{(1-\theta)\rho_1 + \theta\rho_1'}{M_\eta}\right)\right].$$

For ease of notation, let $(\xi_1, \xi_2, \xi_3) = (M_u' - M_u, \beta_0' - \beta_0, \rho_1' - \rho_1)$, $Z_\theta = ((1-\theta)M_u + \theta M_u')g + \epsilon_1 - (1-\theta)\beta_0 - \theta\beta_0'$, and $\rho_\theta = ((1-\theta)\rho_1 + \theta\rho_1')$. By the dominated convergence theorem and a direct calculation using the form of e_{ℓ_τ} (see Lemma C.3), we have

$$\begin{aligned} w'(\theta) = \mathbb{E}\Bigg[&(g\xi_1 - \xi_2)\tau\mathbb{1}\left\{Z_\theta > \tau\frac{\rho_\theta}{M_\eta}\right\} + (g\xi_1 - \xi_2)\frac{Z_\theta M_\eta}{\rho_\theta}\mathbb{1}\left\{-(1-\tau)\frac{\rho_\theta}{M_\eta} \leq Z_\theta \leq \tau\frac{\rho_\theta}{M_\eta}\right\} \\ &- (g\xi_1 - \xi_2)(1-\tau)\mathbb{1}\left\{Z_\theta < -(1-\tau)\frac{\rho_\theta}{M_\eta}\right\} - \frac{\tau^2\xi_3}{2M_\eta}\mathbb{1}\left\{Z_\theta > \tau\frac{\rho_\theta}{M_\eta}\right\} \\ &- \frac{Z_\theta^2 M_\eta \xi_3}{2\rho_\theta^2}\mathbb{1}\left\{-(1-\tau)\frac{\rho_\theta}{M_\eta} \leq Z_\theta \leq \tau\frac{\rho_\theta}{M_\eta}\right\} - \frac{(1-\tau)^2\xi_3}{2M_\eta}\mathbb{1}\left\{Z_\theta < -(1-\tau)\frac{\rho_\theta}{M_\eta}\right\}\Bigg], \end{aligned}$$

and

$$w''(\theta) = \mathbb{E}\left[\left((g\xi_1 - \xi_2)^2 \frac{M_\eta}{\rho_\theta} - (g\xi_1 - \xi_2)\xi_3 \frac{Z_\theta M_\eta}{\rho_\theta^2} + \xi_3^2 \frac{Z_\theta^2 M_\eta}{\rho_\theta^3}\right) \mathbb{1}\left\{-(1-\tau)\frac{\rho_\theta}{M_\eta} \leq Z_\theta \leq \tau \frac{\rho_\theta}{M_\eta}\right\}\right]$$
$$= \frac{M_\theta}{\rho_\theta}\mathbb{E}\left[\left(g\xi_1 - \xi_2 - \xi_3 \frac{Z_\theta}{\rho_\theta}\right)^2 \mathbb{1}\left\{-(1-\tau)\frac{\rho_\theta}{M_\eta} \leq Z_\theta \leq \tau \frac{\rho_\theta}{M_\eta}\right\}\right]$$

Since ϵ_1 has positive support on $\mathbb{R}$ we have that Z_θ must have positive support on $\mathbb{R}$ and that $g\xi_1 - \xi_2 - \xi_3 \frac{Z_\theta}{\rho_\theta}$ must have positive support on $\mathbb{R}$ for $\xi_3 \neq 0$. Moreover, if $\xi_3 = 0$, then $(\xi_1, \xi_2) \neq (0,0)$ and we clearly have that $\mathbb{P}(g\xi_1 - \xi_2 = 0) = 0$. In either case, we conclude that $w''(\theta) > 0$. So, in particular, we find that for any fixed values of $M_\eta \in [c_\eta, C_\eta]$, the function

$$(\beta_0, M_u, \rho_1) \mapsto \mathbb{E}\left[e_{\ell_\tau}\left(M_u g + \epsilon_1 - \beta_0; \frac{\rho_1}{M_\eta}\right)\right],$$

is strictly convex. Since this term does not involve ρ_2 and the remainder of the objective is convex (it is the pointwise limit of a convex function) we conclude that the function

$$(\beta_0, M_u, \rho_1) \mapsto \max_{0 < \rho_2 \leq C_2} A(\beta_0, M_u, \rho_1, M_\eta, \rho_2),$$

is strictly convex. Finally, since A is convex-concave we have that for any $(\beta_0, M_u, \rho_1)\mathbb{R} \times \mathbb{R}_{\geq 0}\mathbb{R}_{>0}$, $M_\eta \mapsto \max_{0<\rho_2\leq C_2} A(\beta_0, M_u, \rho_1, M_\eta, \rho_2)$ is concave on $\mathbb{R}_{>0}$ and thus continuous on $[c_\eta, C_\eta]$. The desired result then follows by Lemma C.2.

Now, consider the case $\rho_1 = 0$. Once again, fix $M_\eta \in [c_\eta, C_\eta]$ and pair of distinct points $(M_u, \beta_0), (M_u', \beta_0') \in \mathbb{R}_{\geq 0} \times \mathbb{R}$. For $\theta \in [0,1]$ consider the function

$$\tilde{w}(\theta) = \mathbb{E}[\ell_\tau(((1-\theta)M_u + \theta M_u')g + \epsilon - (1-\theta)\beta_0 - \theta\beta_0')].$$

Let $Z_\theta := (1-\theta)M_u + \theta M_u')g + \epsilon - (1-\theta)\beta_0 - \theta\beta_0'$. By a direct calculation,

$$\tilde{w}'(\theta) = \mathbb{E}\left[((M_u' - M_u)g - (\beta_0' - \beta_0))\mathbb{1}\{Z_\theta > 0\} - ((M_u' - M_u)g - (\beta_0' - \beta_0))\mathbb{1}\{Z_\theta < 0\}\right]$$
$$= \mathbb{E}\left[2((M_u' - M_u)g - (\beta_0' - \beta_0))\mathbb{1}\{Z_\theta > 0\} - ((M_u' - M_u)g - (\beta_0' - \beta_0))\right]$$

and

$$\begin{aligned}\tilde{w}''(\theta) &= \frac{d}{d\theta}\mathbb{E}\left[\mathbb{E}\left[2((M_u' - M_u)g - (\beta_0' - \beta_0))\mathbb{1}\left\{\epsilon > (1-\theta)(\beta_0 - M_u g) + \theta(\beta_0' - M_u' g)\right\}\right] \mid g\right] \\ &= \mathbb{E}\left[2((M_u' - M_u)g - (\beta_0' - \beta_0))^2 p_\epsilon((1-\theta)(\beta_0 - M_u g) + \theta(\beta_0' - M_u' g))\right],\end{aligned}$$

where p_ϵ denotes the density of ϵ. Since this last term is positive we find that $\tilde{w}$ is strictly convex. The desired result then follows by arguing as above. □

Lemma B.9. *Suppose the assumptions of Theorem 3.1 hold. Then, for any $C_{\beta_0}, C_u, C_1, C_2, C_\eta, c_\eta >$ the asymptotic optimization problem (B.2) admits a unique solution for (β_0, M_u, ρ_1). Moreover, letting $(\beta_0^*, M_u^*, \rho_1^*)$ denote this solution we have that $M_u^* > 0 \implies \rho_1^* > 0$.*

Proof. Since the optimization domain for (β_0, M_u, ρ_1) is compact and A is jointly continuous and convex-concave (Lemma B.7) we the optimization program (B.2) must obtain its minimum in (β_0, M_u, ρ_1) (cf. Proposition 1.26 and Theorem 1.9 of Rockafellar & Wets (1997)). The fact that this minimizer is unique then follows directly from Lemma B.8.

Now, let $(\beta_0^*, M_u^*, \rho_1^*)$ denote this unique solution and suppose that $M_u^* > 0$. Recall the identity (Lemma C.6 below),

$$e_f(x;\rho) + e_{f^*}(x/\rho; 1/\rho) = \frac{x^2}{2\rho}.$$

Applying this to our optimization problem, we have that for any $\rho_2 > 0$ and $0 \le \rho_1 \le C_1$,

$$\begin{aligned}A(\beta_0^*, M_u^*, \rho_1, M_\eta, \rho_2) = \mathbb{E}\left[e_{\ell_\alpha}\left(M_u^* g_1 + \epsilon_1 - \beta_0^*; \frac{\rho_1}{M_\eta}\right)\right] - \frac{M_\eta^2 M_u^* \gamma}{2\rho_2} \\ - \gamma\mathbb{E}\left[e_{\nu^*}\left(M_\eta h_1 + \gamma\frac{\rho_2}{M_u^*}\sqrt{d}\tilde{\beta}_1; \frac{\rho_2}{M_u^*}\right)\right] \\ + \gamma\frac{\rho_2}{2M_u^*}\mathbb{E}\left[\left(\frac{M_\eta M_u^*}{\rho_2}h_1 + \sqrt{d}\tilde{\beta}_1\right)^2\right]\end{aligned}$$

$$
\begin{aligned}
&+ \frac{M_\eta \rho_1}{2} - \frac{M_u^* \rho_2}{2} \\
&= \mathbb{E}\left[e_{\ell_\tau}\left(M_u^* g_1 + \epsilon_1 - \beta_0^*; \frac{\rho_1}{M_\eta} \right) \right] - \gamma \mathbb{E}\left[e_{\nu^*}\left(M_\eta h_1 + \gamma \frac{\rho_2}{M_u} \sqrt{d}\tilde{\beta}_1; \frac{\rho_2}{M_u^*} \right) \right] \\
&\quad + \gamma \frac{\rho_2}{2M_u^*} \mathbb{E}[(\sqrt{d}\tilde{\beta}_1)^2] + \frac{M_\eta \rho_1}{2} - \frac{M_u^* \rho_2}{2}.
\end{aligned}
$$

So,

$$
\begin{aligned}
\max_{(0<\rho_2\le C_2, c_\eta \le M_\eta \le C_\eta)} A(\beta_0^*, M_u^*, 0, M_\eta, \rho_2) &\ge \max_{(0<\rho_2\le C_2)} A(\beta_0^*, M_u^*, 0, 0, \rho_2) \\
&= \mathbb{E}\left[\ell_\tau \left(M_u^* g_1 + \epsilon_1 - \beta_0^* \right) \right] - \gamma \mathbb{E}\left[e_{\nu^*}\left(\gamma \frac{\rho_2}{M_u} \sqrt{d}\tilde{\beta}_1; \frac{\rho_2}{M_u^*} \right) \right] \\
&\quad + \gamma \frac{\rho_2}{2M_u^*} \mathbb{E}[(\sqrt{d}\tilde{\beta}_1)^2] - \frac{M_u^* \rho_2}{2}.
\end{aligned}
$$

Now, fix $\rho_1 > 0$ small and $\rho_2 > 0$. By directly examining the definition of e_{ℓ_τ}, (Lemma C.3) we have the pointwise inequality

$$
\begin{aligned}
&e_{\ell_\tau}\left(M_u^* g_1 + \epsilon_1 - \beta_0^*; \frac{\rho_1}{M_\eta} \right) \\
&\le \ell_\tau(M_u^* g_1 + \epsilon_1 - \beta_0^*) - \min\{\tau^2, (1-\tau)^2\} \frac{\rho_1}{2M_\eta} \mathbb{1}\left\{ M_u^* g_1 + \epsilon_1 - \beta_0^* \notin \left[-\frac{\rho_1}{M_\eta}(1-\tau), \frac{\rho_1}{M_\eta}\tau \right] \right\},
\end{aligned}
$$

and thus for any $M_\eta > 0$,

$$
\begin{aligned}
&\mathbb{E}\left[e_{\ell_\tau}\left(M_u^* g_1 + \epsilon_1 - \beta_0^*; \frac{\rho_1}{M_\eta} \right) \right] \le \mathbb{E}\left[\ell_\tau \left(M_u^* g_1 + \epsilon_1 - \beta_0^*; \frac{\rho_1}{M_\eta} \right) \right] \\
&\qquad - \min\{\tau^2, (1-\tau)^2\} \frac{\rho_1}{2M_\eta} \left(1 - \mathbb{P}\left(M_u^* g_1 + \epsilon_1 - \beta_0^* \in \left[-\frac{\rho_1}{M_\eta}(1-\tau), \frac{\rho_1}{M_\eta}\tau \right] \right) \right) \\
&\le \mathbb{E}\left[e_{\ell_\tau}\left(M_u^* g_1 + \epsilon_1 - \beta_0^*; \frac{\rho_1}{M_\eta} \right) \right] \le \mathbb{E}\left[\ell_\tau \left(M_u^* g_1 + \epsilon_1 - \beta_0^*; \frac{\rho_1}{M_\eta} \right) \right] \\
&\qquad - \min\{\tau^2, (1-\tau)^2\} \frac{\rho_1}{2M_\eta} \left(1 - \mathbb{P}\left(M_u^* g_1 + \epsilon_1 - \beta_0^* \in \left[-\frac{\rho_1}{c_\eta}(1-\tau), \frac{\rho_1}{c_\eta}\tau \right] \right) \right).
\end{aligned}
$$

Now, let ρ_1 be sufficiently small such that $\mathbb{P}\left(M_u^* g_1 + \epsilon_1 - \beta_0^* \in \left[-\frac{\rho_1}{c_\eta}(1-\tau), \frac{\rho_1}{c_\eta}\tau \right] \right) \le 1/2$. Then, the above implies that

$$
\mathbb{E}\left[e_{\ell_\tau}\left(M_u^* g_1 + \epsilon_1 - \beta_0^*; \frac{\rho_1}{M_\eta} \right) \right] \le \mathbb{E}\left[\ell_\tau \left(M_u^* g_1 + \epsilon_1 - \beta_0^*; \frac{\rho_1}{M_\eta} \right) \right] - \min\{\tau^2, (1-\tau)^2\} \frac{\rho_1}{4M_\eta}.
$$

Now, by our assumptions of ν^* we also have that

$$\frac{d}{dM_\eta}\mathbb{E}\left[e_{\nu^*}\left(M_\eta h_1 + \gamma\frac{\rho_2}{M_u}\sqrt{d}\tilde{\beta}_1; \frac{\rho_2}{M_u^*}\right)\right] = \mathbb{E}\left[h_1 \partial_x e_{\nu^*}\left(M_\eta h_1 + \gamma\frac{\rho_2}{M_u}\sqrt{d}\tilde{\beta}_1; \frac{\rho_2}{M_u^*}\right)\right]$$
$$= M_\eta \mathbb{E}\left[\partial_x^2 e_{\nu^*}\left(M_\eta h_1 + \gamma\frac{\rho_2}{M_u}\sqrt{d}\tilde{\beta}_1; \frac{\rho_2}{M_u^*}\right)\right].$$

Let $c = \min_{c_\eta \le M_\eta \le C_\eta} \mathbb{E}\left[\partial_x^2 e_{\nu^*}\left(M_\eta h_1 + \gamma\frac{\rho_2}{M_u}\sqrt{d}\tilde{\beta}_1; \frac{\rho_2}{M_u^*}\right)\right]$. By our assumptions we have that $c > 0$. So, applying the mean value theorem we have that

$$\mathbb{E}\left[e_{\nu^*}\left(M_\eta h_1 + \gamma\frac{\rho_2}{M_u}\sqrt{d}\tilde{\beta}_1; \frac{\rho_2}{M_u^*}\right)\right] \ge \mathbb{E}\left[e_{\nu^*}\left(\gamma\frac{\rho_2}{M_u}\sqrt{d}\tilde{\beta}_1; \frac{\rho_2}{M_u^*}\right)\right] + cM_\eta^2.$$

Putting this all together we find that for ρ_1 sufficiently close to 0,

$$\max_{(0<\rho_2\le C_2, c_\eta \le M_\eta \le C_\eta)} A(\beta_0^*, M_u^*, \rho_1, M_\eta, \rho_2)$$
$$\le \max_{(0<\rho_2\le C_2, c_\eta \le M_\eta \le C_\eta)} \mathbb{E}\left[\ell_\tau\left(M_u^* g_1 + \epsilon_1 - \beta_0^*\right)\right] - \gamma\mathbb{E}\left[e_{\nu^*}\left(M_\eta h_1 + \gamma\frac{\rho_2}{M_u}\sqrt{d}\tilde{\beta}_1; \frac{\rho_2}{M_u^*}\right)\right]$$
$$+ \gamma\frac{\rho_2}{2M_u^*}\mathbb{E}[(\sqrt{d}\tilde{\beta}_1)^2] - \frac{M_u^*\rho_2}{2} + \frac{M_\eta\rho_1}{2} - \min\{\tau^2, (1-\tau)^2\}\frac{\rho_1}{4M_\eta} - cM_\eta^2.$$

Finally, for ρ_1 sufficiently small the term $\frac{M_\eta\rho_1}{2} - \min\{\tau^2, (1-\tau)^2\}\frac{\rho_1}{4M_\eta} - cM_\eta^2$ is always negative. Applying this fact and comparing the above to our bounds in the case $\rho_1 = 0$ we find that

$$\max_{(0<\rho_2\le C_2, c_\eta \le M_\eta \le C_\eta)} A(\beta_0^*, M_u^*, \rho_1, M_\eta, \rho_2) < \max_{(0<\rho_2\le C_2, c_\eta \le M_\eta \le C_\eta)} A(\beta_0^*, M_u^*, 0, M_\eta, \rho_2).$$

Thus, $\rho_1^* \neq 0$, as desired.

□

Lemma B.10. *Suppose the assumptions of Theorem 3.1. Fix any $C_{\beta_0}, C_u, C_\eta, c_\eta C_1, C_2, c_1 > 0$ with $C_\eta > c_\eta$ and let $(\beta_0^*, M_u^*, \rho_1^*)$ denote the unique solution to the asymptotic program defined in Lemma B.9. Suppose that $M_u^* > 0$. Then, the asymptotic program (B.2) admits a unique solution for M_η.*

Proof. Since A is jointly convex-concave (Lemma B.7) we know that the function

$$M_\eta \mapsto \max_{(0<\rho_2\leq C_2)} \min_{(|\beta_0|\leq C_{\beta_0}, 0\leq M_u\leq C_u, 0\leq\rho_1\leq C_1)} A(\beta_0, M_u, \rho_1, M_\eta, \rho_2), \tag{B.5}$$

is concave on $\mathbb{R}_{>0}$ and thus continuous on $[c_\eta, C_\eta]$. Thus, this function obtains its maximum M_η.

It remains to show that the maximizer is unique. For ease of notation let $Z = M_u^* g_1 + \epsilon_1 - \beta_0^*$ and define the function

$$w(M_\eta) = \mathbb{E}\left[e_{\ell_\tau}\left(Z; \frac{\rho_1^*}{M_\eta}\right)\right].$$

We claim that w is strongly concave on $[c_\eta, C_\eta]$. To see this note that by a direct calculation using the form of e_{ℓ_τ} (see Lemma C.3), we have

$$\begin{aligned} w'(M_\eta) = \mathbb{E}\Bigg[&\frac{\tau^2\rho_1}{2M_\eta^2}\mathbb{1}\left\{Z > \tau\frac{\rho_1}{M_\eta}\right\} + \frac{Z^2}{2\rho_1}\mathbb{1}\left\{-(1-\tau)\frac{\rho_1}{M_\eta} \leq Z \leq \tau\frac{\rho_1}{M_\eta}\right\} \\ &+ \frac{(1-\tau)^2\rho_1}{2M_\eta^2}\mathbb{1}\left\{Z < -(1-\tau)\frac{\rho_1}{M_\eta}\right\}\Bigg], \end{aligned}$$

and

$$w''(M_\eta) = \mathbb{E}\left[-\frac{\tau^2\rho_1}{M_\eta^3}\mathbb{1}\left\{Z > \tau\frac{\rho_1}{M_\eta}\right\} - \frac{(1-\tau)^2\rho_1}{M_\eta^3}\mathbb{1}\left\{Z < -(1-\tau)\frac{\rho_1}{M_\eta}\right\}\right],$$

and so in particular,

$$\sup_{c_\eta\leq M_\eta\leq C_\eta} w''(M_\eta) \leq \mathbb{E}\left[-\frac{\tau^2\rho_1}{C_\eta^3}\mathbb{1}\left\{Z > \tau\frac{\rho_1}{c_\eta}\right\} - \frac{(1-\tau)^2\rho_1}{C_\eta^3}\mathbb{1}\left\{Z < -(1-\tau)\frac{\rho_1}{c_\eta}\right\}\right] < 0,$$

where the get the last inequality we have applied the fact that ϵ_1 has support on all of $\mathbb{R}$ (and thus that Z has support on all of $\mathbb{R}$).

Now, assume by contradiction that there exists distinct maximizers M_η^1 and M_η^2 for (B.5) on the domain $[c_\eta, C_\eta]$. Since this is a convex-concave problem we must that that M_η^1 and M_η^2

are maximizers of the function

$$M_\eta \mapsto \max_{0<\rho_2\le C_2} A(\beta_0^*, M_u^*, \rho_q^*, M_\eta, \rho_2),$$

one the same domain. Moreover, we must have that $\max_{c_\eta \le M_\eta \le C_\eta} \max_{0<\rho_2\le C_2} A(\beta_0^*, M_u^*, \rho_q^*, M_\eta, \rho_2) < \infty$. This follows immediately from the fact that

$$\begin{aligned}\max_{\sup_{c_\eta \le M_\eta \le C_\eta}} \max_{0<\rho_2\le C_2} A(\beta_0^*, M_u^*, \rho_q^*, M_\eta, \rho_2) &\le \max_{c_\eta \le M_\eta \le C_\eta} \max_{0<\rho_2\le C_2} A(0,0,0, M_\eta, \rho_2) \\ &= \mathbb{E}[\ell_\alpha(\epsilon_1)] + \gamma \mathbb{E}[e_\nu(\gamma\sqrt{d}\tilde{\beta}_1)] < \infty.\end{aligned}$$

Fix $\delta > 0$ small and let $\rho_2^1, \rho_2^2 \times (0, C_2)$ be any two values such that

$$\min_{k\in\{1,2\}} A(\beta_0^*, M_u^*, \rho_1^*, M_\eta^k, \rho_2^k) \ge \max_{c_\eta \le M_\eta \le C_\eta} \max_{0<\rho_2\le C_2} A(\beta_0^*, M_u^*, \rho_1^*, M_\eta, \rho_2) - \delta.$$

By the strong concavity of $w(\eta)$ and the joint concavity of the remainder of $A(\cdot)$ in (M_η, ρ_2) (Lemma B.7) we have

$$\begin{aligned}&\max_{c_\eta \le M_\eta \le C_\eta} \max_{0<\rho_2\le C_2} A(\beta_0^*, M_u^*, \rho_1^*, M_\eta, \rho_2) \ge A\left(\beta_0^*, M_u^*, \rho_1^*, \frac{1}{2}M_\eta^1 + \frac{1}{2}M_\eta^2, \frac{1}{2}\rho_2^1 + \frac{1}{2}\rho_2^2\right) \\ &\ge \frac{1}{2}A(\beta_0^*, M_u^*, \rho_1^*, M_\eta^1, \rho_2^1) + \frac{1}{2}A(\beta_0^*, M_u^*, \rho_1^*, M_\eta^2, \rho_2^2) + \frac{\sup_{c_\eta \le M_\eta \le C_\eta} w''(M_\eta)}{2}(M_\eta^1 - M_\eta^2)^2 \\ &= \max_{c_\eta \le M_\eta \le C_\eta} \max_{0<\rho_2\le C_2} A(\beta_0^*, M_u^*, \rho_1^*, M_\eta, \rho_2) - 2\delta + \frac{\sup_{c_\eta \le M_\eta \le C_\eta} w''(M_\eta)}{2}(M_\eta^1 - M_\eta^2)^2,\end{aligned}$$

as so rearranging,

$$(M_\eta^1 - M_\eta^2)^2 \le \frac{4\delta}{\sup_{c_\eta \le M_\eta \le C_\eta} w''(M_\eta)}.$$

Sending $\delta \to 0$ gives the desired result. □

Our last result of this section gives a first order condition for ρ_1^*.

Lemma B.11. *Suppose the conditions of Theorem 3.1 hold. Fix any $C_{\beta_0}, C_u, C_\eta, c_\eta, C_1, C_2 > 0$ with $C_\eta > c_\eta$ and $C_1 > \sqrt{C_u^2 + \mathbb{E}[\epsilon_1^2] + C_{\beta_0}^2}$.. Let $(M_u^*, \rho_1^*, \beta_0^*, M_\eta^*)$ denote the unique optimal*

solutions to the asymptotic program (B.2) defined in Lemmas B.9 and B.10. Suppose that $M_u^* > 0$. *Then,* ρ_1^* *satisfies the first order condition*

$$\rho_1^* = \sqrt{\mathbb{E}\left[\left(Z - prox_{\ell_\tau}\left(M_u^* g + \epsilon_1 - \beta_0^*; \frac{\rho_1}{M_\eta^*}\right)\right)^2\right]}.$$

Proof. Under the given assumptions we must have that ρ_1^* minimizes the function

$$w(\rho_1) = \mathbb{E}\left[e_{\ell_\alpha}\left(M_u^* g_1 + \epsilon_1 - \beta_0^*; \frac{\rho_1}{M_\eta^*}\right)\right] + \frac{M_\eta^* \rho_1}{2},$$

on the interval $[0, C_1]$. For ease of notation let $Z = M_u^* g_1 + \epsilon_1 - \beta_0^*$. By, Lemma 15(iii) of Thrampoulidis et al. (2018) we have that $\frac{d}{d\rho} e_{\ell_\tau}(x;\rho) = -\frac{1}{2\rho^2}(x - \text{prox}_{\ell_\tau}(x;\rho))^2$. Applying this fact and the dominated convergence theorem gives

$$w'(\rho_1) = -\frac{M_\eta^*}{2\rho_1^2}\mathbb{E}\left[\left(Z - \text{prox}_{\ell_\tau}\left(Z; \frac{\rho_1}{M_\eta^*}\right)\right)^2\right] + \frac{M_\eta^*}{2}.$$

So,

$$w'(\rho_1) = 0 \iff \rho_1 = \sqrt{\mathbb{E}\left[\left(Z - \text{prox}_{\ell_\tau}\left(Z; \frac{\rho_1}{M_\eta^*}\right)\right)^2\right]}.$$

Finally, recall that the function $h(z) = z - \text{prox}_{\ell_\tau}(z;\rho)$ is 1-Lipschitz (cf. Proposition 12.28 of Bauschke & Combettes (2017)) with $h(z) = 0$. Thus, the right-hand-side above is at most

$$\sqrt{\mathbb{E}\left[\left(Z - \text{prox}_{\ell_\tau}\left(Z; \frac{\rho_1}{M_\eta^*}\right)\right)^2\right]} \leq \sqrt{\mathbb{E}\left[(Z)^2\right]} \leq \sqrt{C_u^2 + \mathbb{E}[\epsilon_1^2] + C_{\beta_0}^2}.$$

By assumption we must have that this last quantity is strictly below C_1. Thus, ρ_1^* must satisfy the given first order condition.

□

B.4 Final steps

In this section we prove Theorem 3.1. We begin by stating a convergence result for the primal variables.

Theorem B.1. *Let $C_u, C_{\beta_0}, C_\eta, c_\eta, C_1, C_2$ be constants satisfying Lemmas B.3, B.2, B.4, and B.5. Then, under the assumptions of Theorem 3.1, we have that for all $\delta > 0$*

$$\mathbb{P}\left(\textit{For all primal solutions to (B.1), } |\|\hat{\beta} - \tilde{\beta}\|_2 - M_u^*| < \delta \textit{ and } |\hat{\beta}_0 - \beta_0^*| < \delta\right) \to 1,$$

where M_u^* and β_0^* denote the unique solutions for M_u and β_0 in the asymptotic program (B.2).

Proof. The proof of this result follows similar steps to the proof of Theorem 3.1. Namely, following similar arguments to those presented in Section B.2 for the dual variables, one can show that to prove this result it is sufficient to bound the value of the program

$$\phi^{\text{primal}}(S) := \max_{(\|s\|_2 \le C_s\sqrt{n}, c_\eta \le M_\eta \in \le C_\eta)} \min_{(\|r\|_2 \le C_r\sqrt{n}, (\beta_0, u) \in S)} \min_{(\eta : \|\eta\|_2 = M_\eta)} \left(\frac{1}{n}\|u\|_2 \eta^\top g + \frac{1}{n}\|\eta\|_2 u^\top h + \frac{1}{n}\eta^\top \epsilon \right.$$
$$\left. - \frac{1}{n}\beta_0 \sum_{i=1}^n \eta_i - \frac{1}{n}\eta^\top r + \frac{1}{n}\sum_{i=1}^n \ell_\tau(r_i) + \frac{1}{\sqrt{n}} s^\top(\tilde{\beta} + u) - \frac{1}{\sqrt{n}}\tilde{\mathcal{R}}_d^*(s) \right),$$

for various choices of S. Arguing as above the values of this program are completely characterized by values of the asymptotic program (B.2). Convergence of $\|\hat{\beta} - \tilde{\beta}\|_2$ and $\hat{\beta}_0$ then follows from similar arguments to those presented in Proposition B.5 where we show a convergence result for $\|\hat{\eta}\|_2$. Since the details of this proof closely mirror our existing arguments, they are omitted. □

We now turn to the proof of Theorem 3.1. The next proposition verifies this result in the case $M_u^* = 0$.

Proposition B.4. *Suppose the conditions of Theorem 3.1 hold. Let (M_u^*, β_0^*) be defined as in Theorem B.1 and suppose that $M_u^* = 0$. Let $p := \mathbb{P}(\epsilon_1 - \beta_0^* < 0)$ and define the distribution $P_\eta = p\delta_{-(1-\tau)} + (1-p)\delta_\tau$. Then, for all $\xi > 0$, we have that with probability tending to one all dual solutions $\hat{\eta}$ to B.1 satisfy*

$$\left|\frac{1}{n}\sum_{i=1}^n \mathbb{1}\{\hat{\eta}_i = -(1-\tau)\} - p\right|, \left|\frac{1}{n}\sum_{i=1}^n \mathbb{1}\{\hat{\eta}_i = \tau\} - (1-p)\right|, \frac{1}{n}\sum_{i=1}^n \mathbb{1}\{\hat{\eta}_i \in (-(1-\tau), \tau)\} \le \xi.$$

In particular, the result Theorem 3.1 goes through with this choice of P_η.

Proof. We will focus on the bound on $\frac{1}{n}\sum_{i=1}^n \mathbb{1}\{\hat{\eta}_i \in (-(1-\tau), \tau)\}$. The bounds on the other two terms are similar. By the first order conditions of the optimization in r we have that for any primal solution $(\hat{\beta}_0, \hat{\beta})$,

$$\hat{\eta}_i \in \begin{cases} \tau, & Y_i > \hat{\beta}_0 + X_i^\top \hat{\beta}, \\ [(1-\tau), \tau], & Y_i = \hat{\beta}_0 + X_i^\top \hat{\beta}, \\ -(1-\tau), & Y_i < \hat{\beta}_0 + X_i^\top \hat{\beta}, \end{cases} = \begin{cases} \tau, & \epsilon_i > \hat{\beta}_0 + X_i^\top(\hat{\beta} - \tilde{\beta}), \\ -[\tau, 1-\tau], & \epsilon_i = \hat{\beta}_0 + X_i^\top(\hat{\beta} - \tilde{\beta}), \\ -(1-\tau), & \epsilon_i < \hat{\beta}_0 + X_i^\top(\hat{\beta} - \tilde{\beta}). \end{cases}$$

Now, by standard results (e.g. Theorem 3.1 of Yin et al. (1988)) we have that $\sigma_{\max}(X)/\sqrt{n}$ is converging in probability to a constant $c > 0$. In particular, this implies that with probability converging to one, $\|X(\hat{\beta} - \tilde{\beta})\|_1 \leq \sqrt{n}\|X(\hat{\beta} - \tilde{\beta})\|_2 \leq n2c\|\hat{\beta} - \tilde{\beta}\|_2$. Now, by the Glivenko-Cantelli theorem the empirical cdf of $\frac{1}{n}\sum_{i=1}^n \delta_{\epsilon_i}$ is converging uniformly in probability to P_ϵ. Thus, for any $\rho > 0$ we have

$$\begin{aligned} \frac{1}{n}\sum_{i=1}^n \mathbb{1}\{\hat{\eta}_i \in (-(1-\tau), \tau)\} &\leq \frac{1}{n}\sum_{i=1}^n \mathbb{1}\{|\epsilon_i - \hat{\beta}_0 + X_i^\top(\hat{\beta} - \tilde{\beta})| \leq \rho\} \\ &\leq \frac{1}{n}\sum_{i=1}^n \mathbb{1}\{|\epsilon_i| \leq 2\rho\} + \frac{1}{n}\sum_{i=1}^n \mathbb{1}\{|\hat{\beta}_0 + X_i^\top(\hat{\beta} - \tilde{\beta})| > \rho\} \\ &\leq \frac{1}{n}\sum_{i=1}^n \mathbb{1}\{|\epsilon_i - \hat{\beta}_0| \leq 2\rho\} + \frac{1}{n\rho}\|\hat{\beta}_0 + X_i^\top(\hat{\beta} - \tilde{\beta})\|_1 \\ &\leq \frac{1}{n}\sum_{i=1}^n \mathbb{1}\{|\epsilon_i - \beta_0^*| \leq 3\rho\} + \mathbb{1}\{|\beta_0^* - \hat{\beta}_0| > \rho\} \\ &\quad + \frac{2c}{\rho}\|\hat{\beta} - \tilde{\beta}\|_2. \end{aligned}$$

Applying the law of large numbers and the results of Lemma B.3 we find that with probability tending to one,

$$\sup_{\hat{\eta}} \frac{1}{n}\sum_{i=1}^n \mathbb{1}\{\hat{\eta}_i \in (-(1-\tau), \tau)\} \leq \sup_{\hat{\beta}_0, \hat{\beta}} \frac{1}{n}\sum_{i=1}^n \mathbb{1}\{|\epsilon_i - \beta_0^*| \leq 3\rho\} + \mathbb{1}\{|\beta_0^* - \hat{\beta}_0| > \rho\} \frac{2c}{\rho}\|\hat{\beta} - \tilde{\beta}\|_2$$

where the supremuma are over all dual solutions for η and all primal solutions for (β_0, β), respectively. The desired bound on the left hand side follows by taking ρ sufficiently small and using the fact that ϵ_1 has a continuous distribution. □

We now turn to the main proof of Theorem 3.1, which focuses on the more difficult case in which $M_u^* > 0$. To begin, we first show that $\|\hat{\eta}_2\|_2$ converges.

Proposition B.5. *Assume the conditions of Theorem 3.1 hold. Let $C_u, C_{\beta_0}, C_\eta, c_\eta, C_1, C_2$ be constants satisfying Lemmas B.3, B.2, B.4, and B.5. Let M_u^* and M_η^* denote the unique solutions for M_u and M_η in the asymptotic program (B.2) and assume that $M_u^* > 0$. Then, for all $\delta > 0$,*

$$\mathbb{P}\left(\textit{For all dual solutions of (B.1)}, \; |\|\hat{\eta}\|_2 - \sqrt{n}M_\eta^*| \leq \delta\right) \to 1.$$

Proof. Let V denote the value of the asymptotic optimization program (B.2). By Propositions B.2 and B.3. It is sufficient to show that there exists $\xi > 0$ such that with probability converging to one,

$$\phi(\{\eta : \sqrt{n}c_\eta \leq \|\hat{\eta}\|_2 \leq \sqrt{n}C_\eta, |\|\hat{\eta}\|_2 - \sqrt{n}M_\eta^*| \geq \delta\}) < V - \xi.$$

Now, by repeating the arguments of Section B.2, we have that

$$\phi(\{\eta : \sqrt{n}c_\eta \leq \|\hat{\eta}\|_2 \leq \sqrt{n}C_\eta, \|\hat{\eta}\|_2 \geq +\sqrt{n}M_\eta^*\delta\}) \xrightarrow{\mathbb{P}} \max_{(\max_{c_\eta, M_\eta^*+\delta} M_\eta \leq \min\{C_\eta, M_\eta^* - \delta, 0 < \rho_2 \leq C_2)} \min_{(|\beta_0| \leq C_{\beta_0}, 0 \leq M_u \leq C_u, 0 \leq \rho_1}$$

By Lemma B.10, the right hand side is strictly less than V, as desired. □

We now prove Theorem 3.1.

Proof of Theorem 3.1. Let $C_u, C_{\beta_0}, C_\eta, c_\eta, C_1, C_2$ be constants satisfying Lemmas B.3, B.2, B.4, B.5 and B.11 and $(M_u^*, \beta_0^*, M_\eta^*, \rho_1^*)$ denote the unique solutions to the associated asymptotic program (B.2) defined in Lemmas B.9 and B.10. If $M_u^* = 0$ the result follows from

Proposition B.4. So, suppose $M_u^* > 0$. Fix any bounded L-lipschitz function ψ. Let V denote the value of the asymptotic optimization program (B.2) and fix any $\kappa > 0$ small. Let $S_{\kappa,\delta}$ denote the set

$$\{\eta : \max\{\sqrt{n}c_\eta, M_\eta^* - \kappa\} \leq \|\hat{\eta}\|_2 \leq \min\{\sqrt{n}C_\eta, M_\eta^* + \kappa\}, |\frac{1}{n}\sum_{i=1}^n \psi(\eta_i) - \mathbb{E}_{Z\sim P_\eta}[\psi(Z)]| \geq \delta\}.$$

By Propositions B.2, B.3, and B.5 it is sufficient to show that there exists $\xi > 0$ such that with probability converging to one,

$$\phi(S_{\kappa,\delta}) < V - \xi.$$

First, note since ψ is Lipchitz we have that for κ sufficiently small, $S_{\kappa,\delta} \subseteq S_{0,\delta/2}$. So,

$$\begin{aligned}
\phi(S_{\kappa,\delta}) &\leq \phi(S_{0,\delta/2}) \\
&= \min_{(\|r\|_2 \leq C_r\sqrt{n}, |\beta_0| \leq C_{\beta_0}, 0 \leq M_u \leq C_u)} \max_{(\|s\|_2 \leq C_s\sqrt{n}, \eta \in S_{0,\delta/2})} \left(\frac{1}{n} M_u \eta^\top g - M_u \left\|\frac{1}{n}\|\eta\|_2 h + \frac{1}{n}s\right\|_2 \right. \\
&\qquad \left. + \frac{1}{n}\eta^\top \epsilon - \frac{1}{n}\beta_0 \sum_{i=1}^n \eta_i - \frac{1}{n}\eta^\top r + \frac{1}{n}\sum_{i=1}^n \ell_\tau(r_i) + \frac{1}{\sqrt{n}} s^\top \tilde{\beta} - \frac{1}{\sqrt{n}}\tilde{\mathcal{R}}_d^*(s)\right) \\
&\leq \min_{(\|r\|_2 \leq C_r\sqrt{n})} \max_{(\|s\|_2 \leq C_s\sqrt{n}, \eta \in S_{0,\delta/2})} \left(\frac{1}{n} M_u^* \eta^\top g - M_u^* \left\|\frac{1}{n} M_\eta^* h + \frac{1}{n}s\right\|_2 + \frac{1}{n}\eta^\top \epsilon \right. \\
&\qquad \left. - \frac{1}{n}\beta_0^* \sum_{i=1}^n \eta_i - \frac{1}{n}\eta^\top r + \frac{1}{n}\sum_{i=1}^n \ell_\tau(r_i) + \frac{1}{\sqrt{n}} s^\top \tilde{\beta} - \frac{1}{\sqrt{n}}\tilde{\mathcal{R}}_d^*(s)\right) \\
&= \min_{(\|r\|_2 \leq C_r\sqrt{n})} \max_{(\eta \in S_{0,\delta/2})} \left(\frac{1}{n}\eta^\top (M_u^* g + \epsilon - \beta_0^* \mathbf{1}_n - r) + \frac{1}{n}\sum_{i=1}^n \ell_\tau(r_i)\right) \\
&\qquad + \max_{\|s\|_2 \leq C_s\sqrt{n}} -M_u^* \left\|\frac{1}{n} M_\eta^* h + \frac{1}{n}s\right\|_2 + \frac{1}{\sqrt{n}} s^\top \tilde{\beta} - \frac{1}{\sqrt{n}}\tilde{\mathcal{R}}_d^*(s).
\end{aligned} \tag{B.6}$$

Arguing as in Section B.2 (and in particular applying Lemmas (B.6 and B.5 along with the law of large numbers), the second term converges as

$$\begin{aligned}
&\max_{\|s\|_2 \leq C_s\sqrt{n}} -M_u^* \left\|\frac{1}{n} M_\eta^* h + \frac{1}{n}s\right\|_2 + \frac{1}{\sqrt{n}} s^\top \tilde{\beta} - \frac{1}{\sqrt{n}}\tilde{\mathcal{R}}_d^*(s) \\
&\xrightarrow{\mathbb{P}} \max_{-0 < \rho_2 \leq C_2} -\frac{(M_\eta^*)^2 M_u^* \gamma}{2} \mathbb{E}\left[e_\nu\left(\frac{M_\eta^* M_u^*}{\rho_2} h + \gamma\tilde{\beta}; \frac{M_u^*}{\rho_2}\right)\right] - \frac{M_u^* \rho_2}{2}.
\end{aligned} \tag{B.7}$$

It remains to consider the first term. Now define the vector $r_i^* = \text{prox}_{\ell_\tau}(M_u^* g_i + \epsilon_i - \beta_0^*; \rho_1^*/M_\eta^*)$. By the law of large numbers we have that with probability converging to one,

$$\frac{1}{n}\sum_{i=1}^n \psi\left(\frac{M_\eta^*(M_u^* g_i + \epsilon - \beta_0^* - r_i^*)}{\rho_1^*}\right) \xrightarrow{\mathbb{P}} \mathbb{E}_{Z\sim P_\eta}[Z],$$

and that $\|M_u^* g_i + \epsilon - \beta_0^* - r_i^*\|_2/\sqrt{n} \xrightarrow{\mathbb{P}} \rho_1^* > 0$ (cf. Lemmas B.9 and B.11). Since ψ is Lipchitz this implies that

$$\begin{aligned}
&\liminf_{n\to\infty} \inf_{\eta\in S_{0,\delta/2}} \left\|\eta - \frac{M_\eta^*(M_u^* g_i + \epsilon - \beta_0^* - r_i^*)}{\rho_1^*}\right|_2 \\
&\geq \liminf_{n\to\infty} \frac{1}{L}\left|\frac{1}{n}\sum_{i=1}^n \psi\left(\frac{M_\eta^*(M_u^* g_i + \epsilon - \beta_0^* - r_i^*)}{\rho_1^*}\right) - \frac{1}{n}\sum_{i=1}^n \psi(\eta_i)\right| \overset{\mathbb{P}}{\geq} \frac{\delta}{2L}
\end{aligned}$$

For ease of notation let $Z^* = M_u^* g + \epsilon - \beta_0^* \mathbf{1}_n - r^*$. Applying these facts we find that the optimization appearing on line (B.6) can be bounded as,

$$\begin{aligned}
&\max_{(\eta\in S_{0,\delta/2})} \left(\frac{1}{n}\eta^\top Z^* + \frac{1}{n}\sum_{i=1}^n \ell_\tau(r_i^*)\right) \\
&= \max_{(\eta\in S_{0,\delta/2})} \left(\frac{1}{n}\frac{\eta^\top Z^*}{M_\eta^*\|Z^*\|_2} M_\eta^* \|Z^*\|_2\right) + \mathbb{E}\left[\ell_\alpha\left(\text{prox}_{\ell_\tau}\left(M_u^* g_1 + \epsilon_1 - \beta_0^*; \frac{\rho_1^*}{M_\eta^*}\right)\right)\right] + o_{\mathbb{P}}(1) \\
&= \max_{(\eta\in S_{0,\delta/2})} \left(1 - \frac{1}{2}\left\|\frac{1}{\sqrt{n}}\frac{\eta}{M_\eta^*} - \frac{Z^*}{\|Z^*\|_2}\right\|_2^2\right) \frac{M_\eta^*\|Z^*\|_2}{\sqrt{n}} \\
&\qquad\qquad + \mathbb{E}\left[\ell_\alpha\left(\text{prox}_{\ell_\tau}\left(M_u^* g_1 + \epsilon_1 - \beta_0^*; \frac{\rho_1^*}{M_\eta^*})\right)\right)\right] + o_{\mathbb{P}}(1) \\
&\leq \left(1 - \frac{\delta^2}{8L^2(M_\eta^*)^2}\right) M_\eta^*\rho_1^* + \mathbb{E}\left[\ell_\alpha\left(\text{prox}_{\ell_\tau}\left(M_u^* g_1 + \epsilon_1 - \beta_0^*; \frac{\rho_1^*}{M_\eta^*})\right)\right)\right] + o_{\mathbb{P}}(1) \\
&= \frac{M_\eta^*\rho_1^*}{2} + \mathbb{E}\left[e_{\ell_\alpha}\left(\text{prox}_{\ell_\tau}\left(M_u^* g_1 + \epsilon_1 - \beta_0^*; \frac{\rho_1^*}{M_\eta^*})\right)\right)\right] - \frac{\delta^2}{8L^2(M_\eta^*)^2} + o_{\mathbb{P}}(1),
\end{aligned}$$

where the last line applies the formula for ρ_1^* given in Lemma B.11. Combining this with (B.7) we conclude that

$$\phi(S_\kappa, \delta) \leq V - \frac{\delta^2}{8L^2(M_\eta^*)^2} + o_{\mathbb{P}}(1).$$

As discussed above, this proves the desired result.

□

B.5 Corollaries of Theorem 3.1

We now prove Corollaries 3.1 and 3.2.

Proof of 3.1. Let $C_u, C_{\beta_0}, C_\eta, c_\eta, C_1, C_2$ be constants satisfying Lemmas B.3, B.2, B.4, B.5 and B.11 and $(M_u^*, \beta_0^*, M_\eta^*, \rho_1^*)$ denote the unique solutions to the associated asymptotic program (B.2) defined in Lemmas B.9 and B.10. We claimed that the unregularized quantile regression program must have $M_u^* > 0$. To see this, let $(\hat{\beta}, \hat{\beta}, \hat{r}, \hat{\eta})$ denote any primal-dual solutions to the quantile regression (B.1). Recall that the first order conditions of this program in r imply that

$$\hat{\eta}_i \in \begin{cases} \tau, \ Y_i > \beta_0 + X_i^\top \hat{\beta} \\ [-(1-\tau), \tau], \ Y_i = \beta_0 + X_i^\top \hat{\beta}, \\ Y_i < \beta_0 + X_i^\top \hat{\beta} \end{cases}$$

By Proposition B.4, $M_u^* = 0$ implies that with probability converging to one

$$\frac{1}{n} \sum_{i=1}^n \mathbb{1}\{\hat{\eta}_i \in (-(1-\tau), \tau)\} < d.$$

We will show that this is not possible. Introduce the notation X_A to denote the submatrix of X consisting of the rows in $A \subseteq \{1, \ldots, n\}$ and $X_{A,B}$ to denote the submatrix with rows in $A \subseteq \{1, \ldots, n\}$ and columns in $B \subseteq \{1, \ldots, d\}$. Similarly, let $\hat{\eta}_A$ denote the corresponding entries of $\hat{\eta}_A$. For any fixed set $S \subseteq \{1, \ldots, n\}$ with $|S| = d - 1$ and vector $v \in \{-(1-\tau), \tau\}^{n-d-1}$ we have that with probability one $v^T X_{S^c}$ is in the kernel of the rowspace of X_S. This follows immediately from the fact that

$$u^\top X_S = v^\top X_{S^c} \implies v^\top X_{S^c,\{1,\ldots,d-1\}} (X_{S,\{1,\ldots,d-1\}}^\top)^{-1} X_{S,\{d\}} = v^\top X_{S^c,\{d\}},$$

which occurs with probability zero since $v^\top X_{S^c,\{d\}}$ is a continuous random variable independent of $v^\top X_{S^c,\{1,\ldots,d-1\}} (X_{S,\{1,\ldots,d-1\}}^\top)^{-1} X_{S,\{d\}}$. Moreover, by first order conditions of the quantile

regression in β we must have that $\hat{\eta}^\top X = 0$. So, putting this all together we find that with probability one all dual solutions must satisfy $\frac{1}{n}\sum_{i=1}^n \mathbb{1}\{\hat{\eta}_i \in (-(1-\tau), \tau)\} \geq d$ and thus that $M_u^* > 0$.

Now, fix any $\delta > 0$ small. Let q^* denote the τ quantile of the asymptotic distribution P_η defined in Theorem 3.1. We will show that the with probability converging to one the empirical quantile of $\hat{\eta}$ lies below q^*. Proof of a matching lower bound is identical. If $q^* = \tau$ then the result is immediate. So, suppose that $q^* < \tau$. Let ψ_δ be the step function

$$\psi_\delta(x) = \begin{cases} 0, \ x > q^* + 2\delta, \\ \frac{q^*+2\delta-x}{\delta}, \ q^* + \delta \leq x \leq q^* + 2\delta \\ 1, \ x < q^* + \delta. \end{cases}$$

By Theorem 3.1 we have that with probability converging to one, all dual solutions satisfy

$$\frac{1}{n}\sum_{i=1}^n \mathbb{1}\{\hat{\eta}_i \leq q^*\} \geq \frac{1}{n}\sum_{i=1}^n \psi_\delta(\hat{\eta}_i) \geq \mathbb{E}_{Z\sim P_\eta}[\psi_\delta(Z)] - \xi \geq \mathbb{P}_{P_\eta}(Z \leq q^*) + \mathbb{P}_{P_\eta}(q^* + \delta \leq Z \leq q^* + 2\delta) - \xi.$$

Since $M_u^* > 0$ we must have that $\rho_1^* > 0$ and thus that P_η has a positive density on $(-(1-\tau), \tau)$ with point masses and $-(1-\tau)$ and τ. In particular, by choosing δ sufficiently small we may guarantee that with probability converging to one, all dual solutions satisfy

$$\frac{1}{n}\sum_{i=1}^n \mathbb{1}\{\hat{\eta}_i \leq q^*\} \geq \mathbb{P}_{P_\eta}(Z \leq q^*) + \mathbb{P}_{P_\eta}(q^* + \delta \leq Z \leq q^* + 2\delta) - \xi \geq \tau,$$

and thus that

$$\text{Quantile}\left(\tau, \frac{1}{n}\sum_{i=1}^n \delta_{\hat{\eta}_i}\right) \leq q^*,$$

as claimed. □

Proof of Corollary 3.2. We claim that is sufficient to show that with probability one all dual solutions satisfy $\hat{\eta}_i \neq 0$ for all i. To see this note that in this case Proposition 2.1 and Lemma

A.1 imply that all dual solutions and all leave-one-out primal solutions satisfy

$$\frac{1}{n}\sum_{i=1}^{n}\mathbb{1}\{\hat{\eta}_i \leq 0\} = \frac{1}{n}\sum_{i=1}^{n}\mathbb{1}\{Y_i \leq \hat{\beta}_0^{-i} + X_i^\top \hat{\beta}^{-i}\},$$

where we use $(\hat{\beta}_0^{-i}, \hat{\beta}^{-i})$ to denote primal solutions obtained when (X_i, Y_i) are excluded from the fit. Notably, these equations imply that $\frac{1}{n}\sum_{i=1}^{n}\mathbb{1}\{\hat{\eta}_i \leq 0\}$ and $\frac{1}{n}\sum_{i=1}^{n}\mathbb{1}\{Y_i \leq \hat{\beta}_0^{-i} + X_i^\top \hat{\beta}^{-i}$ take on the same value no matter which primal and dual solution we pick and so in what follows we may simply discuss these quantities as standard, well-defined random variables. Now, by Theorem 3.1 and the continuity of the distribution of P_η at zero it is straightforward to show that

$$\frac{1}{n}\sum_{i=1}^{n}\mathbb{1}\{Y_i \leq \hat{\beta}_0^{-i} + X_i^\top \hat{\beta}^{-i}\} = \frac{1}{n}\sum_{i=1}^{n}\mathbb{1}\{\hat{\eta}_i \leq 0\} \xrightarrow{\mathbb{P}} \mathbb{P}_{Z\sim P_\eta}(Z \leq 0),$$

and since all these random variables are bounded we moreover have that

$$\mathbb{P}(Y_1 \leq \hat{\beta}_0^{-1} + X_1^\top \hat{\beta}^{-1}) = \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}\mathbb{1}\{Y_i \leq \hat{\beta}_0^{-i} + X_i^\top \hat{\beta}^{-i}\}\right] \to \mathbb{P}_{Z\sim P_\eta}(Z \leq 0),$$

or equivalently that $\mathbb{P}(Y_{n+1} \leq \hat{\beta}_0 + X_{n+1}^\top \hat{\beta}) \to \mathbb{P}_{Z\sim P_\eta}(Z \leq 0)$, which gives us the desired result.

It remains to show that $\hat{\eta}_i \neq 0$. The proof of this fact is essentially identical to the proof of Theorem 2.1 under Assumption 1. We will short proof here emphasizing only the aspects of the arguments that are new and leaving many details to the proof of Theorem 2.1.

Fix any dual solution $\hat{\eta}$ and primal solution $(\hat{\beta}_0, \hat{\beta})$. For ease of notation let $\tilde{X} = [X \mid \mathbf{1}_n]$ denote the matrix obtained by adding a column of all ones to X. For any sets $A \subseteq \{1, \ldots, n\}$ and $B \subseteq \{1, \ldots, d+1\}$ let $\tilde{X}_{A,B}$ denote the sub-matrix of X given by the rows with indices in A and columns with indices in B. Let $\hat{\eta}_A$ be the vector given by the entries with indices in A. Let $I_{\text{int.}}(\hat{\eta}) = \{i \in \{1, \ldots, n\} : -(1-\tau) < \hat{\eta}_i < \tau\}$ and $\Lambda = \text{diag}(\lambda_1, \ldots, \lambda_d)$ be the

diagonal matrix with diagonal entries $\lambda_1, \ldots, \lambda_d$. Assume that $I_{\text{int.}}(\hat{\eta}) \neq \emptyset$ (otherwise there is nothing to prove). Let $J_+ = \{j : \lambda_j > 0\}$. Following the calculations of Theorem 2.1, there exists sets $I_{\text{sub}}(\hat{\eta}) \subseteq I_{\text{int.}}(\hat{\eta})$ and $J_{\text{sub.}}(\hat{\eta}) \subseteq \{j : \lambda_j = 0\}$ such that

$$\hat{\eta}_{I_{\text{int.}}(\hat{\eta})} = \begin{bmatrix} \frac{1}{2}\tilde{X}_{I_{\text{sub.}}(\hat{\eta}),J_+}\Lambda_{J_+}^{-1}X^\top_{I_{\text{int.}}(\hat{\eta}),J_+} \\ \tilde{X}^\top_{I_{\text{int.}}(\hat{\eta}),J_{\text{sub.}}(\hat{\eta})} \end{bmatrix}^{-1} \begin{pmatrix} Y_{I_{\text{sub.}}(\hat{\eta})} - \frac{1}{2}\tilde{X}_{I_{\text{sub.}}(\hat{\eta}),J_+}\Lambda_{J_+}^{-1}X^\top_{I_{\text{int.}}(\hat{\eta})^c,J_+}\hat{\eta}_{I_{\text{int.}}(\hat{\eta})^c} \\ \tilde{X}^\top_{I_{\text{int.}}(\hat{\eta})^c,J_{\text{sub.}}(\hat{\eta})}\hat{\eta}_{I_{\text{int.}}(\hat{\eta})^c} \end{pmatrix}.$$

Now, if $d+1 \notin J_{\text{sub.}}(\hat{\eta})$, then proof of Theorem 2.1 immediately tell us that w.p.1 $\hat{\eta}_{I_{\text{int.}}(\hat{\eta})}$ has no non-zero entries. So, suppose that $d+1 \in J_{\text{sub.}}(\hat{\eta})$. Let $\tilde{J}_{\text{sub.}}(\hat{\eta}) = J_{\text{sub.}}(\hat{\eta}) \setminus \{d+1\}$ and rewrite the above as

$$\hat{\eta}_{I_{\text{int.}}(\hat{\eta})} = \begin{bmatrix} \frac{1}{2}\tilde{X}_{I_{\text{sub.}}(\hat{\eta}),J_+}\Lambda_{J_+}^{-1}X^\top_{I_{\text{int.}}(\hat{\eta}),J_+} \\ \tilde{X}^\top_{I_{\text{int.}}(\hat{\eta}),\tilde{J}_{\text{sub.}}(\hat{\eta})} \\ \mathbf{1}^\top_{|I_{\text{int.}}(\hat{\eta})|} \end{bmatrix}^{-1} \begin{pmatrix} Y_{I_{\text{sub.}}(\hat{\eta})} - \frac{1}{2}\tilde{X}_{I_{\text{sub.}}(\hat{\eta}),J_+}\Lambda_{J_+}^{-1}X^\top_{I_{\text{int.}}(\hat{\eta})^c,J_+}\hat{\eta}_{I_{\text{int.}}(\hat{\eta})^c} \\ \tilde{X}^\top_{I_{\text{int.}}(\hat{\eta})^c,\tilde{J}_{\text{sub.}}(\hat{\eta})}\hat{\eta}_{I_{\text{int.}}(\hat{\eta})^c} \\ \mathbf{1}^\top_{|I_{\text{int.}}(\hat{\eta})^c|}\hat{\eta}_{I_{\text{int.}}(\hat{\eta})^c} \end{pmatrix}.$$

Note that by construction, we may assume without loss of generality that $|I_{\text{sub.}}(\hat{\eta})|+|\tilde{J}_{\text{sub.}}(\hat{\eta})| \neq 0$ (otherwise one may simply make different choices for these sets). Now, take fixed sets $I_{\text{sub.}} \subseteq I_{\text{int.}} \subseteq \{1, \ldots, n\}$, and $\tilde{J}_{\text{sub.}} \subseteq \{1, \ldots, d\}$ with $|I_{\text{sub.}}| + |\tilde{J}_{\text{sub.}}| \neq 0$, $I_{\text{int.}} \neq \emptyset$ and any fixed vector $\eta_{I^c_{\text{int}}} \in \{-(1-\tau), \tau\}^{|I^c_{\text{int}}|}$ and consider the behaviour of the random variables

$$\begin{bmatrix} \frac{1}{2}\tilde{X}_{I_{\text{sub.}},J_+}\Lambda_{J_+}^{-1}X^\top_{I_{\text{int.}},J_+} \\ \tilde{X}^\top_{I_{\text{int.}},\tilde{J}_{\text{sub.}}} \\ \mathbf{1}^\top_{|I_{\text{int.}}|} \end{bmatrix}^{-1} \begin{pmatrix} Y_{I_{\text{sub.}}} - \frac{1}{2}\tilde{X}_{I_{\text{sub.}},J_+}\Lambda_{J_+}^{-1}X^\top_{I^c_{\text{int.}},J_+}\eta_{I^c_{\text{int.}}} \\ \tilde{X}^\top_{I^c_{\text{int.}},\tilde{J}_{\text{sub.}}}\eta_{I^c_{\text{int.}}} \\ \mathbf{1}^\top_{|I^c_{\text{int.}}|}\eta_{I^c_{\text{int.}}} \end{pmatrix}. \tag{B.8}$$

The first $|I_{\text{int.}}| - 1$ entries of the vector above are continuously distributed. Write the matrix inverse above in block form as

$$\begin{bmatrix} \frac{1}{2}\tilde{X}_{I_{\text{sub.}},J_+}\Lambda_{J_+}^{-1}X^\top_{I_{\text{int.}},J_+} \\ \tilde{X}^\top_{I_{\text{int.}},\tilde{J}_{\text{sub.}}} \\ \mathbf{1}_{|I_{\text{int.}}|} \end{bmatrix}^{-1} = \begin{bmatrix} A & B \\ \mathbf{1}^\top_{|I_{\text{int.}}|-1} & 1 \end{bmatrix}^{-1}$$

Using the continuity of X the first $|I_{\text{int.}}| - 1$ columns of the matrix inverse above are equal to

$$\begin{bmatrix} (A - B\mathbf{1}_{|I_{\text{int.}}|-1}^\top)^{-1} \\ -\mathbf{1}_{|I_{\text{int.}}|-1}^\top (A - B\mathbf{1}_{|I_{\text{int.}}|-1}^\top)^{-1} \end{bmatrix}.$$

Notably, each row of this matrix cannot be zero. Since A and B are independent of the vector appearing in (B.8) we conclude that with probability one none of the entries of (B.8) are zero. Taking a union bound over the choices of $I_{\text{sub.}}$, $I_{\text{int.}}$, $\tilde{J}_{\text{sub.}}$, and $\eta_{I_{\text{int}}^c}$ gives the desired result.

□

C Additional technical lemmas

In this section, we state a prove a number of additional results that are useful in the main proofs.

Lemma C.1. *Let* $\{X_i\}_{i=1}^n \overset{i.i.d.}{\sim} \mathcal{N}(0, I_d)$. *Then, as* $d/n \to \gamma < \sqrt{2/\pi}$,

$$\liminf_{d,n\to\infty} \inf_{(\|u\|_2 \le 1, |\beta_0| \le 1, \max\{\|u\|_2, |\beta_0|\} = 1)} \frac{1}{n} \sum_{i=1}^n |X_i^\top u + \beta_0| \overset{\mathbb{P}}{\ge} \sqrt{\frac{2}{\pi}} - \gamma.$$

Proof. Let $X \in \mathbb{R}^{n\times d}$ denote the matrix with rows $X_1, \dots, X_n$. Write

$$\begin{aligned} &\inf_{(\|u\|_2 \le 1, |\beta_0| \le 1, \max\{\|u\|_2, |\beta_0|\} = 1)} \frac{1}{n} \sum_{i=1}^n |X_i^\top u + \beta_0| \\ &= \inf_{(\|u\|_2 \le 1, |\beta_0| \le 1, \max\{\|u\|_2, |\beta_0|\} = 1)} \max_{(v \in \{\pm 1\}^n)} \frac{1}{n} v^\top X u + \frac{1}{n} \beta_0 v^\top \mathbf{1}_n. \end{aligned}$$

By the convex Gaussian min-max theorem (Proposition B.2 above), we have that for any $c > 0$,

$$\begin{aligned} &\mathbb{P}\left(\inf_{(\|u\|_2 \le 1, |\beta_0| \le 1, \max\{\|u\|_2, |\beta_0|\} = 1)} \max_{(v \in \{\pm 1\}^n)} \frac{1}{n} v^\top X u + \frac{1}{n} \beta_0 v^\top \mathbf{1}_n \le c \right) \\ &\le 2\mathbb{P}\left(\inf_{(\|u\|_2 \le 1, |\beta_0| \le 1, \max\{\|u\|_2, |\beta_0|\} = 1)} \max_{(v \in \{\pm 1\}^n)} \frac{1}{n} \|v\|_2 u^\top h + \frac{1}{n} \|u\|_2 v^\top g + \frac{1}{n} \beta_0 v^\top \mathbf{1}_n \le c \right), \end{aligned} \tag{C.1}$$

where $h \sim \mathcal{N}(0, I_d)$ and $g \sim \mathcal{N}(0, I_n)$ are independent. Now,

$$\begin{aligned}
&\inf_{(\|u\|_2 \le 1, |\beta_0| \le 1, \max\{\|u\|_2, |\beta_0|\}=1)} \max_{(v \in \{\pm 1\}^n)} \frac{1}{n}\|v\|_2 u^\top h + \frac{1}{n}\|u\|_2 v^\top g + \frac{1}{n}\beta_0 v^\top \mathbf{1}_n \\
&= \inf_{(\|u\|_2 \le 1, |\beta_0| \le 1, \max\{\|u\|_2, |\beta_0|\}=1)} \frac{1}{\sqrt{n}} u^\top h + \frac{1}{n}\sum_{i=1}^n |\|u\|_2 g_i + \beta_0| \\
&= \inf_{(0 \le c_u \le 1, \beta_0 \le 1, \max\{c_u, \beta_0\}=1)} \frac{1}{n}\sum_{i=1}^n |c_u g_i + \beta_0| - c_u \frac{\|h\|_2}{\sqrt{n}} \\
&\xrightarrow{\mathbb{P}} \inf_{(0 \le c_u \le 1, \beta_0 \le 1, \max\{c_u, \beta_0\}=1)} \mathbb{E}[|c_u g_1 + \beta_0|] - c_u \gamma,
\end{aligned}$$

where the limit follows standard uniform concentration inequalities (e.g. Lemma 7.75 of Miescke & Liese (2008) applied to the set $\{(c_u, \beta_0) : 0 \le c_u \le 1, |\beta_0| \le 1\}$).

Finally note that for any c_u, $\beta_0 \mapsto \mathbb{E}[|c_u g_1 + \beta_0|]$ is a convex, even function and thus obtains its minimum at 0. So,

$$\inf_{(c_u = 1, |\beta_0| \le 1)} \mathbb{E}[|c_u g_1 + \beta_0|] - c_u \gamma = \mathbb{E}[|g_1|] - \gamma = \sqrt{\frac{2}{\pi}} - \gamma.$$

On the other hand, by Jensen's inequality

$$\inf_{(0 \le c_u \le 1, |\beta_0| = 1)} \mathbb{E}[|c_u g_1 + \beta_0|] - c_u \gamma = \mathbb{E}[|g_1|] - \gamma \ge \inf_{(0 \le c_u \le 1)} |c_u \mathbb{E}[g_1] + 1]| - c_u \gamma = 1 - \gamma.$$

Combining the above we conclude that

$$\inf_{(\|u\|_2 \le 1, |\beta_0| \le 1, \max\{\|u\|_2, |\beta_0|\}=1)} \max_{(v \in \{\pm 1\}^n)} \frac{1}{n}\|v\|_2 u^\top h + \frac{1}{n}\|u\|_2 v^\top g + \frac{1}{n}\beta_0 v^\top \mathbf{1}_n \xrightarrow{\mathbb{P}} \sqrt{\frac{2}{\pi}} - \gamma,$$

and applying (C.1) gives the desired result. □

Our next lemma gives sufficient conditions under which partial optimization preserves strict convexity.

Lemma C.2. *[Lemma 19 of Thrampoulidis et al. (2018)] Let $\mathcal{A}$ and $\mathcal{B}$ be convex sets and $\Psi : \mathcal{A} \times \mathcal{B} \to \mathbb{R}$ be strictly convex in its first argument. Assume that $\Psi(a, \cdot)$ obtains its maximum for all $a \in \mathcal{A}$. Then, $a \mapsto \max_{b \in \mathcal{B}} \Psi(a, b)$ is strinctly convex.*

Our next result computes the value of the Moreau envelope of the pinball loss.

Lemma C.3. *For any $x \in \mathbb{R}$ and $\rho \geq 0$ the Moreau envelope of the pinball loss is given by*

$$e_{\ell_\tau}(x;\rho) = \begin{cases} \frac{\tau^2\rho}{2} + \tau(x - \rho\tau), \ x - \rho\tau > 0, \\ \frac{x^2}{2\rho}, \ x \in [-\rho(1-\tau), \rho\tau], \\ \frac{(1-\tau)^2\rho}{2} - (1-\tau)(x + \rho(1-\tau)), \ x + \rho(1-\tau) < 0. \end{cases}$$

Proof. The case $\rho = 0$ is given by Lemmma **??**. Now, consider the case $\rho > 0$. We begin by computing the proximal function. Let $f(v) = \frac{1}{2\rho}(v-x)_2^2 + \ell_\rho(v)$ denote the objective appearing in the definition of the Moreau envelope and the proximal function. We have that

$$\partial f(v) = \begin{cases} \{\frac{v-x}{\rho} + \tau\}, \ v > 0, \\ [\frac{v-x}{\rho} - (1-\tau), \frac{v-x}{\rho} + \tau], \ v = 0, \\ \{\frac{v-x}{\rho} - (1-\tau)\}, \ v < 0. \end{cases}$$

Setting this to zero we find that

$$\text{prox}_{\ell_\tau}(x;\rho) \in \begin{cases} x - \rho\tau, \ x - \rho\tau > 0, \\ 0, \ x \in [-\rho(1-\tau), \rho\tau], \\ x + \rho(1-\tau), \ x + \rho(1-\tau) < 0. \end{cases}$$

Plugging this into the definition of the Moreau envelope gives the result. □

The next three lemmas state a number facts from convex analysis that are useful in the proofs above.

Lemma C.4 (Corollary of Propositions 12.28, and 12.30 in Bauschke & Combettes (2017))**.** *Let $f : \mathbb{R} \to \mathbb{R} \cup \{+\infty\}$ be a proper lower semicontinuous convex function. Then, for all*

$\tau > 0$, $(x, \tau) \to e_f(x; \tau)$ is 1-Lipschitz and differentiable with derivative

$$\frac{d}{dx} e_f(x; \tau) = \frac{1}{\tau}(Id - prox(f; \gamma).$$

Moreover, this latter function is 1-Lipschitz.

Lemma C.5 (Proposition 13.13 in Bauschke & Combettes (2017))**.** *Let $f : \mathbb{R} \to \mathbb{R} \cup \{+\infty\}$. Then f^* is lower semicontinuous.*

Lemma C.6 (Part i) of Theorem 14.3 in Bauschke & Combettes (2017))**.** *Let $f : \mathbb{R} \to \mathbb{R} \cup \{+\infty\}$ be a proper lower semicontinuous convex function. Then, for any $x \in \mathbb{R}$ and $\rho > 0$ we have the identity*

$$e_f(x; \rho) + e_{f^*}(x/\rho; 1/\rho) = \frac{x^2}{2\rho}.$$

Lemma C.7. *[Theorem 1.25 in Rockafellar & Wets (1997)] Let $f : \mathbb{R} \to \mathbb{R}$ be convex, lower semicontinuous and prox-bounded. Then for all $x \in \mathbb{R}$,*

$$\lim_{\rho \downarrow 0} e_f(x; \rho) = f(x).$$