Understanding the Under-Coverage Bias in Uncertainty Estimation

Yu Bai [1] **Song Mei** [2] **Huan Wang** [1] **Caiming Xiong** [1]

Abstract

Estimating the data uncertainty in regression tasks is often done by learning a quantile function or a prediction interval of the true label conditioned on the input. It is frequently observed that quantile regression—a vanilla algorithm for learning quantiles with asymptotic guarantees—tends to *under-cover* than the desired coverage level in reality. While various fixes have been proposed, a more fundamental understanding of why this under-coverage bias happens in the first place remains elusive.

In this paper, we present a rigorous theoretical study on the coverage of uncertainty estimation algorithms in learning quantiles. We prove that quantile regression suffers from an inherent under-coverage bias, in a vanilla setting where we learn a realizable linear quantile function and there is more data than parameters. More quantitatively, for $\alpha > 0.5$ and small d/n, the α-quantile learned by quantile regression roughly achieves coverage $\alpha - (\alpha - 1/2) \cdot d/n$ regardless of the noise distribution, where d is the input dimension and n is the number of training data. Our theory reveals that this under-coverage bias stems from a certain high-dimensional parameter estimation error that is not implied by existing theories on quantile regression. Experiments on simulated and real data verify our theory and further illustrate the effect of various factors such as sample size and model capacity on the under-coverage bias in more practical setups.

1. Introduction

This paper is concerned with the problem of uncertainty estimation in regression problems. Uncertainty estimation is an increasingly important task in modern machine learning applications—Models should not only make high-accuracy predictions, but also have a sense of how much the true label may deviate from the prediction. This capability is crucial for deploying machine learning in the real world, in particular in risk-sensitive domains such as medical AI (Begoli et al., 2019; Jiang et al., 2012), self-driving cars (Michelmore et al., 2018), and so on. A common approach for uncertainty estimation in regression is to learn a *quantile function* or a *prediction interval* of the true label conditioned on the input, which provides useful distributional information about the label. Such learned quantiles are typically evaluated by their *coverage*, i.e., probability that it covers the true label on a new test example. For example, a learned 90% upper quantile function should be an actual upper bound of the true label at least 90% of the time.

Algorithms for learning quantiles date back to the classical quantile regression (Koenker & Hallock, 2001), which estimates the quantile function by solving an empirical risk minimization problem with a suitable loss function that depends on the desired quantile level α. Quantile regression is conceptually simple, and is theoretically shown to achieve asymptotically correct coverage as the sample size goes to infinity (Koenker & Bassett Jr, 1978) or approximately correct coverage in finite samples under specific modeling assumptions (Meinshausen, 2006; Takeuchi et al., 2006; Steinwart et al., 2011). However, it is observed that quantile regression often *under-covers* than the desired coverage level in practice (Romano et al., 2019). Various alternative approaches for constructing quantiles and confidence intervals are proposed in more recent work, for example by aggregating multiple predictions using Bayesian neural networks or ensembles (Gal & Ghahramani, 2016; Lakshminarayanan et al., 2016), or by building on the conformal prediction technique to construct prediction intervals with finite-sample coverage guarantees (Vovk et al., 2005; Vovk, 2012; Lei et al., 2018; Romano et al., 2019). However, despite these advances, a more fundamental understanding on why vanilla quantile regression exhibits this under-coverage bias is still lacking.

This paper revisits quantile regression and presents a first precise theoretical study on its coverage, in a new regime where the number of samples n is proportional to the dimension d, and the ratio d/n is small (so that the problem is under-parametrized). Our main result shows that quantile regression exhibits an inherent under-cover bias under this

[1]Salesforce Research [2]University of California, Berkeley. Correspondence to: Yu Bai <yu.bai@salesforce.com>, Song Mei <songmei@berkeley.edu>.

Presented at the ICML 2021 Workshop on Uncertainty and Robustness in Deep Learning., Copyright 2021 by the author(s).

regime, even in the well-specified setting of learning a linear quantile function when the true data distribution follows a Gaussian linear model. To the best of our knowledge, this is the first rigorous theoretical justification of the under-coverage bias. Our main contributions are summarized as follows.

- *We prove that linear quantile regression exhibits an inherent under-coverage bias* in the well-specified setting where the data is generated from a Gaussian linear model, and the number of samples n is proportional to the feature dimension d with a small d/n (Section 3). More quantitatively, quantile regression at nominal level $\alpha \in (0.5, 1)$ roughly achieves coverage $\alpha - (\alpha - 1/2)d/n$ regardless of the noise distribution. To the best of our knowledge, this is the first rigorous characterization of the under-coverage bias in quantile regression.
- Towards understanding the source of this under-coverage bias, we disentangle the effect of estimating the bias and estimating the linear coefficient on the coverage of the learned linear quantile (Section 4). We show that the estimation error in the bias can have either an under-coverage or over-coverage effect, depending on the noise distribution. In contrast, the estimation error in the linear coefficient always drives the learned quantile to under-cover, and we show this effect is present even on broader classes of data distributions beyond the Gaussian linear model.
- We perform experiments on simulated and real data to test our theory (Appendix B). Our simulations show that the coverage of quantile regression in Gaussian linear models agrees well with our precise theoretical formula as well as the $\alpha-(\alpha-1/2)d/n$ approximation. On real data, we find quantile regression using high-capacity models (such as neural networks) exhibits severe under-coverage biases, while linear quantile regression can also have a mild but non-negligible amount of under-coverage, even after we remove the potential effect of model misspecification.
- On the technical end, our analysis builds on recent understandings of empirical risk minimization problems in the high-dimensional proportional limit with a small d/n, and develops new techniques such as a novel concentration argument to deal with an additional learnable variable in learning linear models with biases, which we believe could be of further interest (Appendix C).

2. Preliminaries

In this paper we focus on the problem of learning quantiles. Suppose we observe a training dataset $\{(\mathbf{x}_i, y_i)\}_{i=1}^n$ drawn i.i.d. from some joint distribution $\mathbb{P}$ on $\mathbb{R}^d \times \mathbb{R}$, where $\mathbf{x}_i \in \mathbb{R}^d$ is the input features and $y \in \mathbb{R}$ is the real-valued response (label). Let $F(t|\mathbf{x}) := \mathbb{P}(Y \le t|\mathbf{X} = \mathbf{x})$ denote the conditional CDF of $Y|\mathbf{X}$. Our goal is to learn the α-(conditional) quantile of $Y|\mathbf{X}$:

$$q^\star_\alpha(\mathbf{x}) := \inf\{t \in \mathbb{R} : F(t|\mathbf{x}) \ge \alpha\}.$$

For example, $q^\star_{0.95}(\mathbf{x})$ is the ground truth 95% quantile of the true conditional distribution $Y|\mathbf{X}$, and can be seen as the "ideal" 95% upper confidence bound for the label y given the features $\mathbf{x}$. Throughout this paper we work with upper quantiles, that is, $\alpha \in (0.5, 1)$ (some typical choices are $\alpha \in \{0.8, 0.9, 0.95\}$); by symmetry our results hold for learning lower quantiles as well.

Coverage For any learned quantile function $\widehat{f} : \mathbb{R}^d \to \mathbb{R}$, the marginal coverage (henceforth "coverage") of $\widehat{f}$ is the probability of $y \le \widehat{f}(\mathbf{x})$ on a new test example $(\mathbf{x}, y)$:

$$\text{Coverage}(\widehat{f}) := \mathbb{P}_{(\mathbf{x},y)}\Big(y \le \widehat{f}(\mathbf{x})\Big) = \mathbb{E}_{\mathbf{x}}\Big[\mathbb{P}\Big(y \le \widehat{f}(\mathbf{x})|\mathbf{x}\Big)\Big]. \tag{1}$$

For learning the α-quantile ($\alpha > 0.5$), we usually expect $\text{Coverage}(\widehat{f}) \approx \alpha$, i.e. $\widehat{f}(\mathbf{x})$ covers the label y on approximately α proportion of the data, under the ground truth data distribution.

We say that $\widehat{f}$ has *under-coverage* if $\text{Coverage}(\widehat{f}) < \alpha$ and *over-coverage* if $\text{Coverage}(\widehat{f}) > \alpha$. Note that these two notions are not symmetric: Over-coverage means that the learned upper quantile $\widehat{f}(\mathbf{x})$ is overly conservative (higher than enough), and is typically tolerable; In contrast, under-coverage means that $\widehat{f}(\mathbf{x})$ fails to cover y with α probability, and is typically considered as a failure.

Quantile regression We consider quantile regression, a standard method for learning quantiles from data (Koenker & Hallock, 2001). Quantile regression estimates the true quantile function $q_\alpha(\cdot)$ via the *pinball loss* (Koenker & Bassett Jr, 1978; Steinwart et al., 2011)

$$\ell^\alpha(t) = -(1-\alpha)t\mathbf{1}\{t \le 0\} + \alpha t\mathbf{1}\{t > 0\}. \tag{2}$$

Note that in the special case of $\alpha = 0.5$, we have $\ell^{0.5}(t) = |t|/2$, and thus the pinball loss strictly generalizes the absolute loss (for learning medians) to learning any quantile. Given the training dataset and any function class $\{f_\theta : \theta \in \Theta\}$ (e.g. linear models or neural networks), quantile regression solves the (unregularized) empirical risk minimization (ERM) problem

$$\widehat{\theta} = \arg\min_{\theta\in\Theta} \widehat{R}_n(\theta) := \frac{1}{n}\sum_{i=1}^n \ell^\alpha(y_i - f_\theta(\mathbf{x}_i)). \tag{3}$$

(We take $\widehat{\theta}$ as any minimizer of $\widehat{R}_n$ when the minimizer is non-unique.) Let $R(\theta) := \mathbb{E}[\widehat{R}_n(\theta)]$ denote the corresponding population risk. It is known that the population risk over

all (measurable) functions is minimized at the true quantile $q^\star_\alpha = \arg\min_f R(f)$ under minimal regularity conditions (for completeness, we provide a proof in Appendix E.1).

3. Quantile regression exhibits under-coverage

We analyze quantile regression in the vanilla setting where the input distribution is a standard Gaussian and y follows a linear model of $\mathbf{x}$:

$$y = \mathbf{w}_\star^\top \mathbf{x} + z, \quad \text{where} \quad \mathbf{x} \sim \mathsf{N}(\mathbf{0}, \mathbf{I}_d), \;\; z \sim P_z. \tag{4}$$

Above, $\mathbf{w}_\star \in \mathbb{R}^d$ is the ground truth coefficient vector, and the noise $z \sim P_z$ is independent of $\mathbf{x}$. The noise distribution P_z is required to satisfy the following smoothness assumption, but can otherwise be arbitrary:

Assumption A (Smooth density)**.** *The noise distribution P_z has a smooth density $\phi_z \in C^\infty(\mathbb{R})$ (with corresponding CDF Φ_z), with bounded derivatives: $\sup_{t\in\mathbb{R}} |\phi_z^{(k)}(t)| < \infty$ for any $k \geq 0$. We further assume that $\phi_z(z_\alpha) > 0$, where $z_\alpha := \inf\{t \in \mathbb{R} : \Phi_z(t) \geq \alpha\}$ is the α-quantile of P_z.*

Under the above model, it is straightforward to see that the true α-conditional quantile of $y|\mathbf{x}$ is also a linear model (with bias):

$$q^\star_\alpha(\mathbf{x}) = \mathbf{w}_\star^\top \mathbf{x} + z_\alpha. \tag{5}$$

Given the training data $\{(\mathbf{x}_i, y_i)\}_{i=1}^n$, we learn a linear quantile function $\widehat{f}(\mathbf{x}) = \widehat{\mathbf{w}}^\top \mathbf{x} + \widehat{b}$ via quantile regression:

$$(\widehat{\mathbf{w}}, \widehat{b}) = \arg\min_{\mathbf{w},b} \widehat{R}_n(\mathbf{w}, b) := \frac{1}{n}\sum_{i=1}^n \ell^\alpha(y_i - (\mathbf{w}^\top \mathbf{x}_i + b)), \tag{6}$$

where ℓ^α is the pinball loss in (2). As our linear function class realizes the true quantile function (5), the population risk is minimized at the true quantile: $\arg\min_{\mathbf{w},b} R(\mathbf{w}, b) = (\mathbf{w}_\star, z_\alpha)$.

We are now ready to state our main result, which shows that quantile regression exhibits an inherent under-coverage bias even in this vanilla realizable setting.

Theorem 1 (Quantile regression exhibits under-coverage bias)**.** *Suppose the data is generated from the linear model (4) and the noise satisfies Assumption A. Let $\widehat{f}(\mathbf{x}) = \widehat{\mathbf{w}}^\top \mathbf{x} + \widehat{b}$ be the output of quantile regression (6) at level $\alpha \in (0.5, 1)$. Then, in the limit of $n, d \to \infty$ and $d/n \to \kappa$ where $\kappa \in (0, \kappa_0]$ for some small $\kappa_0 > 0$, for the coverage (1), we have ($\xrightarrow{p}$ denotes convergence in probability)*

$$\text{Coverage}(\widehat{f}) \xrightarrow{p} \alpha - C_{\alpha,\kappa} \quad \textit{for some } C_{\alpha,\kappa} > 0.$$

That is, the limiting coverage of the learned quantile function is less than α. Further, for small enough κ we have the local linear expansion

$$C_{\alpha,\kappa} = (\alpha - 1/2)\kappa + o(\kappa). \tag{7}$$

Theorem 1 builds on the precise characterization of ERM problems in the high-dimensional proportional limit (Thrampoulidis et al., 2018), along with new techniques over existing work for dealing with the unique challenges in quantile regression (such as analyzing the additional learnable bias b). An overview of the main technical steps is provided in Section C, and the full proof is deferred to Appendix F.

Theorem 1 can be illustrated by the following numeric example. Suppose we perform quantile regression at $\alpha = 0.9$, where the data follows the linear model (4), and our $\kappa = d/n = 0.1$ (so that the sample size is 10x number of parameters). Then Theorem 1 shows that, even in this realizable, under-parametrized setting, the coverage of the learned quantile $\widehat{f}$ is going to be roughly $0.9 - C_{\alpha,\kappa}$ when n, d are large, and further $C_{\alpha,\kappa} \approx (\alpha - 1/2)\kappa = 0.04$. Thus the actual coverage is around $0.9 - 0.04 = 0.86$, and such a 4% under-coverage bias can be rather non-negligible in reality. To the best of our knowledge, this offers a first precise theoretical understanding of why practically trained quantiles or prediction intervals often under-cover than the desired coverage level (Romano et al., 2019).

An important feature of the under-coverage bias shown in Theorem 1 is that it only shows up in the n, d proportional regime, and is not implied by existing theories on quantile regression—Classical asymptotic theory only shows asymptotic normality $\sqrt{n}([\widehat{\mathbf{w}}, \widehat{b}] - [\mathbf{w}_\star, z_\alpha]) \to \mathsf{N}(\mathbf{0}, \mathbf{V})$ in the $n \to \infty$, fixed d limit, (Koenker & Bassett Jr, 1978; Van der Vaart, 2000) in which case the coverage is approximately unbiased at $\alpha \pm O(1/\sqrt{n})$. Christmann & Steinwart (2007); Steinwart et al. (2011) consider the finite n, d setting and establish *self-calibration inequalities* which can be turned into a bound on $|\text{Coverage}(\widehat{f}) - \alpha|$, but does not tell the sign (positive or negative) of the coverage bias.

Extension to over-parametrized learning We additionally prove that the under-coverage bias becomes even more severe in over-parametrized learning, under the same linear model (4): When $d > \widetilde{\Omega}(n)$ and the noise P_z is sub-Gaussian and symmetrically distributed, the convergence point of the gradient descent path on the quantile regression risk $\widehat{R}_n$ is the minimum-norm interpolator of the data, which has coverage $0.5 \pm \widetilde{O}(1/\sqrt{n})$ with high probability (see Appendix G for the formal statement and the proof).

4. Understanding the source of the under-coverage bias

In this section, we take steps towards a deeper understanding of how the under-coverage bias shown in Theorem 1 happens. Recall that the quantile regression returns $\widehat{f}(\mathbf{x}) = \widehat{\mathbf{w}}^\top \mathbf{x} + \widehat{b}$ where $(\widehat{\mathbf{w}}, \widehat{b})$ is a solution to the ERM problem (6) and estimates the true parameters $(\mathbf{w}_\star, z_\alpha)$. Our main approach in this section is to *disentangle the effect of the two sources*—the estimation error in $\widehat{b}$ and the estimation

error in $\mathbf{w}$—on the coverage of $\widehat{f}$.

4.1. Effect of estimation error in $\widehat{b}$

To study the effect of $\widehat{b}$, we use the quantity $\widehat{b} - z_\alpha$ as a measure for its effect on the coverage—Recall that the true quantile is $q^\star_\alpha(\mathbf{x}) = \mathbf{w}_\star^\top + z_\alpha$, thus having $\widehat{b} < z_\alpha$ means that $\widehat{b}$ contributes to under-coverage, whereas $\widehat{b} > z_\alpha$ means $\widehat{b}$ contributes to over-coverage. (This can be seen more straightforwardly in the easier case where we know $\mathbf{w}_\star$ and only output $\widehat{b}$ to estimate z_α.)

The following corollary shows that, under the same settings of Theorem 1, the error $\widehat{b} - z_\alpha$ can be understood precisely. The proof can be found in Appendix H.1.

Corollary 2 (Effect of $\widehat{b}$ on coverage depends on noise distribution)**.** *Under the same settings as Theorem 1, for any $\alpha \in (0.5, 1)$, as $n, d \to \infty$ with $d/n \to \kappa \in (0, \kappa_0]$, we have*

(a) The learned bias $\widehat{b}$ from quantile regression (6) converges to the following limit:

$$\widehat{b} - z_\alpha \xrightarrow{p} C^b_{\alpha,\kappa} = \bar{b}_0 \kappa + o(\kappa),$$

where $\bar{b}_0$ has a closed-form expression:

$$\bar{b}_0 := \frac{-\alpha(1-\alpha)\phi_z'(z_\alpha) - (2\alpha-1)\phi_z^2(z_\alpha)}{2\phi_z^3(z_\alpha)}. \quad (8)$$

(b) We have $\bar{b}_0 < 0$ for several common noise distributions such as Gaussian (with arbitrary scale), in which case $C^b_{\alpha,\kappa} < 0$ for small enough κ.

(c) Conversely, for any $\alpha \in (0.5, 1)$, there exists some noise distribution P_z for which $\bar{b}_0 > 0$, in which case $C^b_{\alpha,\kappa} > 0$ for small enough κ.

Corollary 2 shows that the sign of $C^b_{\alpha,\kappa}$ in the limiting regime (and thus the effect of $\widehat{b}$ on the coverage) depends on $\bar{b}_0$, which in turn depends on the noise distribution P_z. For many common noise distributions we have $C^b_{\alpha,\kappa} < 0$ at small κ, but there exists P_z such that $C^b_{\alpha,\kappa} > 0$.

4.2. Effect of estimation error in $\widehat{\mathbf{w}}$; relaxed data distributions

We now show that the primary source of the under-coverage is the estimation error in $\widehat{\mathbf{w}}$, which happens not only on the linear data distribution assumed in Theorem 1, but also on a broader class of data distributions. We consider the following relaxed data distribution assumption

$$y = \mu_\star(\mathbf{x}) + \sigma_\star(\mathbf{x}) z, \quad (9)$$

where the noise $z \sim P_z$. We do not put structural assumptions on $(\mu_\star, \sigma_\star)$, except that we assume the true α-quantile is still a linear function of $\mathbf{x}$, that is, there exists $(\mathbf{w}_\star, b_\star)$ for which

$$q^\star_\alpha(\mathbf{x}) = \mu_\star(\mathbf{x}) + \sigma_\star(\mathbf{x}) z_\alpha = \mathbf{w}_\star^\top \mathbf{x} + b_\star. \quad (10)$$

Since here we are interested in the effect of estimating $\mathbf{w}_\star$, for simplicity, we assume that we know $b_\star$ and only estimate $\mathbf{w}_\star$ via some estimator $\widehat{\mathbf{w}}$. We now collect our assumptions and state the result.

Assumption B (Relaxed data distribution)**.** *The data is distributed as model (9) with a linear α-quantile function (10). Further, the data distribution satisfies the following regularity conditions:*

(a) The distribution of $\mathbf{x} \in \mathbb{R}^d$ is symmetric about $\mathbf{0}$, has a lower bounded covariance $\mathbb{E}[\mathbf{x}\mathbf{x}^\top] \succeq \underline{\gamma}\mathbf{I}_d$, and is K-sub-Gaussian, for constants $\underline{\gamma}, K > 0$.

(b) The variance function $\sigma_\star(\cdot)$ is bounded and symmetric: For all $\mathbf{x} \in \mathbb{R}^d$ we have $\underline{\sigma} \le \sigma_\star(\mathbf{x}) \le \overline{\sigma}$ for some constants $\underline{\sigma}, \overline{\sigma} > 0$, and $\sigma_\star(\mathbf{x}) = \sigma_\star(-\mathbf{x})$.

(c) The noise density ϕ_z is continuously differentiable and symmetric about 0, i.e. $\phi_z(t) = \phi_z(-t)$ for all $t \in \mathbb{R}$. Further, ϕ_z is uni-modal, i.e. $\phi_z'(t)|_{t<0} > 0$ and $\phi_z'(t)|_{t>0} < 0$.

Theorem 3 (Estimation error in $\widehat{\mathbf{w}}$ leads to under-coverage on a family of data distributions)**.** *Under the relaxed data distribution assumption (Assumption B), for any $\alpha > 3/4$, there exists constants $c, r_0 > 0$ such that for any learned quantile estimate $\widehat{f}(\mathbf{x}) = \widehat{\mathbf{w}}^\top \mathbf{x} + b_\star$ with small estimation error $\|\widehat{\mathbf{w}} - \mathbf{w}_\star\|_2 \le r_0$, we have*

$$\mathrm{Coverage}(\widehat{f}) \le \alpha - c\underline{\gamma}/\overline{\sigma}^2 \cdot \|\widehat{\mathbf{w}} - \mathbf{w}_\star\|_2^2,$$

that is, the learned quantile under-covers by at least $\Omega(\|\widehat{\mathbf{w}} - \mathbf{w}_\star\|_2^2)$. Above, $c > 0$ is an absolute constant, and $r_0 > 0$ depends on $(\underline{\gamma}, \underline{\sigma}, K, \Phi_z, \alpha)$ but not (n, d).

Implications Theorem 3 shows that, for a broad class of data distributions, any estimator $\widehat{\mathbf{w}}$ will under-cover by at least $\Omega(\|\widehat{\mathbf{w}} - \mathbf{w}_\star\|_2^2)$. If particular, any estimator satisfying $\|\widehat{\mathbf{w}} - \mathbf{w}_\star\|_2 \asymp \widetilde{O}(\sqrt{d/n})$ (e.g. from standard generalization theory) will under-cover by $\widetilde{O}(d/n)$. This confirms that the estimation error in the (bulk) regression coefficient $\widehat{\mathbf{w}}$ is the primary source of the under-coverage bias, under assumptions that are more general than Theorem 1 in certain aspects (such as the distribution of $\mathbf{x}$ and $y|\mathbf{x}$). We remark that as opposed to Theorem 1, Therorem 3 does not give an end-to-end characterization of any specific algorithm, but assumes we have an estimator $\widehat{\mathbf{w}}$ with a small error.

5. Additional materials

Due to the space limit, we present our experimental results (both simulations and real data) in Appendix B. A proof overview of Theorem 1 can be found in Appendix C. Additional related work is reviewed in Appendix A.

References

Bike sharing data set. https://archive.ics.uci.edu/ml/datasets/bike+sharing+dataset, Accessed: May, 2021.

Communities and crime data set. http://archive.ics.uci.edu/ml/datasets/communities+and+crime, Accessed: May, 2021.

Medical expenditure panel survey, panel 19, Accessed: May, 2021a. URL https://meps.ahrq.gov/mepsweb/data_stats/download_data_files_detail.jsp?cboPufNumber=HC-181.

Medical expenditure panel survey, panel 20. https://meps.ahrq.gov/mepsweb/data_stats/download_data_files_detail.jsp?cboPufNumber=HC-181, Accessed: May, 2021b.

Medical expenditure panel survey, panel 21. https://meps.ahrq.gov/mepsweb/data_stats/download_data_files_detail.jsp?cboPufNumber=HC-192, Accessed: May, 2021c.

Achilles, C., Bain, H. P., Bellott, F., Boyd-Zaharias, J., Finn, J., Folger, J., Johnston, J., and Word, E. Tennessee's student teacher achievement ratio (star) project. *Harvard Dataverse*, 1:2008, 2008.

Angelopoulos, A., Bates, S., Malik, J., and Jordan, M. I. Uncertainty sets for image classifiers using conformal prediction. *arXiv preprint arXiv:2009.14193*, 2020.

Bai, Y., Mei, S., Wang, H., and Xiong, C. Don't just blame over-parametrization for over-confidence: Theoretical analysis of calibration in binary classification. *arXiv preprint arXiv:2102.07856*, 2021.

Barber, R. F., Candes, E. J., Ramdas, A., and Tibshirani, R. J. Conformal prediction under covariate shift. *arXiv preprint arXiv:1904.06019*, 2019a.

Barber, R. F., Candes, E. J., Ramdas, A., and Tibshirani, R. J. The limits of distribution-free conditional predictive inference. *arXiv preprint arXiv:1903.04684*, 2019b.

Barber, R. F., Candes, E. J., Ramdas, A., and Tibshirani, R. J. Predictive inference with the jackknife+. *The Annals of Statistics*, 49(1):486–507, 2021.

Bates, S., Angelopoulos, A., Lei, L., Malik, J., and Jordan, M. I. Distribution-free, risk-controlling prediction sets. *arXiv preprint arXiv:2101.02703*, 2021.

Bayati, M. and Montanari, A. The dynamics of message passing on dense graphs, with applications to compressed sensing. *IEEE Transactions on Information Theory*, 57 (2):764–785, 2011.

Begoli, E., Bhattacharya, T., and Kusnezov, D. The need for uncertainty quantification in machine-assisted medical decision making. *Nature Machine Intelligence*, 1(1):20–23, 2019.

Candès, E. J., Sur, P., et al. The phase transition for the existence of the maximum likelihood estimate in high-dimensional logistic regression. *The Annals of Statistics*, 48(1):27–42, 2020.

Cauchois, M., Gupta, S., Ali, A., and Duchi, J. C. Robust validation: Confident predictions even when distributions shift. *arXiv preprint arXiv:2008.04267*, 2020.

Cauchois, M., Gupta, S., and Duchi, J. C. Knowing what you know: valid and validated confidence sets in multiclass and multilabel prediction. *Journal of Machine Learning Research*, 22(81):1–42, 2021.

Christmann, A. and Steinwart, I. How svms can estimate quantiles and the median. In *Advances in neural information processing systems*, pp. 305–312, 2007.

Donoho, D. and Montanari, A. High dimensional robust m-estimation: Asymptotic variance via approximate message passing. *Probability Theory and Related Fields*, 166 (3):935–969, 2016.

Donoho, D. L., Maleki, A., and Montanari, A. Message-passing algorithms for compressed sensing. *Proceedings of the National Academy of Sciences*, 106(45):18914–18919, 2009.

Dua, D. and Graff, C. UCI machine learning repository, 2017. URL http://archive.ics.uci.edu/ml.

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

Gal, Y. and Ghahramani, Z. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, pp. 1050–1059. PMLR, 2016.

Geisser, S. The predictive sample reuse method with applications. *Journal of the American statistical Association*, 70(350):320–328, 1975.

Guo, C., Pleiss, G., Sun, Y., and Weinberger, K. Q. On calibration of modern neural networks. *arXiv preprint arXiv:1706.04599*, 2017.

Gupta, C., Podkopaev, A., and Ramdas, A. Distribution-free binary classification: prediction sets, confidence intervals and calibration. *arXiv preprint arXiv:2006.10564*, 2020.

Huang, G., Li, Y., Pleiss, G., Liu, Z., Hopcroft, J. E., and Weinberger, K. Q. Snapshot ensembles: Train 1, get m for free. *arXiv preprint arXiv:1704.00109*, 2017.

Jiang, X., Osl, M., Kim, J., and Ohno-Machado, L. Calibrating predictive model estimates to support personalized medicine. *Journal of the American Medical Informatics Association*, 19(2):263–274, 2012.

Jung, C., Lee, C., Pai, M. M., Roth, A., and Vohra, R. Moment multicalibration for uncertainty estimation. *arXiv preprint arXiv:2008.08037*, 2020.

Karoui, N. E. Asymptotic behavior of unregularized and ridge-regularized high-dimensional robust regression estimators: rigorous results. *arXiv preprint arXiv:1311.2445*, 2013.

Kendall, A. and Gal, Y. What uncertainties do we need in bayesian deep learning for computer vision? *arXiv preprint arXiv:1703.04977*, 2017.

Kivaranovic, D., Johnson, K. D., and Leeb, H. Adaptive, distribution-free prediction intervals for deep networks. In *International Conference on Artificial Intelligence and Statistics*, pp. 4346–4356. PMLR, 2020.

Koenker, R. and Bassett Jr, G. Regression quantiles. *Econometrica: journal of the Econometric Society*, pp. 33–50, 1978.

Koenker, R. and Hallock, K. F. Quantile regression. *Journal of economic perspectives*, 15(4):143–156, 2001.

Kumar, A., Liang, P., and Ma, T. Verified uncertainty calibration. *arXiv preprint arXiv:1909.10155*, 2019.

Lakshminarayanan, B., Pritzel, A., and Blundell, C. Simple and scalable predictive uncertainty estimation using deep ensembles. *arXiv preprint arXiv:1612.01474*, 2016.

Lei, J. Classification with confidence. *Biometrika*, 101(4): 755–769, 2014.

Lei, J., G'Sell, M., Rinaldo, A., Tibshirani, R. J., and Wasserman, L. Distribution-free predictive inference for regression. *Journal of the American Statistical Association*, 113(523):1094–1111, 2018.

Liu, L. T., Simchowitz, M., and Hardt, M. The implicit fairness criterion of unconstrained learning. In *International Conference on Machine Learning*, pp. 4051–4060. PMLR, 2019.

Mackay, D. J. C. *Bayesian methods for adaptive models*. PhD thesis, California Institute of Technology, 1992.

Maddox, W. J., Izmailov, P., Garipov, T., Vetrov, D. P., and Wilson, A. G. A simple baseline for bayesian uncertainty in deep learning. *Advances in Neural Information Processing Systems*, 32:13153–13164, 2019.

Mai, X., Liao, Z., and Couillet, R. A large scale analysis of logistic regression: Asymptotic performance and new insights. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 3357–3361. IEEE, 2019.

Malinin, A. and Gales, M. Predictive uncertainty estimation via prior networks. *arXiv preprint arXiv:1802.10501*, 2018.

Malinin, A., Mlodozeniec, B., and Gales, M. Ensemble distribution distillation. *arXiv preprint arXiv:1905.00076*, 2019.

Meinshausen, N. Quantile regression forests. *Journal of Machine Learning Research*, 7(35):983–999, 2006.

Michelmore, R., Kwiatkowska, M., and Gal, Y. Evaluating uncertainty quantification in end-to-end autonomous driving control. *arXiv preprint arXiv:1811.06817*, 2018.

Orabona, F. Last iterate of sgd converges (even in unbounded domains), 2020; Accessed: May, 2021. URL https://parameterfree.com/2020/08/07/last-iterate-of-sgd-converges-even-in-unbounded-d

Ovadia, Y., Fertig, E., Ren, J., Nado, Z., Sculley, D., Nowozin, S., Dillon, J. V., Lakshminarayanan, B., and Snoek, J. Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift. *arXiv preprint arXiv:1906.02530*, 2019.

Papadopoulos, H. Inductive conformal prediction: Theory and application to neural networks. In *Tools in artificial intelligence*. Citeseer, 2008.

Platt, J. et al. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. 1999.

Quenouille, M. H. Approximate tests of correlation in time-series. *Journal of the Royal Statistical Society: Series B (Methodological)*, 11(1):68–84, 1949.

Romano, Y., Patterson, E., and Candès, E. J. Conformalized quantile regression. *arXiv preprint arXiv:1905.03222*, 2019.

Shabat, E., Cohen, L., and Mansour, Y. Sample complexity of uniform convergence for multicalibration. *arXiv preprint arXiv:2005.01757*, 2020.

Shafer, G. and Vovk, V. A tutorial on conformal prediction. *Journal of Machine Learning Research*, 9(3), 2008.

Steinwart, I., Christmann, A., et al. Estimating conditional quantiles with the help of the pinball loss. *Bernoulli*, 17 (1):211–225, 2011.

Stojnic, M. A framework to characterize performance of lasso algorithms. *arXiv preprint arXiv:1303.7291*, 2013.

Stone, M. Cross-validatory choice and assessment of statistical predictions. *Journal of the Royal Statistical Society: Series B (Methodological)*, 36(2):111–133, 1974.

Sur, P. and Candès, E. J. A modern maximum-likelihood theory for high-dimensional logistic regression. *Proceedings of the National Academy of Sciences*, 116(29):14516–14525, 2019.

Takeuchi, I., Le, Q., Sears, T., Smola, A., et al. Nonparametric quantile estimation. 2006.

Thrampoulidis, C., Oymak, S., and Hassibi, B. Regularized linear regression: A precise analysis of the estimation error. In *Conference on Learning Theory*, pp. 1683–1709. PMLR, 2015.

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

Tukey, J. Bias and confidence in not quite large samples. *Ann. Math. Statist.*, 29:614, 1958.

Van der Vaart, A. W. *Asymptotic statistics*, volume 3. Cambridge university press, 2000.

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

Vovk, V. Conditional validity of inductive conformal predictors. In *Asian conference on machine learning*, pp. 475–490. PMLR, 2012.

Vovk, V. Cross-conformal predictors. *Annals of Mathematics and Artificial Intelligence*, 74(1):9–28, 2015.

Vovk, V., Gammerman, A., and Shafer, G. *Algorithmic learning in a random world*. Springer Science & Business Media, 2005.

Vovk, V., Nouretdinov, I., Manokhin, V., and Gammerman, A. Cross-conformal predictive distributions. In *Conformal and Probabilistic Prediction and Applications*, pp. 37–51. PMLR, 2018.

Wilks, S. S. Determination of sample sizes for setting tolerance limits. *The Annals of Mathematical Statistics*, 12(1):91–96, 1941.

Wilks, S. S. Statistical prediction with special reference to the problem of tolerance limits. *The annals of mathematical statistics*, 13(4):400–409, 1942.

Zadrozny, B. and Elkan, C. Obtaining calibrated probability estimates from decision trees and naive bayesian classifiers. Citeseer.

Zadrozny, B. and Elkan, C. Transforming classifier scores into accurate multiclass probability estimates. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 694–699, 2002.

A. Additional related work

Algorithms for uncertainty estimation in regression The earliest methods for uncertainty estimation in regression adopted subsampling methods (bootstrap) or leave-one-out methods (Jackknife) for assessing or calibrating prediction uncertainty (Quenouille, 1949; Tukey, 1958; Stone, 1974; Geisser, 1975). More recently, a growing line of work builds on the idea of conformal prediction (Shafer & Vovk, 2008) to design uncertainty estimation algorithms for regression. These algorithms provide confidence bounds or prediction intervals by post-processing any predictor, and can achieve distribution-free finite-sample marginal coverage guarantees utilizing exchangeability of the data (Papadopoulos, 2008; Vovk, 2012; Lei et al., 2018; Romano et al., 2019; Kivaranovic et al., 2020; Vovk, 2015; Vovk et al., 2018; 2005; Barber et al., 2021). Further modifications of the conformal prediction technique can yield stronger guarantees such as group coverage (Barber et al., 2019b) or coverage under distribution shift (Barber et al., 2019a) under additional assumptions. Our under-coverage results advocate the necessity of such post-processing techniques, and are complementary in the sense that we provide understandings on the more vanilla quantile regression algorithm. Quantiles and prediction intervals can also be obtained by aggregating multiple predictors, such as using Bayesian neural networks (Mackay, 1992; Gal & Ghahramani, 2016; Kendall & Gal, 2017; Malinin & Gales, 2018; Maddox et al., 2019) or ensembles (Lakshminarayanan et al., 2016; Ovadia et al., 2019; Huang et al., 2017; Malinin et al., 2019). These methods offer an alternative approach for uncertainty estimation, but do not typically come with coverage guarantees.

Theoretical analysis of quantile regression Linear quantile regression with the pinball loss dates back to the late 1970s (Koenker & Bassett Jr, 1978). The same work proved the asymptotic normality of the regression coefficients in the $n \to \infty$, fixed d limit. Takeuchi et al. (2006) studied non-parametric quantile regression using kernel methods, and provided generalization bounds (with the pinball loss) based on the Rademacher complexity. Meinshausen (2006) studied non-parametric quantile regression using random forest and showed its consistency under proper assumptions. Christmann & Steinwart (2007); Steinwart et al. (2011) established a "self-calibration" inequality for the quantile loss, which, when combined with standard generalization bounds, can be translated to an estimation error bound for quantile regression. These works all focus on bounding the parameter or function estimation error, which can be translated to bounds on the coverage bias, but does not tell the sign of this coverage bias as we do in this paper. We also remark that conformalization can be used in conjunction with quantile regression to correct is coverage bias (Romano et al., 2019).

Uncertainty quantification for classification For classification problems, two main types of uncertainty quantification methods have been considered: outputting discrete prediction sets with guarantees of covering the true (discrete) label (Wilks, 1941; 1942; Lei, 2014; Angelopoulos et al., 2020; Bates et al., 2021; Cauchois et al., 2021; 2020), or calibrating the predicted probabilities (Platt et al., 1999; Zadrozny & Elkan; 2002; Lakshminarayanan et al., 2016; Guo et al., 2017). The connection between prediction sets and calibration was discussed in (Gupta et al., 2020). The sample complexity of calibration has been studied in a number of theoretical works (Kumar et al., 2019; Gupta et al., 2020; Shabat et al., 2020; Jung et al., 2020; Liu et al., 2019; Bai et al., 2021). Our work is inspired by the recent work of Bai et al. (2021), which showed that logistic regression is over-confident even if the model is correctly specified and the sample size is larger than the dimension.

High-dimensional behaviors of empirical risk minimization There is a rapidly growing literature on limiting characterizations of convex optimization-based estimators in the $n \propto d$ regime (Donoho et al., 2009; Bayati & Montanari, 2011; El Karoui et al., 2013; Karoui, 2013; Stojnic, 2013; Thrampoulidis et al., 2015; Donoho & Montanari, 2016; Thrampoulidis et al., 2018; Mai et al., 2019; Sur & Candès, 2019; Candès et al., 2020). Our analysis builds on results for unregularized M-estimator derived in (Thrampoulidis et al., 2018) and generalizes theirs in certain aspects (see also (El Karoui et al., 2013; Donoho & Montanari, 2016; Karoui, 2013)).

B. Experiments

B.1. Simulations

Setup We first test our Theorem 1 via simulations. We generate data from the linear model (4) in $d = 100$ dimensions with $\|\mathbf{w}_\star\|_2 = 1$ and noise distribution $P_z = \mathsf{N}(0, 0.25)$. We vary $\kappa = d/n \in \{0.02, 0.04, \ldots, 0.5\}$ where κ determines a sample size n, and vary $\alpha \in \{0.5, 0.52, ..., 0.98\}$.

For each combination of (α, κ), we generate 8 random problem instances, and solve the quantile regression ERM problem (6) on each instance via (sub)-gradient descent. We evaluate the coverage of the learned quantile $\widehat{f}$ (thanks to the linear model (4), the coverage can be computed exactly without needing to introduce a test set). Additional details about the setup can be found in Appendix I.1.

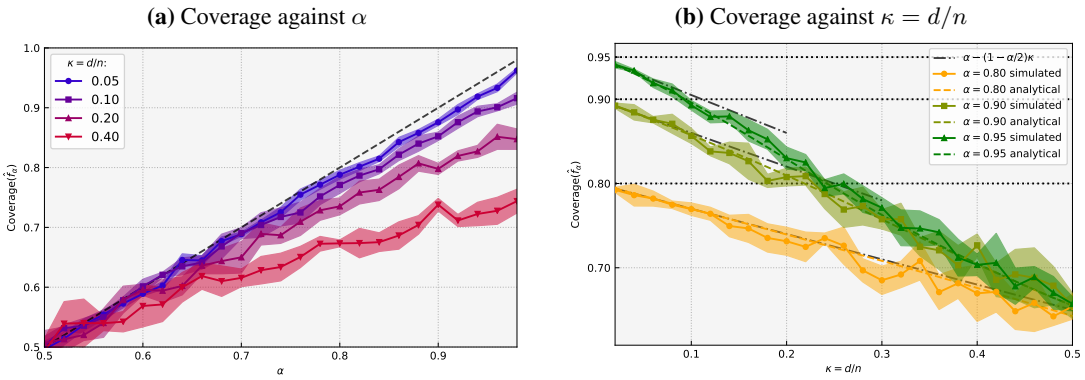

Figure 1: Coverage of quantile regression on simulated data from the realizable linear model (4). **(a)(b)** Each dot represents a combination of (α, κ) and reports the mean and one-std coverage over 8 random problem instances. **(a)** Coverage against the nominal quantile level α for fixed values of $\kappa = d/n$. **(b)** Coverage against κ for fixed $\alpha \in \{0.8, 0.9, 0.95\}$. Here "analytical" refers to our analyitical formula $\alpha - C_{\alpha,\kappa}$ and $\alpha - (1-\alpha/2)\kappa$ is its local linear approximation at small κ (both from Theorem 1).

Table 1: Coverage (%) of quantile regression on real data at nominal level $\alpha = 0.9$. Each entry reports the test-set coverage with mean and std over 8 random seeds. (d, n) denotes the {feature dim, # training examples}.

Dataset	Linear	MLP-3-64	MLP-3-512	MLP-freeze-3-512	d	n
Community	88.63±1.53	76.46±1.41	63.09±2.91	87.85±1.30	100	1599
Bike	89.64±0.44	88.75±0.91	87.67±0.49	89.27±0.57	18	8708
Star	89.48±2.56	83.14±1.76	69.71±1.82	88.05±2.42	39	1728
MEPS_19	90.09±0.72	85.46±0.96	78.55±0.93	89.03±0.51	139	12628
MEPS_20	90.06±0.57	86.52±0.65	80.77±0.72	89.60±0.28	139	14032
MEPS_21	89.99±0.39	83.79±0.52	73.09±0.82	89.15±0.36	139	12524
Nominal (α)	90.00	90.00	90.00	90.00	-	-

Results Figure 1 plots the coverage of the learned quantiles. Observe that quantile regression exhibits under-coverage consistently across different values of (α, κ). Figure 1a shows that at fixed κ, the amount of under-coverage gets more severe at a higher α, which is qualitatively consistent with our approximation formula $(\alpha - 1/2)\kappa$. Figure 1b further compares simulations with our analytical formula $\alpha - C_{\alpha,\kappa}$ (found numerically through solving the system of equations 12), as well as the local linear approximation $\alpha - (\alpha - 1/2)\kappa$ claimed in Theorem 1. Note that the simulations agree extremely well with the analytical formula. The approximation $\alpha - (\alpha - 1/2)\kappa$ is also very accurate for almost all κ at $\alpha = 0.8$, and accurate for small κ at $\alpha = 0.9, 0.95$. These verify our Theorem 1 and suggests it holds at rather realistic values of the dimension ($d = 100$).

B.2. Real data experiments

Datasets and models We take six real-world regression datasets: community and crimes (`Community`) (com, Accessed: May, 2021), bike sharing (`Bike`) (bik, Accessed: May, 2021), Tennessee's student teacher achievement ratio (`STAR`) (Achilles et al., 2008), as well as the medical expenditure survey number 19 (`MEPS_19`) (mep, Accessed: May, 2021a), number 20 (`MEPS_20`) (mep, Accessed: May, 2021b), and number 21 (`MEPS_21`) (mep, Accessed: May, 2021c). All datasets are pre-processed to have standarized features and randomly split into a 80% train set and 20% test set.

To go beyond linear quantile functions, we perform quantile regression with one of the following four models as our f_θ: linear model (`Linear`), a 3-layer MLP (two non-linear layers) with width 64 (`MLP-3-64`), 512 (`MLP-3-512`), and a variant of the width-512 MLP where all representation layers are frozen and only the last linear layer is trained (`MLP-freeze-3-512`). All linear layers include a trainable bias. We minimize the α-quantile loss (3) via momentum SGD with batch size 64. For each setting, we average over 8 random seeds where each seed determines the train-validation split, model initialization, and SGD batching. In our real experiments we fix $\alpha = 0.9$. (Results at $\alpha \in \{0.8, 0.95\}$ as well as additional experimental setups can be found in Appendix I.2).

Results Table 1 reports the coverage of the learned quantile functions (evaluated on the test sets). Observe that all MLPs exhibit under-coverage compared with the nominal level 90%. Additionally, the amount of under-coverage correlates well with model capacity—the two vanilla MLPs under-covers more severely than the MLP-freeze and the linear model. Notice

that the linear model does not have a notable under-coverage on most datasets—we believe this is a consequence of d/n being small on these datasets. The only exception is the `Community` dataset with the highest $d/n \approx 1/16$, on which the linear model does under-cover mildly by roughly 1%.

B.3. Linear quantile regression on pseudo-labels

To further test the coverage of linear quantile regression on real data distributions, we make two modifications: (1) We subset the training data by fixing d and reducing n, so as to test the coverage across differerent values of $\kappa = d/n$; (2) We compare linear quantile regression on both true labels y_i, and *pseudo-labels* y_i^{pseudo} generated from estimated linear models. These pseudo-labels are generated by first fitting a linear model $\widehat{\mathbf{w}} \in \mathbb{R}^d$ (with square loss) on the training data, and then generating a new label using the fitted linear model $\widehat{\mathbf{w}}$:

$$y_i^{\text{pseudo}} = \widehat{\mathbf{w}}^\top \mathbf{x}_i + \widehat{\sigma} z_i,$$

where $\widehat{\sigma}$ is estimated as $\sqrt{\widehat{\mathbb{E}}_{(\mathbf{x},y)}[(y - \widehat{\mathbf{w}}^\top \mathbf{x})^2]}$ on a separate hold-out split, and $z_i \sim \mathsf{N}(0,1)$. The motivation for the pseudo-labels is to make sure that the data comes from a true linear model, removing the potential effect of model misspecification.

Table 2 shows that on the `MEPS_20` dataset, linear quantile regression exhibits under-coverage at relatively large values of κ $(0.1, 0.2, 0.5)$ for both kinds of labels. Also, there is no notable difference between pseudo-labels and true labels. This provides evidence that our theory on *linear* quantile regression may hold broadly on real-world data distributions.

Table 2: Coverage of linear quantile regression on true labels vs. pseudo-labels.

$\kappa = d/n$	0.01	0.02	0.05	0.1	0.2	0.5
MEPS_20	89.83±0.67	89.89±0.81	89.54±0.82	88.74±1.51	87.15±1.52	84.75±1.81
MEPS_20 Pseudo	90.05±0.85	89.95±0.64	89.49±0.64	88.90±1.60	86.96±1.30	83.70±2.98
Nominal (α)	90.00	90.00	90.00	90.00	90.00	90.00

C. Proof overview of Theorem 1

Closed-form expression for coverage Our first step is to obtain a closed-form expression for the coverage. Recall that

$$\text{Coverage}(\widehat{f}) := \mathbb{P}_{(\mathbf{x},y)}(y \le \widehat{f}(\mathbf{x})) = \mathbb{P}_{(\mathbf{x},z)}(\langle \mathbf{w}_\star, \mathbf{x}\rangle + z \le \langle \widehat{\mathbf{w}}, \mathbf{x}\rangle + \widehat{b}).$$

As $\mathbf{x}$ is standard Gaussian, and the random variable z has cumulative distribution function Φ_z, standard calculation then yields the closed form expression (Lemma E.1)

$$\text{Coverage}(\widehat{f}) = \mathbb{E}_{G\sim\mathsf{N}(0,1)}[\Phi_z(\|\widehat{\mathbf{w}} - \mathbf{w}_\star\|_2 G + \widehat{b})].$$

Concentration of $\|\widehat{\mathbf{w}} - \mathbf{w}_\star\|_2$ and $\widehat{b}$ We generalize results from recent advances in high-dimensional M-estimator in linear models (El Karoui et al., 2013; Donoho & Montanari, 2016; Karoui, 2013; Thrampoulidis et al., 2018) to show that $\|\widehat{\mathbf{w}} - \mathbf{w}_\star\|_2$ and $\widehat{b}$ obtained by quantile regression 6 concentrates around fixed values in the high-dimensional limit. We show that, in the limit of $d, n \to \infty$ and $d/n \to \kappa$, the following concentration happens:

$$\|\widehat{\mathbf{w}} - \mathbf{w}_\star\|_2 \xrightarrow{p} \tau_\star(\kappa), \quad \text{and} \quad \widehat{b} \xrightarrow{p} b_\star(\kappa). \tag{11}$$

Above, $\tau_\star$ and $b_\star$ are determined by the solutions of a system of nonlinear equations with three variables (τ, λ, b):

$$\begin{cases} \tau^2\kappa = \lambda^2 \cdot \mathbb{E}_{(G,Z)\sim\mathsf{N}(0,1)\times P_z}[e'_{\ell_b^\alpha}(\tau G + Z; \lambda)^2], \\ \tau\kappa = \lambda \cdot \mathbb{E}_{(G,Z)\sim\mathsf{N}(0,1)\times P_z}[e'_{\ell_b^\alpha}(\tau G + Z; \lambda) G], \\ 0 = \mathbb{E}_{(G,Z)\sim\mathsf{N}(0,1)\times P_z}[e'_{\ell_b^\alpha}(\tau G + Z; \lambda)], \end{cases} \tag{12}$$

where $e_\ell(x;\tau) = \min_v \frac{1}{2\tau}(x-v)^2 + \ell(v)$ and $\ell_b^\alpha = \ell^\alpha(t-b)$ is the shifted pinball loss (2). (See Theorem F.1 for the formal statement.) This is established via two main steps: We first build on the results of Thrampoulidis et al. (2018) to show that a

variant of the risk minimization problem with a fixed bias b concentrates around the solution to the first two equations in (12). We then develop a novel concentration argument to deal with the additional learnable bias b in the minimization problem (6), which introduces the third equation in (12) that will be used in characterizing the limiting value of the minimizer $\widehat{b}$.

The concentration (11) implies that $\mathrm{Coverage}(\widehat{f})$ also converges to the following limiting coverage value (Lemma F.3):

$$\mathrm{Coverage}(\widehat{f}) \xrightarrow{p} \mathbb{E}_{G\sim\mathsf{N}(0,1)}[\Phi_z(\tau_\star(\kappa)G + b_\star(\kappa))] =: \alpha - C_{\alpha,\kappa}. \tag{13}$$

Calculating the limiting coverage via local linear analysis In this final step, as another technical crux of the proof, we further evaluate the small κ approximation of coverage value (13), and determine the sign of $C_{\alpha,\kappa}$. This is achieved by a *local linear analysis* on the solutions of the aforementioned system of equations at small κ (Lemma F.2) in a similar fashion as in (Bai et al., 2021), and a precise analysis on the interplay between the concentration values $\tau_\star$, $b_\star$, and the noise density ϕ_z. Combining these calculations yields that $C_{\alpha,\kappa}/\kappa = -(\alpha - 1/2) + o(1)$ for small enough κ (Lemma F.4). As $\alpha > 1/2$, this establishes Theorem 1. All details on these analyses can be found in our proofs in Appendix F.

D. Technical tools

D.1. The pinball loss

Recall that we took $\ell^\alpha : \mathbb{R} \to \mathbb{R}_{\geq 0}$ to be the pinball loss for the α-quantile, i.e.,

$$\ell^\alpha(t) = -(1-\alpha)t\mathbf{1}\left\{t \leq 0\right\} + \alpha t\mathbf{1}\left\{t > 0\right\}.$$

We denote $\ell_b^\alpha(t) = \ell^\alpha(t-b)$ to be the shifted pinball loss. We will suppress the superscript in $\ell_b = \ell_b^\alpha$ whenever it is clear in the context. The loss function ℓ_b is weakly differentiable, with a weak derivative ℓ_b' given by

$$\ell_b'(t) = -(1-\alpha)\mathbf{1}\left\{t \leq 0\right\} + \alpha\mathbf{1}\left\{t > 0\right\}.$$

D.2. Calculus of the Moreau envelope and prox operator

Given a convex loss function $\ell : \mathbb{R} \to \mathbb{R}$, we define its the Moreau envelope $e_\ell : \mathbb{R} \times \mathbb{R}_{>0} \to \mathbb{R}$ by

$$e_\ell(x;\lambda) = \min_v \Big[\frac{1}{2\lambda}(x-v)^2 + \ell(v)\Big],$$

and the proximal operator $\mathsf{prox}_\ell(x;\lambda) : \mathbb{R} \times \mathbb{R}_{>0} \to \mathbb{R}$ by

$$\mathsf{prox}_\ell(x;\lambda) = \arg\min_v \Big[\frac{1}{2\lambda}(x-v)^2 + \ell(v)\Big].$$

Since ℓ is convex, $\mathsf{prox}_\ell(x;\lambda)$ is well-defined. For $\ell = \ell_b$, we have

$$\begin{aligned}\mathsf{prox}_{\ell_b}(x;\lambda) = b \cdot \mathbf{1}\{x \in [b-(1-\alpha)\lambda, b+\alpha\lambda]\} \\ + (x-\alpha\lambda)\mathbf{1}\{x > b+\alpha\lambda\} + (x+(1-\alpha)\lambda)\mathbf{1}\{x < b-(1-\alpha)\lambda\}.\end{aligned}$$

The function e_{ℓ_b} is differentiable with respect to (x,λ,b), with derivatives

$$\begin{aligned}\partial_x e_{\ell_b}(x;\lambda) &= \frac{x - \mathsf{prox}_{\ell_b}(x;\lambda)}{\lambda}, \\ \partial_\lambda e_{\ell_b}(x;\lambda) &= -\frac{[x-\mathsf{prox}_{\ell_b}(x;\lambda)]^2}{2\lambda^2} = -\frac{1}{2}(\partial_x e_{\ell_b(x;\lambda)})^2, \\ \partial_b e_{\ell_b}(x;\lambda) &= -\partial_x e_{\ell_b}(x;\lambda).\end{aligned} \tag{14}$$

The functions $\partial_x e_{\ell_b}$, $\partial_\lambda e_{\ell_b}$ and $\partial_b e_{\ell_b}$ are weakly-differentiable with respect to (x,λ,b), with the following formulas giving

one (choice of) weak derivative:

$$\begin{aligned}
\partial_x \partial_x e_{\ell_b}(x;\lambda) &= \frac{1}{\lambda}\mathbf{1}\{\mathsf{prox}_{\ell_b}(x;\lambda) = b\} \geq 0,\\
\partial_\lambda \partial_x e_{\ell_b}(x;\lambda) &= -\frac{[x - \mathsf{prox}_{\ell_b}(x;\lambda)]^2}{2\lambda^2} = -\partial_x e_{\ell_b}(x;\lambda)\partial_x\partial_x e_{\ell_b}(x;\lambda),\\
\partial_b \partial_x e_{\ell_b}(x;\lambda) &= -\partial_x\partial_x e_{\ell_b}(x;\lambda),\\
\partial_\lambda \partial_b e_{\ell_b}(x;\lambda) &= -\partial_x e_{\ell_b}(x;\lambda)\partial_b\partial_x e_{\ell_b}(x;\lambda) = \partial_x e_{\ell_b}(x;\lambda)\partial_x\partial_x e_{\ell_b}(x;\lambda),\\
\partial_b \partial_b e_{\ell_b}(x;\lambda) &= \partial_x\partial_x e_{\ell_b}(x;\lambda),\\
\partial_\lambda \partial_\lambda e_{\ell_b}(x;\lambda) &= -\partial_x e_{\ell_b}(x;\lambda)\partial_\lambda\partial_x e_{\ell_b}(x;\lambda) = \partial_x e_{\ell_b}(x;\lambda)^2\partial_x\partial_x e_{\ell_b}(x;\lambda).
\end{aligned} \tag{15}$$

D.3. Implicit function theorem

We state the standard implicit function theorem in the following.

Lemma D.1 (Implicit function theorem)**.** *Let $\boldsymbol{F}(\boldsymbol{p},\kappa) : \mathbb{R}^s \times \mathbb{R}_{\geq 0} \to \mathbb{R}^s$ be a continuously differentiable vector-valued function on $\mathsf{B}(\boldsymbol{p}_0,\varepsilon)\times[0,\bar{\kappa}_0)$ for some $\bar{\kappa}_0 > 0$. Suppose $\boldsymbol{F}(\boldsymbol{p}_0,0) = 0$ and*

$$\sigma_{\min}(\nabla_{\boldsymbol{p}} F(\boldsymbol{p}_0,0)) > 0.$$

Then there exists a constant $\kappa_0 > 0$ and a continuous differentiable path $\boldsymbol{p}_\star(\kappa) \in \mathsf{B}(\boldsymbol{p}_0,\varepsilon)$, such that

$$\boldsymbol{F}(\boldsymbol{p}_\star(\kappa),\kappa) = 0, \quad \forall \kappa \in [0,\kappa_0).$$

D.4. Other technical lemmas

Lemma D.2. *For any vectors $\mathbf{u},\mathbf{v} \in \mathbb{R}^d$ and any positive definite matrix $\mathbf{A} \in \mathbb{R}^{d\times d}$, $\mathbf{A} \succ \mathbf{0}$, we have*

$$\left|\mathbf{u}^\top(\mathbf{A}+\mathbf{v}\mathbf{v}^\top)^{-1}\mathbf{v}\right| \leq \left|\mathbf{u}^\top\mathbf{A}^{-1}\mathbf{v}\right|.$$

Proof. Recall the Sherman-Morrison-Woodbury identity for matrix inversion:

$$(\mathbf{A}+\mathbf{v}\mathbf{v}^\top)^{-1} = \mathbf{A}^{-1} - \frac{\mathbf{A}^{-1}\mathbf{v}\mathbf{v}^\top\mathbf{A}^{-1}}{1+\mathbf{v}^\top\mathbf{A}^{-1}\mathbf{v}}.$$

Applying this, we have

$$\begin{aligned}
\left|\mathbf{u}^\top(\mathbf{A}+\mathbf{v}\mathbf{v}^\top)^{-1}\mathbf{v}\right| &= \left|\mathbf{u}^\top\mathbf{A}^{-1}\mathbf{v} - \mathbf{u}^\top\frac{\mathbf{A}^{-1}\mathbf{v}\mathbf{v}^\top\mathbf{A}^{-1}}{1+\mathbf{v}^\top\mathbf{A}^{-1}\mathbf{v}}\mathbf{v}\right|\\
&= \left|\mathbf{u}^\top\mathbf{A}^{-1}\mathbf{v} - \left(\mathbf{u}^\top\mathbf{A}^{-1}\mathbf{v}\right)\cdot\frac{\mathbf{v}^\top\mathbf{A}^{-1}\mathbf{v}}{1+\mathbf{v}^\top\mathbf{A}^{-1}\mathbf{v}}\right|\\
&= \left|\left(\mathbf{u}^\top\mathbf{A}^{-1}\mathbf{v}\right)\cdot\frac{1}{1+\mathbf{v}^\top\mathbf{A}^{-1}\mathbf{v}}\right| \leq \left|\mathbf{u}^\top\mathbf{A}^{-1}\mathbf{v}\right|.
\end{aligned}$$

Above, the last line used $\mathbf{v}^\top\mathbf{A}^{-1}\mathbf{v} \geq 0$ since $\mathbf{A}^{-1} \succeq \mathbf{0}$. This proves the lemma. □

Lemma D.3. *Let $\boldsymbol{X} \in \mathbb{R}^s$ be a random variable with distribution μ, and let $\boldsymbol{u} : \mathbb{R}^s \to \mathbb{R}^k$ be a continuous function. Assume that there exist $(\boldsymbol{x}_t)_{t\in[k]}$ that are in the support of the distribution of $\boldsymbol{X}$ (i.e., for any $t \in [k]$, we have $\mu(\{\boldsymbol{x} : \|\boldsymbol{x}_t - \boldsymbol{x}\|_2 \leq \varepsilon\}) > 0$ for any $\varepsilon > 0$), such that $[\boldsymbol{u}(\boldsymbol{x}_1),\ldots,\boldsymbol{u}(\boldsymbol{x}_k)] \in \mathbb{R}^{k\times k}$ is full rank. Then we have*

$$\mathbb{E}[\boldsymbol{u}(\boldsymbol{X})\boldsymbol{u}(\boldsymbol{X})^\top] \succ 0.$$

Proof of Lemma D.3. We denote

$$\omega(\varepsilon) = \sup_{t\in[k]}\left[2\sup_{\boldsymbol{x}\in\mathsf{B}(\boldsymbol{x}_t,\varepsilon)}\|\boldsymbol{u}(\boldsymbol{x})-\boldsymbol{u}(\boldsymbol{x}_t)\|_2 \cdot \sup_{\boldsymbol{x}\in\mathsf{B}(\boldsymbol{x}_t,\varepsilon)}\|\boldsymbol{u}(\boldsymbol{x})\|_2 + \sup_{\boldsymbol{x}\in\mathsf{B}(\boldsymbol{x}_t,\varepsilon)}\|\boldsymbol{u}(\boldsymbol{x})-\boldsymbol{u}(\boldsymbol{x}_t)\|_2^2\right].$$

Since $\boldsymbol{u}$ is a continuous function on $\mathbb{R}^s$, we have

$$\lim_{\varepsilon \to 0} \omega(\varepsilon) = 0.$$

We further denote

$$\nu(\varepsilon) = \min_{t \in [k]} \mu(\mathsf{B}(\boldsymbol{x}_t, \varepsilon)).$$

Then by the fact that $(\boldsymbol{x}_t)_{t\in[k]} \subseteq \mathrm{supp}(\mu)$, we have $\nu(\varepsilon) > 0$ for any $\varepsilon > 0$.

Then, for any $\varepsilon > 0$, we have

$$\begin{aligned}
\mathbb{E}[\boldsymbol{u}(\boldsymbol{X})\boldsymbol{u}(\boldsymbol{X})^\top] &\succeq \sum_{t=1}^k \int_{\mathsf{B}(\boldsymbol{x}_t,\varepsilon)} \boldsymbol{u}(\boldsymbol{x})\boldsymbol{u}(\boldsymbol{x})^\top \mu(\mathrm{d}\boldsymbol{x}) \\
&\succeq \sum_{t=1}^k (\boldsymbol{u}(\boldsymbol{x}_t)\boldsymbol{u}(\boldsymbol{x}_t)^\top - \omega(\varepsilon) I_k)\nu(\varepsilon) \\
&= \nu(\varepsilon) \sum_{t=1}^k \boldsymbol{u}(\boldsymbol{x}_t)\boldsymbol{u}(\boldsymbol{x}_t)^\top - \omega(\varepsilon)k\nu(\varepsilon) I_k \\
&\succeq \nu(\varepsilon)\Big[\lambda_{\min}\Big(\sum_{t=1}^k \boldsymbol{u}(\boldsymbol{x}_t)\boldsymbol{u}(\boldsymbol{x}_t)^\top\Big) - \omega(\varepsilon)k\Big] I_k.
\end{aligned}$$

Since $[\boldsymbol{u}(\boldsymbol{x}_1), \ldots, \boldsymbol{u}(x_k)]$ has full rank, we have $\lambda_{\min}(\sum_{t=1}^k \boldsymbol{u}(\boldsymbol{x}_t)\boldsymbol{u}(\boldsymbol{x}_t)^\top) > 0$. We can choose ε sufficiently small, so that $\lambda_{\min}(\sum_{t=1}^k \boldsymbol{u}(\boldsymbol{x}_t)\boldsymbol{u}(\boldsymbol{x}_t)^\top) - \omega(\varepsilon)k > 0$. This gives $\mathbb{E}[\boldsymbol{u}(\boldsymbol{X})\boldsymbol{u}(\boldsymbol{X})^\top] \succ 0$. This proves the lemma. □

E. Properties of quantile regression

E.1. Population minimizer of quantile risk

We can express the population quantile risk as

$$R(f) = \mathbb{E}[\ell^\alpha(y - f(\mathbf{x}))] = \mathbb{E}_{\mathbf{x}}\mathbb{E}[\ell^\alpha(y - f(\mathbf{x}))|\mathbf{x}].$$

Therefore, any function $f(\mathbf{x})$ that minimizes the conditional expectation $\mathbb{E}[\ell^\alpha(y - f(\mathbf{x}))|\mathbf{x}]$ at every $\mathbf{x}$ minimizes the above risk. It is a classical result that for any distribution P on $\mathbb{R}$, a minimizer of $\mathbb{E}_{y\sim P}[\ell^\alpha(y - f)]$ is the α-quantile $q_\alpha = \inf\{t \in \mathbb{R} : F(t) \geq \alpha\}$, where F is the CDF of P (Koenker & Bassett Jr, 1978, Section 3). Therefore, the conditional quantile function $q^\star(\mathbf{x}) = \arg\min_f \mathbb{E}[\ell^\alpha(y - f(\mathbf{x}))|\mathbf{x}]$ is a minimizer of the aforementioned conditional expectation at every $\mathbf{x}$. This proves the claim. □

E.2. Explicit expression of coverage

Lemma E.1. *Under the linear model (4), for any linear quantile function $\widehat{f}(\mathbf{x}) = \widehat{\mathbf{w}}^\top\mathbf{x} + \widehat{b}$, the coverage of $\widehat{f}$ can be expressed as*

$$\mathrm{Coverage}(\widehat{f}) = \mathbb{P}_{(\mathbf{x},y)}\Big(y \leq \widehat{\mathbf{w}}^\top\mathbf{x} + \widehat{b}\Big) = \mathbb{E}_{G\sim\mathsf{N}(0,1)}\Big[\Phi_z\Big(\|\widehat{\mathbf{w}} - \mathbf{w}_\star\|_2 G + \widehat{b}\Big)\Big].$$

Proof. By the linear model (4), we have $y = \mathbf{w}_\star^\top\mathbf{x} + z$ and thus

$$\begin{aligned}
&\mathbb{P}_{(\mathbf{x},y)}\Big(y \leq \widehat{\mathbf{w}}^\top\mathbf{x} + \widehat{b}\Big) = \mathbb{P}_{(\mathbf{x},z)}\Big(\mathbf{w}_\star^\top\mathbf{x} + z \leq \widehat{\mathbf{w}}^\top\mathbf{x} + \widehat{b}\Big) \\
&= \mathbb{P}_{(\mathbf{x},z)}\Big(z \leq (\widehat{\mathbf{w}} - \mathbf{w}_\star)^\top\mathbf{x} + \widehat{b}\Big) \\
&= \mathbb{E}_{\mathbf{x}}\Big[\Phi_z\Big((\widehat{\mathbf{w}} - \mathbf{w}_\star)^\top\mathbf{x} + \widehat{b}\Big)\Big] \\
&= \mathbb{E}_{G\sim\mathsf{N}(0,1)}\Big[\Phi_z\Big(\|\widehat{\mathbf{w}} - \mathbf{w}_\star\|_2 G + \widehat{b}\Big)\Big].
\end{aligned}$$

Above, the last step used the Gaussian input assumption $\mathbf{x} \sim \mathsf{N}(\mathbf{0}, \mathbf{I}_d)$. □

F. Proof of Theorem 1

Recall that $\ell_b^\alpha(t) = \ell^\alpha(t-b)$ where $\ell^\alpha(t)$ is the pinball loss for the α-quantile, i.e.,

$$\ell^\alpha(t) = -(1-\alpha)t\mathbf{1}\left\{t \le 0\right\} + \alpha t \mathbf{1}\left\{t > 0\right\}.$$

We will consider a fixed α, so we often write $\ell_b \equiv \ell_b^\alpha$. We further define

$$e_\ell(x;\lambda) := \min_{v\in\mathbb{R}} \left[\frac{1}{2\lambda}(x-v)^2 + \ell(v)\right].$$

We consider the following system of equations in three variables $(\tau, \lambda, b) \in \mathbb{R}_{>0} \times \mathbb{R}_{>0} \times \mathbb{R}$, which will be key to our analysis of the quantile ERM problem (6):

$$\begin{cases} \tau^2\kappa = \lambda^2 \cdot \mathbb{E}\big[e'_{\ell_b}(\tau G + Z;\lambda)^2\big], \\ \tau\kappa = \lambda \cdot \mathbb{E}\big[e'_{\ell_b}(\tau G + Z;\lambda)G\big], \\ 0 = \mathbb{E}\big[e'_{\ell_b}(\tau G + Z;\lambda)\big]. \end{cases} \tag{16}$$

The following two lemmas show that the system of equations (16) has a unique solution, which further admits a local linear expansion over κ with closed-form coefficients.

Lemma F.1 (Existence of unique solution)**.** *There exists $\kappa_0 > 0$ such that for any $\kappa \in (0, \kappa_0]$, there exists a unique solution $(\tau_\star(\kappa), \lambda_\star(\kappa), b_\star(\kappa))$ of the system of equations (16).*

Define constants

$$\begin{aligned} \bar{\tau}_0^2 &:= \frac{\alpha(1-\alpha)}{\phi_z^2(z_\alpha)}, \\ \bar{\lambda}_0 &:= \frac{1}{\phi_z(z_\alpha)}, \\ \bar{b}_0 &:= \frac{-\alpha(1-\alpha)\phi_z'(z_\alpha) - (2\alpha-1)\phi_z^2(z_\alpha)}{2\phi_z^3(z_\alpha)}. \end{aligned} \tag{17}$$

Lemma F.2 (Local linear expansion of solution at small κ)**.** *Let $(\tau_\star(\kappa), \lambda_\star(\kappa), b_\star(\kappa))$ denote the solutions to (16) for any $\kappa \in (0, \kappa_0]$. The following local linear expansion holds at small κ:*

$$\begin{aligned} \tau_\star^2(\kappa) &= \bar{\tau}_0^2\kappa + o(\kappa), \\ \lambda_\star(\kappa) &= \bar{\lambda}_0\kappa + o(\kappa), \\ b_\star(\kappa) &= z_\alpha + \bar{b}_0\kappa + o(\kappa), \end{aligned} \tag{18}$$

where $z_\alpha = \Phi_z^{-1}(\alpha)$ is the α-quantile of P_z.

We now show that the quantile ERM problem (6) exhibits a sharp concentration in the proportional limit ($n, d \to \infty$, $d/n \to \kappa$) where the concentration values are determined by the solutions $(\tau_\star^2(\kappa), \lambda_\star(\kappa), b_\star(\kappa))$ above. This result is a novel extension of (the unregularized case of) (Thrampoulidis et al., 2018, Theorem 4.1) in that it incorporates—and proves the concentration in presence of—the additional trainable bias parameter b. Recall the ERM problem (6) is

$$(\widehat{\mathbf{w}}, \widehat{b}) \in \arg\min_{\mathbf{w},b} \widehat{R}_n(\mathbf{w}, b) := \frac{1}{n}\sum_{i=1}^n \ell^\alpha(y_i - (\mathbf{w}^\top \mathbf{x}_i + b)). \tag{19}$$

Theorem F.1 (Concentration of quantile ERM)**.** *Under the linear model (4) and Assumption A, consider the limit $n, d \to \infty$ and $d/n \to \kappa \in (0, \kappa_0]$ where $\kappa_0 > 0$ is some constant. Then with probability approaching one, the empirical risk minimizer $(\widehat{\mathbf{w}}, \widehat{b})$ exists (but may not be unique), and for any empirical risk minimizer $(\widehat{\mathbf{w}}, \widehat{b})$, we have*

$$\widehat{b} \xrightarrow{p} b_\star(\kappa), \qquad \|\widehat{\mathbf{w}} - \mathbf{w}_\star\|_2^2 \xrightarrow{p} \tau_\star^2(\kappa).$$

Denote

$$\text{Coverage}_{\alpha,\kappa} \equiv \mathbb{E}_{G\sim\mathsf{N}(0,1)}[\Phi_z(\tau_\star(\kappa)G + b_\star(\kappa))].$$

Combining Theorem F.1, Lemma F.2, and the expression of the coverage in Lemma E.1, the following two lemmas show that $\text{Coverage}(\widehat{f})$ also concentrates around a value $\text{Coverage}_{\alpha,\kappa} = \alpha - C_{\alpha,\kappa}$, where $C_{\alpha,\kappa}$ admits a local linear expansion with a closed-form coefficient.

Lemma F.3. *Under the settings of Theorem 1, we have as $n, d \to \infty$, $d/n \to \kappa \in (0, \kappa_0]$,*

$$\text{Coverage}(\widehat{f}) \xrightarrow{p} \text{Coverage}_{\alpha,\kappa}. \tag{20}$$

Lemma F.4. *Under the same setting as Lemma F.3, we further have*

$$\begin{aligned} &\text{Coverage}_{\alpha,\kappa} = \alpha - C_{\alpha,\kappa} \\ &= \alpha + (\phi_z(z_\alpha)\bar{b}_0 + (1/2)\phi_z'(z_\alpha)\bar{\tau}_0^2)\kappa + o(\kappa). \end{aligned} \tag{21}$$

By Lemma F.4 and the definition of $\bar{b}_0$ and $\bar{\tau}_0^2$ in (17), the above coefficient in front of κ can be simplified as

$$\begin{aligned} &\phi_z(z_\alpha)\bar{b}_0 + (1/2)\phi_z'(z_\alpha)\bar{\tau}_0^2 \\ &= \phi_z(z_\alpha) \cdot \frac{-\alpha(1-\alpha)\phi_z'(z_\alpha) - (2\alpha-1)\phi_z^2(z_\alpha)}{2\phi_z^3(z_\alpha)} + \frac{1}{2}\phi_z'(z_\alpha) \cdot \frac{\alpha(1-\alpha)}{\phi_z^2(z_\alpha)} \\ &= -(\alpha - 1/2). \end{aligned}$$

This shows that $C_{\alpha,\kappa} = (\alpha - 1/2)\kappa + o(\kappa)$, and in particular $C_{\alpha,\kappa} > 0$ for all small κ as $\alpha - 1/2 > 0$. This proves Theorem 1. □

The rest of this section is organized as follows. We prove Lemma F.1 in Section F.1 (which requires analyzing a transformed system of equations and applying the implicit function theorem). In Section F.2, we connect the system of equations to a variational problem over four real variables. We then use this connection to prove Theorem F.1 in Section F.3. Finally, we prove Lemma F.3 and Lemma F.4 in Section F.4.

F.1. Proof of Lemma F.1 and Lemma F.2

F.1.1. Analysis of system of equations (16)

We first perform a change of variables. For any $(\bar{\tau}, \bar{\lambda}, \bar{b}, \kappa) \in \overline{\Omega} \times (0, 1)$ where $\overline{\Omega} = \mathbb{R}_{\geq 0} \times \mathbb{R}_{\geq 0} \times \mathbb{R}$, we rewrite the system of equations (16) as

$$\boldsymbol{F}(\boldsymbol{p}; \kappa) = \boldsymbol{0}, \tag{22}$$

where $\boldsymbol{p} = (\bar{\tau}, \bar{\lambda}, \bar{b})$, $\boldsymbol{F}(\boldsymbol{p}; \kappa) := (F_1(\boldsymbol{p}; \kappa), F_2(\boldsymbol{p}; \kappa), F_3(\boldsymbol{p}; \kappa))$ in which

$$\begin{aligned} F_1(\bar{\tau}, \bar{\lambda}, \bar{b}; \kappa) &:= \bar{\tau}^2 - \bar{\lambda}^2 \cdot \mathbb{E}\Big[e'_{\ell_{\bar{b}\kappa + z_\alpha}}(\bar{\tau}\sqrt{\kappa}G + Z; \bar{\lambda}\kappa)^2\Big], \\ F_2(\bar{\tau}, \bar{\lambda}, \bar{b}; \kappa) &:= \bar{\tau} - \kappa^{-1/2}\bar{\lambda} \cdot \mathbb{E}\Big[e'_{\ell_{\bar{b}\kappa + z_\alpha}}(\bar{\tau}\sqrt{\kappa}G + Z; \bar{\lambda}\kappa)G\Big], \\ F_3(\bar{\tau}, \bar{\lambda}, \bar{b}; \kappa) &:= \kappa^{-1}\mathbb{E}\Big[e'_{\ell_{\bar{b}\kappa + z_\alpha}}(\bar{\tau}\sqrt{\kappa}G + Z; \bar{\lambda}\kappa)\Big]. \end{aligned} \tag{23}$$

Equation (22) and the system (16) are equivalent up to a change of variables: For any fixed κ, any solution $(\tau_\star, \lambda_\star, b_\star)$ of Eq. (16) yields a solution $(\tau_\star/\kappa, \lambda_\star/\kappa, (b_\star - z_\alpha)/\kappa, \kappa)$ of $\boldsymbol{F}(\boldsymbol{p}; \kappa) = \boldsymbol{0}$, and vice versa. Notice that this equivalence allows us to establish Lemma F.1 and Lemma F.2 by considering the transformed equation (22).

The following two auxiliary lemmas, which give a continuity analysis of the function $\boldsymbol{F}$, are key to establishing Lemma F.1 and Lemma F.2. These auxiliary lemmas are required for checking the conditions of the implicit function theorem. The proofs of these two lemmas are deferred to Section F.1.2 and F.1.3 respectively. As a shorthand, we take

$$\boldsymbol{p}_0 = (\bar{\tau}_0, \bar{\lambda}_0, \bar{b}_0),$$

where $\bar{\tau}_0, \bar{\lambda}_0, \bar{b}_0$ are defined in (17).

Lemma F.5. *Let Assumption A hold. Let $\boldsymbol{F}$ be as defined in Eq. (22). Then for any ε such that $\mathsf{B}(\boldsymbol{p}_0, 2\varepsilon) \subseteq \overline{\Omega} = \mathbb{R}_{\geq 0} \times \mathbb{R}_{\geq 0} \times \mathbb{R}$, there exists a continuous matrix function $\boldsymbol{J} : \mathsf{B}(\boldsymbol{p}_0, \varepsilon) \to \mathbb{R}^{3\times 3}$ with*

$$\sigma_{\min}(\boldsymbol{J}(\boldsymbol{p}_0)) > 0, \tag{24}$$

and

$$\lim_{\kappa \to 0} \sup_{\boldsymbol{p} \in \mathsf{B}(\boldsymbol{p}_0, \varepsilon)} \Big\| \nabla_{\boldsymbol{p}} \boldsymbol{F}(\boldsymbol{p}, \kappa) - \boldsymbol{J}(\boldsymbol{p}) \Big\|_{\mathrm{op}} = 0. \tag{25}$$

Lemma F.6. *Let Assumption A hold. Let $\boldsymbol{F}$ be as defined in Eq. (22). Then for any ε such that $\mathsf{B}(\boldsymbol{p}_0, 2\varepsilon) \subseteq \overline{\Omega} = \mathbb{R}_{\geq 0} \times \mathbb{R}_{\geq 0} \times \mathbb{R}$, there exists two continuous vector functions $\boldsymbol{F}_0, \boldsymbol{g} : \mathsf{B}(\boldsymbol{p}_0, \varepsilon) \to \mathbb{R}^3$ such that*

$$\lim_{\kappa \to 0} \sup_{\boldsymbol{p} \in \mathsf{B}(\boldsymbol{p}_0, \varepsilon)} \Big\| \boldsymbol{F}(\boldsymbol{p}, \kappa) - \boldsymbol{F}_0(\boldsymbol{p}) \Big\|_2 = 0,$$
$$\lim_{\kappa \to 0} \sup_{\boldsymbol{p} \in \mathsf{B}(\boldsymbol{p}_0, \varepsilon)} \Big\| \partial_\kappa \boldsymbol{F}(\boldsymbol{p}, \kappa) - \boldsymbol{g}(\boldsymbol{p}) \Big\|_2 = 0.$$

Moreover, we have

$$\lim_{\kappa \to 0+} \boldsymbol{F}(\boldsymbol{p}_0, \kappa) = \boldsymbol{F}_0(\boldsymbol{p}_0) = \boldsymbol{0}.$$

By Lemma F.5 and F.6, we can continuously extend the function $\boldsymbol{F}$ to the region $\mathsf{B}(\boldsymbol{p}_0, \varepsilon) \times [0, \kappa_0)$ for some small κ_0, such that $\boldsymbol{F}(\boldsymbol{p}, \kappa)$ is continuously differentiable in the same region. Moreover, by Lemma F.6, we have $\boldsymbol{F}(\boldsymbol{p}_0, 0) = \lim_{\kappa \to 0} \boldsymbol{F}(\boldsymbol{p}_0, \kappa) = 0$. Finally, by Lemma F.5, we have $\sigma_{\min}(\nabla_{\boldsymbol{p}} \boldsymbol{F}(\boldsymbol{p}_0, 0)) > 0$.

F.1.2. Proof of Lemma F.5

For any $\boldsymbol{p} = (\bar\tau, \bar\lambda, \bar b) \in \overline{\Omega} = \mathbb{R}_{\geq 0} \times \mathbb{R}_{\geq 0} \times \mathbb{R}$, we define a continuous matrix function $\boldsymbol{J} : \overline{\Omega} \to \mathbb{R}^{3\times 3}$ by

$$\boldsymbol{J}(\boldsymbol{p}) = \begin{pmatrix} 2\bar\tau & -2\bar\lambda\alpha(1-\alpha) & 0 \\ 1 - \bar\lambda\phi_z(z_\alpha) & -\bar\tau\phi_z(z_\alpha) & 0 \\ -\bar\tau\phi_z'(z_\alpha) & (1-2\alpha)\phi_z(z_\alpha) & -\phi_z(z_\alpha) \end{pmatrix}.$$

Evaluating $\boldsymbol{J}(\boldsymbol{p}_0)$ (recall $\boldsymbol{p}_0$ is defined in Eq. (17)), we have

$$\boldsymbol{J}(\boldsymbol{p}_0) = \begin{pmatrix} \frac{2\sqrt{\alpha(1-\alpha)}}{\phi_z(z_\alpha)} & -\frac{2\alpha(1-\alpha)}{\phi_z(z_\alpha)} & 0 \\ 0 & -\sqrt{\alpha(1-\alpha)} & 0 \\ -\frac{\sqrt{\alpha(1-\alpha)}}{\phi_z(z_\alpha)}\phi_z'(z_\alpha) & (1-2\alpha)\phi_z(z_\alpha) & -\phi_z(z_\alpha) \end{pmatrix}.$$

Since we have assumed that $\phi_z(z_\alpha) \neq 0$, it is easy to see that $\det(\boldsymbol{J}(\boldsymbol{p}_0)) = -2\alpha(1-\alpha) \neq 0$. This proves Eq. (24).

We next prove Eq. (25). Recall that the definition of $\boldsymbol{F} = (F_1, F_2, F_3)$ as given in Eq. (23), by the calculus of e_{ℓ_b} as in Section D.2, we have

$$F_1(\boldsymbol{p}; \kappa) = \bar\tau^2 - \mathbb{E}_G\Big\{ \frac{1}{\kappa^2} \int_{[\overline{G}_-, \overline{G}_+]} (z - \overline{G})^2 \phi_z(z) \mathrm{d}z + \bar\lambda^2\alpha^2[1 - \Phi_z(\overline{G}_+)] + \bar\lambda^2(1-\alpha)^2\Phi_z(\overline{G}_-) \Big\},$$
$$F_2(\boldsymbol{p}; \kappa) = \bar\tau - \kappa^{-1/2}\mathbb{E}_G\Big\{ \frac{1}{\kappa} \int_{[\overline{G}_-, \overline{G}_+]} (z - \overline{G}) G \phi_z(z) \mathrm{d}z + \bar\lambda\alpha[(1 - \Phi_z(\overline{G}_+))G] - \bar\lambda(1-\alpha)\Phi_z(\overline{G}_-)G \Big\},$$
$$F_3(\boldsymbol{p}; \kappa) = \kappa^{-1}\mathbb{E}_G\Big\{ \frac{1}{\bar\lambda\kappa} \int_{[\overline{G}_-, \overline{G}_+]} (z - \overline{G}) \phi_z(z) \mathrm{d}z + \alpha[1 - \Phi_z(\overline{G}_+)] - (1-\alpha)\Phi_z(\overline{G}_-) \Big\},$$

where

$$\begin{aligned} \overline{G} &\equiv z_\alpha + \kappa\bar b - G\bar\tau\sqrt{\kappa}, \\ \overline{G}_+ &\equiv z_\alpha + \kappa\bar b + \alpha\kappa\bar\lambda - G\bar\tau\sqrt{\kappa}, \\ \overline{G}_- &\equiv z_\alpha + \kappa\bar b - (1-\alpha)\kappa\bar\lambda - G\bar\tau\sqrt{\kappa}. \end{aligned} \tag{26}$$

Using the smoothness property of ϕ_z, with some calculus, we have

$$\begin{aligned}
\lim_{\kappa\to 0} \partial_{\bar\tau} F_1(\boldsymbol{p};\kappa) &= 2\bar\tau,\\
\lim_{\kappa\to 0} \partial_{\bar\tau} F_2(\boldsymbol{p};\kappa) &= 1-\bar\lambda\phi_z(z_\alpha),\\
\lim_{\kappa\to 0} \partial_{\bar\tau} F_3(\boldsymbol{p};\kappa) &= -\bar\tau\phi_z'(z_\alpha),\\
\lim_{\kappa\to 0} \partial_{\bar\lambda} F_1(\boldsymbol{p};\kappa) &= -2\bar\lambda\alpha(1-\alpha),\\
\lim_{\kappa\to 0} \partial_{\bar\lambda} F_2(\boldsymbol{p};\kappa) &= -\bar\tau\phi_z(z_\alpha),\\
\lim_{\kappa\to 0} \partial_{\bar\lambda} F_3(\boldsymbol{p};\kappa) &= (1-2\alpha)\phi_z(z_\alpha),\\
\lim_{\kappa\to 0} \partial_{\bar b} F_1(\boldsymbol{p};\kappa) &= 0,\\
\lim_{\kappa\to 0} \partial_{\bar b} F_2(\boldsymbol{p};\kappa) &= 0,\\
\lim_{\kappa\to 0} \partial_{\bar b} F_3(\boldsymbol{p};\kappa) &= -\phi_z(z_\alpha).
\end{aligned}$$

This proves that $\lim_{\kappa\to 0}\nabla_{\boldsymbol{p}}\boldsymbol{F}(\boldsymbol{p};\kappa)=\boldsymbol{J}(\boldsymbol{p})$. With some more refined analysis, it is easy to see that the convergence above is uniform over $\boldsymbol{p}\in\mathsf{B}(\boldsymbol{p}_0,\varepsilon)$ for small ε. This proves the lemma. □

F.1.3. Proof of Lemma F.6

In this proof, we follow the same notations with the proof of Lemma F.5 as in Section F.1.3.

For any $(\bar\tau,\bar\lambda,\bar b,\kappa)\in\overline{\Omega}\times(0,\kappa_0)$ where $\overline{\Omega}=\mathbb{R}_{\geq 0}\times\mathbb{R}_{\geq 0}\times\mathbb{R}$, we define

$$\begin{aligned}
f_1(\boldsymbol{p},\kappa) &= \mathbb{E}\Big[e'_{\ell_{\bar b\kappa+z_\alpha}}(\bar\tau\sqrt{\kappa}G+Z;\bar\lambda\kappa)^2\Big],\\
f_2(\boldsymbol{p},\kappa) &= \mathbb{E}\Big[e''_{\ell_{\bar b\kappa+z_\alpha}}(\bar\tau\sqrt{\kappa}G+Z;\bar\lambda\kappa)\Big],\\
f_3(\boldsymbol{p},\kappa) &= \mathbb{E}\Big[e'_{\ell_{\bar b\kappa+z_\alpha}}(\bar\tau\sqrt{\kappa}G+Z;\bar\lambda\kappa)\Big].
\end{aligned}\tag{27}$$

By the definition of F_1,F_2,F_3 as in Eq. (23), we have

$$\begin{aligned}
F_1(\boldsymbol{p},\kappa) &= \bar\tau^2-\bar\lambda^2 f_1(\boldsymbol{p},\kappa),\\
F_2(\boldsymbol{p},\kappa) &= \bar\tau-\bar\tau\bar\lambda f_2(\boldsymbol{p},\kappa),\\
F_3(\boldsymbol{p},\kappa) &= \kappa^{-1} f_3(\boldsymbol{p},\kappa).
\end{aligned}\tag{28}$$

Then, Lemma F.6 holds as long as we show that there exists continuous functions $\boldsymbol{T}(\boldsymbol{p})=(T_1(\boldsymbol{p}),T_2(\boldsymbol{p}),T_3(\boldsymbol{p}))$ and $\boldsymbol{g}(\boldsymbol{p})=(g_1(\boldsymbol{p}),g_2(\boldsymbol{p}),g_3(\boldsymbol{p}))$ such that

$$f_1(\boldsymbol{p},\kappa) = T_1(\boldsymbol{p})+o(1),\tag{29}$$
$$\partial_\kappa f_1(\boldsymbol{p},\kappa) = -\bar\lambda^{-2}g_1(\boldsymbol{p})+o(1),\tag{30}$$
$$f_2(\boldsymbol{p},\kappa) = T_2(\boldsymbol{p})+o(1),\tag{31}$$
$$\partial_\kappa f_2(\boldsymbol{p},\kappa) = -(\bar\tau\bar\lambda)^{-1}g_2(\boldsymbol{p})+o(1),\tag{32}$$
$$f_3(\boldsymbol{p},\kappa) = o(1),\tag{33}$$
$$\partial_\kappa f_3(\boldsymbol{p},\kappa) = T_3(\boldsymbol{p})+o(1),\tag{34}$$
$$\partial_\kappa^2 f_3(\boldsymbol{p},\kappa) = g_3(\boldsymbol{p})+o(1),\tag{35}$$

where the $o(1)$ terms convergence to 0 uniformly over $\boldsymbol{p}\in\mathsf{B}(\boldsymbol{p}_0,\varepsilon)$ as $\kappa\to 0+$. Moreover, we need

$$T_1(\boldsymbol{p}_0) = \bar\tau_0^2/\bar\lambda_0^2,\tag{36}$$
$$T_2(\boldsymbol{p}_0) = 1/\bar\lambda_0,\tag{37}$$
$$T_3(\boldsymbol{p}_0) = 0.\tag{38}$$

We first prove Eq. (29), (30) and (36). First, we have (c.f. Eq. (26))

$$\lim_{\kappa\to 0+} f_1(\boldsymbol{p},\kappa) = \lim_{\kappa\to 0+} \mathbb{E}\Big[\frac{1}{\bar\lambda^2\kappa^2}\int_{[\overline{G}_-,\overline{G}_+]}(z-\overline{G})^2\phi_z(z)\mathrm{d}z + \alpha^2[1-\Phi_z(\overline{G}_+)] + (1-\alpha)^2\Phi_z(\overline{G}_-)\Big]$$
$$= \alpha^2(1-\Phi_z(z_\alpha)) + (1-\alpha)^2\Phi_z(z_\alpha) = \alpha(1-\alpha) = \bar\tau_0^2/\bar\lambda_0^2.$$

where the last equality is by the definition in Eq. (17). Further, by smoothness of the density ϕ_z, and the fact that the neighborhood $\mathsf{B}(\boldsymbol{p}_0,\varepsilon)$ is bounded, this convergence is uniform over $\boldsymbol{p}=(\bar\tau,\bar\lambda,\bar b)\in\mathsf{B}(\boldsymbol{p}_0,\varepsilon)$. This proves Eq. (29) and (36).

Moreover, we have

$$\begin{aligned}
&\partial_\kappa f_1(\boldsymbol{p},\kappa)\\
=&\,\mathbb{E}\Big[-\frac{2}{\bar\lambda^2\kappa^3}\int_{[\overline{G}_-,\overline{G}_+]}(z-\overline{G})^2\phi_z(z)\mathrm{d}z\\
&+\frac{1}{\bar\lambda^2\kappa^2}(\overline{G}_+-\overline{G})^2\phi_z(\overline{G}_+)(\bar b+\alpha\bar\lambda-G\bar\tau/(2\sqrt{\kappa}))\\
&-\frac{1}{\bar\lambda^2\kappa^2}(\overline{G}_--\overline{G})^2\phi_z(\overline{G}_-)(\bar b-(1-\alpha)\bar\lambda-G\bar\tau/(2\sqrt{\kappa}))\\
&-\alpha^2\phi_z(\overline{G}_+)(\bar b+\alpha\bar\lambda-G\bar\tau/(2\sqrt{\kappa}))+(1-\alpha)^2\phi_z(\overline{G}_-)(\bar b-(1-\alpha)\bar\lambda-G\bar\tau/(2\sqrt{\kappa}))\Big]\\
=&\,\mathbb{E}\Big[-\frac{2}{\bar\lambda^2\kappa^3}\int_{[\overline{G}_-,\overline{G}_+]}(z-\overline{G})^2\phi_z(z)\mathrm{d}z\Big],
\end{aligned}$$

where the last inequality is by Stein's identity for $Z\sim\mathcal{N}(0,1)$ and a consequence of many cancellation happening. So this gives

$$\lim_{\kappa\to 0+}\partial_\kappa f_1(\boldsymbol{p},\kappa) = -\frac{2}{3\bar\lambda^2}\Big[\alpha^2-(1-\alpha)^3\Big]\phi_z(z_\alpha).$$

Again, by the smoothness of ϕ_z, and the fact that the neighborhood $\mathsf{B}(\boldsymbol{p}_0,\varepsilon)$ is bounded, this convergence is uniform over $\boldsymbol{p}=(\bar\tau,\bar\lambda,\bar b)\in\mathsf{B}(\boldsymbol{p}_0,\varepsilon)$. This proves Eq. (30). The proof of other equations within (29) to (38) follow from similar continuity arguments. This proves Lemma F.6. □

F.1.4. Proof of Lemma F.1 and Lemma F.2

We consider the function $\boldsymbol{F}$ defined in (23). First, by Lemma F.6, we have $\boldsymbol{F}(\boldsymbol{p}_0,0_+)=\mathbf{0}$. Further, by Lemma F.5 and F.6, the conditions in the Implicit Function Theorem (Lemma D.1) are satisfied, from which we can conclude that there exists $\kappa_0>0$ and a continuously differentiable path $\{\boldsymbol{p}(\kappa)=(\bar\tau(\kappa),\bar\lambda(\kappa),\bar b(\kappa)):\kappa\in[0,\kappa_0)\}\subset\mathsf{B}(\boldsymbol{p}_0,\varepsilon)$, such that $\boldsymbol{F}(\boldsymbol{p}(\kappa),\kappa)=0$ for any $\kappa\in[0,\kappa_0)$. Therefore, the set of variables

$$(\tau_\star(\kappa),\lambda_\star(\kappa),b_\star(\kappa)) = \big(\bar\tau(\kappa)\cdot\kappa,\ \bar\lambda(\kappa)\cdot\kappa,\ z_\alpha+\bar b\cdot\kappa\big),$$

is a unique solution to the original system of equations (16) by the equivalence between system (16) and system (22) under this change of variables. This proves Lemma F.1.

In order to prove Lemma F.2 (the local linear expansion), it suffices to prove that $\boldsymbol{p}(\kappa)\to\boldsymbol{p}_0=(\bar\tau_0,\bar\lambda_0,\bar b_0)$. This was already implied by the continuity of $\boldsymbol{p}(\kappa)$ w.r.t. κ as stated above. □

F.2. Connection between system of equations (16) and a variational problem

Define

$$D(\tau,b,\tau_g,\beta)\equiv\Big[\frac{\beta\tau_g}{2}+\frac{1}{\kappa}\mathbb{E}_{(G,Z)\sim\mathcal{N}(0,1)\times P_z}[e_{\ell_b}(\tau G+Z;\tau_g/\beta)]-\tau\beta\Big]. \tag{39}$$

The D defined above is strictly convex-concave as stated in the following lemma.

Lemma F.7 (Strict convexity-concavity). *Suppose $\kappa\in(0,1)$. Then for any $(\tau,b,\tau_g,\beta)\in\mathbb{R}_{>0}\times\mathbb{R}\times\mathbb{R}_{>0}\times\mathbb{R}_{>0}$, the function D defined in (39) is strictly convex in (τ,b,τ_g) ($\nabla^2_{\tau,b,\tau_g}D\succ\mathbf{0}$), and strictly concave in β.*

Proof of Lemma F.7. Define

$$E(\tau, b, \tau_g, \beta) \equiv \mathbb{E}_{(G,Z)\sim\mathcal{N}(0,1)\times P_z}[e_{\ell_b}(\tau G + Z; \tau_g/\beta)].$$

We write in short $\partial_x e = \partial_x e_{\ell_b}(\tau G + Z; \tau_g/\beta)$ and $\partial_x^2 e = \partial_x \partial_x e_{\ell_b}(\tau G + Z; \tau_g/\beta)$. Then by Eq. (14), we have

$$\begin{aligned}
\partial_\tau E(\tau, b, \tau_g, \beta) &\equiv \mathbb{E}[\partial_x e \cdot G], \\
\partial_b E(\tau, b, \tau_g, \beta) &\equiv -\mathbb{E}[\partial_x e], \\
\partial_{\tau_g} E(\tau, b, \tau_g, \beta) &\equiv -\frac{1}{2\beta}\mathbb{E}\big[(\partial_x e)^2\big], \\
\partial_\beta E(\tau, b, \tau_g, \beta) &\equiv \frac{\tau_g}{2\beta^2}\mathbb{E}\big[(\partial_x e)^2\big].
\end{aligned}$$

By Eq. (15), for any $(\tau, b, \tau_g, \beta) \in \mathbb{R}_{>0} \times \mathbb{R} \times \mathbb{R}_{>0} \times \mathbb{R}_{>0}$, we have

$$\partial_\beta^2 E = -\frac{\tau_g}{\beta^3}\mathbb{E}[(\partial_x e)^2] + \frac{\tau_g^2}{\beta^4}\mathbb{E}[(\partial_x e)^2 \partial_x^2 e] = -\frac{\tau_g}{\beta^3}\mathbb{E}[(\partial_x e)^2 \mathbf{1}\{\mathsf{prox}_{\ell_b}(\tau G + Z) \neq b\}] < 0.$$

This gives $\partial_\beta^2 D = \kappa^{-1}\partial_\beta^2 E < 0$, so that D is strictly concave in β (for any fixed (τ, b, τ_g)).

By Eq. (15) again, we have

$$\begin{aligned}
\nabla^2_{(\tau,b,\tau_g)} E &= \mathbb{E}\begin{bmatrix} G^2 \cdot \partial_x^2 e & -G \cdot \partial_x^2 e & -\beta^{-1}\partial_x e \cdot G \cdot \partial_x^2 e \\ -G \cdot \partial_x^2 e & \partial_x^2 e & \beta^{-1}\partial_x e \cdot \partial_x^2 e \\ -\beta^{-1}\partial_x e \cdot G \cdot \partial_x^2 e & \beta^{-1}\partial_x e \cdot \partial_x^2 e & \beta^{-2}(\partial_x e)^2 \cdot \partial_x^2 e \end{bmatrix} \\
&= \frac{\beta}{\tau_g}\mathbb{E}[\mathbf{1}\{\mathsf{prox}_{\ell_b}(\tau G + Z; \tau_g/\beta) \neq b\} \cdot \boldsymbol{u}\boldsymbol{u}^\top].
\end{aligned}$$

where $\boldsymbol{u} = (G, -1, \beta^{-1}\partial_x e)$. Note that there exists (G_1, Z_1), (G_2, Z_2) and (G_3, Z_3) such that $\mathsf{prox}_{\ell_b}(\tau G_1 + Z_1; \tau_g/\beta), \mathsf{prox}_{\ell_b}(\tau G_2 + Z_2; \tau_g/\beta), \mathsf{prox}_{\ell_b}(\tau G_3 + Z_3; \tau_g/\beta) \neq b$, and

$$\begin{bmatrix} G_1 & -1 & \beta^{-1}\partial_x e_{\ell_b}(\tau G_1 + Z_1; \tau_g/\beta) \\ G_2 & -1 & \beta^{-1}\partial_x e_{\ell_b}(\tau G_2 + Z_2; \tau_g/\beta) \\ G_3 & -1 & \beta^{-1}\partial_x e_{\ell_b}(\tau G_3 + Z_3; \tau_g/\beta) \end{bmatrix}$$

is full rank. By Lemma D.3, we have $\nabla^2_{(\tau,b,\tau_g)} E \succ 0$. Note that $\nabla^2_{(\tau,b,\tau_g)} D = \kappa^{-1}\nabla^2_{(\tau,b,\tau_g)} E \succ 0$, so that D is strictly convex in (τ, b, τ_g) (for any fixed β). This proves the lemma. □

We now characterize a min-max variational problem associated with the function D, and show that it has a unique solution for small κ, and the solution is related to the solution of the system of equations (16).

Lemma F.8 (Characterization of variational problem)**.** *Consider the following variational problem in four variables over the function D defined in (39):*

$$\begin{aligned}
&\inf_{\tau>0, b\in\mathbb{R}, \tau_g>0} \sup_{\beta>0} D(\tau, b, \tau_g, \beta) \\
&= \inf_{\tau>0, b\in\mathbb{R}, \tau_g>0} \sup_{\beta>0} \Big[\frac{\beta\tau_g}{2} + \frac{1}{\kappa}\mathbb{E}_{(G,Z)\sim\mathcal{N}(0,1)\times P_z}[e_{\ell_b}(\tau G + Z; \tau_g/\beta)] - \tau\beta\Big].
\end{aligned} \tag{40}$$

For all sufficiently small $\kappa \in (0, \kappa_0]$, there exists a unique solution $(\widetilde{\tau}_\star, \widetilde{b}_\star, \widetilde{\tau}_{g,\star}, \widetilde{\beta}_\star)$ (which depends on κ) to problem (40). This solution is related to the solution $(\tau_\star(\kappa), \lambda_\star(\kappa), b_\star(\kappa))$ of (16) as

$$\widetilde{\tau}_\star = \widetilde{\tau}_{g,\star} = \tau_\star(\kappa), \quad \widetilde{\beta}_\star = \tau_\star(\kappa)/\lambda_\star(\kappa), \quad \widetilde{b}_\star = b_\star(\kappa). \tag{41}$$

Further, for some positive $\varepsilon > 0$, for any $b' \in [b_\star - \varepsilon, b_\star + \varepsilon]$, the following variational problem in three variables

$$\begin{aligned}
&\inf_{\tau>0, \tau_g>0} \sup_{\beta>0} D(\tau, b', \tau_g, \beta) \\
&= \inf_{\tau>0, b\in\mathbb{R}, \tau_g>0} \sup_{\beta>0} \Big[\frac{\beta\tau_g}{2} + \frac{1}{\kappa}\mathbb{E}_{(G,Z)\sim\mathcal{N}(0,1)\times P_z}\big[e_{\ell_{b'}}(\tau G + Z; \tau_g/\beta)\big] - \tau\beta\Big]
\end{aligned} \tag{42}$$

has a unique solution within $\mathbb{R}^3_{>0}$.

Proof of Lemma F.8. Calculating the derivatives of $D(\tau, b, \tau_g, \beta)$, we get

$$\begin{aligned}
\partial_\tau D(\tau, b, \tau_g, \beta) &= \kappa^{-1}\mathbb{E}[G e'_{\ell_b}(\tau G + Z; \tau_g/\beta)] - \beta, \\
\partial_b D(\tau, b, \tau_g, \beta) &= -\kappa^{-1}\mathbb{E}[e'_{\ell_b}(\tau G + Z; \tau_g/\beta)], \\
\partial_{\tau_g} D(\tau, b, \tau_g, \beta) &= \beta/2 - \frac{1}{2\kappa\beta}\mathbb{E}[e'_{\ell_b}(\tau G + Z; \tau_g/\beta)^2], \\
\partial_\beta D(\tau, b, \tau_g, \beta) &= \tau_g/2 - \tau + \frac{\tau_g}{2\kappa\beta^2}\mathbb{E}[e'_{\ell_b}(\tau G + Z; \tau_g/\beta)^2].
\end{aligned}$$

By Lemma F.1, there exists $\kappa_0 > 0$ such that for any $\kappa \in (0, \kappa_0]$, there exists a unique solution $(\tau_\star(\kappa), \lambda_\star(\kappa), b_\star(\kappa))$ of Eq. (16). Plugging in $(\tau, b, \tau_g, \beta) = (\tau_\star(\kappa), b_\star(\kappa), \tau_\star(\kappa), \tau_\star(\kappa)/\lambda_\star(\kappa))$ into the derivatives above and using Eq. (16), we get $\nabla_{(\tau,b,\tau_g,\beta)} D(\tau_\star(\kappa), b_\star(\kappa), \tau_\star(\kappa), \tau_\star(\kappa)/\lambda_\star(\kappa)) = 0$. This proves that $(\widetilde{\tau}_\star, \widetilde{b}_\star, \widetilde{\tau}_{g,\star}, \widetilde{\beta}_\star) = (\tau_\star(\kappa), b_\star(\kappa), \tau_\star(\kappa), \tau_\star(\kappa)/\lambda_\star(\kappa))$ is a stationary point of D.

Since D is jointly strictly convex in (τ, b, τ_g) and strictly concave in β as stated in Lemma F.7, we get

$$\begin{aligned}
\inf_{\tau>0, b\in\mathbb{R}, \tau_g>0} \sup_{\beta>0} D(\tau, b, \tau_g, \beta) &\le \sup_{\beta>0} D(\widetilde{\tau}_\star, \widetilde{b}_\star, \widetilde{\tau}_{g,\star}, \beta) = D(\widetilde{\tau}_\star, \widetilde{b}_\star, \widetilde{\tau}_{g,\star}, \widetilde{\beta}_\star), \\
\inf_{\tau>0, b\in\mathbb{R}, \tau_g>0} \sup_{\beta>0} D(\tau, b, \tau_g, \beta) &\ge \inf_{\tau>0, b\in\mathbb{R}, \tau_g>0} D(\tau, b, \tau_g, \widetilde{\beta}_\star) = D(\widetilde{\tau}_\star, \widetilde{b}_\star, \widetilde{\tau}_{g,\star}, \widetilde{\beta}_\star).
\end{aligned}$$

This proves that $(\widetilde{\tau}_\star, \widetilde{b}_\star, \widetilde{\tau}_{g,\star}, \widetilde{\beta}_\star)$ is a solution of the variational problem (40). By the strict convexity-concavity property of D again, the solution of the variational problem (40) is unique. Finally, the existence and uniqueness of the solution of $\inf_{\tau>0,\tau_g>0} \sup_{\beta>0} D(\tau, b', \tau_g, \beta)$ for $b' \in [b_\star - \varepsilon, b_\star + \varepsilon]$ follows from similar arguments. □

F.3. Proof of Theorem F.1

Preliminary: the asymptotic limit fixed b via CGMT For any convex function $\ell : \mathbb{R} \to \mathbb{R}$, we define notation

$$\ell'_+(v) \equiv \sup_{s\in\partial\ell(v)} |s|.$$

For $\tau > 0$, we define (with some abuse of notation)

$$D(\tau) \equiv \inf_{\tau_g>0} \sup_{\beta>0} \Big[\frac{\beta\tau_g}{2} + \frac{1}{\kappa}\mathbb{E}_{(G,Z)\sim\mathcal{N}(0,1)\times P_z}[e_\ell(\tau G + Z; \tau_g/\beta)] - \tau\beta\Big]. \tag{43}$$

The following proposition is by (**?**)Theorem 4.1]thrampoulidis2018precise, which uses the Convex Gaussian Comparison Theorem (CGMT).

Proposition F.1 (A simplification of Theorem 4.1 in (Thrampoulidis et al., 2018) up to model rescaling)**.** *Let ℓ be a closed proper convex function and P_z be a distribution on the real line satisfying*

- $\mathbb{E}_{(G,Z)\sim\mathcal{N}(0,1)\times P_z}[|\ell'_+(cG + Z)|^2] < \infty$*, for all* $c \in \mathbb{R}$*;*
- $\sup_{v\in\mathbb{R}} |\ell'_+(v)| < \infty$.

Further assume that the set $\arg\min_\tau D(\tau)$ *is bounded for the function D defined in (43). Then D has a unique minimizer $\tau_\star > 0$. Moreover, in the limit $n, d \to \infty$ and $d/n \to \kappa$, we have*

$$\min_{\mathbf{w}} \frac{1}{n}\sum_{i=1}^n \ell(y_i - \langle \mathbf{x}_i, \mathbf{w}\rangle) \xrightarrow{p} \min_\tau D(\tau).$$

Furthermore, for any $\varepsilon > 0$, defining $S_\varepsilon \equiv \{\mathbf{w} : |\|\mathbf{w} - \mathbf{w}_\star\|_2^2 - \tau_\star^2| \le \varepsilon\}$, there exists $\delta > 0$ such that

$$\min_{\mathbf{w}\in S_\varepsilon^c} \frac{1}{n}\sum_{i=1}^n \ell(y_i - \langle \mathbf{x}_i, \mathbf{w}\rangle) \xrightarrow{p} \min_\tau D(\tau) + \delta.$$

As a consequence, for any empirical risk minimizer $\widehat{\mathbf{w}}$ satisfying

$$\widehat{\mathbf{w}} \in \arg\min_{\mathbf{w}} \frac{1}{n}\sum_{i=1}^{n} \ell(y_i - \langle \mathbf{x}_i, \mathbf{w}\rangle),$$

we have

$$\|\widehat{\mathbf{w}} - \mathbf{w}_\star\|_2^2 \xrightarrow{p} \tau_\star^2.$$

We are now ready to prove Theorem F.1.

Proof of Theorem F.1. We define (with some abuse of notation)

$$D(\tau, b) \equiv \inf_{\tau_g>0}\sup_{\beta>0}\Big[\frac{\beta\tau_g}{2} + \frac{1}{\kappa}\mathbb{E}_{(G,Z)\sim\mathcal{N}(0,1)\times P_z}\left[e_{\ell_b^\alpha}(\tau G + Z; \tau_g/\beta)\right] - \tau\beta\Big]. \tag{44}$$

Step 1. Show that $\widehat{b} \xrightarrow{p} b_\star$. For any fixed $b \in \mathbb{R}$, define the associated minimum empirical risk (over $\mathbf{w} \in \mathbb{R}^d$) as

$$L_n(b) \equiv \min_{\mathbf{w}} \widehat{R}_n(\mathbf{w}, b).$$

Notice that $\widehat{b} = \arg\min_{b\in\mathbb{R}} L_n(b)$. Let $(\tau_\star, \kappa_\star, b_\star)$ be defined as in Lemma F.1 (as well as Lemma F.8). By Lemma F.8, there exists some $\varepsilon > 0$ such that for any fixed $b \in [b_\star - \varepsilon, b_\star + \varepsilon]$, we have $\arg\min_\tau D(\tau)$ is a singleton. Therefore the conditions of Proposition F.1 is satisfied, from which we conclude that

$$L_n(b) \xrightarrow{p} \min_\tau D(\tau, b).$$

Now, observe that $\min_\tau D(\tau, b) = \min_{\tau,\tau_g}\max_\beta D(\tau, b, \tau_g, \beta)$ is strictly convex in b (this is because $D(\tau, b, \tau_g, \beta)$ has a positive definite Hessian w.r.t. (τ, b, τ_g) at any (τ, b, τ_g, β) by Lemma F.7). Then for any $\varepsilon > 0$, there exists $\delta > 0$ such that

$$\min_\tau D(\tau, b_\star + \varepsilon) \geq \min_\tau D(\tau, b_\star) + \delta, \qquad \min_\tau D(\tau, b_\star - \varepsilon) \geq \min_\tau D(\tau, b_\star) + \delta.$$

As a consequence, with probability going to 1, we have the event

$$\{L_n(b_\star + \varepsilon) > L_n(b_\star) + \delta/2, \quad L_n(b_\star - \varepsilon) > L_n(b_\star) + \delta/2\}.$$

Furthermore, since $L_n(b)$ is a convex function in b, this implies that, with probability going to 1, we have $|\widehat{b} - b_\star| \leq \varepsilon$. Note that this is for any $\varepsilon > 0$. This proves that $\widehat{b} \xrightarrow{p} b_\star$.

Step 2. Show that $\|\widehat{\mathbf{w}} - \mathbf{w}_\star\|_2^2 \xrightarrow{p} \tau_\star^2$. By Proposition F.1, for any $\varepsilon > 0$, there exists $\delta > 0$ such that

$$\min_{\mathbf{w}\in S_\varepsilon^c} \widehat{R}_n(\mathbf{w}, b_\star) \xrightarrow{p} \min_\tau D(\tau, b_\star) + \delta.$$

where $S_\varepsilon \equiv \{\mathbf{w} : |\|\mathbf{w} - \mathbf{w}_\star\|_2^2 - \tau_\star^2| \leq \varepsilon\}$.

Furthermore, note that $\ell^\alpha(t) = -(1-\alpha)t\mathbf{1}\{t \leq 0\} + \alpha t\mathbf{1}\{t > 0\}$ is a 1-Lipschitz function in t, this gives

$$\sup_{\mathbf{w}} \left|\widehat{R}_n(\mathbf{w}, b_1) - \widehat{R}_n(\mathbf{w}, b_2)\right| \leq |b_1 - b_2|$$

As a consequence, we have

$$\min_{\mathbf{w}\in S_\varepsilon^c} \widehat{R}_n(\mathbf{w}, \widehat{b}) \geq \min_{\mathbf{w}\in S_\varepsilon^c} \widehat{R}_n(\mathbf{w}, b_\star) - |\widehat{b} - b_\star| \xrightarrow{p} \min_\tau D(\tau, b_\star) + \delta.$$

In the mean time, by Proposition F.1, we have

$$\min_{\mathbf{w}} \widehat{R}_n(\mathbf{w}, \widehat{b}) \leq \min_{\mathbf{w}} \widehat{R}_n(\mathbf{w}, b_\star) + |\widehat{b} - b_\star| \xrightarrow{p} \min_\tau D(\tau, b_\star).$$

This implies that, with probability approaching 1, we have

$$\min_{\mathbf{w}\in S_\varepsilon^c} \widehat{R}_n(\mathbf{w}, b_\star) \geq \min_\tau D(\tau, b_\star) + 2\delta/3 \quad \text{and} \quad \min_{\mathbf{w}} \widehat{R}_n(\mathbf{w}, \widehat{b}) \leq \min_\tau D(\tau, b_\star) + \delta/3.$$

On this event we have $\widehat{\mathbf{w}} \in S_\varepsilon$. Note that this is for any $\varepsilon > 0$. This proves that $\|\widehat{\mathbf{w}} - \mathbf{w}_\star\|_2^2 \xrightarrow{p} \tau_\star^2$. □

F.4. Proof of Lemma F.3 and Lemma F.4

Recall that

$$\text{Coverage}(\widehat{f}) = \mathbb{P}_{(\mathbf{x},y)}\Big(y \le \widehat{\mathbf{w}}^\top \mathbf{x} + \widehat{b}\Big) = \mathbb{E}_{G\sim \mathsf{N}(0,1)}\Big[\Phi_z\Big(\|\widehat{\mathbf{w}} - \mathbf{w}_\star\|_2 G + \widehat{b}\Big)\Big].$$

Eq. (20) is simply by the fact that $T(\tau, b; G) \equiv \Phi_z(\tau G + b)$ is a continuous function in (τ, b), by Theorem F.1, and by the dominant convergence theorem. This proves Lemma F.3.

Furthermore, by Taylor expansion, we have

$$\begin{aligned}
&\text{Coverage}_{\alpha,\kappa} = \mathbb{E}[\Phi_z(\tau_\star(\kappa)G + b_\star(\kappa))] \\
&= \Phi_z(z_\alpha) + \phi_z(z_\alpha)\mathbb{E}[(\tau_\star(\kappa)G + b_\star(\kappa) - z_\alpha)] + \frac{1}{2}\phi_z'(z_\alpha)\mathbb{E}[(\tau_\star(\kappa)G + b_\star(\kappa) - z_\alpha)^2] \\
&\quad + \frac{1}{6}\mathbb{E}[\phi_z''(\xi)(\tau_\star(\kappa)G + b_\star(\kappa) - z_\alpha)^3] \\
&= \alpha + \phi_z(z_\alpha)(b_\star(\kappa) - z_\alpha) + \frac{1}{2}\phi_z'(z_\alpha)\tau_\star^2(\kappa) + o(\kappa) \\
&= \alpha + \left(\phi_z(z_\alpha)\bar{b}_0 + \frac{1}{2}\phi_z'(z_\alpha)\bar{\tau}_0^2\right)\kappa + o(\kappa),
\end{aligned}$$

where the last equality is by Lemma F.2 and by the boundedness of ϕ_z''. This proves Eq. (21) and thus Lemma F.4.

G. Extension to over-parametrized learning

In this section we provide a variant of Theorem 1 in the over-parametrized case, i.e. when $d \ge n$, so that the learned quantile functions have the capacity to interpolate the entire training dataset. We still assume that the data are generated from the linear model (4). For notational simplicity, throughout this section we let $\boldsymbol{\theta} := [\mathbf{w}^\top, b]^\top \in \mathbb{R}^{d+1}$ denote the concatenation of $\mathbf{w}$ and b, and let $\widehat{R}_n(\boldsymbol{\theta})$ denote the empirical risk (6). We also let $\widetilde{\mathbf{x}} = [\mathbf{x}^\top, 1]^\top \in \mathbb{R}^{d+1}$ denote the augmented feature so that $\boldsymbol{\theta}^\top\widetilde{\mathbf{x}} = \mathbf{w}^\top\mathbf{x} + b$. We let $\widetilde{\mathbf{X}} \in \mathbb{R}^{n\times(d+1)}$ denote the augmented input matrix and $\mathbf{z} \in \mathbb{R}^n$ denote the noise vector.

In the over-parametrized case, the ERM is no longer well-defined as there are multiple interpolating solutions. We consider instead the quantile functions obtained on the gradient descent path on the empirical risk $\widehat{R}_n$. More precisely, we consider the vanilla (sub)-gradient descent algorithm: Initialize $\boldsymbol{\theta}_1 = \mathbf{0}$, and iterate for all $t \ge 1$

$$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - \eta_t \mathbf{g}_t, \tag{45}$$

where $\mathbf{g}_t \in \partial\widehat{R}_n(\boldsymbol{\theta}_t)$ is any sub-gradient of the empirical risk $\widehat{R}_n$ (6) at $\boldsymbol{\theta}_t$.

Theorem G.1 (Quantile regression under over-parametrization)**.** *Suppose the data is generated from the Gaussian linear model (4) with $\|\mathbf{w}\|_2 = R$, and the nominal quantile level $\alpha \in (0.5, 1)$. Further assume the noise distribution P_z is symmetric about 0 and σ^2-sub-Gaussian. Then, there exists an absolute constant $C_0 > 0$ such that if $n \ge C_0(d + \log(1/\delta))$, the following holds.*

Let $\boldsymbol{\theta}_t$ be the iterates of the sub-gradient descent algorithm (45) with step-size $\eta_t := \beta/\sqrt{t}$ for any $\beta > 0$, and let $\boldsymbol{\theta}_\infty \in \mathbb{R}^{d+1}$ denote any limit point of $\{\boldsymbol{\theta}_t\}_{t\ge1}$, then we have

(a) $\boldsymbol{\theta}_\infty$ is the minimum ℓ_2-norm interpolator of the training data, i.e.

$$\boldsymbol{\theta}_\infty = \operatorname*{arg\,min}_{\boldsymbol{\theta}\in\mathbb{R}^d}\Big\{\|\boldsymbol{\theta}\|_2 : \widetilde{\mathbf{X}}\boldsymbol{\theta} = \mathbf{y}\Big\}.$$

(b) With probability at least $1-\delta$ (over the training data), the coverage of the limiting quantile function $\widehat{f}_\infty := \boldsymbol{\theta}_\infty^\top\widetilde{\mathbf{x}} = \mathbf{w}_\infty^\top\mathbf{x} + b_\infty$ concentrates around 0.5:

$$\Big|\text{Coverage}(\widehat{f}_\infty) - 0.5\Big| \le C(R+\sigma)\cdot\sqrt{\frac{\log(1/\delta)}{d}} \le C(R+\sigma)\cdot\sqrt{\frac{\log(1/\delta)}{n}},$$

where $C > 0$ is a constant that only depends on $\sup_{t\in\mathbb{R}}|\phi_z(t)|$.

Implications Theorem G.1 shows that a severe under-coverage bias in the over-parametrized case: The coverage of the limiting quantile function (of the gradient descent path) is $0.5 \pm \widetilde{O}(1/\sqrt{d})$, *regardless of the nominal quantile level* $\alpha \in (0.5, 1)$. Therefore $\widehat{f}_\infty$ under-covers by $\alpha - 0.5 = \Theta(1)$, and this under-coverage bias does not diminish as we increase n, d.

The proof of Theorem G.1 is established in the following two subsections.

G.1. Proof of Part (a)

We begin by observing that the sub-gradients of the quantile risk (6) takes the form

$$\mathbf{g}_t = \frac{1}{n}\sum_{i=1}^{n} (\ell^\alpha)'(y_i - \boldsymbol{\theta}_t^\top \widetilde{\mathbf{x}}_i) \cdot \widetilde{\mathbf{x}}_i \in \mathrm{span}\{\widetilde{\mathbf{x}}_1, \ldots, \widetilde{\mathbf{x}}_n\}, \tag{46}$$

where $(\ell^\alpha)'(t)$ is the sub-gradient of ℓ^α, which takes value $-(1-\alpha)$ at $t < 0$, α at $t > 0$, and any value within $[-(1-\alpha), \alpha]$ at $t = 0$. As we initialized at $\boldsymbol{\theta}_1 = \mathbf{0}$, this implies that

$$\boldsymbol{\theta}_t \in \mathrm{span}\{\widetilde{\mathbf{x}}_1, \ldots, \widetilde{\mathbf{x}}_n\}$$

for all $t \geq 1$. Also, by (46) we have $\|\mathbf{g}_t\|_2 \leq M := \max_{i\in[n]} \|\widetilde{\mathbf{x}}_i\|_2$, since $|(\ell^\alpha)'| \leq \max\{\alpha, 1-\alpha\} \leq 1$.

Also, let $\boldsymbol{\theta}_{\ell_2}$ denote the minimum ℓ_2-norm interpolator of the dataset:

$$\boldsymbol{\theta}_{\ell_2} := \operatorname*{arg\,min}_{\boldsymbol{\theta}\in\mathbb{R}^d} \left\{ \|\boldsymbol{\theta}\|_2 : \widetilde{\mathbf{X}}\boldsymbol{\theta} = \mathbf{y} \right\} = \widetilde{\mathbf{X}}^\dagger \mathbf{y} = \widetilde{\mathbf{X}}^\top (\widetilde{\mathbf{X}}\widetilde{\mathbf{X}}^\top)^{-1}\mathbf{y}. \tag{47}$$

This $\boldsymbol{\theta}_{\ell_2}$ exists whenever $d+1 \geq n$ (so that $\widetilde{\mathbf{x}}_i \in \mathbb{R}^{d+1}$ are linearly independent with probability one and thus $\widetilde{\mathbf{X}}\widetilde{\mathbf{X}}^\top \in \mathbb{R}^{n\times n}$ is invertible). It further satisfies

- $\widehat{R}_n(\boldsymbol{\theta}_{\ell_2}) = 0$ (since $\boldsymbol{\theta}_{\ell_2}^\top \widetilde{\mathbf{x}}_i = y_i$). Therefore $\boldsymbol{\theta}_{\ell_2}$ is a minimizer of $\widehat{R}_n$ since $\widehat{R}_n \geq 0$.
- $\boldsymbol{\theta}_{\ell_2} \in \mathrm{span}\{\widetilde{\mathbf{x}}_1, \ldots, \widetilde{\mathbf{x}}_n\}$.
- $\boldsymbol{\theta}_{\ell_2}$ is the only point within $\mathrm{span}\{\widetilde{\mathbf{x}}_1, \ldots, \widetilde{\mathbf{x}}_n\}$ that satisfies $\widehat{R}_n(\boldsymbol{\theta}_{\ell_2}) = 0$, as any such point $\boldsymbol{\theta} \in \mathbb{R}^{d+1}$ must satisfy $\widetilde{\mathbf{X}}\boldsymbol{\theta} = \mathbf{y}$, and there is only one such point in the span because of the linear independence of $\{\widetilde{\mathbf{x}}_i\}_{i=1}^n$.

We now use the following lemma on the last-iterate convergence of sub-gradient descent, adapted from (Orabona, 2020; Accessed: May, 2021, Corollary 3):

Lemma G.1 (Last-iterate convergence of sub-gradient descent)**.** *Suppose $F : \mathbb{R}^D \to \mathbb{R}$ is a convex function with bounded sub-gradients: $\|\mathbf{g}\|_2 \leq M$ for all $\mathbf{g} \in \partial F(\boldsymbol{\theta})$ and any $\boldsymbol{\theta} \in \mathbb{R}^D$. Let $\boldsymbol{\theta}_\star \in \mathbb{R}^D$ be any minimizer of F with $F_\star = F(\boldsymbol{\theta}_\star) > -\infty$. Consider the sub-gradient descent algorithm*

$$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - \eta_t \mathbf{g}_t,$$

where $\mathbf{g}_t \in \partial F(\boldsymbol{\theta}_t)$, and $\eta_t = \beta/\sqrt{t}$ for some $\beta > 0$. Then, we have for all $T \geq 3$ that

$$F(\boldsymbol{\theta}_T) - F_\star \leq \frac{\|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_\star\|_2^2 + 4M^2\beta^2 \log T}{2\beta\sqrt{T}}.$$

Applying Lemma G.1 with on the quantile risk $\widehat{R}_n$ the associated minimizer $\boldsymbol{\theta}_{\ell_2}$, we get that (for $T \geq 3$)

$$\widehat{R}_n(\boldsymbol{\theta}_T) \leq \frac{\|\boldsymbol{\theta}_{\ell_2}\|_2^2 + 4M^2\beta^2 \log T}{2\beta\sqrt{T}}.$$

This implies that $\widehat{R}_n(\boldsymbol{\theta}_T) \to 0$ as $T \to \infty$.

The above implies that any limit point $\boldsymbol{\theta}_\infty$ of the sequence $\{\boldsymbol{\theta}_t\}_{t\geq 1}$ must satisfy

- $\widehat{R}_n(\boldsymbol{\theta}_\infty) = 0$, by continuity of $\widehat{R}_n$;
- $\boldsymbol{\theta}_\infty \in \mathrm{span}(\widetilde{\mathbf{x}}_1, \ldots, \widetilde{\mathbf{x}}_n)$, by the closedness of the span.

Combined with the above assertions on $\boldsymbol{\theta}_{\ell_2}$, this shows that $\boldsymbol{\theta}_\infty = \boldsymbol{\theta}_{\ell_2}$, establishing part (a) of the theorem. □

G.2. Proof of part (b)

We first establish a covariance lower bound useful for the subsequent analyses. As $\mathbf{x}_i \sim \mathsf{N}(\mathbf{0}, \mathbf{I}_d)$, the input matrix $\mathbf{X} \in \mathbb{R}^{n\times d}$ has i.i.d. $\mathsf{N}(0,1)$ entries, and thus $\mathbf{X}$'s columns are also i.i.d. $\mathsf{N}(\mathbf{0}, \mathbf{I}_n)$. By standard sub-Gaussian covariance concentration, we have with probability at least $1-\delta$ that

$$\left\|\frac{1}{d}\mathbf{X}\mathbf{X}^\top - \mathbf{I}_n\right\|_{\mathrm{op}} \le C\left(\sqrt{\frac{n+\log(1/\delta)}{d}} + \frac{n+\log(1/\delta)}{d}\right)$$

for some absolute constant $C > 0$ (this can be found in e.g. (Vershynin, 2018, Example 4.7.3)). In particular, we have $\left\|\mathbf{X}\mathbf{X}^\top/d - \mathbf{I}_n\right\|_{\mathrm{op}} \le 1/4$ provided $d \ge C(n+\log(1/\delta))$. On this event, we have

$$\mathbf{X}\mathbf{X}^\top \succeq \frac{3d}{4}\mathbf{I}_n.$$

We will apply a small variant of this result: as long as $d-1 \ge C(n+\log(1/\delta))$, we also have for any fixed matrix $\mathbf{V}_\star \in \mathbb{R}^{d\times(d-1)}$ with orthogonal columns that

$$\mathbf{X}\mathbf{V}_\star\mathbf{V}_\star^\top\mathbf{X}^\top \succeq \frac{3(d-1)}{4}\mathbf{I}_n \succeq \frac{d}{2}\mathbf{I}_n. \tag{48}$$

Bounding $|b_\infty|$ By (47), we have

$$\begin{aligned}\begin{bmatrix}\mathbf{w}_\infty\\ b_\infty\end{bmatrix} &= \boldsymbol{\theta}_\infty = \boldsymbol{\theta}_{\ell_2} = \widetilde{\mathbf{X}}^\top(\widetilde{\mathbf{X}}\widetilde{\mathbf{X}}^\top)^{-1}\mathbf{y} = \widetilde{\mathbf{X}}^\top(\widetilde{\mathbf{X}}\widetilde{\mathbf{X}}^\top)^{-1}(\mathbf{X}\mathbf{w}_\star + \mathbf{z})\\ &= \begin{bmatrix}\mathbf{X}^\top\\ \mathbf{1}_n^\top\end{bmatrix}(\widetilde{\mathbf{X}}\widetilde{\mathbf{X}}^\top)^{-1}(\mathbf{X}\mathbf{w}_\star + \mathbf{z}).\end{aligned}$$

Therefore

$$b_\infty = \mathbf{1}_n^\top(\widetilde{\mathbf{X}}\widetilde{\mathbf{X}}^\top)^{-1}(\mathbf{X}\mathbf{w}_\star + \mathbf{z}) = \underbrace{\mathbf{1}_n^\top\left(\mathbf{X}\mathbf{X}^\top + \mathbf{1}_n\mathbf{1}_n^\top\right)^{-1}\mathbf{X}\mathbf{w}_\star}_{\mathrm{I}} + \underbrace{\mathbf{1}_n^\top\left(\mathbf{X}\mathbf{X}^\top + \mathbf{1}_n\mathbf{1}_n^\top\right)^{-1}\mathbf{z}}_{\mathrm{II}}.$$

We now bound terms I and II separately.

For term I, let us assume for the moment that $\|\mathbf{w}_\star\|_2 = 1$. Let $\mathbf{V}_\star \in \mathbb{R}^{d\times d-1}$ denote the orthogonal complement to the matrix $\mathbf{w}_\star$ (i.e. so that $[\mathbf{w}_\star,\ \mathbf{V}_\star] \in \mathbb{R}^{d\times d}$ is an orthogonal matrix). We have

$$\mathrm{I} = \mathbf{1}_n^\top\left(\mathbf{X}\mathbf{V}_\star\mathbf{V}_\star^\top\mathbf{X}^\top + \mathbf{X}\mathbf{w}_\star\mathbf{w}_\star^\top\mathbf{X}^\top + \mathbf{1}_n\mathbf{1}_n^\top\right)^{-1}\mathbf{X}\mathbf{w}_\star.$$

As $\mathbf{X}\mathbf{V}_\star\mathbf{V}_\star^\top\mathbf{X}^\top$ is an positive definite matrix with probability one whenever $d-1 \ge n$, applying Lemma D.2 twice, we get

$$|\mathrm{I}| \le \left|\mathbf{1}_n^\top\left(\mathbf{X}\mathbf{V}_\star\mathbf{V}_\star^\top\mathbf{X}^\top + \mathbf{1}_n\mathbf{1}_n^\top\right)^{-1}\mathbf{X}\mathbf{w}_\star\right| \le \left|\mathbf{1}_n^\top\left(\mathbf{X}\mathbf{V}_\star\mathbf{V}_\star^\top\mathbf{X}^\top\right)^{-1}\mathbf{X}\mathbf{w}_\star\right|.$$

Now, notice that $\mathbf{X}\mathbf{V}_\star \in \mathbb{R}^{n\times d-1}$ and $\mathbf{X}\mathbf{w}_\star \in \mathbb{R}^n$ have i.i.d. $\mathsf{N}(0,1)$ entries and are independent of each other. Further, $\mathbf{X}\mathbf{w}_\star \sim \mathsf{N}(\mathbf{0}, \mathbf{I}_n)$, and thus the random variable $\mathbf{1}_n^\top\left(\mathbf{X}\mathbf{V}_\star\mathbf{V}_\star^\top\mathbf{X}^\top\right)^{-1}\mathbf{X}\mathbf{w}_\star$ (conditional on $\mathbf{X}\mathbf{V}_\star$) is $\|\mathbf{v}_\mathrm{I}\|_2^2$-sub-Gaussian (due to the independence between $\mathbf{X}\mathbf{V}_\star$ and $\mathbf{X}\mathbf{w}_\star$), where

$$\|\mathbf{v}_\mathrm{I}\|_2^2 = \mathbf{1}_n^\top\left(\mathbf{X}\mathbf{V}_\star\mathbf{V}_\star^\top\mathbf{X}^\top\right)^{-2}\mathbf{1}_n \le \frac{4}{d^2}\|\mathbf{1}_n\|_2^2 = \frac{4n}{d^2},$$

where the inequality used the covariance lower bound (48). This shows that

$$|\mathrm{I}| \le C\sqrt{4n/d^2\cdot\log(1/\delta)} \le C\sqrt{\log(1/\delta)/d}$$

with probability at least $1-\delta$, where the last step used $n \le d$. It is straightforward to see that, for general $\|\mathbf{w}_\star\|_2 = R$, we have

$$|\mathrm{I}| \le CR\sqrt{4n/d^2\cdot\log(1/\delta)} \le CR\sqrt{\log(1/\delta)/d}. \tag{49}$$

For term II, As $\mathbf{X}$ and $\mathbf{z}$ are independent, the random variable $\mathrm{II} = \mathbf{1}_n^\top(\mathbf{X}\mathbf{X}^\top + \mathbf{1}_n\mathbf{1}_n^\top)^{-1}\mathbf{z}$ (conditional on $\boldsymbol{X}$) is $\|\mathbf{v}_{\mathrm{II}}\|_2^2\,\sigma^2$-sub-Gaussian, where

$$\|\mathbf{v}_{\mathrm{II}}\|_2^2 = \mathbf{1}_n^\top(\mathbf{X}\mathbf{X}^\top + \mathbf{1}_n\mathbf{1}_n^\top)^{-2}\mathbf{1}_n \le \frac{4}{d^2}\,\|\mathbf{1}_n\|_2^2 = \frac{4n}{d^2} \le \frac{4}{d}.$$

Similar as above, we have with probability at least $1-\delta$ that

$$|\mathrm{II}| \le C\sigma\sqrt{\log(1/\delta)/d}. \tag{50}$$

Combining (49) and (50), we get with probability at least $1-\delta$ that (rescaling $3\delta \to \delta$)

$$|b_\infty| \le C(R+\sigma)\sqrt{\log(1/\delta)/d}. \tag{51}$$

Bounding the coverage bias We now translate the bound on $|b_\infty|$ to a bound on the coverage error $\left|\mathrm{Coverage}(\widehat{f}_\infty) - 0.5\right|$. First, note that by symmetry of the distribution of $(\mathbf{w}_\infty - \mathbf{w}_\star)^\top\mathbf{x}$ and the fact that $\Phi_z(t)+\Phi_z(-t) = 1$ (due to the symmetry of P_z), we have

$$\mathbb{E}\big[\Phi_z\big((\mathbf{w}_\infty - \mathbf{w}_\star)^\top\mathbf{x}\big)\big] = \mathbb{E}\left[\frac{1}{2}\Big(\Phi_z\big((\mathbf{w}_\infty - \mathbf{w}_\star)^\top\mathbf{x}\big) + \Phi_z\big(-(\mathbf{w}_\infty - \mathbf{w}_\star)^\top\mathbf{x}\big)\Big)\right] = 0.5.$$

Therefore we have

$$\begin{aligned}
&\left|\mathrm{Coverage}(\widehat{f}_\infty) - 0.5\right| = \left|\mathbb{E}\big[\Phi_z\big((\mathbf{w}_\infty - \mathbf{w}_\star)^\top\mathbf{x} + b_\infty\big) - \Phi_z\big((\mathbf{w}_\infty - \mathbf{w}_\star)^\top\mathbf{x}\big)\big]\right| \\
&\le \sup_{t\in\mathbb{R}} |\phi_z(t)| \cdot |b_\infty| \\
&\le C\sup_{t\in\mathbb{R}} |\phi_z(t)| \cdot |b_\infty| \le C\sup_{t\in\mathbb{R}} |\phi_z(t)| \cdot (R+\sigma)\sqrt{\log(1/\delta)/d}.
\end{aligned}$$

Notably the bound is also upper bounded by $C\sup_{t\in\mathbb{R}} |\phi_z(t)| \cdot (R+\sigma)\sqrt{\log(1/\delta)/n}$ as we assumed $d \ge n$. This proves part (b) of the theorem. □

H. Proofs for Section 4

H.1. Proof of Corollary 2

First, part (a) is a direct consequence of Lemma F.2 which was established within the proof of Theorem 1.

We now prove part (b) and (c). We show $\bar{b}_0 < 0$ for P_z being any Gaussian distribution. We first observe that to determine the sign of $\bar{b}_0$, it suffices to consider the standard Gaussian: The value of $\bar{b}_0$ does not depend on the location parameter (since ϕ_z and z_α shifts together with a location shift). Also, scalings won't change the sign of $\bar{b}_0$ (although it scales the numerator and the denominator by a different amount).

We next calculate $\bar{b}_0$ for $P_z = \mathsf{N}(0,1)$. We have $\phi_z'(z_\alpha) = -z_\alpha\phi_z(z_\alpha)$ for $\phi_z(t) = \exp(-t^2/2)/\sqrt{2\pi}$. Therefore the numerator of $\bar{b}_0$ is

$$-\alpha(1-\alpha)\phi_z'(z_\alpha) - (2\alpha-1)\phi_z^2(z_\alpha) = (\alpha(1-\alpha)z_\alpha - (2\alpha-1)\phi_z(z_\alpha))\phi_z(z_\alpha).$$

Consider the change of variable $t := z_\alpha$ so that $\alpha = \Phi_z(t)$. To show the above quantity is negative, it suffices to show that

$$\begin{aligned}
&\Phi(t)(1-\Phi(t))t - (2\Phi(t)-1)\phi(t) < 0 \\
\iff &\underbrace{\frac{t(1-\Phi(t))}{\phi(t)} - 2 + \frac{1}{\Phi(t)}}_{:=F(t)} < 0
\end{aligned}$$

for all $t > 0$, where $\Phi(t) = \Phi_z(t)$ is shorthand for the standard Gaussian CDF. To show this, we first observe that $F(0) = -2 + 1/\Phi(0) = 0$, and further

$$F'(t) = \frac{(1+t^2)(1-\Phi(t))}{\phi(t)} - t - \frac{\phi(t)}{\Phi(t)^2}.$$

We can numerically check that $F'(t) < -0.03$ for $t \in [0,1]$, within which range we have $F(t) < -0.03t < 0$. On the other hand, using the Gaussian CDF approximation bound

$$1 - \frac{1}{t^2} \le \frac{t(1-\Phi(t))}{\phi(t)} \le 1 - \frac{1}{t^2} + \frac{3}{t^4} \quad \text{for all } t > 0,$$

we have

$$\begin{aligned} F(t) &\le 1 - \frac{1}{t^2} + \frac{3}{t^4} - 2 + \frac{1}{1-(t^{-1}-t^{-3})\phi(t)} \\ &\overset{(i)}{\le} -\frac{1}{t^2} + \frac{3}{t^4} + 2(t^{-1}-t^{-3})\phi(t) \le \frac{3 + 2t^3\phi(t) - t^2}{t^4} \overset{(ii)}{<} 0, \end{aligned}$$

where (i) happens when $(t^{-1}-t^{-3})\phi(t) < 1/2$, which happens for all $t \ge 1$, and (ii) happens when $t \ge 2$. This shows that $F(t) < 0$ for $t \ge 2$. For $t \in [1,2]$, one can check numerically that $F(t) < -0.1 < 0$. This shows $F(t) < 0$ for all $t > 0$, which establishes $\bar{b}_0 < 0$ for $P_z = \mathsf{N}(0,1)$, showing part (b).

Finally, for any $\alpha \in (0.5, 1)$, we show that there exists a noise distributions $\widetilde{P}_z$ for which $\bar{b}_0 > 0$. Indeed, simply take any smooth density ϕ_z (such as standard Gaussian density), and modify ϕ_z locally around z_α into some new smooth density $\widetilde{\phi}_z$ such that both the new α-quantile $\widetilde{z}_\alpha \approx z_\alpha$ and $\widetilde{\phi}_z(\widetilde{z}_\alpha) \approx \phi_z(z_\alpha)$ (with arbitrarily small differences), but $\widetilde{\phi}'_z(\widetilde{z}_\alpha) < 0$ is negative with a high magnitude $|\widetilde{\phi}'_z(\widetilde{z}_\alpha)|$. Taking this magnitude high enough, we can always make $-\alpha(1-\alpha)\widetilde{\phi}'_z(\widetilde{z}_\alpha) - (2\alpha-1)\widetilde{\phi}_z(\widetilde{z}_\alpha)^2 > 0$, which gives $\bar{b}_0 > 0$ for the noise distribution $\widetilde{P}_z$ defined by the density $\widetilde{\phi}_z$. This shows part (c).

H.2. Proof of Theorem 3

For any $\widehat{f}(\mathbf{x}) = \widehat{\mathbf{w}}^\top \mathbf{x} + b_\star$, the coverage can be expressed as

$$\begin{aligned} \text{Coverage}(\widehat{f}) &= \mathbb{P}\big(y \le \widehat{\mathbf{w}}^\top \mathbf{x} + b_\star\big) \overset{(i)}{=} \mathbb{P}\big(\mu_\star(\mathbf{x}) + \sigma_\star(\mathbf{x}) z \le \widehat{\mathbf{w}}^\top \mathbf{x} + b_\star\big) \\ &\overset{(ii)}{=} \mathbb{P}\big(\sigma_\star(\mathbf{x})(z - z_\alpha) \le (\widehat{\mathbf{w}} - \mathbf{w}_\star)^\top \mathbf{x}\big) = \mathbb{P}\left(z \le z_\alpha + \frac{(\widehat{\mathbf{w}} - \mathbf{w}_\star)^\top \mathbf{x}}{\sigma_\star(\mathbf{x})}\right) \\ &= \mathbb{E}\left[\Phi_z\left(z_\alpha + \frac{(\widehat{\mathbf{w}} - \mathbf{w}_\star)^\top \mathbf{x}}{\sigma_\star(\mathbf{x})}\right)\right]. \end{aligned}$$

Above, (i) used the data distribution assumption (9), and (ii) follows by subtracting both sides by $\mu_\star(\mathbf{x}) + \sigma_\star(\mathbf{x}) = \mathbf{w}_\star^\top \mathbf{x} + b_\star$ by the linear true quantile assumption (10).

Now, by assumption $\alpha \ge 3/4$, we have $z_\alpha > z_{1/2} = 0$. We claim the following holds for all $t \in \mathbb{R}$:

$$\frac{1}{2}(\Phi_z(z_\alpha + t) + \Phi_z(z_\alpha - t)) \le \Phi_z(z_\alpha) - ct^2 \mathbf{1}\left\{|t| \le z_\alpha\right\}, \tag{52}$$

where $c > 0$ is a constant that only depends on Φ_z and z_α. To see this, notice that $\Phi''_z(t) = \phi'_z(t) < 0$ for $t > 0$ and thus Φ_z is concave for $t \ge 0$. Further, Φ_z is c-strongly concave on $[z_\alpha/2, 3z_\alpha/2]$ for some $c > 0$ as $\Phi''_z(t) = \phi'_z(t)$ is continuous and negative on this compact interval. This shows that

$$\frac{1}{2}(\Phi_z(z_\alpha + t) + \Phi_z(z_\alpha - t)) \le \Phi_z(z_\alpha) - ct^2$$

for $|t| \le z_\alpha/2$, and further by the concavity of Φ_z on $[0, 2z_\alpha]$ that

$$\frac{1}{2}(\Phi_z(z_\alpha + t) + \Phi_z(z_\alpha - t)) \le \frac{1}{2}(\Phi_z(z_\alpha + t_0) + \Phi_z(z_\alpha - t_0)) \le \Phi_z(z_\alpha) - ct_0^2 \le \Phi_z(z_\alpha) - ct^2/4$$

for $|t| \in (z_\alpha/2, z_\alpha]$ (where $t_0 := z_\alpha/2$). This verifies claim (52) for $|t| \le z_\alpha$. On the other hand, if $|t| \ge z_\alpha$, we have (taking $t > 0$ w.l.o.g.) $\Phi_z(z_\alpha + t) \le 1$ always and $\Phi(z_\alpha - t) \le \Phi_z(0) = 1/2$. Therefore

$$\frac{1}{2}(\Phi_z(z_\alpha + t) + \Phi_z(z_\alpha - t)) \le \frac{1}{2}(1 + 1/2) = 3/4 \le \Phi(z_\alpha).$$

This verifies claim (52) for $|t| > z_\alpha$.

Now, note that $(\widehat{\mathbf{w}} - \mathbf{w}_\star)^\top \mathbf{x}/\sigma_\star(\mathbf{x})$ is symmetric about 0 by our assumption that $\mathbf{x}$ has a symmetric distribution and $\sigma_\star(\mathbf{x}) = \sigma_\star(-\mathbf{x})$. Therefore, we can rewrite and upper bound the coverage using (52):

$$\begin{aligned}
&\text{Coverage}(\widehat{f}) = \mathbb{E}\left[\frac{1}{2}\left(\Phi_z\left(z_\alpha + \frac{(\widehat{\mathbf{w}} - \mathbf{w}_\star)^\top \mathbf{x}}{\sigma_\star(\mathbf{x})}\right) + \Phi_z\left(z_\alpha - \frac{(\widehat{\mathbf{w}} - \mathbf{w}_\star)^\top \mathbf{x}}{\sigma_\star(\mathbf{x})}\right)\right)\right] \\
&\overset{(i)}{\leq} \mathbb{E}\left[\Phi_z(z_\alpha) - c\left(\frac{(\widehat{\mathbf{w}} - \mathbf{w}_\star)^\top \mathbf{x}}{\sigma_\star(\mathbf{x})}\right)^2 \mathbf{1}\left\{\left|\frac{(\widehat{\mathbf{w}} - \mathbf{w}_\star)^\top \mathbf{x}}{\sigma_\star(\mathbf{x})}\right| \leq z_\alpha\right\}\right] \\
&= \alpha - c\mathbb{E}\left[\left(\frac{(\widehat{\mathbf{w}} - \mathbf{w}_\star)^\top \mathbf{x}}{\sigma_\star(\mathbf{x})}\right)^2 \mathbf{1}\left\{\left|\frac{(\widehat{\mathbf{w}} - \mathbf{w}_\star)^\top \mathbf{x}}{\sigma_\star(\mathbf{x})}\right| \leq z_\alpha\right\}\right] \\
&\overset{(ii)}{\leq} \alpha - \frac{c}{\overline{\sigma}^2}\mathbb{E}\Big[\big((\widehat{\mathbf{w}} - \mathbf{w}_\star)^\top \mathbf{x}\big)^2 \mathbf{1}\left\{\left|(\widehat{\mathbf{w}} - \mathbf{w}_\star)^\top \mathbf{x}\right| \leq z_\alpha \underline{\sigma}\right\}\Big] \\
&= \alpha - \frac{c}{\overline{\sigma}^2}\Big((\widehat{\mathbf{w}} - \mathbf{w}_\star)^\top \mathbb{E}[\mathbf{x}\mathbf{x}^\top](\widehat{\mathbf{w}} - \mathbf{w}_\star) - \mathbb{E}\Big[\big((\widehat{\mathbf{w}} - \mathbf{w}_\star)^\top \mathbf{x}\big)^2 \mathbf{1}\left\{\left|(\widehat{\mathbf{w}} - \mathbf{w}_\star)^\top \mathbf{x}\right| > z_\alpha \underline{\sigma}\right\}\Big]\Big) \\
&\overset{(iii)}{\leq} \alpha - \frac{c}{\overline{\sigma}^2}\Bigg(\underline{\gamma}\left\|\widehat{\mathbf{w}} - \mathbf{w}_\star\right\|_2^2 - \underbrace{\mathbb{E}\Big[\big((\widehat{\mathbf{w}} - \mathbf{w}_\star)^\top \mathbf{x}\big)^2 \mathbf{1}\left\{\left|(\widehat{\mathbf{w}} - \mathbf{w}_\star)^\top \mathbf{x}\right| > z_\alpha \underline{\sigma}\right\}\Big]}_{(\star)}\Bigg).
\end{aligned}$$

Above, (i) used (52); (ii) used the bound $\underline{\sigma} \leq \sigma_\star(\mathbf{x}) \leq \overline{\sigma}$; (iii) used the covariance lower bound $\mathbb{E}[\mathbf{x}\mathbf{x}^\top] \succeq \underline{\gamma}\mathbf{I}_d$. Further, letting $r := \left\|\widehat{\mathbf{w}} - \mathbf{w}_\star\right\|_2$, the random variable $(\widehat{\mathbf{w}} - \mathbf{w}_\star)^\top \mathbf{x}$ (with randomness only in $\mathbf{x}$) is Kr^2-sub-Gaussian, since $\mathbf{x}$ is K-sub-Gaussian by our assumption. Therefore the term $(\star)$ can be further upper bounded as

$$\begin{aligned}
(\star) &\leq \left(\mathbb{E}\Big[\big((\widehat{\mathbf{w}} - \mathbf{w}_\star)^\top \mathbf{x}\big)^4\Big] \cdot \mathbb{P}\big(\left|(\widehat{\mathbf{w}} - \mathbf{w}_\star)^\top \mathbf{x}\right| > z_\alpha \underline{\sigma}\big)\right)^{1/2} \\
&\leq \left(CK^2 r^4 \cdot 2\exp(-z_\alpha^2 \underline{\sigma}^2 / Kr^2)\right)^{1/2} \\
&\leq CKr^2 \cdot \exp(-z_\alpha^2 \underline{\sigma}^2 / 2Kr^2) \overset{(i)}{\leq} \frac{1}{2}\underline{\gamma} r^2,
\end{aligned}$$

where (i) happens if $r \leq r_0$ for some $r_0 = r_0(\underline{\gamma}, \underline{\sigma}, K, z_\alpha)$. Plugging this back into the preceding bound yields

$$\text{Coverage}(\widehat{f}) \leq \alpha - \frac{c\underline{\gamma}}{2\overline{\sigma}^2} \cdot r^2 = \alpha - \frac{c\underline{\gamma}}{2\overline{\sigma}^2} \cdot \left\|\widehat{\mathbf{w}} - \mathbf{w}_\star\right\|_2^2$$

for any $\widehat{\mathbf{w}}$ such that $\left\|\widehat{\mathbf{w}} - \mathbf{w}_\star\right\|_2 \leq r_0$. This proves the desired result. □

I. Additional experimental details and ablations

I.1. Simulations

We provide additional details about our simulations in Section B.1. In each problem instance, we generate $(\mathbf{x}_i, y_i)$ from the Gaussian linear model (4): $\mathbf{x}_i \sim \mathsf{N}(\mathbf{0}, \mathbf{I}_d)$, $y_i = \mathbf{w}_\star^\top \mathbf{x}_i + z_i$ where $z_i \overset{\text{iid}}{\sim} P_z = \mathsf{N}(0, 0.25)$. We choose $\left\|\mathbf{w}_\star\right\|_2 = 1$. We run the (sub)-gradient descent algorithm on the full empirical risk $\widehat{R}_n$ (note the risk also depends on the quantile level α) for 50k steps, with initial learning rate 0.01 and a 10x learning rate decay at the 25k-th step. For all our settings (choice of n, d, α), this optimization schedule ensures that the training loss changes by less than 10^{-5} between consecutive iterations at the final iteration.

Each problem instance yields a solution $(\widehat{\mathbf{w}}, \widehat{b})$ which specifies a linear quantile function $\widehat{f}(\mathbf{x}) = \widehat{\mathbf{w}}^\top \mathbf{x} + \widehat{b}$. We evaluate its coverage *exactly* using the closed-form formula (cf. Section C)

$$\text{Coverage}(\widehat{f}) = \mathbb{E}_{G \sim \mathsf{N}(0,1)}\Big[\Phi_z(\left\|\widehat{\mathbf{w}} - \mathbf{w}_\star\right\|_2 G + \widehat{b})\Big].$$

We compute this by using numerical integration (over the gaussian random variable G). The entire set of experiments (for producing Figure 1) is done on a single CPU machine in roughly 6 hours.

I.2. Real data experiments

We provide additional details about our real data experiments in Section B.2 and B.3. All models (linear, MLP, MLP-freeze) in Section B.2 are trained by minimizing the quantile risk (3). We use SGD with momentum 0.9, initial learning rate 10^{-3} for 1500 epochs, and apply a 10x learning rate decay at epoch $\{500, 1000\}$. For each dataset and each random seed, we perform a train-validation split where we use 80% of the data as the train set and 20% of the data as the test set. The coverage of the trained model is evaluated on the test split. For all datasets and all models, we repeat the same experiment across 8 random seeds, and report the mean and standard deviation of the coverage in Table 1.

For our pseudo-label experiments in Section B.3, we train the linear model $\widehat{\mathbf{w}}$ first by minimizing the square loss and using the same optimization schedule above. After $\widehat{\mathbf{w}}$ is learned, we generate the pseudo-labels y_i^{pseudo} using $\widehat{\mathbf{w}}$ and the estimated standard deviation $\widehat{\sigma}$ as described in Section B.3. This is done for both the train and test sets for which we obtain a "pseudo" train set and a "pseudo" test set. We then perform linear quantile regression on these pseudo datasets in a same fashion as in Section B.3.

The experiments for Sections B.2 and B.3 are done on a 8-GPU machine (with Tesla V-100 GPUs) in roughly a day.

Ablations on α Table 3 and 4 report coverage results on the real data with $\alpha \in \{0.8, 0.95\}$ respectively, in the same settings as in Section B.2. These tables also show that under-coverage happens consistently across different datasets and different models, with patterns similar as in Table 1 (which uses $\alpha = 0.9$).

Table 3: Coverage $(\%)$ of quantile regression on real data at nominal level $\alpha = 0.8$. Each entry reports the test-set coverage with mean and std over 8 random seeds. (d, n) denotes the {feature dim, # training examples}.

Dataset	Linear	MLP-3-64	MLP-3-512	MLP-freeze-3-512	d	n
Community	78.25±1.75	66.07±1.48	56.17±2.81	77.45±1.76	100	1599
Bike	79.95±0.66	78.07±1.00	78.66±0.86	79.46±0.83	18	8708
Star	79.97±2.37	72.95±1.83	59.26±1.41	78.42±2.04	39	1728
MEPS_19	80.11±1.12	76.47±0.93	70.04±0.75	79.02±1.28	139	12628
MEPS_20	79.84±0.75	77.11±0.73	71.88±0.87	79.29±0.53	139	14032
MEPS_21	79.57±0.72	74.58±0.70	65.55±0.69	79.29±0.73	139	12524
Nominal (α)	80.00	80.00	80.00	80.00	-	-

Table 4: Coverage $(\%)$ of quantile regression on real data at nominal level $\alpha = 0.95$. Each entry reports the test-set coverage with mean and std over 8 random seeds. (d, n) denotes the {feature dim, # training examples}.

Dataset	Linear	MLP-3-64	MLP-3-512	MLP-freeze-3-512	d	n
Community	93.82±0.98	86.23±1.43	74.38±1.86	93.58±1.33	100	1599
Bike	94.56±0.45	93.77±0.63	93.16±0.80	94.19±0.65	18	8708
Star	94.08±1.73	90.96±1.91	81.58±1.82	93.39±1.68	39	1728
MEPS_19	94.69±0.41	90.71±0.72	85.32±1.23	94.19±0.42	139	12628
MEPS_20	94.84±0.30	92.06±0.43	87.32±0.77	94.58±0.32	139	14032
MEPS_21	94.97±0.34	89.55±0.39	80.70±0.79	94.42±0.29	139	12524
Nominal (α)	95.00	95.00	95.00	95.00	-	-

I.3. License of datasets

The `Community` (com, Accessed: May, 2021) and `Bike` (bik, Accessed: May, 2021) datasets are retrieved from the publicly available UCI machine learning repository (Dua & Graff, 2017) and subject to the license of the repository. The `STAR` dataset (Achilles et al., 2008) is also a public access dataset. The three mediecal expenditure survey datasets `MEPS_19`, `MEPS_20`, `MEPS_21` contain a data use agreement section in their documentation (cf. the "documentation" link in (mep, Accessed: May, 2021a;A;A)) which our use case (train quantile functions and report coverages) comply with. All the datasets are anonymized and to the best of our knowledge do not contain personally identifiable information or offensive contents.