Parallel Inference for Quantile Regression Using Stochastic Subgradient Descent

Zhiyuan Chen

October 3, 2024

Abstract

1 Introduction

2 Background

2.1 Quantile Regression

Quantile regression is a widely used technique in statistics and econometrics. Unlike ordinary least squares (OLS), which estimates the conditional mean of the response variable based on predictor variables, quantile regression estimates conditional quantiles (e.g. median). This makes it a valuable extension of linear regression, allowing for a more comprehensive analysis of the relationships between variables at different points in the response distribution.

Consider that the data $\left\{y_i \in \mathbb{R}, x_i \in \mathbb{R}^d\right\}_{i=1,\dots,n}$ are generated from

$$y_i = x_i^\top \beta^* + \varepsilon_i, \tag{1}$$

where $\beta^* \in \mathbb{R}^d$ is a vector of unknown parameters and the error ε_i satisfies $\mathbb{P}(\varepsilon_i \leq |x_i) = \tau$ of a fixed quantile $\tau \in (0,1)$. Define the check function

$$q(\beta; y_i, x_i) = (y_i - x_i^\top \beta)(\tau - \mathbb{I}\{y_i - x_i^\top \beta \leq 0\}), \tag{2}$$

where $(\cdot)$ is the indicator function. Let

$$Q(\beta) := \mathbb{E}[q(\beta; y_i, x_i)], \tag{3}$$

then β^* is characterized by

$$\beta^* := \arg\min_{\beta \in \mathbb{R}^d} Q(\beta). \tag{4}$$

The standard approach to estimate β^* is the M-estimator proposed by Koenker and Bassett [2]:

$$\hat{\beta}_n := \arg\min_{\beta \in \mathbb{R}^d} \frac{1}{n} \sum_{i=1}^{n} q(\beta; y_i, x_i) \tag{5}$$

2.2 Stochastic Subgradient Descent

To tackle ultra-large quantile regression problems, which scale up to $(n, d) \sim (10^7, 10^3)$, we estimate β^* by stochastic (sub)gradient descent (S-subGD).

Suppose that we have i.i.d. data $\{y_i, x_i, i = 1, \ldots, n\}$, where the ordering of these observations is randomized. S-subGD produces a sequence of estimators, denoted by β_i, a solution path updated as

$$\beta_i = \beta_{i-1} - \gamma_i \nabla q(\beta_{i-1}; x_i, y_i), \tag{6}$$

where $\nabla q(\beta_{i-1}, x_i, y_i)$ is a subgradient of the check function with respect to the $i-1$-th update, and γ_i is a predetermined learning rate. Next, consider the average of the sequence

$$\bar{\beta}_n := \frac{1}{n} \sum_{i=1}^{n} \beta_i, \tag{7}$$

known as the Polyak-Ruppert average. It has been shown that $\sqrt{n}(\bar{\beta}_n - \beta^*)$ is asymptotically normal, and in the remainder of this paper, we will present an alternative method to make inferences about $\bar{\beta}_n$ using parallel computation.

2.3 Inference for Parallel Stochastic Optimization

This method proposed by Zhu et al. [3] focuses on constructing confidence intervals with efficient computation and rapid convergence to the nominal level. It involves performing K parallel runs of a predetermined stochastic algorithm, calculating the sample variance of the linear functional of interest from these K runs, and employing self-normalization to derive asymptotically pivotal t-statistics and the corresponding confidence interval.

Consider a general stochastic algorithm characterized by the update rule h_i at the i-th step and K parallel runs. For the k-th run in parallels $1, \ldots, K$, begining with a proper initialization $\hat{\beta}_0^{(k)}$, the estimate in the k-th run at the i-th iteration is denoted by $\hat{\beta}_i^{(k)}$. The recursive update is

$$\hat{\beta}_i^{(k)} = h_i(\xi_i^{(k)}, \mathcal{F}_{i-1}^{(k)}), \tag{8}$$

where $\mathcal{F}_{i-1}^{(k)} = \sigma(\xi_{i-1}^{(k)}, \xi_{i-2}^{(k)}, \ldots)$ encapsulates information from the previous step, such as $\hat{\beta}_{i-1}^{(k)}, \hat{\beta}_{i-2}^{(k)}, \ldots$, or other intermediate estimates according to the algorithm. For inference at the last step, the results can be aggregated from the K parallel estimates' average and studentization.

3 Methodology

To implement parallel run inference in this offline setting, we randomly divide these data $\left\{y_i \in \mathbb{R}, x_i \in \mathbb{R}^d\right\}_{i=1,\ldots,N}$ into K batches. We define $N = nK$ with N as the total of samples, K as the number of parallel processes, and n as the number of samples processed in each parallel. For the k-th parallel, the i-th update step is given by

$$\beta_i^{(k)} = \beta_{i-1}^{(k)} - \gamma_i \nabla q\left(\beta_{i-1}^{(k)}; y_i^{(k)}, x_i^{(k)}\right), \tag{9}$$

$$\bar{\beta}_i^{(k)} = \bar{\beta}_{i-1}^{(k)} \frac{i-1}{i} + \frac{1}{i} \beta_i^{(k)}. \tag{10}$$

After performing n updates in all K parallels, we average these estimates to obtain the overall sample average

$$\bar{\beta}_{K,n} = \frac{1}{K} \Sigma_{k=1}^{K} \bar{\beta}_n^{(k)}. \tag{11}$$

For any $v \in \mathbb{R}^d$, we consider inference for the linear functional $v^\top x^*$. Define the sample variance $\hat{\sigma}_v^2$ as

$$\hat{\sigma}_v^2 = \frac{1}{K-1} \sum_{k=1}^{K} \left(v^\top \bar{\beta}_n^{(k)} - v^\top \bar{\beta}_{K,n}\right)^2. \tag{12}$$

With the i.i.d. property of $\left\{\bar{\beta}_n^{(k)}\right\}_{k=1,\ldots,K}$, we can studentize $\sqrt{K}\left(v^\top \bar{\beta}_{K,n} - v^\top \beta^*\right)$ with $\hat{\sigma}_v$ to obtain a t-statistic which asymptotically follows a t-distribution

$$\hat{t}_v = \frac{\sqrt{K}\left(v^\top \bar{\beta}_{K,n} - v^\top \beta^*\right)}{\hat{\sigma}_v} \to t_{k-1}. \tag{13}$$

Based on this pivotal t-statistics $\hat{t}_v$, we can construct a $(1-\alpha)\times 100\%$ confidence interval as

$$\mathrm{CI}_v = \left[v^\top \bar{\beta}_{K,n} - \frac{t_{1-\frac{\alpha}{2},K-1}\cdot \hat{\sigma}_v}{\sqrt{K}}, v^\top \bar{\beta}_{K,n} + \frac{t_{1-\frac{\alpha}{2},K-1}\cdot \hat{\sigma}_v}{\sqrt{K}}\right]. \tag{14}$$

Algorithm 1 S-subGD Parallel Inference for Quantile Regression

Input: data $\{y_i, x_i\}_{i=1,\ldots,N}$, check function $q(\beta, y_i, x_i)$, initial learning rate γ_0 and constant a
for $k = 1, \ldots, K$ **do**
  for $i = 1, \ldots, n$ **do**
    Update $\gamma_i = \gamma_0 i^{-a}$
    Update $\beta_i^{(k)} = \beta_{i-1}^{(k)} - \gamma_i \nabla q\left(\beta_{i-1}^{(k)}; y_i^{(k)}, x_i^{(k)}\right)$
    Update $\bar{\beta}_i^{(k)} = \bar{\beta}_{i-1}^{(k)} \frac{i-1}{i} + \frac{1}{i}\beta_i^{(k)}$
  end for
end for
Output:
$\bar{\beta}_{K,n} = \frac{1}{K}\sum_{k=1}^K \bar{\beta}_n^{(k)}$
$\hat{\sigma}_v^2 = \frac{1}{K-1}\sum_{k=1}^K \left(v^\top \bar{\beta}_n^{(k)} - v^\top \bar{\beta}_{K,n}\right)^2$
$\mathrm{CI}_v = \left[v^\top \bar{\beta}_{K,n} - t_{1-\frac{\alpha}{2},K-1}\cdot \hat{\sigma}_v/\sqrt{K},\ v^\top \bar{\beta}_{K,n} + t_{1-\frac{\alpha}{2},K-1}\cdot \hat{\sigma}_v/\sqrt{K}\right]$

4 Asymptotic Theory

Zhu et al. [3] propose a novel Gaussian approximation, from which the asymptotic normality of the average stochastic gradient descent estimator (ASGD) follows as a direct consequence. Based on their assumptions and theorem, we can get the corollary in our situation.

Corollary 1. $\left\{\beta_i^{(k)}\right\}_{i=1,\ldots,n}$ *is a SGD sequence defined by:*

$$\beta_i^{(k)} = \beta_{i-1}^{(k)} - \gamma_i \nabla q\left(\beta_{i-1}^{(k)}; y_i^{(k)}, x_i^{(k)}\right), \quad i = 1, 2, \ldots, n,$$

where $\gamma_i = \gamma_0 \times i^{-a}$ *for some constant* $a \in (1/2, 1)$. *Let* $\bar{\beta}_n^{(k)} = \frac{1}{n}\sum_{i=1}^n \beta_i^{(k)}$. *Under Assumptions 1-3, on a sufficiently rich probability space, there exists a random vector* $W_n \overset{\mathcal{D}}{=} \sqrt{n}\left(\bar{\beta}_n^{(k)} - \beta^*\right)$ *and a centered Gaussian random vector* $Z_n \sim \mathcal{N}(0, \Gamma_n)$, *such that*

$$\mathbb{E}\left|W_n - Z_n\right|^2 \lesssim \max\left(n^{1-2a}, \frac{\log n}{n^{1-2/p}}, \frac{\left|\beta_0 - \beta^*\right|^2}{n}\right),$$

where for $n \geq 1$, *we define*

$$\Gamma_n = \frac{1}{n}\sum_{k=1}^n U_k S U_k^\top,\ U_k = \sum_{i=k}^n Y_k^i \gamma_k,$$

with $Y_k^k = \mathbf{I}_d$, $Y_k^i = \prod_{l=k+1}^i \left(\mathbf{I}_d - \gamma_l \nabla^2 Q\left(\beta^*\right)\right)$, $i > k$.

Corollary 1 demonstrates that $\sqrt{n}\left(\bar{\beta}_n^{(k)} - \beta^*\right)$ converges to normality:

$$\left(\mathbb{E}|\sqrt{n}\left(\bar{\beta}_n^{(k)} - \beta^*\right) - Z_n|^2\right)^{1/2} \lesssim \delta(n),$$

with the approximation rate $\delta(n) = \max\left(n^{1-2a}, \frac{\log n}{n^{1-2/p}}\right) \to 0$. With this corollary, we can show that the statistic in (13) is asymptotically pivotal.

Corollary 2. *Suppose we run Algorithm 3 and Corollary 1 holds. For any v and $\widehat{t}_v$ defined in (13) we have*

$$\sup_{z\in\mathbb{R}} \left|\mathbb{P}\left(\widehat{t}_v \geq z\right) - \mathbb{P}\left(T_{K-1} \geq z\right)\right| \lesssim (\delta(N/K))^{1/4}$$

where T_{K-1} is a random variable following t distribution with degree of freedom $K-1$, N is the total sample size and K is the number of parallel runs. Consequently, for any confidence level $\alpha \in (0,1)$,

$$\left|\frac{\mathbb{P}\left(|\widehat{t}_\nu| \geq t_{1-\alpha/2,K-1}\right)}{\alpha} - 1\right| \lesssim \alpha^{-1}\delta(N/K)^{1/4}$$

where $t_{1-\alpha/2,K-1}$ is the $(1-\alpha/2)\times 100\%$ percentile for the t_{K-1} distribution and the constant in $\lesssim$ does not depend on α. For $\alpha(N)$ goes to zero with $\delta(N/K)^{1/4} \ll \alpha(N)$, the relative error of coverage goes to zero when $\alpha \geq \alpha(N)$, i.e.,

$$\Delta_N = \sup_{\alpha(N)\leq\alpha<1} \left|\frac{\mathbb{P}\left(v^\top x^* \in \widehat{\mathrm{CI}}\right) - (1-\alpha)}{\alpha}\right| \to 0$$

5 Experiments

5.1 Monte Carlo Simulation

The simulation is based on the data generating process (1), where first element of $x_i \in \mathbb{R}^{(d+1)}$ is 1 and the remaining d elements are generated from $N(0,1)$. The error term ε_i is generated from $N(0,1)$, and the true value of parameter β^* is set to be $(1,\ldots,1)$.

We set the total sample size $N = 10^5$, dimension d varies from 10 to 1500. The initial value β_0 is estimated by convolution-type smoothed quantile regression in He et al. [1], with R package *conquer*. The learning rate is set to be $\gamma = \gamma_0 t^{-a}$ with $\gamma_0 = 1$ and $a = 0.501$.

Table 5.1 demonstrates the computation time, the coverage rate, the length of 95% confidence interval and relative error

$$\Delta_\alpha := \left|\frac{\mathbb{P}(v^\top\beta^* \in \widehat{\mathrm{CI}}) - (1-\alpha)}{\alpha}\right| = \left|\frac{\mathbb{P}(v^\top\beta^* \notin \widehat{\mathrm{CI}})}{\alpha} - 1\right|,$$

which can help construct a fair CI with a high level of confidence, i.e., $\alpha \approx 0$.

Table 1: Performance of S-subGD with $N = 10^5$.

d	Time (s)	Coverage rate	CI length	Relative Error
10	0.94	0.9982	0.0784	0.9636
20	0.99	0.9971	0.0732	0.9429
40	1.27	0.9995	0.0815	0.9902
80	1.79	0.9988	0.0839	0.9753
160	2.67	0.9991	0.0892	0.9814
320	4.33	0.9985	0.0983	0.9695
500	6.01	0.9986	0.1208	0.9725
750	8.70	0.9986	0.1393	0.9720
1000	11.37	0.9989	0.1631	0.9772
1200	13.42	0.9987	0.1823	0.9744
1500	18.67	0.9988	0.2147	0.9757

5.2 Inference for The College Wage Premium

To further explore the application of parallel inference for quantile regression, we consider studying the gender gap in the college wage premium. The literature has pointed out a stylized fact that the higher college wage premium for women is the main cause for attracting more women to attend and graduate from colleges than men.

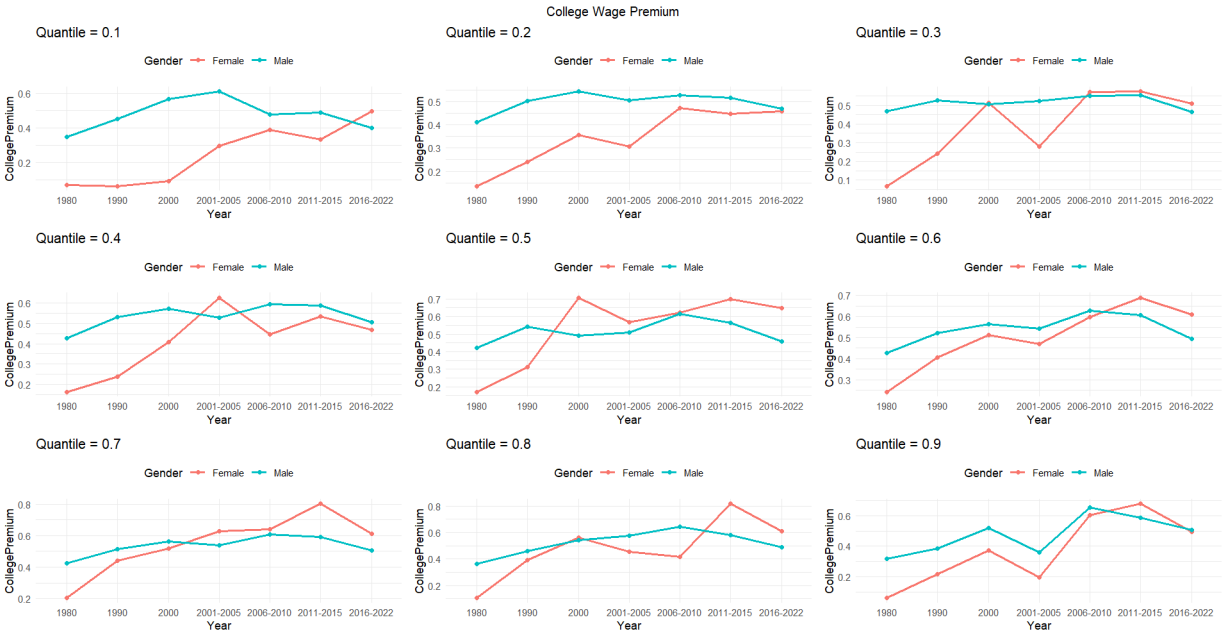

Figure 1: College wage premium of female and male of quantile from 0.1 to 0.9.

References

[1] Xuming He, Xiaoou Pan, Kean Ming Tan, and Wen-Xin Zhou. Smoothed quantile regression with large-scale inference. *Journal of Econometrics*, 232(2):367–388, 2023.

[2] Roger Koenker and Gilbert Bassett. Regression quantiles. *Econometrica*, 46(1):33–50, 1978.

[3] Wanrong Zhu, Zhipeng Lou, Ziyang Wei, and Wei Biao Wu. High confidence level inference is almost free using parallel stochastic optimization, 2024.

A Appendix

In this section, we present the proof of our asymptotic theory, as stated in Corollary 1. We begin by introducing some regularity assumptions, under which Zhu et al. [3] established a Gaussian approximation result, Theorem 1, for the ASGD estimator.

Assumption 1. *$\exists\, l > 0$ and $L > 0$ s.t.*

$$(x - x')^\top (\nabla F(x) - \nabla F(x')) \geq l\|x - x'\|^2,$$
$$\|\nabla F(x) - \nabla F(x')\| \leq L\|x - x'\|.$$

Assumption 2. *Denote $\Delta(x, \xi) = \nabla F(x) - \nabla f(x, \xi)$ for parameter $x \in \mathbb{R}^d$ and data $\xi \sim \Pi$. Given $p > 4$, we have $\mathbb{E}_\xi \|\Delta(x^*, \xi)\|^p < \infty$ and $\exists\, \mathfrak{L} > 0$ s.t. $\forall x, x' \in \mathbb{R}^d$,*

$$(\mathbb{E}_\xi \|\Delta(x, \xi) - \Delta(x', \xi)\|^p)^{1/p} \leq \mathfrak{L}\|x - x'\|$$

Assumption 3. *$\exists\, \mathcal{L}$ s.t. for $x \in \mathbb{R}^d$,*

$$\|\nabla F(x) - \nabla^2 F(x^*)(x - x^*)\| \leq \mathcal{L}\|x - x^*\|^2$$

$F(x)$ is an objective function defined as $F(x) = \mathbb{E}_{\xi\sim\Pi} f(x, \xi)$, where $f(x, \xi)$ represents a noisy measurement of $F(x)$. Assumptions 1-3 are common and mild in the context of convex optimization based on the ASGD algorithm.

Theorem 1. *Assume that $\{x_i\}_{i=1}^n$ is a SGD sequence defined by:*

$$x_i = x_{i-1} - \eta_i \nabla f\left(x_{i-1}, \xi_i\right), \quad i = 1, 2, \ldots, n,$$

where $\eta_i = \eta \times i^{-\beta}$ for some constant $\beta \in (1/2, 1)$. Let $\bar{x}_n = n^{-1}\sum_{i=1}^n x_i$. Under Assumptions 1-3, on a sufficiently rich probability space, there exists a random vector $W_n \overset{\mathcal{D}}{=} \sqrt{n}\left(\bar{x}_n - x^*\right)$ and a centered Gaussian random vector $Z_n \sim \mathcal{N}\left(0, \Gamma_n\right)$, such that

$$\mathbb{E}\left\|W_n - Z_n\right\|^2 \lesssim \max\left(n^{1-2\beta}, \frac{\log n}{n^{1-2/q}}, \frac{\left\|x_0 - x^*\right\|^2}{n}\right),$$

where for $n \geq 1$, we define

$$\Gamma_n = \frac{1}{n}\sum_{k=1}^n U_k S U_k^\top,\ U_k = \sum_{i=k}^n Y_k^i \eta_k,$$

with $Y_k^k = \mathbf{I}_d$, $Y_k^i = \prod_{l=k+1}^i \left(\mathbf{I}_d - \eta_l \nabla^2 F\left(x^*\right)\right)$, $i > k$.

In the context of quantile regression, as long as the check function $q(\beta; y_i, x_i)$ in (2) and $Q(\beta)$ in (3) satisfy Assumptions 1-3, corollary 1 follows directly from Theorem 1.

A.1 Verification of Assumption 1

In the situation of Alsorithm 3, we replace objective function $F(x)$ with $Q(\beta)$ in (3), parameter x with β and ξ with data $\left\{y_i \in \mathbb{R}, x_i \in \mathbb{R}^d\right\}_{i=1,\ldots,n}$ in (1)

$$\begin{aligned}
\nabla Q(\beta) &= \nabla\, \mathbb{E}_{x_i, y_i}\left[q\left(\beta; y_i, x_i\right)\right] \\
&= \mathbb{E}_{x_i, y_i}\left[\nabla q\left(\beta; x_i, x_i\right)\right] \\
&= \mathbb{E}_{x_i, y_i}\left[x_i\left(\mathbb{I}\left\{y_i \leq x_i^\top \beta\right\} - \tau\right)\right] \\
&= \mathbb{E}_{x_i, \varepsilon_i}\left[x_i\left(\mathbb{I}\left\{x_i^\top \beta^* + \varepsilon_i \leq x_i^\top \beta\right\} - \tau\right)\right] \\
&= \mathbb{E}_{x_i}\left\{x_i \cdot \mathbb{E}_{\varepsilon_i | x_i}\left[\mathbb{I}\left\{\varepsilon_i \leq x_i^\top\left(\beta - \beta^*\right)\right\} \,\middle|\, x_i\right] - x_i \tau\right\} \\
&= \mathbb{E}_{x_i}\left\{x_i \cdot F_{\varepsilon_i}\left(x_i^\top\left(\beta - \beta^*\right)\right)\right\} - \tau \mathbb{E}\left(x_i\right),
\end{aligned}$$

where $F_{\varepsilon_i}(z)$ is the CDF of ε_i. Then

$$\begin{aligned}
(\beta-\beta')^\top(\nabla Q(\beta)-\nabla Q(\beta')) &= (\beta-\beta')^\top\left\{\mathbb{E}_{x_i}\left[x_i F_{\varepsilon_i}\left(x_i^\top(\beta-\beta^*)\right)\right]-\mathbb{E}_{x_i}\left[x_i F_{\varepsilon_i}\left(x_i^\top(\beta'-\beta^*)\right)\right]\right\}\\
&= (\beta-\beta')^\top\mathbb{E}_{x_i}\left\{x_i\mathbb{P}\left[x_i^\top(\beta'-\beta^*)\le\varepsilon_i\le x_i^\top(\beta-\beta^*)\right]\right\}\\
&= (\beta-\beta')^\top\mathbb{E}_{x_i}\left[x_i\int_{x_i^\top(\beta'-\beta^*)}^{x_i^\top(\beta-\beta^*)} f_{\varepsilon_i}(z)\,dz\right]\\
&\ge (\beta-\beta')^\top\mathbb{E}_{x_i}\left[x_i\cdot m\cdot x_i^\top(\beta-\beta')\right]\\
&= m\cdot(\beta-\beta')^\top\mathbb{E}\left[x_i x_i^\top\right](\beta-\beta')\\
&= m\cdot\mathrm{tr}\left\{(\beta-\beta')^\top\mathbb{E}\left[x_i x_i^\top\right](\beta-\beta')\right\}\\
&= m\cdot\mathrm{tr}\left\{\mathbb{E}\left[x_i x_i^\top\right]\right\}\|\beta-\beta'\|^2.
\end{aligned}$$

m exists beacause the PDF $f_{\varepsilon_i}(z)$ for ε_i is bounded. Let $l=m\cdot\mathrm{tr}\left\{\mathbb{E}_{x_i}\left[x_i x_i^\top\right]\right\}$, then $\exists\, l>0$ s.t.

$$(\beta-\beta')^\top(\nabla Q(\beta)-\nabla Q(\beta'))\ge l\|\beta-\beta'\|^2.$$

Similarly, there exists M s.t.

$$\begin{aligned}
\|\nabla Q(\beta)-\nabla Q(\beta')\| &= \|\mathbb{E}_{x_i}\left[x_i\int_{x_i^\top(\beta'-\beta^*)}^{x_i^\top(\beta-\beta^*)} f_{\varepsilon_i}(z)\,dz\right]\|\\
&\le \|\mathbb{E}_{x_i}\left[x_i\cdot M\cdot x_i^\top(\beta-\beta')\right]\|\\
&= M\cdot\|\mathbb{E}\left[x_i x_i^\top\right](\beta-\beta')\|\\
&\le M\cdot\|\mathbb{E}\left[x_i x_i^\top\right]\|\cdot\|\beta-\beta'\|
\end{aligned}$$

Let $L=M\cdot\|\mathbb{E}\left[x_i x_i^\top\right]\|$, then $\exists\, L>0$ s.t.

$$\|\nabla Q(\beta)-\nabla Q(\beta')\|\le L\|\beta-\beta'\|.$$

A.2 Verification of Assumption 2

$\nabla Q(\beta)=\mathbb{E}_{x_i,y_i}\nabla q(\beta;y_i,x_i)$, $\Delta(\beta,x_i,y_i)$ is the negative bias of $\nabla q(\beta;y_i,x_i)$, which is

$$\begin{aligned}
\Delta(\beta,x_i,y_i) &= \nabla Q(\beta)-\nabla q(\beta;y_i,x_i)\\
&= \mathbb{E}_{x_i}\left[x_i\cdot F_{\varepsilon_i}\left(x_i^\top(\beta-\beta^*)\right)\right]-\tau\mathbb{E}(x_i)-x_i\left[\mathbb{I}\left(y_i\le x_i^\top\beta\right)-\tau\right]\\
&= \mathbb{E}_{x_i}\left[x_i\cdot F_{\varepsilon_i}\left(x_i^\top(\beta-\beta^*)\right)\right]-x_i\mathbb{I}\left\{\varepsilon_i\le x_i^\top(\beta-\beta^*)\right\}-\tau\mathbb{E}(x_i)+x_i\tau.
\end{aligned}$$

So $\mathbb{E}_{x_i,y_i}\|\Delta(\beta^*,x_i,y_i)\|^p$ is the p-th central moment of $\nabla q(\beta^*;y_i,x_i)$. Given $p>4$, we want to prove that all the p-th central moment exist.

$$\begin{aligned}
\Delta(\beta^*,x_i,y_i) &= \mathbb{E}_{x_i}\left[x_i\cdot F_{\varepsilon_i}(0)\right]-x_i\mathbb{I}\left\{\varepsilon_i\le 0\right\}-\tau\mathbb{E}(x_i)+x_i\tau\\
&= \mathbb{P}\left(\varepsilon_i\le 0\mid x_i\right)\mathbb{E}(x_i)-x_i\mathbb{I}\left\{\varepsilon_i\le 0\right\}-\tau\mathbb{E}(x_i)+x_i\tau\\
&= \tau\mathbb{E}(x_i)-x_i\mathbb{I}\left\{\varepsilon_i\le 0\right\}-\tau\mathbb{E}(x_i)+x_i\tau\\
&= x_i\tau-x_i\mathbb{I}\left\{\varepsilon_i\le 0\right\},\\
\mathbb{E}_{x_i,y_i}\|\Delta(\beta^*,x_i,y_i)\|^p &= \mathbb{E}_{x_i,\varepsilon_i}\|x_i\tau-x_i\mathbb{I}\left\{\varepsilon_i\le 0\right\}\|^p\\
&\le \mathbb{E}_{x_i,\varepsilon_i}\|x_i\|^p\cdot|\tau-\mathbb{I}\{\varepsilon_i\le 0\}|^p\\
&= \mathbb{E}_{x_i}\mathbb{E}_{\varepsilon_i|x_i}\left[\|x_i\|^p\cdot|\tau-\mathbb{I}\left\{\varepsilon_i\le 0\right\}|^p\middle| x_i\right]\\
&= \mathbb{E}_{x_i}\|x_i\|^p\cdot\mathbb{E}_{\varepsilon_i|x_i}|\tau-\mathbb{I}\left\{\varepsilon_i\le 0\mid x_i\right\}|^p\\
&= \mathbb{E}_{x_i}\|x_i\|^p\cdot(|\tau-1|^p\tau+|\tau|^p(1-\tau))\\
&\le \infty,
\end{aligned}$$

since $\tau \in (0,1)$ is a fixed quantile. For any $\beta, \beta' \in \mathbb{R}^d$,

$$
\begin{aligned}
&\left(\mathbb{E}_{x_i,y_i} \left\|\Delta\left(\beta, x_i, y_i\right) - \Delta\left(\beta', x_i, y_i\right)\right\|^p\right)^{\frac{1}{p}} \\
&= \left(\mathbb{E}_{x_i,\varepsilon_i} \left\|\mathbb{E}_{x_i}\left[x_i \mathbb{P}\left(x_i^\top\left(\beta' - \beta^*\right) \le \varepsilon_i \le x_i^\top\left(\beta - \beta^*\right)\right)\right] - x_i \mathbb{I}\left\{\varepsilon_i \le x_i^\top\left(\beta - \beta^*\right)\right\} + x_i \mathbb{I}\left\{\varepsilon_i \le x_i^\top\left(\beta' - \beta^*\right)\right\}\right\|^p\right)^{\frac{1}{p}} \\
&\leqslant \left(\left\|\mathbb{E}_{x_i} x_i \mathbb{P}\left(x_i^\top\left(\beta' - \beta^*\right) \le \varepsilon_i \le x_i^\top\left(\beta - \beta^*\right)\right)\right\|^p\right)^{\frac{1}{p}} + \left(\mathbb{E}_{x_i,\varepsilon_i} \left\|x_i\left[\mathbb{I}\left\{\varepsilon_i \le x_i^\top\left(\beta - \beta^*\right)\right\} - \mathbb{I}\left\{\varepsilon_i \le x_i^\top\left(\beta' - \beta^*\right)\right\}\right]\right\|^p\right)^{\frac{1}{p}}.
\end{aligned}
$$

The first term is

$$
\begin{aligned}
\left(\left\|\mathbb{E}_{x_i} x_i \mathbb{P}\left(x_i^\top\left(\beta' - \beta^*\right) \le \varepsilon_i \le x_i^\top\left(\beta - \beta^*\right)\right)\right\|^p\right)^{\frac{1}{p}} &= \left\|\mathbb{E}_{x_i} x_i \mathbb{P}\left(x_i^\top\left(\beta' - \beta^*\right) \le \varepsilon_i \le x_i^\top\left(\beta - \beta^*\right)\right)\right\| \\
&= \left\|\mathbb{E}_{x_i} x_i \int_{x_i^\top(\beta' - \beta^*)}^{x_i^\top(\beta - \beta^*)} f_{\varepsilon_i}(z)\, dz\right\| \\
&\le \left\|\mathbb{E}_{x_i} x_i \cdot M \cdot x_i^\top(\beta - \beta')\right\| \\
&\le M \|\mathbb{E}\left[x_i x_i^\top\right]\| \cdot \|\beta - \beta'\|.
\end{aligned}
$$

Without loss of generality, assume $x_i^\top(\beta - \beta^*) \ge x_i^\top(\beta' - \beta^*)$. Then the second term equals to

$$
\begin{aligned}
&\left(\mathbb{E}_{x_i,\varepsilon_i} \left\|x_i\left[\mathbb{I}\left\{\varepsilon_i \le x_i^\top(\beta - \beta^*)\right\} - \mathbb{I}\left\{\varepsilon_i \le x_i^\top(\beta' - \beta^*)\right\}\right]\right\|^p\right)^{\frac{1}{p}} \\
&= \left(\mathbb{E}_{x_i} \mathbb{E}_{\varepsilon_i|x_i} \left\|x_i\left[\mathbb{I}\left\{\varepsilon_i \le x_i^\top(\beta - \beta^*)\right\} - \mathbb{I}\left\{\varepsilon_i \le x_i^\top(\beta' - \beta^*)\right\}\right]\right\|^p \,\middle|\, x_i\right)^{\frac{1}{p}} \\
&= \left(\mathbb{E}_{x_i} \|x_i\|^p \,\mathbb{P}\left(x_i^\top(\beta' - \beta^*) \le \varepsilon_i \le x_i^\top(\beta - \beta^*)\right)\right)^{\frac{1}{p}} \\
&\le \left(\mathbb{E}_{x_i} \|x_i\|^p \cdot M \cdot x_i^\top(\beta - \beta')\right)^{\frac{1}{p}} \\
&\le \left(\mathbb{E}_{x_i} \|x_i\|^{p+1} \cdot M \cdot \|\beta - \beta'\|\right)^{\frac{1}{p}} \\
&= \left(\mathbb{E}_{x_i} \|x_i\|^{p+1}\right)^{\frac{1}{p}} M^{\frac{1}{p}} \cdot \|\beta - \beta'\|^{\frac{1}{p}}
\end{aligned}
$$

A.3 Verification of Assumption 3

$$
\begin{aligned}
\nabla^2 Q(\beta) &= \nabla \cdot \nabla Q(\beta) \\
&= \nabla_\beta \mathbb{E}_{x_i}\left[x_i F_{\varepsilon_i}\left(x_i^\top(\beta - \beta^*)\right)\right] \\
&= \mathbb{E}_{x_i}\left[x_i \nabla_\beta F_{\varepsilon_i}\left(x_i^\top(\beta - \beta^*)\right)\right] \\
&= \mathbb{E}_{x_i}\left[x_i x_i^\top f_{\varepsilon_i}\left(x_i^\top(\beta - \beta^*)\right)\right],
\end{aligned}
$$

$$
\begin{aligned}
\|\nabla Q(\beta) - \nabla^2 Q(\beta^*)(\beta - \beta^*)\| &= \left\|\mathbb{E}_{x_i}\left[x_i F_{\varepsilon_i}(x_i^\top(\beta - \beta^*))\right] - \tau \mathbb{E}(x_i) - \mathbb{E}\left[x_i x_i^\top f_{\varepsilon_i}(0)(\beta - \beta^*)\right]\right\| \\
&= \left\|\mathbb{E}_{x_i}\left\{x_i\left[\mathbb{P}(\varepsilon_i \le x_i^\top(\beta - \beta^*)) - \mathbb{P}(\varepsilon_i \le 0)\right]\right\} - \mathbb{E}[x_i x_i^\top] f_{\varepsilon_i}(0)(\beta - \beta^*)\right\| \\
&\le \left\|\mathbb{E}_{x_i}\left[x_i \mathbb{P}(0 \le \varepsilon_i \le x_i^\top(\beta - \beta^*))\right]\right\| + \left\|\mathbb{E}[x_i x_i^\top] f_{\varepsilon_i}(0)(\beta - \beta^*)\right\| \\
&\le \|\mathbb{E}_{x_i}[x_i x_i^\top] \cdot M \cdot x_i^\top(\beta - \beta^*)\| + \|\mathbb{E}[x_i x_i^\top] f_{\varepsilon_i}(0)(\beta - \beta^*)\| \\
&\le \|\mathbb{E}[x_i x_i^\top]\| \cdot (M + f_{\varepsilon_i}(0)) \cdot \|\beta - \beta^*\|
\end{aligned}
$$