Appendix

Table of Contents

A Basic facts

This section collects some basic facts concerning the loss function. First, as we state in Section 2, the pseudo-Huber loss (2.1) exhibits behavior similar to the Huber loss (Huber, 1964), approximating $x^2/2$ when $x^2 \lesssim \tau^2$ and resembling a straight line with slope τ when $x^2 \gtrsim \tau^2$. To see this, some algebra yields

$$\begin{cases} \frac{\epsilon^2-2(1+\epsilon)}{2\epsilon^2}x^2 \le \ell_\tau(x) \le \frac{x^2}{2}, & \text{if } x^2 \le \tau^2\cdot 4(1+\epsilon)/\epsilon^2, \\ \frac{\tau|x|}{1+\epsilon} \le \ell_\tau(x) \le \tau|x|, & \text{if } x^2 > \tau^2\cdot 4(1+\epsilon)/\epsilon^2. \end{cases}$$

Second, we give the first-order derivatives and the Hessian matrix for the empirical loss function. Let $\tau = v\sqrt{n}/z$ throughout the appendix. Recall that our empirical loss function is

$$\begin{aligned} L_n(\mu, v) &= \frac{1}{n}\sum_{i=1}^n \ell(y_i-\mu, v) = \frac{1}{n}\sum_{i=1}^n\left\{\frac{\sqrt{n}}{z}\sqrt{\frac{nv^2}{z^2}+(y_i-\mu)^2} - \left(\frac{n}{z^2}-a\right)v\right\} \\ &= \frac{1}{n}\sum_{i=1}^n\left\{\frac{\sqrt{n}}{z}\left(\sqrt{\tau^2+(y_i-\mu)^2}-\tau\right) + a\cdot\frac{\tau}{z\sqrt{n}}\right\}. \end{aligned}$$

Algorithm 1 An alternating gradient descent algorithm.

Input: $\mu_{\text{init}}, v_{\text{init}}, v_0, V_0, \eta_1, \eta_2, (y_1, \dots, y_n)$
for $k = 0, 1, \dots$ until convergence **do**
$\mu_{k+1} = \mu_k - \eta_1 \nabla_\mu L_n(\mu_k, v_k)$
$\widetilde{v}_{k+1} = v_k - \eta_2 \nabla_\tau L_n(\mu_{k+1}, v_k)$ and $v_{k+1} = \min\{\max\{\widetilde{v}_{k+1}, v_0\}, V_0\}$
end for
Output: $\widehat{\mu} = \mu_{k+1}$, $\widehat{v} = v_{k+1}$

The first-order and second-order derivatives of $L_n(\mu, v)$ are

$$\nabla_\mu L_n(\mu, v) = -\frac{1}{n}\sum_{i=1}^n \frac{y_i - \mu}{v\sqrt{1 + z^2(y_i-\mu)^2/(nv^2)}} = -\frac{\sqrt{n}}{z}\cdot\frac{1}{n}\sum_{i=1}^n \frac{y_i - \mu}{\sqrt{\tau^2 + (y_i-\mu)^2}},$$

$$\nabla_v L_n(\mu, v) = \frac{1}{n}\sum_{i=1}^n \frac{n/z^2}{\sqrt{1 + z^2(y_i-\mu)^2/(nv^2)}} - \left(\frac{n}{z^2} - a\right) = \frac{n}{z^2}\cdot\frac{1}{n}\sum_{i=1}^n \left(\frac{\tau}{\sqrt{\tau^2 + (y_i-\mu)^2}} - 1\right) + a$$

where $a = 1/2$. The Hessian matrix is

$$H(\mu, v) = \begin{bmatrix} \frac{\sqrt{n}}{z}\frac{1}{n}\sum_{i=1}^n \frac{\tau^2}{\left(\tau^2+(y_i-\mu)^2\right)^{3/2}} & \frac{n}{z^2}\frac{1}{n}\sum_{i=1}^n \frac{\tau(y_i-\mu)}{(\tau^2+(y_i-\mu)^2)^{3/2}} \\ \frac{n}{z^2}\frac{1}{n}\sum_{i=1}^n \frac{\tau(y_i-\mu)}{(\tau^2+(y_i-\mu)^2)^{3/2}} & \frac{n^{3/2}}{z^3}\frac{1}{n}\sum_{i=1}^n \frac{(y_i-\mu)^2}{(\tau^2+(y_i-\mu)^2)^{3/2}} \end{bmatrix}.$$

B An alternating gradient descent algorithm

This section presents an alternating gradient descent algorithm to optimize (3.1). The algorithm generates the solution sequence $\{(\mu_k, v_k) : k \geq 0\}$ with the initialization $(\mu_0, v_0) = (\mu_{\text{init}}, v_{\text{init}})$. At the working solution (μ_k, v_k) for any $k \geq 0$, the $(k+1)$-th iteration involves the following two steps:

1. $\mu_{k+1} = \mu_k - \eta_1 \nabla_\mu L_n(\mu_k, v_k)$,
2. $\widetilde{v}_{k+1} = v_k - \eta_2 \nabla_\tau L_n(\mu_{k+1}, v_k)$ and $v_{k+1} = \min\{\max\{\widetilde{v}_{k+1}, v_0\}, V_0\}$,

where η_1 and η_2 are the learning rates and

$$\nabla_\mu L_n(\mu, v) = -\frac{1}{n}\sum_{i=1}^n \frac{y_i - \mu}{v\sqrt{1 + z^2(y_i-\mu)^2/(nv^2)}},$$

$$\nabla_v L_n(\mu, v) = \frac{1}{n}\sum_{i=1}^n \frac{n/z^2}{\sqrt{1 + z^2(y_i-\mu)^2/(nv^2)}} - \left(\frac{n}{z^2} - a\right).$$

The above two steps are repeated until convergence. The algorithm routine is summarized in Algorithm 1. The learning rates η_1 and η_2 can be chosen adaptively in practice. In our experiments, we utilize alternating gradient descent with the Barzilai and Borwein method and backtracking line search.

C Comparing with Lepski's method

We compare our method with Lepski's method. Specifically, we employ Lepski's method to tune the robustification parameter v and, consequently $\tau = v\sqrt{n}/z$, in the empirical pseudo-Huber loss:

$$L_n^h(\mu, v) := \frac{1}{n}\sum_{i=1}^n \left(\tau\sqrt{\tau^2 + (y_i - \mu)^2} - \tau^2\right).$$

Lepski's method proceeds as follows. Let $v_{\max}$ be an upper bound for σ, and $\tau_{\max} = v_{\max}\sqrt{n}/z$ with $z = \sqrt{\log(1/\delta)}$. Let n be sufficiently large. Then with probability at least $1 - \delta$, we have

$$|\widetilde{\mu}(v_{\max}) - \mu^*| \leq 6 v_{\max}\sqrt{\frac{\log(4/\delta)}{n}} =: \epsilon(v_{\max}, \delta),$$

where $\widetilde{\mu}(v_{\max}) = \operatorname{argmin}_{\mu} L_n(\mu, v_{\max})$. Let us by convention set $\epsilon(v_{\max}, 0) = +\infty$. Clearly, $\epsilon(v_{\max}, \delta)$ is homogeneous in the sense that

$$\epsilon(v_{\max}, \delta) = B(\delta) v_{\max}, \quad \text{where } B(\delta) = 6\sqrt{\frac{\log(4/\delta)}{n}}.$$

For some parameters $V \in \mathbb{R}$, $\rho > 1$, and $s \in \mathbb{N}$, we choose the following probability measure $\mathcal{V}$ for $v_{\max}$

$$\mathcal{V}(v_{\max}) = \begin{cases} 1/(2s+1), & \text{if } v_{\max} = V\rho^k,\ k \in \mathbb{Z},\ |k| \leq s, \\ 0, & \text{otherwise.} \end{cases}$$

Let us consider for any $v_{\max}$ such that $\epsilon(v_{\max}, \delta\mathcal{V}(v_{\max})) < \infty$ the confidence interval

$$I(v_{\max}) = \widetilde{\mu}(v_{\max}) + \epsilon(v_{\max}, \delta\, \mathcal{V}(v_{\max})) \times [-1, 1],$$

where

$$\epsilon(v_{\max}, \delta\, \mathcal{V}(v_{\max})) = 6 v_{\max} \sqrt{\frac{\log(4/\delta) + \log(2s+1)}{n}}$$

if $v_{\max} = V\rho^k$ for any $k \in \mathbb{Z}$ and $|k| \leq s$. We set $I(v_{\max}) = \mathbb{R}$ when $\epsilon(v_{\max}, \delta\mathcal{V}(v_{\max})) = +\infty$. Let us consider the non-decreasing family of closed intervals

$$J(v_1) = \bigcap \left\{ I(v_{\max}) : v_{\max} \geq v_1 \right\},\ v_1 \in \mathbb{R}_+.$$

In this definition, we can restrict the intersection to the support of $\mathcal{V}$, since otherwise $I(v_{\max}) = \mathbb{R}$. Lepski's method picks the center point of the intersection

$$\bigcap \{ J(v_1) : v_1 \in \mathbb{R}_+,\ J(v_1) \neq \emptyset \}$$

to be the final estimator $\widehat{\mu}_{\text{Lepski}}$. Then the following result is due to Catoni (2012).

Proposition C.1. Suppose $|\log(\sigma/V)| \leq 2s\log(\rho)$. Then with probability at least $1 - \delta$

$$|\widehat{\mu}_{\text{Lepski}} - \mu^*| \leq 12\rho\sigma\sqrt{\frac{\log(4/\delta) + \log(2s+1)}{n}}.$$

If we take the grid fine enough such that $s = n$, then the upper bound above reduces to

$$12\rho\sigma\sqrt{\frac{\log(4/\delta) + \log(2n+1)}{n}},$$

which agrees with deviation bound for our proposed estimator, up to a constant multiplier. Therefore, our proposed estimator is comparable to Lepski's method in terms of the deviation upper ound. Computationally, our estimator is self-tuned and thus computationally more efficient than Lepski's method; detailed numerical results can be found in Section 4.

D PROOFS FOR SECTION 2

D.1 PROOFS FOR THEOREM 2.3

Proof of Theorem 2.3. We prove first the finite-sample result and then the asymptotic result. Recall that $\tau_* = v_*\sqrt{n}/z$.

Proving the finite-sample result. On one side, if $v_* = 0$ and by the definition of v_*, v_* satisfies

$$1 - \frac{az^2}{n} = \mathbb{E}\frac{\sqrt{n}v_*}{\sqrt{nv_*^2 + z^2\varepsilon^2}} = 0,$$

which is a contradiction. Thus $v_* > 0$. Using the convexity of $1/\sqrt{1+x}$ for $x > -1$ and Jensen's inequality acquires

$$1 - \frac{az^2}{n} = \mathbb{E}\frac{\sqrt{n}v_*}{\sqrt{nv_*^2 + z^2\varepsilon^2}} = \mathbb{E}\frac{1}{\sqrt{1 + z^2\varepsilon^2/(nv_*^2)}} \geq \frac{1}{\sqrt{1 + z^2\sigma^2/(nv_*^2)}} \geq 1 - \frac{z^2\sigma^2}{2nv_*^2},$$

where the last inequality uses the inequality $(1+x)^{-1/2} \geq 1 - x/2$, i.e., Lemma H.4 (i) with $r = -1/2$. This implies

$$v_*^2 \leq \frac{\sigma^2}{2a}.$$

On the other side, using the concavity of $\sqrt{x}$, we obtain, for any $\gamma \in [0, 1)$, that

$$\begin{aligned}
1 - \frac{az^2}{n} &= \mathbb{E}\frac{\sqrt{n}v_*}{\sqrt{nv_*^2 + z^2\varepsilon^2}} = \mathbb{E}\frac{1}{\sqrt{1 + \sigma^2 z^2\varepsilon^2/(nv_*^2)}} \\
&\leq \sqrt{\mathbb{E}\left(\frac{1}{1 + z^2\varepsilon^2/(nv_*^2)}\right)} \\
&\leq \sqrt{\mathbb{E}\left\{\left(1 - (1-\gamma)\frac{z^2\varepsilon^2}{nv_*^2}\right)1\left(\frac{z^2\varepsilon^2}{nv_*^2} \leq \frac{\gamma}{1-\gamma}\right) + \frac{1}{1 + z^2\varepsilon^2/(nv_*^2)}1\left(\frac{z^2\varepsilon^2}{nv_*^2} > \frac{\gamma}{1-\gamma}\right)\right\}} \\
&\leq \sqrt{1 - (1-\gamma)\,\mathbb{E}\left\{\frac{z^2\varepsilon^2}{nv_*^2}1\left(\frac{z^2\varepsilon^2}{nv_*^2} \leq \frac{\gamma}{1-\gamma}\right)\right\}} \\
&\leq \sqrt{1 - (1-\gamma)\,\frac{\mathbb{E}\left\{\varepsilon^2 1\left(\varepsilon^2 \leq \gamma\tau_*^2/(1-\gamma)\right)\right\}}{nv_*^2/z^2}},
\end{aligned} \tag{D.1}$$

where the second inequality uses Lemma D.1, that is,

$$(1+x)^{-1} \leq 1 - (1-\gamma)x, \text{ for any } x \in \left[0, \frac{\gamma}{1-\gamma}\right].$$

Taking square on both sides of inequality (D.1) and using the fact that $n \geq az^2$ together with Lemma H.4 (i) with $r = 2$, aka $(1+x)^2 \geq 1 + 2x$ for $x \geq -1$, we obtain

$$1 - \frac{2az^2}{n} \leq \left(1 - \frac{az^2}{n}\right)^2 \leq 1 - (1-\gamma)\,\frac{\mathbb{E}\{\varepsilon^2 1(\varepsilon^2 \leq \gamma\tau_*^2/(1-\gamma))\}}{nv_*^2/z^2},$$

or equivalently

$$v_*^2 \geq \frac{\sigma^2_{\varphi\tau_*^2}}{2a},$$

where $\varphi = \gamma/(1-\gamma)$. Combining the upper bound and the lower bound for v_*^2 completes the proof for the finite-sample result.

Proving the asymptotic result. The above derivation implies that $v_* < \infty$ for any $a > 0$. By the definition of v_*, we obtain

$$\frac{az^2}{n} = 1 - \mathbb{E}\frac{1}{\sqrt{1 + z^2\varepsilon^2/(nv_*^2)}}. \tag{D.2}$$

We must have $nv_*^2/z^2 \to \infty$. Otherwise assume

$$\limsup_{n\to\infty} nv_*^2/z^2 \leq M < \infty.$$

Taking $n \to \infty$, the left hand side of the above equality goes to 0 while the right hand is lower bounded as

$$1 - \mathbb{E}\frac{1}{\sqrt{1+\varepsilon^2/M}} \geq 1 - \sqrt{\mathbb{E}\left(\frac{1}{1+\varepsilon^2/M}\right)}$$
$$\geq 1 - \sqrt{1 - \frac{\mathbb{E}\left\{\varepsilon^2 1(\varepsilon^2 \leq M)\right\}}{2M}}$$
$$\geq 1 - \sqrt{\frac{1}{2}} > 0,$$

where the first two inequalities follow from the same arguments in deriving (D.1) but with $\gamma = 1/2$, and the third inequality uses the fact that

$$\mathbb{E}\{\varepsilon^2 1(\varepsilon^2 \leq M)\} \leq M.$$

This is a contradiction. Thus $nv_*^2/z^2 \to \infty$. Multiplying both sides of the above equality by n, taking $n \to \infty$, and using the dominated convergence theorem, we obtain

$$az^2 = \lim_{n\to\infty} \mathbb{E}\left(n \cdot \frac{\sqrt{1+z^2\varepsilon^2/(nv_*^2)} - 1}{\sqrt{1+z^2\varepsilon^2/(nv_*^2)}}\right)$$
$$= \lim_{n\to\infty} \mathbb{E}\left(n \cdot \frac{1}{\sqrt{1+z^2\varepsilon^2/(nv_*^2)}} \cdot \frac{\sqrt{1+z^2\varepsilon^2/(nv_*^2)} - 1}{z^2\varepsilon^2/(2nv_*^2)} \cdot \frac{z^2\varepsilon^2}{2nv_*^2}\right)$$
$$= \frac{\mathbb{E}z^2\varepsilon^2}{2\lim_{n\to\infty} v_*^2},$$

and thus $\lim_{n\to\infty} v_*^2 = \sigma^2/(2a)$. This proves the asymptotic result.

□

D.2 Proof of Proposition 2.4

Proof of Proposition 2.4. The convexity proof consists of two steps: (1) proving that $L_n(\mu, v)$ is jointly convex in μ and v; (2) proving that $L_n(\mu, v)$ is strictly convex, provided that there are at least two distinct data points.

To show that $L_n(\mu, v) = n^{-1}\sum_{i=1}^n \ell^{\mathrm{p}}(y_i - \mu, v)$ in (2.5) is jointly convex in μ and v, it suffices to show that each $\ell^{\mathrm{p}}(y_i - \mu, v)$ is jointly convex in μ and v. Recall that $\tau = v\sqrt{n}/z$. The Hessian matrix of $\ell^{\mathrm{p}}(y_i - \mu, v)$ is

$$H_i(\mu, v) = \frac{\sqrt{n}}{z} \cdot \frac{1}{\left(\tau^2 + (y_i - \mu)^2\right)^{3/2}} \begin{bmatrix} \tau^2 & (\sqrt{n}/z)\,\tau(y_i - \mu) \\ (\sqrt{n}/z)\,\tau(y_i - \mu) & (\sqrt{n}/z)^2\,(y_i - \mu)^2 \end{bmatrix} \succeq 0,$$

and thus positive semi-definite. Therefore, $L_n(\mu, v)$ is jointly convex in μ and v.

We proceed to show (2). Because the Hessian matrix $H(\mu, v)$ of $L_n(\mu, v)$ satisfies $H(\mu, v) = n^{-1}\sum_{i=1}^n H_i(\mu, v)$ and each $H_i(\mu, v)$ is positive semi-definite, we only need to show that $H(\mu, v)$ is of full rank. Without generality, assume that $y_1 \neq y_2$. Then

$$H_1(\mu, v) + H_2(\mu, v) = \frac{\sqrt{n}}{z} \cdot \sum_{i=1}^{2} \frac{1}{\left(\tau^2 + (y_i - \mu)^2\right)^{3/2}} \begin{bmatrix} \tau^2 & (\sqrt{n}/z)\,\tau(y_i - \mu) \\ (\sqrt{n}/z)\,\tau(y_i - \mu) & (\sqrt{n}/z)^2\,(y_i - \mu)^2 \end{bmatrix}.$$

Some algebra yields

$$\det\left(H_1(\mu, v) + H_2(\mu, v)\right) = \frac{n^2\tau^2}{z^4} \cdot \frac{(y_1 - y_2)^2}{(\tau^2 + (y_1 - \mu)^2)^{3/2}(\tau^2 + (y_2 - \mu)^2)^{3/2}} \neq 0$$

for any $\tau > 0$ ($v > 0$), and $\mu \in \mathbb{R}$, provided that $y_1 \neq y_2$. Therefore, $H_1(\mu, v) + H_2(\mu, v)$ is of full rank and thus is $H(\mu, \tau)$, provided $v > 0$, $\mu \in \mathbb{R}$, and $y_1 \neq y_2$. □

D.3 SUPPORTING LEMMAS

Lemma D.1. Let $0 \le \gamma < 1$. For any $0 \le x \le \gamma/(1-\gamma)$, we have

$$(1+x)^{-1} \le 1 - (1-\gamma)x.$$

Proof of Lemma D.1. To prove the lemma, it suffices to show, for any $\gamma \in [0,1)$, that

$$1 \le (1+x) - (1-\gamma)x(1+x), \quad \forall\, 0 \le x \le \frac{\gamma}{1-\gamma},$$

which is equivalently to

$$x\left(x - \frac{\gamma}{1-\gamma}\right) \le 0, \quad \forall\, 0 \le x \le \frac{\gamma}{1-\gamma}.$$

The above inequality always holds, and this completes the proof. □

E RESULTS AND PROOFS FOR THE FIXED v CASE

This section presents the theoretical results concerning the minimizer of the empirical penalized pseudo-Huber loss in (2.5) with v fixed, aka Theorem E.2 and Corollary E.4, and their proofs. Corollary E.4 is a rigorous version of the informal result, aka Theorem 2.1, in Section 2.

E.1 RESULTS FOR THE FIXED v CASE

With an abuse of notation, we use $\widehat{\mu}(v)$ to denote the minimizer of the empirical penalized pseudo-Huber loss in (2.5) with v fixed. Recall that we have used $\widehat{\mu}(\tau)$ to denote the minimizer of the empirical pseudo-Huber loss in (2.2), and $\widehat{\mu}(v)$ is equivalent to $\widehat{\mu}(\tau)$ with $\tau = v\sqrt{n}/z$. We begin by examining the theoretical properties of $\widehat{\mu}(v)$. We require the following locally strong convexity assumption, which will be verified later in this subsection.

Assumption E.1 (Locally strong convexity in μ)**.** The empirical Hessian matrix is locally strongly convex with respect to μ such that, for any $\mu \in \mathbb{B}_r(\mu^*) := \{\mu : |\mu - \mu^*| \le r\}$,

$$\inf_{\mu \in \mathbb{B}_r(\mu^*)} \frac{\langle \nabla_\mu L_n(\mu, v) - \nabla_\mu L_n(\mu^*, v), \mu - \mu^* \rangle}{|\mu - \mu^*|^2} \ge \kappa_\ell > 0$$

where $r > 0$ is a local radius parameter.

Theorem E.2. For any $0 < \delta < 1$, let $v > 0$ be fixed and $z^2 = \log(1/\delta)$. Assume Assumption E.1 holds with any $r \ge r_0(\kappa_\ell) := \kappa_\ell^{-1}\left(\sigma/(\sqrt{2}v) + 1\right)^2 \sqrt{\log(2/\delta)/n}$. Then, with probability at least $1-\delta$, we have

$$|\widehat{\mu}(v) - \mu^*| < \frac{1}{\kappa_\ell}\left(\frac{\sigma}{\sqrt{2}v} + 1\right)^2 \sqrt{\frac{\log(2/\delta)}{n}} = \frac{C}{\kappa_\ell}\sqrt{\frac{\log(2/\delta)}{n}},$$

where $C = (\sigma/(\sqrt{2}v) + 1)^2$ only depends on v and σ.

The above theorem states that under the assumption of locally strong convexity, $\widehat{\mu}(v)$ achieves a sub-Gaussian deviation bound when the data have only bounded variances. In particular, if we choose $v = \sigma$ in the theorem, we obtain

$$|\widehat{\mu}(\sigma) - \mu^*| \le \frac{1}{\kappa_\ell}\left(\frac{\sigma}{\sigma} + 1\right)^2 \sqrt{\frac{\log(2/\delta)}{n}} \le \frac{4}{\kappa_\ell}\sqrt{\frac{\log(2/\delta)}{n}}.$$

Assumption E.1 essentially requires the loss function to exhibit curvature in a small neighborhood $\mathbb{B}_r(\mu^*)$, while the penalized loss (2.4) transitions from a quadratic function to a linear function roughly at $|x| = \tau \propto \sqrt{n}$. Quadratic functions always have curvature, so intuitively, Assumption E.1 holds as long as

$$\sqrt{n} \gtrsim r \ge r_0(\kappa_\ell) \propto \sqrt{\frac{1}{n}}.$$

The condition above is automatically guaranteed when n is sufficiently large. Choosing r to be the smallest $r_0(\kappa_\ell)$ results in Assumption E.1 being at its weakest. In other words, in this scenario, the empirical loss function only needs to exhibit curvature in a diminishing neighborhood of μ^*, approximately with a radius of $\sqrt{1/n}$. The following lemma rigorously proves this claim.

Lemma E.3. Suppose $v \geq v_0$. For any $0 < \delta < 1$, let $n \geq C \max\left\{z^2(\sigma^2 + r^2)/v_0^2,\, \log(1/\delta)\right\}$ for some absolute constant C. Then, with probability at least $1 - \delta$, Assumption E.1 with $\kappa_\ell = 1/(2v)$ and any local radius $r \geq r_0(\kappa_\ell) = r_0(1/(2v))$ holds uniformly over $v \geq v_0 > 0$.

The first sample complexity condition, $n \geq Cz^2(\sigma^2 + r^2)/v_0^2$, arises from the requirement that $\tau_{v_0}^2 := v_0^2 n/z^2 \geq C(\sigma^2 + r^2)$. Because the robustification parameter $\tau_{v_0}^2 = v_0^2 n/z^2$ determines the size of the quadratic region, this requirement is minimal in the sense that Assumption E.1 can hold only when τ_v^2 is larger than r^2 plus the noise variance σ^2. As argued before, Assumption E.1 holds with any r such that $\sqrt{n} \gtrsim r \gtrsim \sqrt{1/n}$. For example, we can take $r \propto \sigma$ to be a constant, and this will not worsen the sample complexity condition. Finally, by combining Lemma E.3 and Theorem E.2, we obtain the following result.

Corollary E.4. Suppose $v \geq v_0$. For any $0 < \delta < 1$, let $n \geq C \max\left\{(r^2 + \sigma^2)/v_0^2, 1\right\} \log(1/\delta)$ for some universal constant C, where $r \geq 2r_0(1/(2v))$. Take $z^2 = \log(1/\delta)$. Then, for any $v \geq v_0$, with probability at least $1 - \delta$, we have

$$|\widehat{\mu}(v) - \mu^*| \leq 2v\left(\frac{\sigma}{\sqrt{2}v} + 1\right)^2 \sqrt{\frac{\log(4/\delta)}{n}} \lesssim v\sqrt{\frac{1 + \log(1/\delta)}{n}}.$$

This section collects proofs for Theorem E.2, Lemma E.3, and Corollary E.4. Recall that $\tau = v\sqrt{n}/z$, and the gradients with respect to μ and v are

$$\nabla_\mu L_n(\mu, v) = -\frac{1}{n}\sum_{i=1}^n \frac{y_i - \mu}{v\sqrt{1 + z^2(y_i - \mu)^2/(nv^2)}} = -\frac{\sqrt{n}}{z} \cdot \frac{1}{n}\sum_{i=1}^n \frac{y_i - \mu}{\sqrt{\tau^2 + (y_i - \mu)^2}},$$

$$\nabla_v L_n(\mu, v) = \frac{1}{n}\sum_{i=1}^n \frac{n/z^2}{\sqrt{1 + z^2(y_i - \mu)^2/(nv^2)}} - \left(\frac{n}{z^2} - a\right) = \frac{n}{z^2} \cdot \frac{1}{n}\sum_{i=1}^n \left(\frac{\tau}{\sqrt{\tau^2 + (y_i - \mu)^2}} - 1\right) + a.$$

E.2 Proof of Theorem E.2

Proof of Theorem E.2. Because $\widehat{\mu}(v)$ is the stationary point of $L_n(\mu, v)$, we have

$$\frac{\partial}{\partial \mu} L_n(\widehat{\mu}(v), v) = -\frac{1}{n}\sum_{i=1}^n \frac{y_i - \widehat{\mu}(v)}{v\sqrt{1 + z^2(y_i - \widehat{\mu}(v))^2/(nv^2)}} = -\frac{\sqrt{n}}{z} \cdot \frac{1}{n}\sum_{i=1}^n \frac{y_i - \widehat{\mu}(v)}{\sqrt{\tau^2 + (y_i - \widehat{\mu}(v))^2}} = 0.$$

Let $\Delta = \widehat{\mu}(v) - \mu$. We first assume that $|\Delta| := |\widehat{\mu}(v) - \mu^*| \leq r_0 \leq r$. Using Assumption E.1 obtains

$$\begin{aligned}
\kappa_\ell |\widehat{\mu}(v) - \mu^*|^2 &\leq \left\langle \frac{\partial}{\partial \mu} L_n(\widehat{\mu}(v), v) - \frac{\partial}{\partial \mu} L_n(\mu^*, v), \widehat{\mu}(v) - \mu^* \right\rangle \\
&\leq \left| \frac{1}{\sqrt{n}} \sum_{i=1}^n \frac{\varepsilon_i}{z\sqrt{\tau^2 + \varepsilon_i^2}} \right| |\widehat{\mu}(v) - \mu^*|,
\end{aligned}$$

or equivalently

$$\kappa_\ell |\widehat{\mu}(v) - \mu^*| \leq \left| \frac{1}{\sqrt{n}} \sum_{i=1}^n \frac{\varepsilon_i}{z\sqrt{\tau^2 + \varepsilon_i^2}} \right|.$$

Applying Lemma E.5 with the fact that $\left|\mathbb{E}\left(\tau\varepsilon_i/(\tau^2 + \varepsilon_i^2)^{1/2}\right)\right| \leq \sigma^2/(2\tau)$, we obtain with probability at least $1 - 2\delta$ that

$$\kappa_\ell |\widehat{\mu}(v) - \mu^*| \leq \left| \frac{\sqrt{n}}{\tau} \frac{1}{n} \sum_{i=1}^n \frac{\tau \varepsilon_i}{z\sqrt{\tau^2 + \varepsilon_i^2}} \right| \leq \frac{\sqrt{n}}{z\tau}\left(\sigma\sqrt{\frac{2\log(1/\delta)}{n}} + \frac{\tau \log(1/\delta)}{3n} + \frac{\sigma^2}{2\tau}\right),$$

or equivalently

$$\kappa_\ell|\widehat{\mu}(v)-\mu^*| \le \sqrt{\frac{2\log(1/\delta)}{z^2\tau^2/\sigma^2}} + \frac{\log(1/\delta)}{3z\sqrt{n}} + \frac{\sqrt{n}\sigma^2}{2z\tau^2}.$$

Since $\tau = v\sqrt{n}/z$, we have

$$\kappa_\ell|\widehat{\mu}(v)-\mu^*| \le \left(\frac{\sqrt{2}\sigma}{v} + \frac{\sqrt{\log(1/\delta)}}{3z}\right)\sqrt{\frac{\log(1/\delta)}{n}} + \frac{1}{2}\cdot\frac{\sigma^2}{v^2}\cdot\frac{z}{\sqrt{n}}.$$

Taking $z = \sqrt{\log(1/\delta)}$ then yields

$$\begin{aligned}
\kappa_\ell|\widehat{\mu}(v)-\mu^*| &\le \left(\frac{\sqrt{2}\sigma}{v} + \frac{\sqrt{\log(1/\delta)}}{3\sqrt{\log(1/\delta)}}\right)\sqrt{\frac{\log(1/\delta)}{n}} + \frac{1}{2}\cdot\frac{\sigma^2}{v^2}\cdot\sqrt{\frac{\log(1/\delta)}{n}} \\
&\le \left(\frac{\sqrt{2}\sigma}{v} + \frac{1}{3} + \frac{1}{2}\cdot\frac{\sigma^2}{v^2}\right)\sqrt{\frac{\log(1/\delta)}{n}} \\
&< \left(1 + \frac{\sigma}{\sqrt{2}v}\right)^2\sqrt{\frac{\log(1/\delta)}{n}}
\end{aligned}$$

for any $\delta \in (0, 1/2)$. Moving κ_ℓ to the right hand side and using a change of variable $2\delta \to \delta$, we obtain

$$\begin{aligned}
|\widehat{\mu}(v)-\mu^*| &< \frac{1}{\kappa_\ell}\cdot\left(1 + \frac{\sigma}{\sqrt{2}v}\right)^2\sqrt{\frac{\log(2/\delta)}{n}} \\
&= r_0 \le r.
\end{aligned}$$

This completes the proof, provided that $|\Delta| \le r_0$.

Lasty, we show that $|\Delta| \le r_0$ must hold. If not, we shall construct an intermediate solution between μ^* and $\widehat{\mu}(v)$, denoted by $\mu_\eta = \mu^* + \eta(\widehat{\mu}(v)-\mu^*)$, such that $|\mu_\eta - \mu^*| = r_0$. Specifically, we can choose some $\eta \in (0, 1)$ such that $|\mu_\eta - \mu^*| = r_0$. We then repeat the above calculation and obtain

$$\begin{aligned}
|\widehat{\mu}(v)-\mu^*| &\le \frac{1}{\kappa_\ell}\cdot\left(\frac{\sqrt{2}\sigma}{v} + \frac{1}{3} + \frac{1}{2}\cdot\frac{\sigma^2}{v^2}\right)\sqrt{\frac{\log(2/\delta)}{n}} \\
&< r_0 = \frac{1}{\kappa_\ell}\cdot\left(1 + \frac{\sigma}{\sqrt{2}v}\right)^2\sqrt{\frac{\log(2/\delta)}{n}}
\end{aligned}$$

which is a contradiction. Therefore, it must hold that $|\Delta| \le r_0$. □

E.3 PROOF OF LEMMA E.3

Proof of Lemma E.3. We first prove that, with probability at least $1-\delta$, Assumption E.1 with $\kappa_\ell = 1/(2v)$ and radius r holds for any fixed $v \ge v_0$. Recall that $\tau = v\sqrt{n}/z$. For notational simplicity, let $\Delta = \mu - \mu^*$ and $\tau_{v_0} = v_0\sqrt{n}/z$. It follows that

$$\begin{aligned}
\langle \nabla_\mu L_n(\mu, v) - \nabla_\mu L_n(\mu^*, v),\, \Delta\rangle &= \left\langle \frac{1}{\sqrt{n}}\sum_{i=1}^n \frac{\varepsilon_i}{z\sqrt{\tau^2+\varepsilon_i^2}} - \frac{1}{\sqrt{n}}\sum_{i=1}^n \frac{y_i-\mu}{z\sqrt{\tau^2+(y_i-\mu)^2}},\, \Delta\right\rangle \\
&= \frac{1}{\sqrt{n}}\sum_{i=1}^n \frac{\tau^2}{z(\tau^2+(y_i-\widetilde{\mu})^2)^{3/2}}\,\Delta^2,
\end{aligned}$$

where $\widetilde{\mu}$ is some convex combination of μ^* and μ, that is, $\widetilde{\mu} = (1-\lambda)\mu^* + \lambda\mu$ for some $\lambda \in [0, 1]$. Obviously, we have $|\widetilde{\mu} - \mu^*| = \lambda|\Delta| \le |\Delta| \le r$. Since $(y_i - \widetilde{\mu})^2 \le 2\varepsilon_i^2 + 2\lambda^2\Delta^2 \le 2\varepsilon_i^2 + 2\Delta^2 \le$

$2\varepsilon_i^2 + 2r^2$ the above displayed equality implies that, with probability at least $1-\delta$,

$$
\begin{aligned}
&\inf_{\mu\in\mathbb{B}_r(\mu^*)} \frac{\langle \nabla_\mu L_n(\mu,v) - \nabla_\mu L_n(\mu^*,v), \mu-\mu^*\rangle}{|\mu-\mu^*|^2} \\
&\geq \frac{\sqrt{n}}{z}\cdot\frac{1}{n}\sum_{i=1}^n \frac{\tau^2}{(\tau^2+2r^2+2\varepsilon_i^2)^{3/2}} \\
&= \frac{\sqrt{n}}{z}\cdot\frac{\tau^2}{(\tau^2+2r^2)^{3/2}}\cdot\frac{1}{n}\sum_{i=1}^n \frac{(\tau^2+2r^2)^{3/2}}{(\tau^2+2r^2+2\varepsilon_i^2)^{3/2}} \\
&\geq \frac{\sqrt{n}}{z}\cdot\frac{\tau^2}{(\tau^2+2r^2)^{3/2}}\cdot\left(\mathbb{E}\frac{(\tau_{v_0}^2+2r^2)^{3/2}}{(\tau_{v_0}^2+2r^2+2\varepsilon_i^2)^{3/2}} - \sqrt{\frac{\log(1/\delta)}{2n}}\right) \\
&= \frac{\sqrt{n}}{z}\cdot\frac{\tau^2}{(\tau^2+2r^2)^{3/2}}\cdot\left(\mathrm{I} - \sqrt{\frac{\log(1/\delta)}{2n}}\right),
\end{aligned}
\tag{E.1}
$$

where the last inequality uses Lemma E.6.

It remains to lower bound I. Using the convexity of $1/(1+x)^{3/2}$ and Jensen's inequality, we obtain

$$
\begin{aligned}
\frac{1}{n}\sum_{i=1}^n \mathbb{E}\frac{(\tau_{v_0}^2+2r^2)^{3/2}}{(\tau_{v_0}^2+2r^2+2\varepsilon_i^2)^{3/2}} &= \mathbb{E}\frac{(\tau_{v_0}^2+2r^2)^{3/2}}{(\tau_{v_0}^2+2r^2+2\varepsilon_i^2)^{3/2}} \\
&= \mathbb{E}\frac{1}{(1+2\varepsilon_i^2/(\tau_{v_0}^2+2r^2))^{3/2}} \\
&\geq \frac{1}{(1+2\sigma^2/(\tau_{v_0}^2+2r^2))^{3/2}} \\
&= \frac{(\tau_{v_0}^2+2r^2)^{3/2}}{(\tau_{v_0}^2+2r^2+2\sigma^2)^{3/2}}.
\end{aligned}
$$

Plugging the above lower bound into (E.1) and using the facts

$$
\frac{\tau^3}{(\tau^2+2r^2)^{3/2}} \geq \frac{\tau_{v_0}^3}{(\tau_{v_0}^2+2r^2)^{3/2}} \quad \text{for } \tau_{v_0}\geq\tau \quad \text{and} \quad \frac{\tau^3}{(\tau^2+2r^2)^{3/2}} \leq 1,
$$

we obtain with probability at least $1-\delta$

$$
\begin{aligned}
&\inf_{\mu\in\mathbb{B}_r(\mu^*)} \frac{\langle \nabla_\mu L_n(\mu) - \nabla_\mu L_n(\mu^*), \mu-\mu^*\rangle}{|\mu-\mu^*|^2} \\
&\geq \frac{\sqrt{n}}{z}\cdot\frac{\tau^2}{(\tau^2+2r^2)^{3/2}}\cdot\left(\frac{(\tau_{v_0}^2+2r^2)^{3/2}}{(\tau_{v_0}^2+2r^2+2\sigma^2)^{3/2}} - \sqrt{\frac{\log(1/\delta)}{2n}}\right) \\
&\geq \frac{\sqrt{n}}{z\tau}\cdot\frac{\tau^3}{(\tau^2+2r^2)^{3/2}}\cdot\left(\frac{(\tau_{v_0}^2+2r^2)^{3/2}}{(\tau_{v_0}^2+2r^2+2\sigma^2)^{3/2}} - \sqrt{\frac{\log(1/\delta)}{2n}}\right) \\
&= \frac{\sqrt{n}}{z\tau}\left(\frac{\tau^3}{(\tau^2+2r^2)^{3/2}}\cdot\frac{(\tau_{v_0}^2+2r^2)^{3/2}}{(\tau_{v_0}^2+2r^2+2\sigma^2)^{3/2}} - \frac{\tau^3}{(\tau^2+2r^2)^{3/2}}\cdot\sqrt{\frac{\log(1/\delta)}{2n}}\right) \\
&\geq \frac{\sqrt{n}}{z\tau}\left(\frac{1}{(1+(2r^2+2\sigma^2)/\tau_{v_0}^2)^{3/2}} - \sqrt{\frac{\log(1/\delta)}{2n}}\right) \\
&= \frac{1}{v}\left(\frac{1}{(1+(2r^2+2\sigma^2)/\tau_{v_0}^2)^{3/2}} - \sqrt{\frac{\log(1/\delta)}{2n}}\right) \\
&\geq \frac{1}{2v}
\end{aligned}
$$

provided $\tau_{v_0}^2 \geq 4r^2+4\sigma^2$ and $n \geq C\log(1/\delta)$ for some large enough absolute constant C.

Lastly, the above result holds uniformly over $v \geq v_0$ with probability at least $1 - \delta$ since the probability event does not depend on v.

□

E.4 Proof of Corollary E.4

Proof of Corollary E.4. Recall $z = \sqrt{\log(1/\delta)}$ and

$$r \geq 2v \left(\frac{\sigma}{\sqrt{2v}} + 1 \right)^2 \sqrt{\frac{\log(2/\delta)}{n}}.$$

If $n \geq C \max\left\{(r^2 + \sigma^2)/v_0^2, 1\right\} \log(1/\delta)$, which is guaranteed by the conditions of the corollary, then Lemma E.3 implies that, with probability at least $1 - \delta$, Assumption E.1 holds with $\kappa_\ell = 1/(2v)$ and radius r uniformly over $v \geq v_0$. Denote this probability event by $\mathcal{E}$. If Assumption E.1 holds, then by Theorem E.2, we have

$$\mathbb{P}\left(|\widehat{\mu}(v) - \mu^*| \leq 2v \left(\frac{\sigma}{\sqrt{2v}} + 1 \right)^2 \sqrt{\frac{\log(2/\delta)}{n}} \,\middle|\, \mathcal{E} \right) \geq 1 - \delta.$$

Thus

$$\begin{aligned}
&\mathbb{P}\left(|\widehat{\mu}(v) - \mu^*| > 2v \left(\frac{\sigma}{\sqrt{2v}} + 1 \right)^2 \sqrt{\frac{\log(2/\delta)}{n}} \right) \\
&= \mathbb{P}\left(|\widehat{\mu}(v) - \mu^*| > 2v \left(\frac{\sigma}{\sqrt{2v}} + 1 \right)^2 \sqrt{\frac{\log(2/\delta)}{n}}, \mathcal{E} \right) \\
&\quad + \mathbb{P}\left(|\widehat{\mu}(v) - \mu^*| > 2v \left(\frac{\sigma}{\sqrt{2v}} + 1 \right)^2 \sqrt{\frac{\log(2/\delta)}{n}}, \mathcal{E}^c \right) \\
&\leq \mathbb{P}\left(|\widehat{\mu}(v) - \mu^*| > 2v \left(\frac{\sigma}{\sqrt{2v}} + 1 \right)^2 \sqrt{\frac{\log(2/\delta)}{n}} \,\middle|\, \mathcal{E} \right) + \mathbb{P}\left(\mathcal{E}^c\right) \\
&\leq 2\delta.
\end{aligned}$$

Then with probability at least $1 - 2\delta$, we have

$$|\widehat{\mu}(v) - \mu^*| \leq 2v \left(\frac{\sigma}{\sqrt{2v}} + 1 \right)^2 \sqrt{\frac{\log(2/\delta)}{n}}.$$

Using a change of variable $2\delta \to \delta$ finishes the proof. □

E.5 Supporting lemmas

This subsection collects two supporting lemmas that are used earlier in this section.

Lemma E.5. Let ε_i be i.i.d. random variables such that $\mathbb{E}\varepsilon_i = 0$ and $\mathbb{E}\varepsilon_i^2 = 1$. For any $0 < \delta < 1$, with probability at least $1 - 2\delta$, we have

$$\left| \frac{1}{n} \sum_{i=1}^n \frac{\tau \varepsilon_i}{\sqrt{\tau^2 + \varepsilon_i^2}} - \mathbb{E} \frac{\tau \varepsilon_i}{\sqrt{\tau^2 + \varepsilon_i^2}} \right| \leq \sigma \sqrt{\frac{2\log(1/\delta)}{n}} + \frac{\tau \log(1/\delta)}{3n}.$$

Proof of Lemma E.5. The random variables $Z_i := \tau\psi_\tau(\varepsilon_i) = \tau\varepsilon_i/(\tau^2+\varepsilon_i^2)^{1/2}$ with $\mu_z = \mathbb{E}Z_i$ and $\sigma_z^2 = \mathrm{var}(Z_i)$ are bounded i.i.d. random variables such that

$$\begin{aligned}
|Z_i| &= \left|\tau\varepsilon_i/(\tau^2+\varepsilon_i^2)^{1/2}\right| \le |\varepsilon_i| \wedge \tau \le \tau,\\
|\mu_z| = |\mathbb{E}Z_i| &= \left|\mathbb{E}\left(\tau\varepsilon_i/(\tau^2+\varepsilon_i^2)^{1/2}\right)\right| \le \frac{\sigma^2}{2\tau},\\
\mathbb{E}Z_i^2 &= \mathbb{E}\left(\frac{\tau^2\varepsilon_i^2}{\tau^2+\varepsilon_i^2}\right) \le \sigma^2,\\
\sigma_z^2 := \mathrm{var}(Z_i) &= \mathbb{E}\big(\tau\varepsilon_i/(\tau^2+\varepsilon_i^2)^{1/2} - \mu_z\big)^2\\
&= \mathbb{E}\left(\frac{\tau^2\varepsilon_i^2}{\tau^2+\varepsilon_i^2}\right) - \mu_z^2 \le \sigma^2.
\end{aligned}$$

For third and higher order absolute moments, we have

$$\mathbb{E}|Z_i|^k = \mathbb{E}\left|\frac{\tau\varepsilon_i}{\sqrt{\tau^2+\varepsilon_i^2}}\right|^k \le \sigma^2\tau^{k-2} \le \frac{k!}{2}\sigma^2(\tau/3)^{k-2}, \text{ for all integers } k \ge 3.$$

Using Lemma H.2 with $v = n\sigma^2$ and $c = \tau/3$, we have for any $t > 0$

$$\mathbb{P}\left(\left|\sum_{i=1}^n \frac{\tau\varepsilon_i}{\sqrt{\tau^2+\varepsilon_i^2}} - \sum_{i=1}^n \mathbb{E}\frac{\tau\varepsilon_i}{\sqrt{\tau^2+\varepsilon_i^2}}\right| \ge \sqrt{2n\sigma^2 t} + \frac{\tau t}{3}\right) \le 2\exp(-t).$$

Taking $t = \log(1/\delta)$ acquires that for any $0 < \delta < 1$

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^n \frac{\tau\varepsilon_i}{\sqrt{\tau^2+\varepsilon_i^2}} - \frac{1}{n}\sum_{i=1}^n \mathbb{E}\frac{\tau\varepsilon_i}{\sqrt{\tau^2+\varepsilon_i^2}}\right| \le \sigma\sqrt{\frac{2\log(1/\delta)}{n}} + \frac{\tau\log(1/\delta)}{3n}\right) \ge 1 - 2\delta.$$

This completes the proof. □

Lemma E.6. For any $0 < \delta < 1$, with probability at least $1-\delta$,

$$\frac{1}{n}\sum_{i=1}^n \frac{\tau^3}{(\tau^2+\varepsilon_i^2)^{3/2}} - \mathbb{E}\frac{\tau^3}{(\tau^2+\varepsilon_i^2)^{3/2}} \ge -\sqrt{\frac{\log(1/\delta)}{2n}}.$$

Moreover, with probability at least $1-\delta$, it holds uniformly over $\tau \ge \tau_{v_0} \ge 0$ that

$$\frac{1}{n}\sum_{i=1}^n \frac{\tau^3}{(\tau^2+\varepsilon_i^2)^{3/2}} \ge \mathbb{E}\frac{\tau_{v_0}^3}{(\tau_{v_0}^2+\varepsilon_i^2)^{3/2}} - \sqrt{\frac{\log(1/\delta)}{2n}}.$$

Proof of Lemma E.6. The random variables $Z_i = Z_i(\tau) := \tau^3/(\tau^2+\varepsilon_i^2)^{3/2}$ with $\mu_z = \mathbb{E}Z_i$ and $\sigma_z^2 = \mathrm{var}(Z_i)$ are bounded i.i.d. random variables such that

$$0 \le Z_i = \tau^3/(\tau^2+\varepsilon_i^2)^{3/2} \le 1.$$

Therefore, using Lemma H.1 with $v = n$ acquires that for any $t > 0$

$$\mathbb{P}\left(\sum_{i=1}^n \frac{\tau^3}{(\tau^2+\varepsilon_i^2)^{3/2}} - \sum_{i=1}^n \mathbb{E}\left(\frac{\tau^3}{(\tau^2+\varepsilon_i^2)^{3/2}}\right) \le -\sqrt{\frac{nt}{2}}\right) \le \exp(-t).$$

Taking $t = \log(1/\delta)$ acquires that for any $0 < \delta < 1$

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^n \frac{\tau^3}{(\tau^2+\varepsilon_i^2)^{3/2}} - \frac{1}{n}\sum_{i=1}^n \mathbb{E}\left(\frac{\tau^3}{(\tau^2+\varepsilon_i^2)^{3/2}}\right) > -\sqrt{\frac{\log(1/\delta)}{2n}}\right) > 1 - \delta.$$

The second result follows from the fact that $Z_i(\tau)$ is an increasing function of τ. Specifically, we have with probability at least $1-\delta$

$$
\begin{aligned}
\frac{1}{n}\sum_{i=1}^n \frac{\tau^3}{(\tau^2+\varepsilon_i^2)^{3/2}} &\geq \frac{1}{n}\sum_{i=1}^n \frac{\tau_{v_0}^3}{(\tau_{v_0}^2+\varepsilon_i^2)^{3/2}} \\
&\geq \mathbb{E}\left(\frac{\tau_{v_0}^3}{(\tau_{v_0}^2+\varepsilon_i^2)^{3/2}}\right) + \frac{1}{n}\sum_{i=1}^n \frac{\tau_{v_0}^3}{(\tau_{v_0}^2+\varepsilon_i^2)^{3/2}} - \mathbb{E}\left(\frac{\tau_{v_0}^3}{(\tau_{v_0}^2+\varepsilon_i^2)^{3/2}}\right) \\
&\geq \mathbb{E}\left(\frac{\tau_{v_0}^3}{(\tau_{v_0}^2+\varepsilon_i^2)^{3/2}}\right) - \sqrt{\frac{\log(1/\delta)}{2n}}.
\end{aligned}
$$

This finishes the proof.

□

F PROOFS FOR THE SELF-TUNED CASE

This section collects the proofs for Theorems 3.1 and 3.2.

F.1 PROOF OF THEOREM OF 3.1

Proof of Theorem of 3.1. Recall that $\tau = v\sqrt{n}/z$. For simplicity, let $\widehat{\tau} = \widehat{v}\sqrt{n}/z$. Define the profile loss $L_n^{\text{pro}}(v)$ as

$$
L_n^{\text{pro}}(v) := L_n(\widehat{\mu}(v), v) = \min_{\mu} L_n(\mu, v).
$$

Then it is convex and its first-order gradient is

$$
\nabla L_n^{\text{pro}}(v) = \nabla L_n(\widehat{\mu}(v), v) = \frac{\partial}{\partial v}\widehat{\mu}(v)\cdot \frac{\partial}{\partial v}L_n(\mu, v)\Big|_{\mu=\widehat{\mu}(v)} + \frac{\partial}{\partial v}L_n(\mu, v)\Big|_{\mu=\widehat{\mu}(v)} = \frac{\partial}{\partial v}L_n(\widehat{\mu}(v), v), \tag{F.1}
$$

where we use the fact that $\partial/\partial\mu\, L_n(\mu, v)|_{\mu=\widehat{\mu}(v)} = 0$, implied by the stationarity of $\widehat{\mu}(v)$.

Assuming that the constraint is inactive. We first assume that the constraint is not active for any stationary point $\widehat{v}$, that is, any stationary point $\widehat{v}$ is an interior point of $[v_0, V_0]$, aka $\widehat{v} \in (v_0, V_0)$. By the joint convexity of $L_n(\mu, v)$ and the convexity of $L_n^{\text{pro}}(v)$, $(\widehat{\mu}(\widehat{v}), \widehat{v})$ and $\widehat{v}$ are stationary points of $L_n(\mu, v)$ and $L_n(\widehat{\mu}(v), v)$, respectively. Thus we have

$$
\begin{aligned}
\frac{\partial}{\partial \mu}L_n(\mu, v)\Big|_{(\mu,v)=(\widehat{\mu}(\widehat{v}),\widehat{v})} &= -\frac{\sqrt{n}}{z}\cdot\frac{1}{n}\sum_{i=1}^n \frac{y_i - \widehat{\mu}(\widehat{v})}{\sqrt{\widehat{\tau}^2 + (y_i-\widehat{\mu}(\widehat{v}))^2}} = 0, \\
\frac{\partial}{\partial v}L_n(\mu, v)\Big|_{(\mu,v)=(\widehat{\mu}(\widehat{v}),\widehat{v})} &= \frac{n}{z^2}\cdot\frac{1}{n}\sum_{i=1}^n \frac{\widehat{\tau}}{\sqrt{\widehat{\tau}^2 + (y_i-\widehat{\mu}(\widehat{v}))^2}} - \left(\frac{n}{z^2} - a\right) = 0, \\
\nabla L_n^{\text{pro}}(v)\Big|_{v=\widehat{v}} = \nabla L_n(\widehat{\mu}(\widehat{v}), \widehat{v})\Big|_{v=\widehat{v}} &= \frac{\partial}{\partial v}L_n(\widehat{\mu}(v), v)\Big|_{v=\widehat{v}} = \frac{\partial}{\partial v}L_n(\mu, v)\Big|_{(\mu,v)=(\widehat{\mu}(\widehat{v}),\widehat{v})} = 0,
\end{aligned}
$$

where the first two equalities are on partial derivatives of $L_n(\mu, v)$ and the last one is on the derivative of the profile loss $L_n^{\text{pro}}(v) \equiv L_n(\widehat{\mu}(v), v)$.

Recall that $\tau = \sqrt{n}v/z$. Let $f(\tau) = z^2\nabla L_n^{\text{pro}}(v)/n$, that is,

$$
f(\tau) = \frac{1}{n}\sum_{i=1}^n \frac{\tau}{\sqrt{\tau^2 + (y_i - \widehat{\mu}(v))^2}} - \left(1 - \frac{az^2}{n}\right).
$$

In other words, $\widehat{\tau} = \sqrt{n}\widehat{v}/z$ satisfies $f(\widehat{\tau}) = 0$. Assuming that the conststraint is inactive, we split the proof into two steps.

Step 1: Proving $\widehat{v} \le C_0\sigma$ for some universal constant C_0. We will employ the method of proof by contradiction. Assume there exists some v such that

$$v > (1+\epsilon)\sqrt{r^2+\sigma^2} \quad \text{and} \quad \nabla L_v^{\text{pro}}(v) = 0;$$

or equivalently, there exists some τ such that

$$\tau > (1+\epsilon)\sqrt{r^2+\sigma^2}\sqrt{n}/z =: \bar{\tau} \ \text{ and } \ f(\tau) = 0, \tag{F.2}$$

where ϵ and r are to be determined later. Let $\tau_{v_0} = v_0\sqrt{n}/z$. Then, provided n is large enough, Lemma E.3 implies that Assumption E.1 with $\kappa_\ell = 1/(2v)$ and local radius $r \ge r_0(\kappa_\ell)$ holds uniformly over $v \ge v_0$ conditional on the following event

$$\mathcal{E}_1 := \left\{ \frac{1}{n}\sum_{i=1}^n \frac{(\tau_{v_0}^2 + 2r^2)^{3/2}}{(\tau_{v_0}^2 + 2r^2 + 2\varepsilon_i^2)^{3/2}} - \frac{1}{n}\sum_{i=1}^n \mathbb{E}\frac{(\tau_{v_0}^2 + 2r^2)^{3/2}}{(\tau_{v_0}^2 + 2r^2 + 2\varepsilon_i^2)^{3/2}} \ge -\sqrt{\frac{\log(1/\delta)}{2n}} \right\}.$$

Conditional on the intersection of event $\mathcal{E}_1$ and the following event

$$\mathcal{E}_2 := \left\{ \sup_{v\in[v_0,V_0]} \left| \frac{1}{n}\sum_{i=1}^n \frac{\varepsilon_i}{\sqrt{\tau^2+\varepsilon_i^2}} \right| \le C\cdot\frac{V_0}{v_0}\cdot\frac{\log(n/\delta)}{n} \right\},$$

where $z \lesssim \sqrt{\log(n/\delta)}$ and C is some constant, and following the proof of Theorem E.2, for any fixed v and thus fixed $\tau = v\sqrt{n}/z$, we have

$$\kappa_\ell|\widehat{\mu}(v) - \mu^*| \le \left| \frac{1}{\sqrt{n}}\sum_{i=1}^n \frac{\varepsilon_i}{z\sqrt{\tau^2+\varepsilon_i^2}} \right|.$$

Thus, for any v such that $v_0 \vee \bar{v}_0 := v_0 \vee (1+\epsilon)\sqrt{r^2+\sigma^2} < v < V_0$, we have on $\mathcal{E}_2$ that

$$\begin{aligned}
\sup_{v_0\vee\bar{v}_0<v<V_0} \kappa_\ell(v)\,|\widehat{\mu}(v)-\mu^*| &\le \sup_{v\in[v_0,V_0]} \kappa_\ell(v)\,|\widehat{\mu}(v)-\mu^*| \\
&\le \sup_{v\in[v_0,V_0]} \left| \frac{1}{\sqrt{n}}\sum_{i=1}^n \frac{\varepsilon_i}{z\sqrt{\tau^2+\varepsilon_i^2}} \right| \\
&\le C\cdot\frac{V_0}{v_0}\cdot\frac{\log(n/\delta)}{z\sqrt{n}},
\end{aligned}$$

which, by Lemma E.3, yields

$$\sup_{v\in[v_0,V_0]} |\widehat{\mu}(v)-\mu^*| \le 2C\cdot\frac{V_0^2}{v_0}\cdot\frac{\log(n/\delta)}{z\sqrt{n}} =: r. \tag{F.3}$$

The above r can be further refined by using the finer lower bound $\bar{v}_0$ of v instead of v_0, but we use v_0 for simplicity. Let $\Delta = \mu^* - \widehat{\mu}(v)$, and we have $|\Delta| \le r$. Let the event $\mathcal{E}_3$ be

$$\mathcal{E}_3 := \left\{ \frac{1}{n}\sum_{i=1}^n \frac{\sqrt{\bar{\tau}^2 + 2(r^2+\varepsilon_i^2)} - \bar{\tau}}{\sqrt{\bar{\tau}^2 + 2(r^2+\varepsilon_i^2)}} - \mathbb{E}\left(\frac{\sqrt{\bar{\tau}^2 + 2(r^2+\varepsilon_i^2)} - \bar{\tau}}{\sqrt{\bar{\tau}^2 + 2(r^2+\varepsilon_i^2)}} \right) \le \sqrt{\frac{\log(1/\delta)2(r^2+\sigma^2)}{n\bar{\tau}^2}} + \frac{\log(1/\delta)}{3n} \right\}.$$

Thus on the event $\mathcal{E}_1 \cap \mathcal{E}_2 \cap \mathcal{E}_3$ and using the fact that $1 - 1/\sqrt{1+x}$ is an increasing function, we have

$$
\begin{aligned}
f(\tau) &= \frac{az^2}{n} - \frac{1}{n}\sum_{i=1}^{n} \frac{\sqrt{\tau^2 + (\Delta+\varepsilon_i)^2} - \tau}{\sqrt{\tau^2+(\Delta+\varepsilon_i)^2}} \geq \frac{az^2}{n} - \frac{1}{n}\sum_{i=1}^{n} \frac{\sqrt{\tau^2 + 2(r^2+\varepsilon_i^2)} - \tau}{\sqrt{\tau^2+2(r^2+\varepsilon_i^2)}} \\
&> \frac{az^2}{n} - \frac{1}{n}\sum_{i=1}^{n} \frac{\sqrt{\bar{\tau}^2 + 2(r^2+\varepsilon_i^2)} - \bar{\tau}}{\sqrt{\bar{\tau}^2+2(r^2+\varepsilon_i^2)}} && (\tau < \bar{\tau}) \\
&\geq \frac{az^2}{n} - \left\{ \mathbb{E}\left(\frac{\sqrt{\bar{\tau}^2 + 2(r^2+\varepsilon_i^2)} - \bar{\tau}}{\sqrt{\bar{\tau}^2+2(r^2+\varepsilon_i^2)}} \right) + \frac{1}{n}\sum_{i=1}^{n} \frac{\sqrt{\bar{\tau}^2 + 2(r^2+\varepsilon_i^2)} - \bar{\tau}}{\sqrt{\bar{\tau}^2+2(r^2+\varepsilon_i^2)}} - \mathbb{E}\left(\frac{\sqrt{\bar{\tau}^2 + 2(r^2+\varepsilon_i^2)} - \bar{\tau}}{\sqrt{\bar{\tau}^2+2(r^2+\varepsilon_i^2)}} \right) \right\} \\
&\geq \frac{az^2}{n} - \left(\frac{r^2+\sigma^2}{\bar{\tau}^2} + \sqrt{\frac{\log(1/\delta)\cdot 2(r^2+\sigma^2)}{n\bar{\tau}^2}} + \frac{\log(1/\delta)}{3n} \right) \\
&= \frac{z^2}{n}\left(a - \frac{\log(1/\delta)}{3z^2} \right) - \left(\frac{r^2+\sigma^2}{r^2+\sigma^2}\frac{z^2}{(1+\epsilon)^2 n} + \sqrt{\frac{r^2+\sigma^2}{r^2+\sigma^2}\frac{2z^2\log(1/\delta)}{(1+\epsilon)^2 n^2}} \right) \\
& && (\text{Definition of } \bar{\tau}) \\
&\geq \frac{(a-1/3)z^2}{n} - \left(\frac{r^2+\sigma^2}{r^2+\sigma^2}\frac{z^2}{(1+\epsilon)^2 n} + \sqrt{\frac{r^2+\sigma^2}{r^2+\sigma^2}\frac{2z^4}{(1+\epsilon)^2 n^2}} \right) && (z^2 \geq \log(1/\delta)) \\
&\geq \frac{(a-1/3)z^2}{n} - \frac{z^2}{n}\cdot\left(\frac{1}{(1+\epsilon)^2} + \sqrt{\frac{2}{(1+\epsilon)^2}} \right) \\
&= \frac{z^2}{n}\left(a - \frac{1}{3} - \frac{1}{(1+\epsilon)^2} - \sqrt{\frac{2}{(1+\epsilon)^2}} \right) \\
&\geq 0,
\end{aligned}
$$

provided that

$$
\frac{1}{1+\epsilon} \leq \frac{\sqrt{1+2(a-1/3)} - 1}{\sqrt{2}},
$$

or equivalently

$$
\epsilon \geq \frac{\sqrt{4a+2/3} + 2/3 + \sqrt{2} - 2a}{2(a-1/3)} =: \epsilon(a).
$$

In other words, conditional on the event $\mathcal{E}_1 \cap \mathcal{E}_2 \cap \mathcal{E}_3$ and taking $\epsilon \geq \epsilon(a)$, $f(\tau) > 0$ for $\tau > \bar{\tau} := (1+\epsilon)\sqrt{r^2+\sigma^2}\sqrt{n}/z$. This contradicts with (F.2), and thus

$$
\widehat{\tau} \leq (1+\epsilon)\sqrt{r^2+\sigma^2}\sqrt{n}/z.
$$

If $a = 1/2$ and conditional on the same event, the above holds with

$$
\epsilon = 9 \geq \epsilon\,(1/2)\,.
$$

If n is large enough such that $12\sigma \geq 10\sqrt{r^2+\sigma^2}$, then conditional on the event $\mathcal{E}_1 \cap \mathcal{E}_2 \cap \mathcal{E}_3$, we have

$$
v_0 \leq \widehat{v} \leq C_0\sigma,
$$

where $C_0 = 12$.

Step 2: Proving $\widehat{v} \geq c_0 \left(\frac{\sigma_{\tau_{v_0}^2/2-1}}{\sigma_{\tau_{v_0}^2/2}} \wedge 1 \right) \sigma_{\tau_{v_0}^2-1}$ for some universal constant c_0. We will again employ the method of proof by contradiction. Let

$$
g(\tau) := \left(\frac{1}{n}\sum_{i=1}^{n} \frac{\tau^2}{\sqrt{\tau^2+(\Delta+\varepsilon_i)^2}} \right)^2 - \left(1 - \frac{az^2}{n} \right)^2.
$$

Assume there exists some v such that

$$v < c \quad \text{and} \quad \frac{\partial}{\partial v} L_n(\widehat{\mu}(v), v) = 0;$$

or equivalently, assume there exists some τ such that

$$\tau < c\sqrt{n}/z =: \underline{\tau} \;\text{ and }\; g(\tau) = 0. \tag{F.4}$$

It is impossible that $c \leq v_0$ because any stationary point v is in (v_0, V_0). Thus $c > v_0$. Let $\Delta = \widehat{\mu}(v) - \mu^*$. Then on the event $\mathcal{E}_1 \cap \mathcal{E}_2$, using the facts that $\sqrt{x}$ is a concave function and $1/\sqrt{1 + y/x}$ is an increasing function of x, we have

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^{n} \frac{\tau^2}{\sqrt{\tau^2 + (\Delta + \varepsilon_i)^2}} &= \frac{1}{n}\sum_{i=1}^{n} \frac{1}{\sqrt{1 + (\Delta + \varepsilon_i)^2/\tau^2}} \\
&\leq \frac{1}{n}\sum_{i=1}^{n} \frac{1}{\sqrt{1 + (\Delta + \varepsilon_i)^2/\underline{\tau}^2}} \\
&\leq \sqrt{\frac{1}{n}\sum_{i=1}^{n} \frac{1}{1 + (\Delta + \varepsilon_i)^2/\underline{\tau}^2}} \\
&\leq \sqrt{\frac{1}{n}\sum_{i=1}^{n} \frac{1}{1 + \underline{\tau}^{-2}(\Delta + \varepsilon_i)^2 \cdot 1\left((\Delta + \varepsilon_i)^2 \leq \underline{\tau}^2\right)}} \\
&\leq \sqrt{1 - \frac{1}{n} \cdot \frac{1}{2\underline{\tau}^2} \sum_{i=1}^{n} (\Delta + \varepsilon_i)^2 \cdot 1\left((\Delta + \varepsilon_i)^2 \leq \underline{\tau}^2\right)}.
\end{aligned}$$

By the proof from step 1, we have on the event $\mathcal{E}_1 \cap \mathcal{E}_2$ that

$$\sup_{v \in [v_0, V_0]} |\widehat{\mu}(v) - \mu^*| \leq r,$$

where r is defined in (F.3). Then

$$\begin{aligned}
g(\tau) &\leq 1 - \frac{1}{n} \cdot \frac{1}{2\underline{\tau}^2} \sum_{i=1}^{n} (\Delta + \varepsilon_i)^2 \cdot 1\left((\Delta + \varepsilon_i)^2 \leq \underline{\tau}^2\right) - \left(1 - \frac{az^2}{n}\right)^2 \\
&< \frac{2az^2}{n} - \frac{1}{n} \cdot \frac{1}{2\underline{\tau}^2} \sum_{i=1}^{n} (\Delta + \varepsilon_i)^2 \cdot 1\left((\Delta + \varepsilon_i)^2 \leq \underline{\tau}^2\right) \qquad \text{(as long as } az^2/n > 0\text{)} \\
&\leq \frac{2az^2}{n} - \frac{1}{n} \cdot \frac{1}{2\underline{\tau}^2} \sum_{i=1}^{n} \left(\varepsilon_i^2 + 2\Delta\varepsilon_i\right) \cdot 1\left(\varepsilon_i^2 \leq \frac{\underline{\tau}^2}{2} - r^2\right) \\
&\leq \frac{2az^2}{n} - \frac{1}{2\underline{\tau}^2} \left(\frac{1}{n}\sum_{i=1}^{n} \varepsilon_i^2 1\left(\varepsilon_i^2 \leq \frac{\underline{\tau}^2}{2} - r^2\right) - \frac{2}{n}\sum_{i=1}^{n} r|\varepsilon_i| 1\left(\varepsilon_i^2 \leq \frac{\underline{\tau}^2}{2} - r^2\right)\right) \\
&= \frac{2az^2}{n} - \frac{1}{2\underline{\tau}^2} \left(\mathrm{I} - 2r \cdot \mathrm{II}\right).
\end{aligned}$$

Define the probability event $\mathcal{E}_4$ as

$$\mathcal{E}_4 := \mathcal{E}_{41} \cap \mathcal{E}_{42},$$

where

$$\mathcal{E}_{41} =: \left\{\frac{1}{n}\sum_{i=1}^{n} \varepsilon_i^2 1\left(\varepsilon_i^2 \leq \frac{\underline{\tau}^2}{2} - r^2\right) \geq \mathbb{E}\varepsilon_i^2 1\left(\varepsilon_i^2 \leq \frac{\underline{\tau}^2}{2} - r^2\right) - \sigma_{\frac{\underline{\tau}^2}{2}} \sqrt{\frac{\underline{\tau}^2 \log(1/\delta)}{n}} - \frac{\underline{\tau}^2 \log(1/\delta)}{6n}\right\} \quad \text{and}$$

$$\mathcal{E}_{42} =: \left\{\frac{1}{n}\sum_{i=1}^{n} |\varepsilon_i| 1\left(\varepsilon_i^2 \leq \frac{\underline{\tau}^2}{2} - r^2\right) \leq \mathbb{E}|\varepsilon_i| 1\left(\varepsilon_i^2 \leq \frac{\underline{\tau}^2}{2} - r^2\right) + \sqrt{\frac{2\sigma^2_{\underline{\tau}^2/2} \log(1/\delta)}{n}} + \frac{\underline{\tau} \log(1/\delta)}{3\sqrt{2}n}\right\}.$$

If n is sufficiently large such that

$$r^2 \le \epsilon_0 \lesssim \left(\frac{\log n + \log(1/\delta)}{z\sqrt{n}}\right)^2 \le 1 \quad \text{and}$$
$$\frac{r}{\underline{\tau}^2}\left(\sigma^2_{\underline{\tau}^2/2} + \sqrt{\frac{2\sigma^2_{\underline{\tau}^2/2}\log(1/\delta)}{n}} + \frac{\underline{\tau}\log(1/\delta)}{3\sqrt{2}n}\right) \le \frac{1}{12}\frac{\log(1/\delta)}{n},$$

then conditional on $\mathcal{E}_4$, we have

$$\mathrm{I} \ge \mathbb{E}\varepsilon_i^2 1\left(\varepsilon_i^2 \le \frac{\underline{\tau}^2}{2} - r^2\right) - \sigma_{\frac{\underline{\tau}^2}{2}}\sqrt{\frac{\underline{\tau}^2\log(1/\delta)}{n}} - \frac{\underline{\tau}^2\log(1/\delta)}{6n} \quad \text{and}$$
$$\mathrm{II} \le \mathbb{E}|\varepsilon_i| 1\left(\varepsilon_i^2 \le \frac{\underline{\tau}^2}{2} - r^2\right) + \sqrt{\frac{2\sigma^2_{\underline{\tau}^2/2}\log(1/\delta)}{n}} + \frac{\underline{\tau}\log(1/\delta)}{3\sqrt{2}n}.$$

Thus conditional on $\mathcal{E}_4$ we have

$$\begin{aligned}
g(\tau) &< \frac{2az^2}{n} - \frac{1}{2\underline{\tau}^2}(\mathrm{I} - 2r\cdot\mathrm{II})\\
&\le \frac{2az^2}{n} - \frac{1}{2\underline{\tau}^2}\left(\mathbb{E}\varepsilon_i^2 1\left(\varepsilon_i^2 \le \frac{\underline{\tau}^2}{2} - r^2\right) - \sigma_{\underline{\tau}^2/2}\sqrt{\frac{\underline{\tau}^2\log(1/\delta)}{n}} - \frac{\underline{\tau}^2\log(1/\delta)}{6n}\right)\\
&\qquad + \frac{r}{\underline{\tau}^2}\left(\mathbb{E}|\varepsilon_i| 1\left(\varepsilon_i^2 \le \frac{\underline{\tau}^2}{2} - r^2\right) + \sqrt{\frac{2\sigma^2_{\underline{\tau}^2/2}\log(1/\delta)}{n}} + \frac{\underline{\tau}\log(1/\delta)}{3\sqrt{2}n}\right)\\
&\le \frac{2az^2}{n} - \frac{\sigma^2_{\underline{\tau}^2/2-\epsilon_0}}{2\underline{\tau}^2} + \frac{\sigma_{\underline{\tau}^2/2}\sqrt{\log(1/\delta)}}{2\underline{\tau}\sqrt{n}} + \frac{\log(1/\delta)}{12n} + \frac{r}{\underline{\tau}^2}\left(\sigma^2_{\underline{\tau}^2/2} + \sqrt{\frac{2\sigma^2_{\underline{\tau}^2/2}\log(1/\delta)}{n}} + \frac{\underline{\tau}\log(1/\delta)}{3\sqrt{2}n}\right)\\
&\le \frac{z^2}{n}\left(2a + \frac{\log(1/\delta)}{z^2}\cdot\frac{1}{6}\right) - \frac{\sigma^2_{\underline{\tau}^2/2-\epsilon_0}}{2\underline{\tau}^2} + \frac{\sigma_{\underline{\tau}^2/2}\sqrt{\log(1/\delta)}}{2\underline{\tau}\sqrt{n}}\\
&= \frac{z^2}{2n}\left(4a + \frac{\log(1/\delta)}{z^2}\cdot\frac{1}{3} - \frac{\sigma^2_{\underline{\tau}^2/2-\epsilon_0}}{c^2} + \frac{\sigma_{\underline{\tau}^2/2}}{c}\cdot\frac{\sqrt{\log(1/\delta)}}{z}\right) && (\underline{\tau} = c\sqrt{n}/z)\\
&\le \frac{z^2}{2n}\left(4a + \frac{1}{3} - \frac{\sigma^2_{\underline{\tau}^2/2-\epsilon_0}}{c^2} + \frac{\sigma_{\underline{\tau}^2/2}}{c}\right) && (z^2 \ge \log(1/\delta))\\
&\le 0,
\end{aligned}$$

for any c such that

$$c \le \frac{\sigma_{\underline{\tau}^2/2}}{2(4a+1/3)}\left(\sqrt{1 + \frac{4(4a+1/3)\sigma^2_{\underline{\tau}^2/2-\epsilon_0}}{\sigma^2_{\underline{\tau}^2/2}}} - 1\right),$$

In other words, conditional on the event $\mathcal{E}_1 \cap \mathcal{E}_2 \cap \mathcal{E}_4$ and taking any c satisfying the above inequality, we have

$$g(\tau) < 0 \text{ for any } \tau < \underline{\tau} = c\sqrt{n}/z.$$

This is a contradiction. Thus, $\widehat{\tau} \ge \underline{\tau} = c\sqrt{n}/z$, or equivalently $\widehat{v} \ge c > v_0$. Using the inequality

$$\sqrt{1+x} - 1 \ge 1(x \ge 3) + \frac{x}{3}1(0 \le x < 3) \ge \frac{x}{3} \wedge 1 \quad \forall\, x \ge 0,$$

we obtain

$$
\begin{aligned}
&\frac{\sigma_{\underline{\tau}^2/2}}{2(4a+1/3)}\left(\sqrt{1+\frac{4(4a+1/3)\sigma^2_{\underline{\tau}^2/2-\epsilon_0}}{\sigma^2_{\underline{\tau}^2/2}}}-1\right)\\
&=\frac{3\sigma_{\tau_{v_0}^2/2}}{14}\left(\sqrt{1+\frac{28\sigma^2_{\underline{\tau}^2/2-\epsilon_0}}{3\sigma^2_{\underline{\tau}^2/2}}}-1\right) && (a=1/2)\\
&\geq \frac{3\sigma_{\underline{\tau}^2/2}}{14}\left(\frac{28\sigma^2_{\underline{\tau}^2/2-\epsilon_0}}{9\sigma^2_{\underline{\tau}^2/2}}\wedge 1\right)\\
&=\frac{2\sigma^2_{\underline{\tau}^2/2-\epsilon_0}}{3\sigma_{\underline{\tau}^2/2}}\wedge\frac{3\sigma_{\underline{\tau}^2/2}}{14}\\
&\geq \frac{1}{5}\left(\frac{\sigma_{\underline{\tau}^2/2-1}}{\sigma_{\underline{\tau}^2/2}}\wedge 1\right)\sigma_{\underline{\tau}^2/2-1}\\
&\geq \frac{1}{5}\left(\frac{\sigma_{\tau^2_{v_0}/2-1}}{\sigma_{\tau^2_{v_0}/2}}\wedge 1\right)\sigma_{\tau^2_{v_0}/2-1}.
\end{aligned}
$$

Therefore we can take $c = 5^{-1}(\sigma_{\tau^2_{v_0}/2-1}/\sigma_{\tau^2_{v_0}/2}\wedge 1)\sigma_{\tau^2_{v_0}/2-1}$. Thus on the event $\mathcal{E}_1\cap\mathcal{E}_2\cap\mathcal{E}_4$, we have

$$\widehat{v}\geq c := c_0\left(\frac{\sigma_{\tau^2_{v_0}/2-1}}{\sigma_{\tau^2_{v_0}/2}}\wedge 1\right)\sigma_{\tau^2_{v_0}/2-1},$$

where $c_0 = 1/5$ is a universal constant. This finishes the proof of step 2.

Proving that the constraint is inactive. If $\widehat{v}\notin(v_0,V_0)$, then $\widehat{v}\in\{v_0,V_0\}$. Suppose $\widehat{v}=v_0$, then $\widehat{v}=v_0<c$. Recall that $\tau_{v_0}=v_0\sqrt{n}/z$. Then we must have $f(\tau_{v_0})\geq 0$, and thus $g(\tau_{v_0})\geq 0$. However, conditional on the probability event $\mathcal{E}_1\cap\mathcal{E}_2\cap\mathcal{E}_4$, repeating the above analysis in step 2 obtains $g(\tau_{v_0})<0$. This is a contradiction. Therefore $\widehat{v}\neq v_0$. Similarly, conditional on probability event $\mathcal{E}_1\cap\mathcal{E}_2\cap\mathcal{E}_3$, we can obtain $\widehat{v}\neq V_0$. Therefore, conditional on the probability event $\mathcal{E}_1\cap\mathcal{E}_2\cap\mathcal{E}_3\cap\mathcal{E}_4$, the constraint must be inactive, aka $\widehat{v}\in(v_0,V_0)$.

Using the first result of Lemma E.6 with τ^2 and ε_i^2 replaced by $\tau^2_{v_0}+2r^2$ and $2\varepsilon_i^2$ respectively, Lemma F.1, Lemma F.2 with τ^2 and w_i^2 replaced by $\bar{\tau}^2$ and $2(r^2+\varepsilon_i^2)$ respectively, and Lemma F.3, we obtain

$$\mathbb{P}(\mathcal{E}_1)\geq 1-\delta,\ \mathbb{P}(\mathcal{E}_2)\geq 1-\delta,\ \mathbb{P}(\mathcal{E}_3)\geq 1-\delta,\ \mathbb{P}(\mathcal{E}_4)\geq 1-2\delta,$$

and thus

$$\mathbb{P}(\mathcal{E}_1\cap\mathcal{E}_2\cap\mathcal{E}_3\cap\mathcal{E}_4)\geq 1-5\delta.$$

Putting the above results together, and using Lemmas F.1 and F.3, we obtain with probability at least $1-5\delta$ that

$$c_0(\sigma_{\tau^2_{v_0}/2-1}/\sigma_{\tau^2_{v_0}/2}\wedge 1)\sigma_{\tau^2_{v_0}/2-1}\leq\widehat{v}\leq C_0\sigma.$$

Using a change of variable $5\delta\to\delta$ completes the proof. □

F.2 Proof of Theorem 3.2

Proof of Theorem 3.2. On the probability event $\mathcal{E}_1\cap\mathcal{E}_2\cap\mathcal{E}_3\cap\mathcal{E}_4$ where $\mathcal{E}_k$'s are defined the same as in the proof of Theorem 3.1, we have

$$c_0(\sigma_{\tau^2_{v_0}/2-1}/\sigma_{\tau^2_{v_0}/2}\wedge 1)\sigma_{\tau^2_{v_0}/2-1}\leq\widehat{v}\leq C_0\sigma.$$

Following the proof of Theorem E.2, for any fixed v and thus τ, we have

$$\kappa_\ell|\widehat{\mu}(v)-\mu^*|\leq\left|\frac{1}{\sqrt{n}}\sum_{i=1}^n\frac{\varepsilon_i}{z\sqrt{\tau^2+\varepsilon_i^2}}\right|.$$

For any v such that $c_0'\sigma_{\tau_{v_0}^2/2-1} \le v \le C_0\sigma$ where $c_0' = c_0(\sigma_{\tau_{v_0}^2/2-1}/\sigma_{\tau_{v_0}^2/2} \wedge 1)$ and any $z > 0$, using Lemma F.1 but with v_0 and V_0 replaced by $c_0'\sigma_{\tau_{v_0}^2/2-1}$ and $C_0\sigma$ respectively, we obtain with probability at least $1-\delta$

$$\begin{aligned}\sup_{v\in[c_0'\sigma_{\tau_{v_0}^2/2-1},\, C_0\sigma]} \kappa_\ell(v)\,|\widehat{\mu}(v)-\mu^*| &\le \sup_{v\in[c_0'\sigma_{\tau_{v_0}^2/2-1},\, C_0\sigma]} \kappa_\ell(v)\,|\widehat{\mu}(v)-\mu^*| \\ &\le \sup_{v\in[c_0'\sigma_{\tau_{v_0}^2/2-1},\, C_0\sigma]} \left|\frac{1}{\sqrt{n}}\sum_{i=1}^n \frac{\varepsilon_i}{z\sqrt{\tau^2+\varepsilon_i^2}}\right| \\ &\le \frac{\sigma}{c_0'\sigma_{\tau_{v_0}^2/2-1}}\sqrt{\frac{2\log(n/\delta)}{n}} + \frac{1}{z}\frac{\log(n/\delta)}{\sqrt{n}} \\ &\quad + \frac{\sigma^2}{2c_0'^2\sigma_{\tau_{v_0}^2/2-1}^2}\frac{z}{\sqrt{n}} + \frac{3(C_0\sigma - c_0'\sigma_{\tau_{v_0}^2/2-1})}{\sigma_{\tau_{v_0}^2/2-1}}\frac{1}{z\sqrt{n}},\end{aligned}$$

which yields

$$\sup_{v\in[c_0'\sigma_{\tau_{v_0}^2/2-1},\, C_0\sigma]} |\widehat{\mu}(v)-\mu^*| \le C\sigma\,\frac{\log(n/\delta)\vee z^2 \vee 1}{z\sqrt{n}},$$

where C is some constant only depending on $\sigma/\sigma_{\tau_{v_0}^2/2-1}$, c_0', and C_0. Putting the above pieces together and if $\log(1/\delta) \le z^2 \le \log(n/\delta)$, we obtain with probability at least $1-6\delta$ that

$$|\widehat{\mu}(\widehat{v})-\mu^*| \le \sup_{v\in[c_0'\sigma_{\tau_{v_0}^2/2-1},\, C_0\sigma]} |\widehat{\mu}(v)-\mu^*| \le C\cdot\sigma\,\frac{\log(n/\delta)\vee 1}{z\sqrt{n}}.$$

Using a change of variable $6\delta \to \delta$ and then setting $z = \log(n/\delta)$ gives

$$|\widehat{\mu}(\widehat{v})-\mu^*| \le \sup_{v\in[c_0'\sigma_{\tau_{v_0}^2/2-1},\, C_0\sigma]} |\widehat{\mu}(v)-\mu^*| \le C\cdot\sigma\sqrt{\frac{\log(n/\delta)}{n}}$$

with a lightly different constant C, provided that $\log(n/\delta) \ge 1$, aka $n \ge e\delta$. This completes the proof. □

F.3 Supporting lemmas

We collect supporting lemmas, aka Lemmas F.1, F.2, and F.3, in this subsection.

Lemma F.1. Let $0 < \delta < 1$. Suppose $\sigma \lesssim V_0$ and $z \lesssim \sqrt{\log(n/\delta)}$. Then, with probability at least $1-\delta$, we have

$$\sup_{v\in[v_0,V_0]} \left|\frac{1}{n}\sum_{i=1}^n \frac{\varepsilon_i}{\sqrt{\tau^2+\varepsilon_i^2}}\right| \le C\cdot\frac{V_0}{v_0}\cdot\frac{\log(n/\delta)}{n}$$

where C is some constant.

Proof of Lemma F.1. To prove the uniform bound over $[v_0, V_0]$, we adopt a covering argument. For any $0 < \epsilon \le 1$, there exists an ϵ-cover $\mathcal{N}$ of $[v_0, V_0]$ such that $|\mathcal{N}| \le 3(V_0 - v_0)/\epsilon$. Let $\tau_w = w\sqrt{n}/z$.

Then for every $v \in [v_0, V_0]$, there exists a $w \in \mathcal{N} \subset [v_0, V_0]$ such that $|w - \tau| \le \epsilon$ and

$$
\begin{aligned}
\left| \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{\varepsilon_i}{z\sqrt{\tau^2 + \varepsilon_i^2}} \right| &\le \left| \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{\varepsilon_i}{z\sqrt{\tau_w^2 + \varepsilon_i^2}} \right| \\
&\quad + \left| \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{\varepsilon_i}{z\sqrt{\tau_w^2 + \varepsilon_i^2}} - \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{\varepsilon_i}{z\sqrt{\tau^2 + \varepsilon_i^2}} \right| \\
&\le \left| \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{\varepsilon_i}{z\sqrt{\tau_w^2 + \varepsilon_i^2}} - \mathbb{E}\left[\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{\varepsilon_i}{z\sqrt{\tau_w^2 + \varepsilon_i^2}} \right] \right| \\
&\quad + \left| \mathbb{E}\left[\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{\varepsilon_i}{z\sqrt{\tau_w^2 + \varepsilon_i^2}} \right] \right| \\
&\quad + \left| \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{\varepsilon_i}{z\sqrt{\tau_w^2 + \varepsilon_i^2}} - \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{\varepsilon_i}{z\sqrt{\tau^2 + \varepsilon_i^2}} \right| \\
&= \mathrm{I} + \mathrm{II} + \mathrm{III}.
\end{aligned}
$$

For II, we have

$$
\mathrm{II} \le \frac{\sqrt{n}}{z} \cdot \frac{\sigma^2}{2\tau_w^2} \le \frac{z\sigma^2}{2v_0^2\sqrt{n}}.
$$

For III, using the inequality

$$
\left| \frac{x}{\sqrt{\tau_w^2 + x^2}} - \frac{x}{\sqrt{\tau^2 + x^2}} \right| \le \frac{|\tau_w - \tau|}{2\,|\tau_w| \wedge |\tau|},
$$

we obtain

$$
\mathrm{III} \le \frac{\sqrt{n}}{z} \cdot \frac{\epsilon}{2(w \wedge v)} \le \frac{\sqrt{n}}{z} \cdot \frac{\epsilon}{2v_0}.
$$

We then bound I. For any fixed τ_w, applying Lemma E.5 with the fact that $\left|\mathbb{E}\left(\tau_w \varepsilon_i / (\tau_w^2 + \varepsilon_i^2)^{1/2}\right)\right| \le \sigma^2/(2\tau_w)$, we obtain with probability at least $1 - 2\delta$

$$
\begin{aligned}
\left| \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{\varepsilon_i}{z\sqrt{\tau_w^2 + \varepsilon_i^2}} - \mathbb{E}\left[\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{\varepsilon_i}{z\sqrt{\tau_w^2 + \varepsilon_i^2}} \right] \right| &\le \frac{\sqrt{n}}{z\tau_w} \left(\sigma \sqrt{\frac{2\log(1/\delta)}{n}} + \frac{\tau_w \log(1/\delta)}{n} \right) \\
&\le \frac{\sigma}{z\tau_{v_0}} \sqrt{2\log(1/\delta)} + \frac{1}{z} \frac{\log(1/\delta)}{\sqrt{n}}
\end{aligned}
$$

where $\tau_{v_0} = v_0\sqrt{n}/z$. Therefore, putting above pieces together and using the union bound, we obtain with probability at least $1 - 6\epsilon^{-1}(V_0 - v_0)\delta$

$$
\begin{aligned}
\sup_{v \in [v_0, V_0]} \left| \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{\varepsilon_i}{z\sqrt{\tau^2 + \varepsilon_i^2}} \right| &\le \sup_{w \in \mathcal{N}} \left| \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{\varepsilon_i}{z\sqrt{\tau_w^2 + \varepsilon_i^2}} - \mathbb{E}\left[\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{\varepsilon_i}{z\sqrt{\tau_w^2 + \varepsilon_i^2}} \right] \right| \\
&\quad + \frac{z\sigma^2}{2v_0^2\sqrt{n}} + \frac{\sqrt{n}}{z} \cdot \frac{\epsilon}{2v_0} \\
&\le \frac{\sigma}{v_0} \sqrt{\frac{2\log(1/\delta)}{n}} + \frac{1}{z} \frac{\log(1/\delta)}{\sqrt{n}} + \frac{\sigma^2}{2v_0^2} \frac{z}{\sqrt{n}} + \frac{\sqrt{n}}{z} \cdot \frac{\epsilon}{2v_0}.
\end{aligned}
$$

Taking $\epsilon = 6(V_0 - v_0)/n$, we obtain with probability at least $1 - n\delta$

$$
\sup_{v \in [v_0, V_0]} \left| \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{\varepsilon_i}{z\sqrt{\tau^2 + \varepsilon_i^2}} \right| \le \frac{\sigma}{v_0} \sqrt{\frac{2\log(1/\delta)}{n}} + \frac{1}{z} \frac{\log(1/\delta)}{\sqrt{n}} + \frac{\sigma^2}{2v_0^2} \frac{z}{\sqrt{n}} + \frac{3(V_0 - v_0)}{v_0} \frac{1}{z\sqrt{n}}.
$$

Thus with probability at least $1-\delta$, we have

$$\sup_{v\in[v_0,V_0]}\left|\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{\varepsilon_i}{z\sqrt{\tau^2+\varepsilon_i^2}}\right| \le \frac{\sigma}{v_0}\sqrt{\frac{2\log(n/\delta)}{n}}+\frac{1}{z}\frac{\log(n/\delta)}{\sqrt{n}}+\frac{\sigma^2}{2v_0^2}\frac{z}{\sqrt{n}}+\frac{3(V_0-v_0)}{v_0}\frac{1}{z\sqrt{n}}$$
$$\le C\cdot\frac{V_0}{v_0}\cdot\frac{\log(n/\delta)}{z\sqrt{n}}$$

provided $z \lesssim \sqrt{\log(n/\delta)}$, where C is a constant only depending on $\sigma^2/(v_0V_0)$. When v_0 and V_0 are taken symmetrically around 1, v_0V_0 is close to 1. Multiplying both sides by $z/\sqrt{n}$ finishes the proof. □

Lemma F.2. Let w_i be i.i.d. copies of w. For any $0<\delta<1$, with probability at least $1-\delta$

$$\frac{1}{n}\sum_{i=1}^{n}\frac{\sqrt{\tau^2+w_i^2}-\tau}{\sqrt{\tau^2+w_i^2}}-\mathbb{E}\left(\frac{\sqrt{\tau^2+w_i^2}-\tau}{\sqrt{\tau^2+w_i^2}}\right)\le\sqrt{\frac{\log(1/\delta)\,\mathbb{E}w_i^2}{n\tau^2}}+\frac{\log(1/\delta)}{3n}.$$

Proof of Lemma F.2. The random variables

$$Z_i=Z_i(\tau):=\frac{\sqrt{\tau^2+w_i^2}-\tau}{\sqrt{\tau^2+w_i^2}}=\frac{\sqrt{1+w_i^2/\tau^2}-1}{\sqrt{1+w_i^2/\tau^2}}$$

with $\mu_z=\mathbb{E}Z_i$ and $\sigma_z^2=\mathrm{var}(Z_i)$ are bounded i.i.d. random variables such that

$$0\le Z_i\le 1\wedge\frac{w_i^2}{2\tau^2}.$$

Moreover we have

$$\mathbb{E}Z_i^2\le\frac{\mathbb{E}w_i^2}{2\tau^2},\ \sigma_z^2:=\mathrm{var}(Z_i)\le\frac{\mathbb{E}w_i^2}{2\tau^2}.$$

For third and higher order absolute moments, we have

$$\mathbb{E}|Z_i|^k\le\frac{\mathbb{E}w_i^2}{2\tau^2}\le\frac{k!}{2}\cdot\frac{\mathbb{E}w_i^2}{2\tau^2}\cdot\left(\frac{1}{3}\right)^{k-2},\ \text{for all integers } k\ge 3.$$

Therefore, using Lemma H.2 with $v=n\,\mathbb{E}w_i^2/(2\tau^2)$ and $c=1/3$ acquires that for any $t>0$

$$\mathbb{P}\left(\sum_{i=1}^{n}\frac{(1+w_i^2/\tau^2)^{1/2}-1}{(1+w_i^2/\tau^2)^{1/2}}-\sum_{i=1}^{n}\mathbb{E}\left(\frac{(1+w_i^2/\tau^2)^{1/2}-1}{(1+w_i^2/\tau^2)^{1/2}}\right)\ge-\sqrt{\frac{tn\,\mathbb{E}w_i^2}{\tau^2}}-\frac{t}{3}\right)\le\exp(-t).$$

Taking $t=\log(1/\delta)$ acquires that for any $0<\delta<1$

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n}\frac{(1+w_i^2/\tau^2)^{1/2}-1}{(1+w_i^2/\tau^2)^{1/2}}-\mathbb{E}\left(\frac{(1+w_i^2/\tau^2)^{1/2}-1}{(1+w_i^2/\tau^2)^{1/2}}\right)>-\sqrt{\frac{\log(1/\delta)\,\mathbb{E}w_i^2}{n\tau^2}}-\frac{\log(1/\delta)}{3n}\right)>1-\delta.$$

This finishes the proof.

□

Lemma F.3. For any $0<\delta<1$, we have with probability at least $1-\delta$ that

$$\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i^2 1\left(\varepsilon_i^2\le\frac{\tau^2}{2}-r^2\right)\ge\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\varepsilon_i^2 1\left(\varepsilon_i^2\le\frac{\tau^2}{2}-r^2\right)-\sigma_{\tau^2/2}\sqrt{\frac{\tau^2\log(1/\delta)}{n}}-\frac{\tau^2\log(1/\delta)}{6n}.$$

For any $0<\delta<1$, we have with probability at least $1-\delta$ that

$$\frac{1}{n}\sum_{i=1}^{n}|\varepsilon_i|1\left(\varepsilon_i^2\le\frac{\tau^2}{2}-r^2\right)\le\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}|\varepsilon_i|1\left(\varepsilon_i^2\le\frac{\tau^2}{2}-r^2\right)+\sqrt{\frac{2\sigma_{\tau^2/2}^2\log(1/\delta)}{n}}+\frac{\tau\log(1/\delta)}{3\sqrt{2}n}.$$

Consequently, we have, with probability at least $1-2\delta$, the above two inequalities hold simultaneously.

Proof of Lemma F.3. We prove the first two results and the last result directly follows from first two.

First result. Let $Z_i = \varepsilon_i^2 1\left(\varepsilon_i^2 \le \underline{\tau}^2/2 - r^2\right)$. The random variables Z_i with $\mu_z = \mathbb{E}Z_i$ and $\sigma_z^2 = \mathrm{var}(Z_i)$ are bounded i.i.d. random variables such that

$$\begin{aligned}
|Z_i| &= \left|\varepsilon_i^2 1\left(\varepsilon_i^2 \le \underline{\tau}^2/2 - r^2\right)\right| \le \underline{\tau}^2/2,\\
|\mu_z| &= |\mathbb{E}Z_i| = \left|\mathbb{E}\left(\varepsilon_i^2 1\left(\varepsilon_i^2 \le \underline{\tau}^2/2 - r^2\right)\right)\right| \le \sigma_{\underline{\tau}^2/2}^2,\\
\mathbb{E}Z_i^2 &= \mathbb{E}\left(\varepsilon_i^4 1\left(\varepsilon_i^2 \le \underline{\tau}^2/2 - r^2\right)\right) \le \underline{\tau}^2\sigma_{\underline{\tau}^2/2}^2/2,\\
\sigma_z^2 &:= \mathrm{var}(Z_i) = \mathbb{E}\big(Z_i - \mu_z\big)^2 \le \underline{\tau}^2\sigma_{\underline{\tau}^2/2}^2/2.
\end{aligned}$$

For third and higher order absolute moments, we have

$$\mathbb{E}|Z_i|^k = \mathbb{E}\left|\varepsilon_i^2 1\left(\varepsilon_i^2 \le \underline{\tau}^2/2 - r^2\right)\right|^k \le \frac{\underline{\tau}^2\sigma_{\underline{\tau}^2/2}^2}{2}\left(\frac{\underline{\tau}^2}{2}\right)^{k-2} \le \frac{k!}{2}\frac{\underline{\tau}^2\sigma_{\underline{\tau}^2/2}^2}{2}\left(\frac{\underline{\tau}^2}{6}\right)^{k-2}, \text{ for all integers } k \ge 3.$$

Using Lemma H.2 with $v = n\underline{\tau}^2\sigma_{\underline{\tau}^2/2}^2/2$ and $c = \underline{\tau}^2/6$, we have for any $t > 0$

$$\mathbb{P}\left(\sum_{i=1}^n \varepsilon_i^2 1\left(\varepsilon_i^2 \le \frac{\underline{\tau}^2}{2} - r^2\right) - \sum_{i=1}^n \mathbb{E}\varepsilon_i^2 1\left(\varepsilon_i^2 \le \frac{\underline{\tau}^2}{2} - r^2\right) \le -\sqrt{n\underline{\tau}^2\sigma_{\underline{\tau}^2/2}^2 t} - \frac{\underline{\tau}^2 t}{6}\right) \le \exp\left(-t\right).$$

Taking $t = \log(1/\delta)$ acquires the desired result.

Second result. With an abuse of notation, let $Z_i = |\varepsilon_i| 1\left(\varepsilon_i^2 \le \underline{\tau}^2/2 - r^2\right)$. The random variables Z_i with $\mu_z = \mathbb{E}Z_i$ and $\sigma_z^2 = \mathrm{var}(Z_i)$ are bounded i.i.d. random variables such that

$$\begin{aligned}
|Z_i| &= \left|\varepsilon_i 1\left(\varepsilon_i^2 \le \underline{\tau}^2/2 - r^2\right)\right| \le \underline{\tau}/\sqrt{2},\\
|\mu_z| &= |\mathbb{E}Z_i| = \left|\mathbb{E}\left(|\varepsilon_i| 1\left(\varepsilon_i^2 \le \underline{\tau}^2/2 - r^2\right)\right)\right| \le \sqrt{2}\sigma_{\underline{\tau}^2/2}^2/\underline{\tau},\\
\mathbb{E}Z_i^2 &= \mathbb{E}\left(\varepsilon_i^2 1\left(\varepsilon_i^2 \le \underline{\tau}^2/2 - r^2\right)\right) \le \sigma_{\underline{\tau}^2/2}^2,\\
\sigma_z^2 &:= \mathrm{var}(Z_i) = \mathbb{E}\big(Z_i - \mu_z\big)^2 \le \sigma_{\underline{\tau}^2/2}^2.
\end{aligned}$$

For third and higher order absolute moments, we have

$$\mathbb{E}|Z_i|^k = \mathbb{E}\left||\varepsilon_i| 1\left(\varepsilon_i^2 \le \underline{\tau}^2/2 - r^2\right)\right|^k \le \sigma_{\underline{\tau}^2/2}^2\left(\frac{\underline{\tau}}{\sqrt{2}}\right)^{k-2} \le \frac{k!}{2}\sigma_{\underline{\tau}^2/2}^2\left(\frac{\underline{\tau}}{3\sqrt{2}}\right)^{k-2}, \text{ for all integers } k \ge 3.$$

Using Lemma H.2 with $v = n\sigma_{\underline{\tau}^2/2}^2$ and $c = \underline{\tau}/(3\sqrt{2})$, we have for any $t > 0$

$$\mathbb{P}\left(\sum_{i=1}^n |\varepsilon_i| 1\left(\varepsilon_i^2 \le \frac{\underline{\tau}^2}{2} - r^2\right) - \sum_{i=1}^n \mathbb{E}|\varepsilon_i| 1\left(\varepsilon_i^2 \le \frac{\underline{\tau}^2}{2} - r^2\right) \ge \sqrt{2n\sigma_{\underline{\tau}^2/2}^2 t} + \frac{\underline{\tau}t}{3\sqrt{2}}\right) \le \exp\left(-t\right).$$

Taking $t = \log(1/\delta)$ acquires the desired result. □

G PROOFS FOR SECTION 3.2

This section collects proofs for results in Section 3.2.

G.1 PROOF OF THEOREM 3.5

Proof of Theorem 3.5. First, the MoM estimator $\widehat{\mu}^{\mathrm{MoM}} = M(z_1, \ldots, z_k)$ is equivalent to

$$\mathrm{argmin} \sum_{j=1}^k |z_j - \mu|.$$

For any $x \in \mathbb{R}$, let $\ell(x) = |x|$ and define $L(x) = \mathbb{E}\ell'(x + Z)$ where $Z \sim \mathcal{N}(0,1)$ and

$$\ell'(x) = \begin{cases} 1, & \text{if } x > 0,\\ 0, & \text{if } x = 0,\\ -1, & \text{otherwise.}\end{cases}$$

If the assumptions of Theorem 4 of Minsker (2019) are satisfied, we obtain, after some algebra, that

$$\sqrt{n}\left(\widehat{\mu}^{\mathrm{MoM}}-\mu^*\right) \rightsquigarrow \mathcal{N}\left(0, \frac{\mathbb{E}(\ell'(Z))^2}{(L'(0))^2}\right).$$

Some algebra derives that

$$\frac{\mathbb{E}(\ell'(Z))^2}{(L'(0))^2}=\frac{\pi\sigma^2}{2}.$$

It remains to check the assumptions there. Assumptions (1), (4), and (5) trivially hold. Assumption (2) can be verified by using the following Berry-Esseen bound.

Fact G.1. Let $y_1,\dots,y_m$ be i.i.d. random copies of y with mean μ, variance σ^2 and $\mathbb{E}|y-\mu|^{2+\iota}<\infty$ for some $\iota\in(0,1]$. Then there exists an absolute constant C such that

$$\sup_{t\in\mathbb{R}}\left|\mathbb{P}\left(\sqrt{m}\frac{\bar{y}-\mu}{\sigma}\le t\right)-\Phi(t)\right|\le C\,\frac{\mathbb{E}|y-\mu|^{2+\iota}}{\sigma^{2+\iota}m^{\iota/2}}.$$

It remains to check Assumption (3). Because $g(m)\lesssim m^{-\iota/2}$, $\sqrt{k}g(m)\lesssim\sqrt{k}m^{-\iota/2}\to 0$ if $k=o(n^{\iota/(1+\iota)})$ as $n\to\infty$. Thus Assumption (3) holds if $k=o(n^{\iota/(1+\iota)})$ and $k\to\infty$. This completes the proof. □

G.2 Proof of Theorem 3.3

In this subsection, we state and prove a stronger result of Theorem 3.3, aka Theorem G.2. Theorem 3.3 can then be proved following the same proof under the assumption that $\mathbb{E}|\varepsilon_i|^{2+\iota}<\infty$ for any prefixed $0<\iota\le 1$.

Theorem G.2. Assume the same assumptions as in Theorem 3.1. Take $z^2\ge 2\log(n)$. If $\mathbb{E}\varepsilon_i^4<\infty$, then

$$\sqrt{n}\begin{bmatrix}\widehat{\mu}-\mu^*\\ \widehat{v}-v_*\end{bmatrix}\rightsquigarrow\mathcal{N}\left(0,\Sigma\right),\ \text{where } \Sigma=\begin{bmatrix}\sigma^2 & \sigma\,\mathbb{E}\varepsilon_i^3/2\\ \sigma\,\mathbb{E}\varepsilon_i^3/2 & (\sigma^2\mathbb{E}\varepsilon_i^4-\sigma^6)/4\end{bmatrix}.$$

Proof of Theorem G.2. Now we are ready to analyze the self-tuned mean estimator $\widehat{\mu}=\widehat{\mu}(\widehat{v})$. For any $\delta\in(0,1)$, following the proof of Theorem 3.1, we obtain with probability at least $1-\delta$ that

$$|\widehat{\mu}(\widehat{v})-\mu^*|\le\sup_{v\in[v_0,V_0]}|\widehat{\mu}(v)-\mu^*|\le 2C\cdot\frac{V_0^2}{v_0}\cdot\frac{\log(n/\delta)}{z\sqrt{n}}.$$

Taking $z^2\ge\log(n/\delta)$ with $\delta=1/n$ in the above inequality, we obtain $\widehat{\mu}\to\mu^*$ in probability. Theorem G.3 implies that $\widehat{v}\to\sigma$ in probability. Thus we have $\|\widehat{\theta}-\theta^*\|_2\to 0$ in probability, where

$$\widehat{\theta}=(\widehat{\mu},\widehat{v})^{\mathrm{T}},\ \text{and } \theta^*=(\mu^*,\sigma)^{\mathrm{T}}.$$

Using the Taylor's theorem for vector-valued functions, we obtain

$$\nabla L_n(\widehat{\theta})=0=\nabla L_n(\theta^*)+H_n(\theta^*)(\widehat{\theta}-\theta^*)\,+\,\frac{R_2(\theta)}{2}\big(\widehat{\theta}-\theta^*\big)^{\otimes 2},$$

where $\otimes$ indicates the tensor product. Let $\tau_\sigma=\sigma\sqrt{n}/z$. We say that X_n and Y_n are asymptotically equivalent, denoted as $X_n\simeq Y_n$, if both X_n and Y_n converge in distribution to some same random variable/vector Z. Rearranging, we obtain

$$\begin{aligned}
\sqrt{n}\left(\widehat{\theta}-\theta^*\right) &\simeq [H_n(\theta^*)]^{-1}\left(-\sqrt{n}\,\nabla L_n(\theta^*)\right)\\
&= \begin{bmatrix} \frac{\sqrt{n}}{z}\cdot\frac{1}{n}\sum_{i=1}^n \frac{\tau_\sigma^2}{(\tau_\sigma^2+\varepsilon_i^2)^{3/2}} & \frac{n}{z^2}\cdot\frac{1}{n}\sum_{i=1}^n \frac{\tau_\sigma\varepsilon_i}{(\tau_\sigma^2+\varepsilon_i^2)^{3/2}} \\ \frac{n}{z^2}\cdot\frac{1}{n}\sum_{i=1}^n \frac{\tau_\sigma\varepsilon_i}{(\tau_\sigma^2+\varepsilon_i^2)^{3/2}} & \frac{n^{3/2}}{z^3}\cdot\frac{1}{n}\sum_{i=1}^n \frac{\varepsilon_i^3}{(\tau_\sigma^2+\varepsilon_i^2)^{3/2}} \end{bmatrix}^{-1} \\
&\quad \begin{bmatrix} \sqrt{n}\cdot\frac{1}{n}\sum_{i=1}^n \frac{\tau_\sigma\varepsilon_i}{\sigma\sqrt{\tau_\sigma^2+\varepsilon_i^2}} \\ \sqrt{n}\cdot\frac{n}{z^2}\frac{1}{n}\sum_{i=1}^n \frac{\sqrt{1+\varepsilon_i^2/\tau_\sigma^2}-1}{\sqrt{1+\varepsilon_i^2/\tau_\sigma^2}} - \sqrt{n}\cdot a \end{bmatrix} \\
&\simeq \begin{bmatrix} \sigma & 0 \\ 0 & \sigma^3 \end{bmatrix} \begin{bmatrix} \sqrt{n}\cdot\frac{1}{n}\sum_{i=1}^n \frac{\tau_\sigma\varepsilon_i}{\sigma\sqrt{\tau_\sigma^2+\varepsilon_i^2}} \\ \sqrt{n}\cdot\frac{n}{z^2}\frac{1}{n}\sum_{i=1}^n \frac{\sqrt{1+\varepsilon_i^2/\tau_\sigma^2}-1}{\sqrt{1+\varepsilon_i^2/\tau_\sigma^2}} - \sqrt{n}\cdot a \end{bmatrix} \\
&= \begin{bmatrix} \sigma & 0 \\ 0 & \sigma^3 \end{bmatrix} \begin{bmatrix} \mathrm{I} \\ \mathrm{II} \end{bmatrix},
\end{aligned}$$

where the second $\simeq$ uses the fact that

$$H_n(\theta^*) \xrightarrow{\text{a.s.}} \begin{bmatrix} \frac{1}{\sigma} & 0 \\ 0 & \frac{1}{\sigma^3} \end{bmatrix}.$$

We proceed to derive the asymptotic property of $(\mathrm{I},\mathrm{II})^{\mathrm{T}}$. For I, we have

$$\begin{aligned}
\mathrm{I} &= \sqrt{n}\cdot\left(\frac{1}{n}\sum_{i=1}^n \frac{\tau_\sigma\varepsilon_i}{\sigma\sqrt{\tau_\sigma^2+\varepsilon_i^2}} - \mathbb{E}\left[\frac{\tau_\sigma\varepsilon_i}{\sigma\sqrt{\tau_\sigma^2+\varepsilon_i^2}}\right]\right) + \sqrt{n}\cdot\mathbb{E}\left[\frac{\tau_\sigma\varepsilon_i}{\sigma\sqrt{\tau_\sigma^2+\varepsilon_i^2}}\right] \\
&\rightsquigarrow \mathcal{N}\left(0, \lim_{n\to\infty}\operatorname{var}\left[\frac{\tau_\sigma\varepsilon_i}{\sigma\sqrt{\tau_\sigma^2+\varepsilon_i^2}}\right]\right) + \lim_{n\to\infty}\sqrt{n}\cdot\mathbb{E}\left[\frac{\tau_\sigma\varepsilon_i}{\sigma\sqrt{\tau_\sigma^2+\varepsilon_i^2}}\right].
\end{aligned}$$

It remains to calculate

$$\lim_{n\to\infty}\mathbb{E}\left(\frac{\sqrt{n}\tau_\sigma\varepsilon_i}{\sqrt{\tau_\sigma^2+\varepsilon_i^2}}\right) \quad\text{and}\quad \lim_{n\to\infty}\operatorname{var}\left[\frac{\tau\varepsilon_i}{\sqrt{\tau_\sigma^2+\varepsilon_i^2}}\right].$$

For the former term, if there exists some $0<\iota\le 1$ such that $\mathbb{E}|\varepsilon_i|^{2+\iota}<\infty$, using the fact that $\mathbb{E}\varepsilon_i = 0$, we have

$$\begin{aligned}
\left|\mathbb{E}\left(\frac{\sqrt{n}\tau_\sigma\varepsilon_i}{\sqrt{\tau_\sigma^2+\varepsilon_i^2}}\right)\right| &= \sqrt{n}\tau_\sigma\cdot\left|\mathbb{E}\left\{\frac{-\varepsilon_i/\tau_\sigma}{\sqrt{1+\varepsilon_i^2/\tau_\sigma^2}}\right\}\right| = \sqrt{n}\tau_\sigma\cdot\left|\mathbb{E}\left\{\frac{\tau_\sigma^{-1}\varepsilon_i\left(\sqrt{1+\varepsilon_i^2/\tau_\sigma^2}-1\right)}{\sqrt{1+\varepsilon_i^2/\tau_\sigma^2}}\right\}\right| \\
&\le \frac{\sqrt{n}\tau_\sigma}{2}\cdot\mathbb{E}\left|\frac{\varepsilon_i^3/\tau_\sigma^3}{\sqrt{1+\varepsilon_i^2/\tau_\sigma^2}}\right| \le \frac{\sqrt{n}\tau_\sigma}{2}\cdot\frac{\mathbb{E}|\varepsilon_i|^{2+\iota}}{\tau_\sigma^{2+\iota}} \\
&\le \frac{\sqrt{n}\,\mathbb{E}|\varepsilon_i|^{2+\iota}}{2\tau_\sigma^{1+\iota}} \to 0, \qquad \text{(G.1)}
\end{aligned}$$

where the first inequality uses Lemma H.4 (ii) with $r=1/2$, that is, $\sqrt{1+x}\le 1+x/2$ for $x\ge -1$. For the second term, we have

$$\lim_{n\to\infty}\operatorname{var}\left[\frac{\tau_\sigma\varepsilon_i}{\sqrt{\tau_\sigma^2+\varepsilon_i^2}}\right] = \lim_{n\to\infty}\mathbb{E}\left[\frac{\tau_\sigma^2\varepsilon_i^2}{\tau_\sigma^2+\varepsilon_i^2}\right] = \sigma^2,$$

by the dominated convergence theorem. Thus

$$\mathrm{I} \rightsquigarrow \mathcal{N}(0,1).$$

For II, recall $a = 1/2$ and using the facts that

$$\begin{aligned}
&\lim_{n\to\infty}\frac{n}{z^2}\cdot\mathbb{E}\left(\frac{\sqrt{1+\varepsilon_i^2/\tau_\sigma^2}-1}{\sqrt{1+\varepsilon_i^2/\tau_\sigma^2}}\right) = \lim_{n\to\infty}\frac{n}{2\tau_\sigma^2 z^2}\cdot\mathbb{E}\left(\frac{1}{\sqrt{1+\varepsilon_i^2/\tau_\sigma^2}}\cdot\frac{\sqrt{1+\varepsilon_i^2/\tau_\sigma^2}-1}{1/(2\tau_\sigma^2)}\right) = \frac{1}{2}, \\
&\lim_{n\to\infty}\sqrt{n}\cdot\left(\frac{n}{z^2}\cdot\mathbb{E}\left(\frac{\sqrt{1+\varepsilon_i^2/\tau_\sigma^2}-1}{\sqrt{1+\varepsilon_i^2/\tau_\sigma^2}}\right)-\frac{1}{2}\right) = 0,
\end{aligned}$$

we have

$$\begin{aligned}
\mathrm{II} &= \sqrt{n} \cdot \frac{n}{z^2} \cdot \frac{1}{n} \sum_{i=1}^{n} \frac{\sqrt{1+\varepsilon_i^2/\tau_\sigma^2}-1}{\sqrt{1+\varepsilon_i^2/\tau_\sigma^2}} - \sqrt{n} \cdot \frac{1}{2} \\
&\simeq \sqrt{n} \cdot \frac{1}{n} \sum_{i=1}^{n} \left(\frac{n}{z^2} \cdot \frac{\sqrt{1+\varepsilon_i^2/\tau_\sigma^2}-1}{\sqrt{1+\varepsilon_i^2/\tau_\sigma^2}} - \mathbb{E}\left(\frac{n}{z^2} \cdot \frac{\sqrt{1+\varepsilon_i^2/\tau_\sigma^2}-1}{\sqrt{1+\varepsilon_i^2/\tau_\sigma^2}} \right) \right) \\
&\simeq \mathcal{N}\left(0, \lim_{n\to\infty} \operatorname{var}\left(\frac{n}{z^2} \cdot \frac{\sqrt{1+\varepsilon_i^2/\tau_\sigma^2}-1}{\sqrt{1+\varepsilon_i^2/\tau_\sigma^2}} \right)\right).
\end{aligned}$$

If $\mathbb{E}\varepsilon_i^4 < \infty$, then

$$\lim_{n\to\infty} \operatorname{var}\left(\frac{n}{z^2} \cdot \frac{\sqrt{1+\varepsilon_i^2/\tau_\sigma^2}-1}{\sqrt{1+\varepsilon_i^2/\tau_\sigma^2}} \right) = \frac{\mathbb{E}\varepsilon_i^4}{4\sigma^4} - \frac{1}{4},$$

and thus $\mathrm{II} \simeq \mathcal{N}\left(0, (\mathbb{E}\varepsilon_i^4/\sigma^4 - 1)/4\right)$. For the cross covariance, we have

$$\begin{aligned}
&\lim_{n\to\infty} \operatorname{cov}\left(\frac{\tau_\sigma \varepsilon_i}{\sigma\sqrt{\tau_\sigma^2+\varepsilon_i^2}}, \frac{n}{z^2} \cdot \frac{\sqrt{1+\varepsilon_i^2/\tau_\sigma^2}-1}{\sqrt{1+\varepsilon_i^2/\tau_\sigma^2}} \right) \\
&= \lim_{n\to\infty} \mathbb{E}\left(\frac{\tau_\sigma \varepsilon_i}{\sigma\sqrt{\tau_\sigma^2+\varepsilon_i^2}} \cdot \frac{n}{z^2} \cdot \frac{\sqrt{1+\varepsilon_i^2/\tau_\sigma^2}-1}{\sqrt{1+\varepsilon_i^2/\tau_\sigma^2}} \right) \\
&= \frac{\mathbb{E}\varepsilon_i^3}{2\sigma^3}.
\end{aligned}$$

Thus

$$\sqrt{n}\,(\widehat{\theta} - \theta^*) \rightsquigarrow \mathcal{N}(0, \Sigma),$$

where

$$\Sigma = \begin{bmatrix} \sigma & 0 \\ 0 & \sigma^3 \end{bmatrix} \begin{bmatrix} 1 & \mathbb{E}\varepsilon_i^3/(2\sigma^3) \\ \mathbb{E}\varepsilon_i^3/(2\sigma^3) & (\mathbb{E}\varepsilon_i^4/\sigma^4 - 1)/4 \end{bmatrix} \begin{bmatrix} \sigma & 0 \\ 0 & \sigma^3 \end{bmatrix} = \begin{bmatrix} \sigma^2 & \sigma\mathbb{E}\varepsilon_i^3/2 \\ \sigma\mathbb{E}\varepsilon_i^3/2 & (\sigma^2\mathbb{E}\varepsilon_i^4 - \sigma^6)/4 \end{bmatrix}.$$

Therefore, for $\widehat{\mu}$ only, we have

$$\sqrt{n}\,(\widehat{\mu} - \mu^*) \rightsquigarrow \mathcal{N}(0, \sigma^2).$$

□

G.3 Consistency of $\widehat{v}$

This subsection proves that $\widehat{v}$ is a consistent estimator of σ. Recall that

$$\nabla_v L_n(\mu, v) = \frac{n}{z^2} \cdot \frac{1}{n} \sum_{i=1}^{n} \left(\frac{\tau}{\sqrt{\tau^2 + (y_i - \mu)^2}} - 1 \right) + a$$

where $a = 1/2$. We emphasize that the following proof only needs the second moment assumption $\sigma^2 = \mathbb{E}\varepsilon_i^2 < \infty$.

Theorem G.3 (Consistency of $\widehat{v}$)**.** Assume the same assumptions as in Theorem 3.1. Take $z^2 \geq \log(n)$. Then

$$\widehat{v} \longrightarrow \sigma \text{ in probability.}$$

Proof of Theorem G.3. By the proof of Theorem 3.1, we obtain with probability at least $1 - \delta$ that the following two results hold simultaneously:

$$\sup_{v \in [v_0, V_0]} |\widehat{\mu}(v) - \mu^*| \leq 2C \cdot \frac{V_0^2}{v_0} \cdot \frac{\log(n/\delta)}{z\sqrt{n}} =: r, \tag{G.2}$$

$$v_0 < c_0 \sigma_{\tau_{v_0}^2 - 1} \leq \widehat{v} \leq C_0 \sigma < V_0, \tag{G.3}$$

provided that $z^2 \geq \log(5/\delta)$ and n is large enough. Therefore, the constraint in the optimization problem (3.1) is not active, and thus

$$\nabla_v L_n(\widehat{\mu}, \widehat{v}) = 0.$$

Using Lemma G.4 together with the equality above, we obtain with probability at least $1-\delta$ that

$$\begin{aligned}
\frac{c_0}{V_0^3}|\widehat{v}-\sigma|^2 \leq \frac{c_0}{\widehat{v}^3 \vee \sigma^3}|\widehat{v}-\sigma|^2 &\leq \rho_\ell |\widehat{v}-\sigma|^2 \\
&\leq \langle \nabla_v L_n(\widehat{\mu}, \widehat{v}) - \nabla_v L_n(\widehat{\mu}, \sigma), \widehat{v}-\sigma \rangle \\
&\leq |\nabla_v L_n(\widehat{\mu}, \sigma)|\, |\widehat{v}-\sigma| \\
&\leq \left| \frac{n}{z^2} \cdot \frac{1}{n} \sum_{i=1}^n \left(\frac{\tau_\sigma}{\sqrt{\tau_\sigma^2 + (y_i - \widehat{\mu})^2}} - 1 \right) + a \right| |\widehat{v}-\sigma| .
\end{aligned}$$

Plugging (G.2) into the above inequality and canceling $|\widehat{v}-\sigma|$ on both sides, we obtain with probability at least $1-2\delta$ that

$$\begin{aligned}
\frac{c_0}{V_0^3}|\widehat{v}-\sigma| &\leq \left| \frac{n}{z^2} \cdot \frac{1}{n} \sum_{i=1}^n \left(\frac{\tau_\sigma}{\sqrt{\tau_\sigma^2 + (y_i - \widehat{\mu})^2}} - 1 \right) + a \right| \\
&\leq \sup_{\mu \in \mathbb{B}_r(\mu^*)} \left| \frac{n}{z^2} \cdot \frac{1}{n} \sum_{i=1}^n \left(\frac{\tau_\sigma}{\sqrt{\tau_\sigma^2 + (y_i - \mu)^2}} - 1 \right) + a \right| \\
&= \frac{n}{z^2} \cdot \sup_{\mu \in \mathbb{B}_r(\mu^*)} \left| \frac{1}{n} \sum_{i=1}^n \left(\frac{\tau_\sigma}{\sqrt{\tau_\sigma^2 + (y_i - \mu)^2}} - 1 \right) + \frac{az^2}{n} \right| \\
&\leq \frac{n}{z^2} \cdot \sup_{\mu \in \mathbb{B}_r(\mu^*)} \left| \frac{1}{n} \sum_{i=1}^n \left(1 - \frac{\tau_\sigma}{\sqrt{\tau_\sigma^2 + (y_i - \mu)^2}} \right) - \mathbb{E}\left(1 - \frac{\tau_\sigma}{\sqrt{\tau_\sigma^2 + (y_i - \mu)^2}} \right) \right| \\
&\qquad + \frac{n}{z^2} \cdot \sup_{\mu \in \mathbb{B}_r(\mu^*)} \left| \mathbb{E}\left(1 - \frac{\tau_\sigma}{\sqrt{\tau_\sigma^2 + (y_i - \mu)^2}} \right) - \frac{az^2}{n} \right| \\
&=: \mathrm{I} + \mathrm{II}.
\end{aligned}$$

It remains to bound terms I and II. We start with term II. Let $r_i^2 = (y_i - \mu)^2$. We have

$$\begin{aligned}
\mathrm{II} &= \frac{n}{z^2} \cdot \sup_{\mu \in \mathbb{B}_r(\mu^*)} \left| \mathbb{E}\left(1 - \frac{\tau_\sigma}{\sqrt{\tau_\sigma^2 + (y_i - \mu)^2}} \right) - \frac{az^2}{n} \right| \\
&= \max\left\{ \sup_{\mu \in \mathbb{B}_r(\mu^*)} \left(\frac{n}{z^2} \cdot \mathbb{E}\frac{\sqrt{1 + r_i^2/\tau_\sigma^2} - 1}{\sqrt{1 + r_i^2/\tau_\sigma^2}} - a \right), \sup_{\mu \in \mathbb{B}_r(\mu^*)} \left(a - \frac{n}{z^2} + \mathbb{E}\frac{1}{\sqrt{1 + r_i^2/\tau_\sigma^2}} \right) \right\} \\
&=: \mathrm{II}_1 \vee \mathrm{II}_2.
\end{aligned}$$

In order to bound II, we bound II_1 and II_2 respectively. For term II_1, using Lemma H.4 (ii), aka $(1+x)^r \leq 1 + rx$ for $x \geq -1$ and $r \in (0,1)$, and $a = 1/2$, we have

$$\begin{aligned}
\mathrm{II}_1 &= \sup_{\mu \in \mathbb{B}_r(\mu^*)} \left(\frac{n}{z^2} \cdot \mathbb{E}\frac{\sqrt{1 + r_i^2/\tau_\sigma^2} - 1}{\sqrt{1 + r_i^2/\tau_\sigma^2}} - a \right) \\
&\leq \sup_{\mu \in \mathbb{B}_r(\mu^*)} \left\{ \frac{n}{z^2} \cdot \left(1 + \mathbb{E}\frac{r_i^2}{2\tau_\sigma^2} - 1 \right) - a \right\} \\
&\leq \frac{n}{z^2} \cdot \mathbb{E}\frac{\varepsilon_i^2 + 2r|\varepsilon_i| + r^2|}{2\tau_\sigma^2} - \frac{1}{2} && (a = 1/2) \\
&\leq \frac{r}{\sigma}\left(1 + \frac{r}{2\sigma} \right) \\
&\leq \frac{2r}{\sigma}
\end{aligned}$$

if n is large enough such that $r \le 2\sigma$. To bound II_2, we need Lemma D.1. Specifically, for any $0 \le \gamma < 1$, we have

$$(1+x)^{-1} \le 1-(1-\gamma)x, \text{ for any } 0 \le x \le \frac{\gamma}{1-\gamma}.$$

Using this result, we obtain

$$\begin{aligned}
\mathbb{E}\frac{1}{\sqrt{1+r_i^2/\tau_\sigma^2}} &\le \sqrt{\mathbb{E}\frac{1}{1+r_i^2/\tau_\sigma^2}} && (\text{concavity of } \sqrt{x}) \\
&\le \sqrt{\mathbb{E}\left\{\left(1-\frac{(1-\gamma)r_i^2}{\tau_\sigma^2}\right)1\left(\frac{r_i^2}{\tau_\sigma^2} \le \frac{\gamma}{1-\gamma}\right) + \frac{1}{1+r_i^2/\tau_\sigma^2}1\left(\frac{r_i^2}{\tau_\sigma^2} > \frac{\gamma}{1-\gamma}\right)\right\}} \\
&\le \sqrt{1-(1-\gamma)\,\mathbb{E}\left(\frac{r_i^2}{\tau_\sigma^2}1\left(\frac{r_i^2}{\tau_\sigma^2} \le \frac{\gamma}{1-\gamma}\right)\right)} && (\text{Lemma D.1}) \\
&\le \sqrt{1-(1-\gamma)\,\mathbb{E}\left(\frac{r_i^2}{\tau_\sigma^2}1\left(\frac{r_i^2}{\tau_\sigma^2} \le \frac{\gamma}{1-\gamma}\right)\right)} \\
&\le \sqrt{1-(1-\gamma)\,\mathbb{E}\left(\frac{\varepsilon_i^2-2r|\varepsilon_i|+r^2}{\tau_\sigma^2}1\left(\frac{2(\varepsilon_i^2+r^2)}{\tau_\sigma^2} \le \frac{\gamma}{1-\gamma}\right)\right)} \\
& && (\forall\, \mu \in \mathbb{B}_r(\mu^*)) \\
&\le 1-\frac{1-\gamma}{2}\,\mathbb{E}\left(\frac{\varepsilon_i^2-2r|\varepsilon_i|+r^2}{\tau_\sigma^2}1\left(\frac{2(\varepsilon_i^2+r^2)}{\tau_\sigma^2} \le \frac{\gamma}{1-\gamma}\right)\right),
\end{aligned}$$

where the first inequality uses the concavity of $\sqrt{x}$, the third inequality uses Lemma D.1, and the last inequality uses the inequality that $(1+x)^{-1} \le 1-x/2$ for $x \in [0,1]$, aka Lemma H.4 (iii) with $r=-1$, provided that

$$(1-\gamma)\,\mathbb{E}\left(\frac{\varepsilon_i^2-2r|\varepsilon_i|-r^2}{\tau_\sigma^2}1\left(\frac{2(\varepsilon_i^2+r^2)}{\tau_\sigma^2} \le \frac{\gamma}{1-\gamma}\right)\right) \le (1-\gamma)\frac{\sigma^2-2r\sigma-r^2}{\tau_\sigma^2} \le 1.$$

Thus term II_2 can be bounded as

$$\begin{aligned}
\mathrm{II}_2 &= \sup_{\mu \in \mathbb{B}_r(\mu^*)}\left(a-\frac{n}{z^2}+\frac{n}{z^2}\cdot\mathbb{E}\frac{1}{\sqrt{1+r_i^2/\tau_\sigma^2}}\right) \\
&\le a-\frac{n}{z^2}+\frac{n}{z^2}\cdot\left\{1-\frac{1-\gamma}{2}\,\mathbb{E}\left(\frac{\varepsilon_i^2-2r|\varepsilon_i|+r^2}{\tau_\sigma^2}1\left(\frac{2(\varepsilon_i^2+r^2)}{\tau_\sigma^2} \le \frac{\gamma}{1-\gamma}\right)\right)\right\} \\
&\le a-\frac{1-\gamma}{2\sigma^2}\cdot\mathbb{E}\varepsilon_i^2+\frac{1-\gamma}{2\sigma^2}\cdot 2r\cdot\mathbb{E}\left(|\varepsilon_i|\right) \\
&\le a-\frac{1-\gamma}{2}+\frac{r(1-\gamma)}{\sigma} \\
&= \frac{\gamma}{2}+\frac{r(1-\gamma)}{\sigma}. && (a=1/2)
\end{aligned}$$

Combining the upper bound for II_1 and II_2 and using the fact that, we obtain

$$\mathrm{II} \le \max\{\mathrm{II}_1, \mathrm{II}_2\} \le \frac{\gamma}{2}+\frac{2r}{\sigma} \to 0,$$

if $\gamma = \gamma(n) \to 0$.

We proceed to bound I. Recall that

$$\mathrm{I} = \frac{n}{z^2}\cdot\sup_{\mu \in \mathbb{B}_r(\mu^*)}\left|\frac{1}{n}\sum_{i=1}^n\left(1-\frac{\tau_\sigma}{\sqrt{\tau_\sigma^2+(y_i-\mu)^2}}\right)-\mathbb{E}\left(1-\frac{\tau_\sigma}{\sqrt{\tau_\sigma^2+(y_i-\mu)^2}}\right)\right|.$$

For any $0 < \epsilon \le 2r$, there exists an ϵ-cover $\mathcal{N} \subseteq \mathbb{B}_r(\mu^*)$ of $\mathbb{B}_r(\mu^*)$ such that $|\mathcal{N}| \le 6r/\epsilon$. Then for any $\mu \in \mathbb{B}_r(\mu^*)$ there exists a $\omega \in \mathcal{N}$ such that $|\omega - \mu| \le \gamma$, and

$$\begin{aligned}
&\left|\frac{1}{n}\sum_{i=1}^n \left(1 - \frac{\tau_\sigma}{\sqrt{\tau_\sigma^2 + (y_i - \mu)^2}}\right) - \mathbb{E}\left(1 - \frac{\tau_\sigma}{\sqrt{\tau_\sigma^2 + (y_i - \mu)^2}}\right)\right| \\
&= \left|\frac{1}{n}\sum_{i=1}^n \frac{\sqrt{1 + (y_i - \mu)^2/\tau_\sigma^2} - 1}{\sqrt{1 + (y_i - \mu)^2/\tau_\sigma^2}} - \mathbb{E}\frac{\sqrt{1 + (y_i - \mu)^2/\tau_\sigma^2} - 1}{\sqrt{1 + (y_i - \mu)^2/\tau_\sigma^2}}\right| \\
&\le \left|\frac{1}{n}\sum_{i=1}^n \frac{\sqrt{1 + (y_i - \omega)^2/\tau_\sigma^2} - 1}{\sqrt{1 + (y_i - \omega)^2/\tau_\sigma^2}} - \mathbb{E}\frac{\sqrt{1 + (y_i - \omega)^2/\tau_\sigma^2} - 1}{\sqrt{1 + (y_i - \omega)^2/\tau_\sigma^2}}\right| \\
&\quad + \left|\frac{1}{n}\sum_{i=1}^n \frac{\sqrt{1 + (y_i - \mu)^2/\tau_\sigma^2} - 1}{\sqrt{1 + (y_i - \mu)^2/\tau_\sigma^2}} - \frac{1}{n}\sum_{i=1}^n \frac{\sqrt{1 + (y_i - \omega)^2/\tau_\sigma^2} - 1}{\sqrt{1 + (y_i - \omega)^2/\tau_\sigma^2}}\right| \\
&\quad + \left|\mathbb{E}\frac{\sqrt{1 + (y_i - \mu)^2/\tau_\sigma^2} - 1}{\sqrt{1 + (y_i - \mu)^2/\tau_\sigma^2}} - \mathbb{E}\frac{\sqrt{1 + (y_i - \omega)^2/\tau_\sigma^2} - 1}{\sqrt{1 + (y_i - \omega)^2/\tau_\sigma^2}}\right| \\
&= \mathrm{I}_1 + \mathrm{I}_2 + \mathrm{I}_3.
\end{aligned}$$

For I_1, using Lemma F.2 acquires with probability at least $1 - 2\delta$ that

$$\begin{aligned}
\mathrm{I}_1 &\le \sqrt{\frac{\mathbb{E}(y_i - \omega)^2\ \log(1/\delta)}{n\tau_\sigma^2}} + \frac{\log(1/\delta)}{3n} \\
&\le \sqrt{\frac{2(\sigma^2 + r^2)\log(1/\delta)}{n\tau_\sigma^2}} + \frac{\log(1/\delta)}{3n} \\
&\le \frac{2z\sqrt{\log(1/\delta)}}{n} + \frac{\log(1/\delta)}{3n}
\end{aligned}$$

provided $r^2 \le \sigma^2$. Let

$$g(x) = -\frac{1}{n}\sum_{i=1}^n \frac{\tau}{\sqrt{\tau^2 + (x + \varepsilon_i)^2}}.$$

Using the mean value theorem and the inequality that $|x/(1+x^2)^{3/2}| \le 1/2$, we obtain

$$|g(x) - g(y)| = \left|\frac{1}{n}\sum_{i=1}^n \frac{(\widetilde{x} + \varepsilon_i)/\tau_\sigma}{(1 + (\widetilde{x} + \varepsilon_i)^2/\tau_\sigma^2)^{3/2}} \cdot \frac{x - y}{\tau_\sigma}\right| \le \frac{|x - y|}{2\tau_\sigma},$$

where $\widetilde{x}$ is some convex combination of x and y. Then we have

$$\mathrm{I}_2 = \left|\frac{1}{n}\sum_{i=1}^n \frac{(\widetilde{\Delta} + \varepsilon_i)/\tau_\sigma}{(1 + (\widetilde{\Delta} + \varepsilon_i)^2/\tau_\sigma^2)^{3/2}} \cdot \frac{\Delta_\mu - \Delta_\omega}{\tau_\sigma}\right| \le \frac{\epsilon}{2\tau_\sigma}$$

where $\widetilde{\Delta}$ is some convex combination of $\Delta_w = \mu^* - w$ and $\Delta_\mu = \mu^* - \mu$. For II_3, a similar argument for bounding II_2 yields

$$\begin{aligned}
\mathrm{I}_3 &= \left|\mathbb{E}\left(\frac{(\widetilde{\Delta} + \varepsilon_i)/\tau_\sigma}{(1 + (\widetilde{\Delta} + \varepsilon_i)^2/\tau_\sigma^2)^{3/2}}\right) \cdot \frac{\Delta_\mu - \Delta_\omega}{\tau_\sigma}\right| \\
&\le \mathbb{E}|\widetilde{\Delta} + \varepsilon_i| \cdot \frac{\epsilon}{\tau_\sigma^2} \\
&\le \frac{\epsilon\sqrt{2(r^2 + \sigma^2)}}{\tau_\sigma^2},
\end{aligned}$$

where the last inequality uses Jensen's inequality, i.e. $\mathbb{E}|\widetilde{\Delta} + \varepsilon_i| \le \sqrt{\mathbb{E}(\widetilde{\Delta} + \varepsilon_i^2)} \le \sqrt{2(r^2 + \sigma^2)}$. Putting the above pieces together and using the union bound, we obtain with probability at least

$1 - 12\epsilon^{-1}r\delta$

$$\begin{aligned}\mathrm{I} &\leq \frac{n}{z^2} \cdot \sup_{\omega \in \mathcal{N}} \left| \frac{1}{n} \sum_{i=1}^n \frac{\sqrt{1 + (y_i - \omega)^2/\tau_\sigma^2} - 1}{\sqrt{1 + (y_i - \omega)^2/\tau_\sigma^2}} - \mathbb{E} \frac{\sqrt{1 + (y_i - \omega)^2/\tau_\sigma^2} - 1}{\sqrt{1 + (y_i - \omega)^2/\tau_\sigma^2}} \right| \\ &\quad + \frac{n}{z^2} \cdot \frac{\epsilon}{2\tau_\sigma} \left(1 + \frac{2\sqrt{2(r^2 + \sigma^2)}}{\tau_\sigma} \right) \\ &\leq \frac{2\sqrt{\log(1/\delta)}}{z} + \frac{\log(1/\delta)}{3z^2} + \frac{\epsilon\sqrt{n}}{\sigma z},\end{aligned}$$

provided that

$$2\sqrt{2(r^2 + \sigma^2)} \leq \tau_\sigma.$$

Putting above results together, we obtain with probability at least $1 - (12r/\epsilon + 2)\delta$ that

$$\begin{aligned}|\widehat{v} - \sigma| &\lesssim \mathrm{I} + \mathrm{II} \\ &\leq \frac{2\sqrt{\log(1/\delta)}}{z} + \frac{\log(1/\delta)}{3z^2} + \frac{\epsilon\sqrt{n}}{\sigma z} + \frac{\gamma}{2} + \frac{2r}{\sigma}.\end{aligned}$$

Let $C' = 24CV_0^2/v_0$. Therefore, taking $\epsilon = 1/\sqrt{n}$, $\delta = 1/\log n$, and $z^2 \geq \log(n)$, we obtain with probability at least

$$1 - \frac{C'\left(\sqrt{\log n} + \log\log n/\sqrt{\log n}\right) + 2}{\log n}$$

that

$$|\widehat{v} - \sigma| \lesssim \sqrt{\frac{\log\log n}{\log n}} + \frac{\log\log n}{\log n} + \frac{1}{\sqrt{\log n}} + \gamma + r \to 0.$$

Therefore $\widehat{v} \to \sigma$ in probability. This finishes the proof.

□

G.4 Local strong convexity in v

In this section, we first present the local strong convexity of the empirical loss function with respect to v uniformly over a neighborhood of μ^*.

Lemma G.4 (Local strong convexity in v)**.** Let $\mathbb{B}_r(\mu^*) = \{\mu : |\mu - \mu^*| \leq r\}$. Assume $r = r(n) = o(1)$. Let $0 < \delta < 1$ and n is sufficiently large. Take ϖ such that $\max\{\varpi r\sqrt{n}, \varpi\} \to 0$ and $\varpi\sqrt{n} \to \infty$. Then, with probability at least $1 - \delta$, we have

$$\inf_{\mu \in \mathbb{B}_r(\mu^*)} \frac{\langle \nabla_v L_n(\mu, v) - \nabla_v L_n(\mu, v_*), v - \sigma \rangle}{|v - \sigma|^2} \geq \rho_\ell = \frac{\sigma^2_{c\varpi^2 n/(4z^2)}}{2(v^3 \vee \sigma^3)} \geq \frac{c_0}{v^3 \vee \sigma^3},$$

where c and c_0 are some constants.

Proof of Lemma G.4. Recall $\tau = v\sqrt{n}/z$. For notational simplicity, write $\tau_\sigma = \sigma\sqrt{n}/z$, $\tau_{v_0} = v_0\sqrt{n}/z$, $\tau_\varpi = \varpi\sqrt{n}/z$, and $\Delta = \mu^* - \mu$. It follows that

$$\begin{aligned}\langle \nabla_v L_n(\mu, v) - \nabla_v L_n(\mu, \sigma), v - \sigma \rangle &= \frac{n}{z^2} \left\langle \frac{1}{n} \sum_{i=1}^n \frac{\tau}{\sqrt{\tau^2 + (y_i - \mu)^2}} - \frac{1}{n} \sum_{i=1}^n \frac{\tau_\sigma}{\sqrt{\tau_\sigma^2 + (y_i - \mu)^2}}, v - \sigma \right\rangle \\ &= \frac{n^{3/2}}{z^3} \cdot \frac{1}{n} \sum_{i=1}^n \frac{(y_i - \mu)^2}{(\widetilde{\tau}^2 + (y_i - \mu)^2)^{3/2}} |v - \sigma|^2 \\ &\geq \frac{n^{3/2}}{z^3} \cdot \frac{1}{n} \sum_{i=1}^n \frac{(y_i - \mu)^2}{((\tau \vee \tau_\sigma)^2 + (y_i - \mu)^2)^{3/2}} |v - \sigma|^2\end{aligned}$$

where $\widetilde{\tau}$ is some convex combination of τ and τ_σ, that is $\widetilde{\tau} = (1-\lambda)\tau_\sigma + \lambda\tau$ for some $\lambda \in [0,1]$. Because $\tau^3 x^2/(\tau^2 + x^2)^{3/2}$ is an increasing function of τ, if $\tau_\varpi \le \tau \vee \tau_\sigma$, we have

$$\frac{\langle \nabla_v L_n(\mu, v) - \nabla_v L_n(\mu,\sigma), v - v_*\rangle}{|v-\sigma|^2} \ge \frac{n^{3/2}}{z^3(\tau\vee\tau_\sigma)^3}\cdot\frac{1}{n}\sum_{i=1}^n \frac{(\tau\vee\tau_\sigma)^3(y_i-\mu)^2}{(\tau^2\vee\tau_\sigma^2 + (y_i-\mu)^2)^{3/2}}$$
$$\ge \frac{n^{3/2}}{z^3(\tau\vee\tau_\sigma)^3}\cdot\frac{1}{n}\sum_{i=1}^n \frac{\tau_\varpi^3(y_i-\mu)^2}{(\tau_\varpi^2 + (y_i-\mu)^2)^{3/2}}.$$

Thus

$$\begin{aligned}
&\inf_{\mu\in\mathbb{B}_r(\mu^*)} \frac{\langle \nabla_v L_n(\mu, v) - \nabla_v L_n(\mu,\sigma), v - v_*\rangle}{|v-\sigma|^2}\\
&\ge \frac{n^{3/2}}{z^3(\tau\vee\tau_*)^3}\cdot \inf_{\mu\in\mathbb{B}_r(\mu^*)}\frac{1}{n}\sum_{i=1}^n \frac{\tau_\varpi^3(y_i-\mu)^2}{(\tau_\varpi^2 + (y_i-\mu)^2)^{3/2}}\\
&= \frac{n^{3/2}}{z^3(\tau\vee\tau_\sigma)^3}\cdot\left(\inf_{\mu\in\mathbb{B}_r(\mu^*)}\left(\mathbb{E}\frac{\tau_\varpi^3(y_i-\mu)^2}{(\tau_\varpi^2 + (y_i-\mu)^2)^{3/2}}\right)\right.\\
&\qquad\left. - \sup_{\mu\in\mathbb{B}_r(\mu^*)}\left|\frac{1}{n}\sum_{i=1}^n \frac{\tau_\varpi^3(y_i-\mu)^2}{(\tau_\varpi^2 + (y_i-\mu)^2)^{3/2}} - \mathbb{E}\frac{\tau_\varpi^3(y_i-\mu)^2}{(\tau_\varpi^2 + (y_i-\mu)^2)^{3/2}}\right|\right)\\
&= \frac{n^{3/2}}{z^3(\tau\vee\tau_\sigma)^3}\cdot(\mathrm{I}-\mathrm{II}).
\end{aligned}$$

It remains to lower bound I and upper bound II. We start with I. Let $f(x) = x/(1+x)^{3/2}$ which satisfies

$$f(x) \ge \begin{cases}\epsilon x & x \le c_\epsilon\\ 0 & x > c_\epsilon,\end{cases}$$

and $Z = (y-\mu)^2/\tau_\varpi^2$ in which $y \sim y_i$. Suppose $r^2 \le c_\epsilon \tau_\varpi^2/4$, then we have

$$\begin{aligned}
\inf_{\mu\in\mathbb{B}_r(\mu^*)}\left(\mathbb{E}\frac{\tau_\varpi^3(y_i-\mu)^2}{(\tau_\varpi^2 + (y_i-\mu)^2)^{3/2}}\right) &= \inf_{\mu\in\mathbb{B}_r(\mu^*)}\mathbb{E}\left(\frac{\tau_\varpi^2 Z}{(1+Z)^{3/2}}\right)\\
&\ge \epsilon\cdot\inf_{\mu\in\mathbb{B}_r(\mu^*)}\mathbb{E}\left[(y-\mu)^2 1((y-\mu)^2 \le c_\epsilon\tau_\varpi^2)\right]\\
&\ge \epsilon\cdot\inf_{\mu\in\mathbb{B}_r(\mu^*)}\mathbb{E}\left[(y-\mu)^2 1(\varepsilon^2 \le c_\epsilon\tau_\varpi^2/2 - r^2)\right]\\
&\ge \epsilon\cdot\inf_{\mu\in\mathbb{B}_r(\mu^*)}\left(\mathbb{E}\left[(\Delta^2+\varepsilon^2)1\left(\varepsilon^2 \le \frac{c_\epsilon\tau_\varpi^2}{4}\right)\right] - \frac{8\Delta\sigma^2}{c_\epsilon\tau_\varpi^2}\right)\\
&\ge \epsilon\cdot\left(\mathbb{E}\left[\varepsilon^2 1\left(\varepsilon^2 \le \frac{c_\epsilon\tau_\varpi^2}{4}\right)\right] - \frac{8r\sigma^2}{c_\epsilon\tau_\varpi^2}\right).
\end{aligned}$$

We then proceed with II. For any $0 < \gamma \le 2r$, there exists an γ-cover $\mathcal{N}$ of $\mathbb{B}_r(\mu^*)$ such that $|\mathcal{N}| \le 6r/\gamma$. Then for any $\mu \in \mathbb{B}_r(\mu^*)$ there exists an $\omega \in \mathcal{N}$ such that $|\omega - \mu| \le \gamma$, and thus by Lemma G.5 we have

$$\begin{aligned}
&\left|\frac{1}{n}\sum_{i=1}^n \frac{\tau_\varpi^3(y_i-\mu)^2}{(\tau_\varpi^2 + (y_i-\mu)^2)^{3/2}} - \mathbb{E}\frac{\tau_\varpi^3(y_i-\mu)^2}{(\tau_\varpi^2 + (y_i-\mu)^2)^{3/2}}\right|\\
&\le \left|\frac{1}{n}\sum_{i=1}^n \frac{\tau_\varpi^3(y_i-\omega)^2}{(\tau_\varpi^2 + (y_i-\omega)^2)^{3/2}} - \mathbb{E}\frac{\tau_\varpi^3(y_i-\omega)^2}{(\tau_\varpi^2 + (y_i-\omega)^2)^{3/2}}\right|\\
&\quad + \left|\frac{1}{n}\sum_{i=1}^n \frac{\tau_\varpi^3(y_i-\omega)^2}{(\tau_\varpi^2 + (y_i-\omega)^2)^{3/2}} - \frac{1}{n}\sum_{i=1}^n \frac{\tau_\varpi^3(y_i-\mu)^2}{(\tau_\varpi^2 + (y_i-\mu)^2)^{3/2}}\right|\\
&\quad + \left|\mathbb{E}\frac{\tau_\varpi^3(y_i-\omega)^2}{(\tau_\varpi^2 + (y_i-\omega)^2)^{3/2}} - \mathbb{E}\frac{\tau_\varpi^3(y_i-\mu)^2}{(\tau_\varpi^2 + (y_i-\mu)^2)^{3/2}}\right|\\
&= \mathrm{II}_1 + \mathrm{II}_2 + \mathrm{II}_3.
\end{aligned}$$

For II_1, Lemma G.5 implies with probability at least $1-2\delta$

$$\mathrm{II}_1 \le \sqrt{\frac{2\tau_\varpi^2 \mathbb{E}(y_i-\omega)^2\log(1/\delta)}{3n}} + \frac{\tau_\varpi^2\log(1/\delta)}{3\sqrt{3}n} \le \sqrt{\frac{2\tau_\varpi^2(\sigma^2+r^2)\log(1/\delta)}{3n}} + \frac{\tau_\varpi^2\log(1/\delta)}{3\sqrt{3}n}.$$

Let

$$g(x) = \frac{1}{n}\sum_{i=1}^n \frac{\tau^3(x+\varepsilon_i)^2}{(\tau^2+(x+\varepsilon_i)^2)^{3/2}}.$$

Using the mean value theorem and the inequality that $|\tau^2 x/(\tau^2+x^2)^{3/2}| \le 1/\sqrt{3}$, we obtain

$$|g(x)-g(y)| = \left|\frac{1}{n}\sum_{i=1}^n \frac{\tau^3(\widetilde{x}+\varepsilon_i)\left(\tau^2-(\widetilde{x}+\varepsilon_i)^2\right)}{(\tau^2+(\widetilde{x}+\varepsilon_i)^2)^{5/2}}(x-y)\right| \le \frac{\tau}{\sqrt{3}}|x-y|.$$

Then we have

$$\mathrm{II}_2 = \left|\frac{1}{n}\sum_{i=1}^n \frac{\tau_\varpi^3(\widetilde{\Delta}+\varepsilon_i)\left(\tau_\varpi^2-(\widetilde{\Delta}+\varepsilon_i)^2\right)}{(\tau_\varpi^2+(\widetilde{\Delta}+\varepsilon_i)^2)^{5/2}}(\Delta_w-\Delta_\mu)\right| \le \frac{\tau_\varpi\gamma}{\sqrt{3}}$$

where $\widetilde{\Delta}$ is some convex combination of $\Delta_w=\mu^*-w$ and $\Delta_\mu=\mu^*-\mu$. For II_3, we have

$$\mathrm{II}_3 = \left|\mathbb{E}\left(\frac{\tau_\varpi^3(\widetilde{\Delta}+\varepsilon_i)\left(\tau_\varpi^2-(\widetilde{\Delta}+\varepsilon_i)^2\right)}{(\tau_\varpi^2+(\widetilde{\Delta}+\varepsilon_i)^2)^{5/2}}\right)(\Delta_w-\Delta_\mu)\right| \le \gamma\mathbb{E}|\widetilde{\Delta}+\varepsilon_i| \le \gamma\sqrt{\mathbb{E}\left(\widetilde{\Delta}+\varepsilon_i\right)^2},$$

where the last inequality uses Jensen's inequality. Putting the above pieces together and using the union bound, we obtain with probability at least $1-12\gamma^{-1}r\delta$

$$\begin{aligned}
\mathrm{II} &\le \sup_{\omega\in\mathcal{N}}\left|\frac{1}{n}\sum_{i=1}^n \frac{\tau_\varpi^3(y_i-\omega)^2}{(\tau_\varpi^2+(y_i-\omega)^2)^{3/2}} - \mathbb{E}\frac{\tau_\varpi^3(y_i-\omega)^2}{(\tau_\varpi^2+(y_i-\omega)^2)^{3/2}}\right| + \frac{\tau_\varpi\gamma}{\sqrt{3}} + \gamma\sqrt{r^2+\sigma^2} \\
&\le \sqrt{\frac{2\tau_\varpi^2(r^2+\sigma^2)\log(1/\delta)}{3n}} + \frac{\tau_\varpi^2\log(1/\delta)}{3\sqrt{3}n} + \frac{\tau_\varpi\gamma}{\sqrt{3}} + \gamma\sqrt{r^2+\sigma^2} \\
&= \sqrt{r^2+\sigma^2}\left(\sqrt{\frac{2\varpi^2\log(1/\delta)}{3z^2}}+\gamma\right) + \frac{\varpi^2\log(1/\delta)}{3\sqrt{3}z^2} + \frac{\varpi\gamma\sqrt{n}}{\sqrt{3}}.
\end{aligned}$$

Combining the bounds for I and II yields with probability at least $1-\delta$

$$\begin{aligned}
&\inf_{\mu\in\mathbb{B}_r(\mu^*)} \frac{\langle \nabla_v L_n(\mu,v)-\nabla_v L_n(\mu,\sigma), v-\sigma\rangle}{|v-\sigma|^2} \\
&\ge \frac{n^{3/2}}{z^3(\tau\vee\tau_\sigma)^3}\Bigg\{\epsilon\left(\mathbb{E}\left[\varepsilon^2 1\left(\varepsilon^2\le\frac{c_\epsilon\tau_\varpi^2}{4}\right)\right]-\frac{8r\sigma^2}{c_\epsilon\tau_\varpi^2}\right) \\
&\qquad - \sqrt{r^2+\sigma^2}\left(\sqrt{\frac{2\varpi^2\log(1/\delta)}{3z^2}}+\gamma\right) - \frac{\varpi^2\log(1/\delta)}{3\sqrt{3}z^2} - \frac{\varpi\gamma\sqrt{n}}{\sqrt{3}}\Bigg\} \\
&\ge \frac{1}{2(v\vee\sigma)^3}\mathbb{E}\left[\varepsilon^2 1\left(\varepsilon^2\le\frac{c_\epsilon\tau_\varpi^2}{4}\right)\right]
\end{aligned}$$

where ϵ,ϖ,γ,n are picked such that $\epsilon=3/4$, $\gamma=12r$, and

$$\begin{aligned}
&\epsilon\left(\mathbb{E}\left[\varepsilon^2 1\left(\varepsilon^2\le\frac{c_\epsilon\tau_\varpi^2}{4}\right)\right]-\frac{8r\sigma^2z^2}{c_\epsilon\varpi^2 n}\right) - \sqrt{r^2+\sigma^2}\left(\sqrt{\frac{2\varpi^2\log(1/\delta)}{3z^2}}+\gamma\right) - \frac{\varpi^2\log(1/\delta)}{3\sqrt{3}z^2} - \frac{\varpi\gamma\sqrt{n}}{\sqrt{3}} \\
&\ge \frac{1}{2}\mathbb{E}\left[\varepsilon^2 1\left(\varepsilon^2\le\frac{c_\epsilon\tau_\varpi^2}{4}\right)\right] \ge \frac{1}{4}\sigma.
\end{aligned}$$

For example, we can pick ϖ such that

$$\max\{\varpi r\sqrt{n},\varpi\}\to 0 \text{ and } \varpi\sqrt{n}\to\infty$$

as $n\to\infty$. This completes the proof.

□

G.5 SUPPORTING LEMMAS

This subsection proves a supporting lemma that is used prove Lemma G.4.

Lemma G.5. Let w_i be i.i.d. copies of w. For any $0 < \delta < 1$, we have

$$\frac{1}{n}\sum_{i=1}^{n}\frac{\tau^3 w_i^2}{(\tau^2+w_i^2)^{3/2}} - \mathbb{E}\frac{\tau^3 w_i^2}{(\tau^2+w_i^2)^{3/2}} \geq -\sqrt{\frac{2\tau^2\mathbb{E}w_i^2\log(1/\delta)}{3n}} - \frac{\tau^2\log(1/\delta)}{3\sqrt{3}n}, \text{ with prob. } 1-\delta,$$

$$\left|\frac{1}{n}\sum_{i=1}^{n}\frac{\tau^3 w_i^2}{(\tau^2+w_i^2)^{3/2}} - \mathbb{E}\frac{\tau^3 w_i^2}{(\tau^2+w_i^2)^{3/2}}\right| \leq \sqrt{\frac{2\tau^2\mathbb{E}w_i^2\log(1/\delta)}{3n}} + \frac{\tau^2\log(1/\delta)}{3\sqrt{3}n}, \text{ with prob. } 1-2\delta.$$

Proof of Lemma G.5. We only prove the first result and the second result follows similarly. The random variables $Z_i = Z_i(\tau) := \tau^3 w_i^2/(\tau^2+w_i^2)^{3/2}$ with $\mu_z = \mathbb{E}Z_i$ and $\sigma_z^2 = \mathrm{var}(Z_i)$ are bounded i.i.d. random variables such that

$$0 \leq Z_i = \tau^3 w_i^2/(\tau^2+w_i^2)^{3/2} \leq w_i^2 \wedge \frac{\tau^2}{\sqrt{3}} \wedge \frac{\tau|w_i|}{\sqrt{3}}.$$

Moreover we have

$$\mathbb{E}Z_i^2 = \mathbb{E}\left(\frac{\tau^6 w_i^4}{(\tau^2+\varepsilon_i^2)^3}\right) \leq \frac{\tau^2\mathbb{E}w_i^2}{3},\ \sigma_z^2 := \mathrm{var}(Z_i) \leq \frac{\tau^2\mathbb{E}w_i^2}{3}.$$

For third and higher order absolute moments, we have

$$\mathbb{E}|Z_i|^k = \mathbb{E}\left|\frac{\tau^3 w_i^2}{(\tau^2+\varepsilon_i^2)^{3/2}}\right|^k \leq \frac{\tau^2\mathbb{E}w_i^2}{3}\cdot\left(\frac{\tau^2}{\sqrt{3}}\right)^{k-2} \leq \frac{k!}{2}\cdot\frac{\tau^2\mathbb{E}w_i^2}{3}\cdot\left(\frac{\tau^2}{3\sqrt{3}}\right)^{k-2}, \text{ for all integers } k\geq 3.$$

Therefore, using Lemma H.2 with $v = n\tau^2\,\mathbb{E}w_i^2/3$ and $c = \tau^2/(3\sqrt{3})$ acquires that for any $t \geq 0$

$$\mathbb{P}\left(\sum_{i=1}^{n}\frac{\tau^3 w_i^2}{(\tau^2+\varepsilon_i^2)^{3/2}} - \sum_{i=1}^{n}\mathbb{E}\left(\frac{\tau^3 w_i^2}{(\tau^2+\varepsilon_i^2)^{3/2}}\right) \geq -\sqrt{\frac{2n\tau^2\mathbb{E}w_i^2 t}{3}} - \frac{\tau^2 t}{3\sqrt{3}}\right) \leq \exp(-t).$$

Taking $t = \log(1/\delta)$ acquires that for any $0 < \delta < 1$

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n}\frac{\tau^3 w_i^2}{(\tau^2+w_i^2)^{3/2}} - \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left(\frac{\tau^3 w_i^2}{(\tau^2+\varepsilon_i^2)^{3/2}}\right) > -\sqrt{\frac{2\tau^2\mathbb{E}w_i^2\log(1/\delta)}{3n}} - \frac{\tau^2\log(1/\delta)}{3\sqrt{3}n}\right) > 1-\delta.$$

This finishes the proof. □

H PRELIMINARY LEMMAS

This section collects preliminary lemmas that are frequently used in the proofs for the main results and supporting lemmas. We first collect the Hoeffding's inequality and then present a form of Bernstein's inequality. We omit their proofs and refer interested readers to Boucheron et al. (2013).

Lemma H.1 (Hoeffding's inequality)**.** Let $Z_1, \ldots, Z_n$ be independent real-valued random variables such that $a \leq Z_i \leq b$ almost surely. Let $S_n = \sum_{i=1}^n (Z_i - \mathbb{E}Z_i)$ and $v = n(b-a)^2$. Then for all $t \geq 0$,

$$\mathbb{P}\left(S_n \geq \sqrt{vt/2}\right) \leq e^{-t},\ \mathbb{P}\left(S_n \leq -\sqrt{vt/2}\right) \leq e^{-t},\ \mathbb{P}\left(|S_n| \geq \sqrt{vt/2}\right) \leq 2e^{-t}.$$

Lemma H.2 (Bernstein's inequality)**.** Let $Z_1, \ldots, Z_n$ be independent real-valued random variables such that

$$\sum_{i=1}^{n}\mathbb{E}Z_i^2 \leq v,\ \sum_{i=1}^{n}\mathbb{E}|Z_i|^k \leq \frac{k!}{2}vc^{k-2} \text{ for all } k\geq 3.$$

If $S_n = \sum_{i=1}^n (Z_i - \mathbb{E}Z_i)$, then for all $t \geq 0$,

$$\mathbb{P}\left(S_n \geq \sqrt{2vt}+ct\right) \leq e^{-t},\ \mathbb{P}\left(S_n \leq -(\sqrt{2vt}+ct)\right) \leq e^{-t},\ \mathbb{P}\left(|S_n| \geq \sqrt{2vt}+ct\right) \leq 2e^{-t}.$$

Proof of Lemma H.2. This lemma involves a two-sided extension of Theorem 2.10 by Boucheron et al. (2013). The proof follows from a similar argument used in the proof of Theorem 2.10, and thus is omitted. □

Our third lemma concerns the localized Bregman divergence for convex functions. It was first established in Fan et al. (2018). For any loss function L, define the Bregman divergence and the symmetric Bregman divergence as

$$D_L(\beta_1, \beta_2) = L(\beta_1) - L(\beta_2) - \langle \nabla L(\beta_2), \beta_1 - \beta_2 \rangle,$$
$$D_L^s(\beta_1, \beta_2) = D_L(\beta_1, \beta_2) + D_L(\beta_2, \beta_1).$$

Lemma H.3. For any $\beta_\eta = \beta^* + \eta(\beta - \beta^*)$ with $\eta \in (0, 1]$ and any convex loss function L, we have

$$D_L^s(\beta_\eta, \beta^*) \leq \eta D_L^s(\beta, \beta^*).$$

Our forth lemma in this section concerns three basic inequalities that are frequently used in the proofs.

Lemma H.4. The following inequalities hold:

(i) $(1 + x)^r \geq 1 + rx$ for $x \geq -1$ and $r \in \mathbb{R} \setminus (0, 1)$;

(ii) $(1 + x)^r \leq 1 + rx$ for $x \geq -1$ and $r \in (0, 1)$;

(iii) $(1 + x)^r \leq 1 + (2^r - 1)x$ for $x \in [0, 1]$ and $r \in \mathbb{R} \setminus (0, 1)$.