A Technical Details

Throughout the proofs, we use C, c to denote positive constants, and they are not necessarily the same at each occurrence. Given a matrix A, we use $\mathcal{R}(A)$ to denote the column space (range) of A and A^+ to denote the Moore-Penrose inverse of A. Denote by $P_A = A(A^TA)^+A^T$ the orthogonal projection on $\mathcal{R}(A)$.

A.1 Basics

The conventional definition of Orlicz ψ-norms goes as follows: Given a strictly increasing convex function ψ on $\mathbb{R}_+ := [0, +\infty)$ with $\psi(0) = 0$, then the Orlicz ψ-norm of a random variable Y is defined as $\|Y\|_\psi = \inf\{t > 0 : \mathbb{E}\psi(|Y|/t) \leq 1\}$.

Some well-known examples of Orlicz ψ-norms are the L_q-norms: $\|Y\|_q = (\mathbb{E}|Y|^q)^{1/q}$ associated with $\psi(x) = x^q$ $(q \geq 1)$ (e.g., the Pareto distribution), and the ψ_q-norms $(q \geq 1)$ with

$$\psi_q(x) = \exp(x^q) - 1. \tag{A.1}$$

(A.1) encompasses both sub-Gaussian and sub-Exponential type random variables for $q = 2, 1$, without the requirement for the random variables to be centered.

Our upcoming theorems often relax the strict requirements of *strict* monotonicity and *convexity* for ψ. This flexibility allows us to handle random variables with much heavier tails. For instance, we consider the extension of (A.1), known as *sub-Weibull* random variables, which have finite ψ_q-norms for $q > 0$ (cf. Kuchibhotla and Chakrabortty (2022)). As $0 < q < 1$, ψ_q is nonconvex, and these random variables, including Weibull, exhibit heavier tails compared to sub-Exponential ones (Götze et al., 2021).

Throughout the paper, when we refer to the Orlicz norm of a random variable, denoted as $\|\cdot\|_\psi$ or sometimes $\|\cdot\|_\varphi$, it is always understood that $\psi(\cdot)$ (or $\varphi(\cdot)$) is an nondecreasing function defined on $\mathbb{R}_+$ with $\psi(0) = 0$. Theorem 1 also requires ψ to satisfy the regularity condition $\limsup_{x,y\to\infty} \psi(x)\psi(y)/\psi(cxy) < \infty$ for some constant $c > 0$ (van der Vaart and Wellner, 2013). It is easy to verify that all these conditions are met by L_q with $q > 0$ and ψ_q with $0 < q \leq 2$. Orlicz norms provide a useful framework for analyzing skewed random variables, including those without zero mean.

In this paper, we define that a random vector $\epsilon \in \mathbb{R}^n$ has its Orlicz ψ-norm $\|\epsilon\|_\psi$ bounded above by ω if

$$\|\langle \epsilon, \alpha\rangle\|_\psi \leq \omega\|\alpha\|_2, \ \forall \alpha \in \mathbb{R}^n. \tag{A.2}$$

Note that (A.2) is defined using the Euclidean norm $\|\cdot\|_2$, and the function ψ may not necessarily be convex. Furthermore, the components of ϵ are not required to be independent or centered. However, if ϵ does have centered, independent components, its vector Orlicz ψ-norm is bounded by the largest Orlicz ψ-norm among its components (up to a multiplicative constant)..

Lemma A.1. *Let $\epsilon_1, \ldots, \epsilon_n$ be centered, independent random variables satisfying $\|\epsilon_i\|_{\psi_q} \leq \omega$ for some $q \in (0, 2]$. Given any $\alpha \in \mathbb{R}^n$, we have $\|\langle \alpha, \epsilon\rangle\|_{\psi_q} \leq C\|\alpha\|_2\omega$, where C is a constant depending on q only.*

To prove the lemma, we first introduce two lemmas. The first is Theorem 1.5 in Götze et al. (2021).

Lemma A.2. *Let $\epsilon_1, \dots, \epsilon_n$ be independent random variables satisfying $\|\epsilon_i\|_{\psi_q} \le \omega$ for some $q \in (0, 2]$. Let $f(\epsilon) : \mathbb{R}^n \to \mathbb{R}$ be a polynomial of degree $D \in \mathbb{N}$ and denote by $f^{(d)}$ the d-tensor of its d-th order partial derivatives for $1 \le d \le D$. Then for all $t > 0$, we have*

$$\mathbb{P}(|f(\epsilon) - \mathbb{E}f(\epsilon)| \ge t) \le 2\exp\Big(-C \min_{1\le d\le D} \big(\frac{t}{\omega^d \|\mathbb{E}f^{(d)}(\epsilon)\|_{HS}}\big)^{\frac{q}{d}}\Big), \tag{A.3}$$

where C is a constant that depends on D and q and $\|\cdot\|_{HS}$ denotes the Hilbert-Schmidt norm (or the Frobenius norm in the case of a matrix).

The second fact is a slight modification of Lemma 2.2.1 in van der Vaart and Wellner (2013).

Lemma A.3. *Let X be a random variable such that for some $q > 0$,*

$$P(|X| > t) \le c\exp\Big(-\big(\frac{t}{\omega}\big)^q\Big), \quad \forall t > 0, \tag{A.4}$$

where $\omega > 0$ and c is constant, then we have $\|X\|_{\psi_q} \lesssim \omega$.

The proof is straightforward:

$$\begin{aligned}
\mathbb{E}\big(\exp(M|X|^q) - 1\big) &= \mathbb{E}\int_0^{|X|^q} M\exp(Mu)\,\mathrm{d}u \\
&= \int_0^{\infty} \mathbb{P}(|X| > u^{\frac{1}{q}}) M\exp(Mu)\,\mathrm{d}u \\
&= \int_0^{\infty} cM\exp(Mu - \frac{u}{\omega^q})\,\mathrm{d}u \le \frac{cM}{\omega^{-q} - M}.
\end{aligned}$$

It suffices to take $M^{-1/q} \ge c'\omega$ to have $\mathbb{E}(\exp(M|X|^q) - 1) \le 1$. Therefore, $\|X\|_{\psi_q} \lesssim \omega$.

Now, given any $\alpha \in \mathbb{R}^n$, let $f_\alpha(\epsilon) = \langle \alpha, \epsilon\rangle = \sum_{i=1}^n \alpha_i\epsilon_i$. Then (i) $f_\alpha(\epsilon)$ is an 1-degree polynomial of $\epsilon_1, \dots, \epsilon_n$, (ii) $\|\mathbb{E}[\nabla f_\alpha(\epsilon)]\|_2 = \|\alpha\|_2$ and $\mathbb{E}[\sum_{i=1}^n \alpha_i\epsilon_i] = 0$, and (iii) $\epsilon_1, \dots, \epsilon_n$ are independent and satisfy $\|\epsilon_i\|_{\psi_q} \le \omega$. Applying Lemma A.2 with $D = 1, d = 1$ yields

$$\mathbb{P}\big(|\sum_{i=1}^n \alpha_i\epsilon_i| > t\big) = \mathbb{P}\big(|f_\alpha(\epsilon)| > t\big) \le 2\exp\Big(-C\big(\frac{t}{\omega\|\alpha\|_2}\big)^q\Big), \tag{A.5}$$

where C is a constant depending on q only. By Lemma A.3, (A.5) implies $\|\langle\alpha, \epsilon\rangle\|_{\psi_q} \le C\|\alpha\|_2\omega$, where C is a constant depending on q only. The proof of Lemma A.1 is complete.

The following lemma is useful for stating the assumptions on effective noises.

Lemma A.4. *Let ψ, φ be any two nondecreasing nonzero functions defined on $\mathbb{R}_+$ with $\psi(0) = \varphi(0) = 0$ (not necessarily convex). Define $\varphi^{-1}(t) := \sup\{x \in \mathbb{R}_+ : \varphi(x) \le t\}$. and ψ^{-1} similarly. (i) Suppose that $\psi(\varphi^{-1}(t)/c_0)$ is concave in t on $\mathbb{R}_+$ for some $c_0 > \varphi^{-1}(1)/\psi^{-1}(1)$. Then for any random variable X, we have $\|X\|_\psi \le c_0\|X\|_\varphi$. (ii) Suppose that $\psi(\varphi^{-1}(t)/c_0) \le t$ for some $c_0 > 0$, then $\|X\|_\psi \le c_0\|X\|_\varphi$.*

We remark that the condition for c_0 in part (i) can be replaced by $c_0 \geq \varphi^{-1}(1)/\psi^{-1}(1)$ when ψ is continuous at 1. For completeness, we provide the proof below.

First, by the definition of φ^{-1}, $u \leq \varphi^{-1}(\varphi(u))$ for any $u \geq 0$. Therefore, $X/\|X\|_\varphi \leq \varphi^{-1}\big\{\varphi(X/\|X\|_\varphi)\big\}$, from which it follows that

$$\psi\Big(\frac{X}{c\|X\|_\varphi}\Big) \leq \psi\Big[\frac{1}{c}\varphi^{-1}\big\{\varphi(\frac{X}{\|X\|_\varphi})\big\}\Big]. \tag{A.6}$$

To prove part (i), let $f(u) = \psi\big(\varphi^{-1}(u)/c\big)$ with $c > 0$ to be determined. Then $f(u)$ is an increasing function on $\mathbb{R}_+$. (In fact, for any $u \geq u' \geq 0$, $\varphi^{-1}(u') \leq \varphi^{-1}(u)$, and so $\psi\{\varphi^{-1}(u')/c\} \leq \psi\{\varphi^{-1}(u)/c\}$.) From (A.6), picking $t = \varphi(X/\|X\|_\varphi)$ gives

$$\psi\Big(\frac{X}{c\|X\|_\varphi}\Big) \leq \psi\Big[\frac{1}{c}\varphi^{-1}\big\{\varphi(\frac{X}{\|X\|_\varphi})\big\}\Big] = f(t). \tag{A.7}$$

With $c > \varphi^{-1}(1)/\psi^{-1}(1)$ (or $c \geq \varphi^{-1}(1)/\psi^{-1}(1)$ when ψ is continuous at 1), we can use Jensen's inequality to get

$$\mathbb{E}\Big\{\psi\Big(\frac{X}{c\|X\|_\varphi}\Big)\Big\} \leq \mathbb{E}\Big(\psi\Big[\frac{1}{c}\varphi^{-1}\big\{\varphi(\frac{X}{\|X\|_\varphi})\big\}\Big]\Big) = \mathbb{E}\big(f(t)\big) \leq f\big(\mathbb{E}(t)\big) \leq f(1) \leq 1. \tag{A.8}$$

To prove part (ii), we still set $f(u) = \psi\big(\varphi^{-1}(u)/c\big)$ with $c > 0$ and $t = \varphi(X/\|X\|_\varphi)$. Based on (A.6), we get

$$\psi\big(\frac{X}{c\|X\|_\varphi}\big) \leq f(t) \leq t = \varphi\big(\frac{X}{\|X\|_\varphi}\big). \tag{A.9}$$

The proof of Lemma A.4 is complete.

A.2 An Excess Risk Bound

This part establishes an excess risk bound for location estimation using pivot-blend, which is uniform in scale parameters, shedding light on the impact of asymmetrical scales (skewness) and the presence of an unknown pivotal point. This result is non-asymptotic and non-parametric, making it applicable to various scenarios.

Let $y_i \in \mathbb{R}$ and $X_i \in \mathbb{R}^p$ which satisfy satisfy $(X_i, y_i) \overset{i.i.d.}{\sim} F^*$, where F^* is a distribution that depends on $\beta^*, m^*, \sigma^*, \nu^*$, where we use superscript * to denote the statistical truth. As discussed in Section 2.3, we consider the practice of estimating the scales beforehand and then minimizing a criterion over all location parameters $\beta_j (1 \leq j \leq p), m$. For ease of presentation, given (σ, ν), let $l_{\sigma,\nu}(\cdot; X_i, y_i)$, imposed on $\theta := (\beta, m)$, denote a loss motivated by skewed pivot-blend:

$$\begin{aligned} l_{\sigma,\nu}(\theta; X_i, y_i) := {} & \rho\big(\frac{r_i - m}{\sigma} + m\big)1_{r_i - m \leq 0} + \rho\big(\frac{r_i - m}{\nu} + m\big)1_{r_i - m > 0} \\ & + \chi_0 \log\big[\sigma\Phi(m) + \nu\{1 - \Phi(m)\}\big] \quad \text{with} \quad r_i = y_i - X_i^T\beta, \end{aligned} \tag{A.10}$$

where the calibration parameter $\chi_0 > 0$ and $0 \leq \Phi(m) \leq 1$. Note that Φ is not necessarily directly associated with ρ. Define $\hat{\theta}_{\sigma,\nu}$ by empirical risk minimization

$$\hat{\theta}_{\sigma,\nu} = \arg\min_\theta \sum_{i=1}^n l_{\sigma,\nu}(\theta; X_i, y_i). \tag{A.11}$$

Certainly, opting for different values of σ and ν produces diverse asymmetric losses and influences the overall risk. In our analysis, *no* restrictions will be placed on σ and ν. The values of (σ, ν) can be specified based on domain knowledge or determined in a data-dependent manner. For notational simplicity, we sometimes abbreviate $\hat{\theta}_{\sigma,\nu}$ as $\hat{\theta}$ when there is no ambiguity.

To evaluate the generalization performance of $\hat{\theta}$, let (X_0, y_0) be a new observation that follows F^* but is independent of the training data $(X_i, y_i), (1 \le i \le n)$, and define the population risk of $\hat{\theta}$ by

$$R_{\sigma,\nu}(\hat{\theta}) := \mathbb{E}_{(X_0,y_0)}\big[l_{\sigma,\nu}(\hat{\theta}; X_0, y_0)\big], \tag{A.12}$$

where the expectation is taken with respect to the new observation (X_0, y_0) only. Due to the finite number of observations in estimation, $R_{\sigma,\nu}(\hat{\theta}_{\sigma,\nu})$ is always greater than the population risk of the ideal $\theta^*_{\sigma,\nu} = \arg\min_\theta R_{\sigma,\nu}(\theta)$, or $R_{\sigma,\nu}(\theta^*_{\sigma,\nu}) = \inf_{\theta\in\Omega} R_{\sigma,\nu}(\theta)$. In such a setup, the notion of **excess risk** $\mathcal{E}(\hat{\theta}_{\sigma,\nu}; \sigma, \nu)$ is helpful (Devroye et al., 1996):

$$\text{Excess Risk:} \quad \mathcal{E}(\hat{\theta}_{\sigma,\nu}; \sigma, \nu) \;:=\; R_{\sigma,\nu}(\hat{\theta}_{\sigma,\nu}) - R_{\sigma,\nu}(\theta^*_{\sigma,\nu}). \tag{A.13}$$

Our main objective is to establish a non-asymptotic bound for $\mathcal{E}(\hat{\theta}_{\sigma,\nu}; \sigma, \nu)$ regardless of the data distribution for a broad range of ρ that satisfy the following assumption.

ASSUMPTION $\mathcal{A}$: Assume that the loss ρ satisfies (i) ρ is bounded: $\rho \in [0, B]$ for some $B > 0$, and (ii) ρ is regular in the sense that ρ is piecewise polynomial on $K \ge 1$ intervals $\rho(t) = P_i(t)$, $\forall t \in [u_{i-1}, u_i)$, $i = 1, \ldots, K$, where $u_0 = -\infty, u_K = \infty$, and each polynomial function P_i has degree at most $D \ge 0$.

Assumption $\mathcal{A}$ encompasses a wide range of practically used loss functions in robust regression and classification, specifically designed to handle extreme outliers. For example, some loss functions like $\rho(t) = \int_0^{|t|} \psi(s)\,\mathrm{d}s$, with a redescending ψ, such as Tukey's bisquare $\psi(t) = t\{1 - (t/c)^2\}^2$ if $|t| \le c$, and 0 otherwise (Hampel et al., 2011), fit in the category. Some other ρ functions can be effectively approximated by piecewise polynomial functions.

Theorem A.1. *As long as the loss ρ satisfies Assumption $\mathcal{A}$, the estimator $\hat{\theta}_{\sigma,\nu}$ defined in (A.11) satisfies the following probabilistic bound for all $\sigma, \nu > 0$,*

$$\mathbb{P}\left\{\sup_{\sigma,\nu>0} \mathcal{E}(\hat{\theta}_{\sigma,\nu}; \sigma, \nu) - C\left[\frac{B\sqrt{p\log\{(K+1)(D+1)\}}}{\sqrt{n}} + \frac{(B\vee\chi_0)\{\log\frac{\sigma\vee\nu}{\sigma\wedge\nu} + |\log(\sigma\wedge\nu)|\}}{\sqrt{n}} \times \right.\right.$$
$$\left.\left.\left\{\sqrt{\log\Big(\log\frac{\sigma\vee\nu}{\sigma\wedge\nu} \vee |\log(\sigma\wedge\nu)| \vee 1\Big)} + \sqrt{\log\frac{1}{\epsilon}}\right\}\right] \le 0\right\} \ge 1 - \epsilon.$$

The theorem provides a bound for the excess risk $\mathcal{E}(\hat{\theta}_{\sigma,\nu}; \sigma, \nu)$ that holds uniformly in σ and ν with probability at least $1 - \epsilon$, as characterized by the following rate

$$\frac{B\sqrt{p\log\{(K+1)(D+1)\}}}{\sqrt{n}} + \frac{B\vee\chi_0\{\log\frac{\sigma\vee\nu}{\sigma\wedge\nu} + |\log(\sigma\wedge\nu)|\}}{\sqrt{n}} \times \left\{\sqrt{\log\frac{1}{\epsilon}} + \right.$$
$$\left.\sqrt{\log\Big(\log\frac{\sigma\vee\nu}{\sigma\wedge\nu} \vee |\log(\sigma\wedge\nu)| \vee 1\Big)}\right\}. \tag{A.14}$$

The first term in (A.14) illustrates the influence of loss complexity and problem dimensions on the excess risk. The second term, which incorporates both σ and ν, arises due to the pivot estimation. Clearly, when there is no skewness, $\sigma = \nu \Rightarrow \log \frac{\sigma \vee \nu}{\sigma \wedge \nu} = 0$ and the rate becomes

$$\frac{B\sqrt{p \log\{(K+1)(D+1)\}}}{\sqrt{n}} + \frac{(B+\chi_0)|\log(\sigma \wedge \nu)|}{\sqrt{n}} \times \left\{ \sqrt{\log(1+|\log(\sigma \wedge \nu)|)} + \sqrt{\log \frac{1}{\epsilon}} \right\}.$$

In the more practical scenario of unequal scales, the risk for location estimation can significantly increase, and the provided bound quantitatively characterizes how skewness inflates the risk non-asymptotically.

To prove the theorem, we first introduce a basic excess-risk bound for fixed $\sigma, \nu > 0$.

Lemma A.5. *Suppose that the loss ρ satisfies Assumption $\mathcal{A}$. Fixing the values of $\sigma, \nu > 0$ in (A.10), the corresponding estimator $\hat{\theta}_{\sigma,\nu}$ from (A.11) satisfies*

$$\mathcal{E}(\hat{\theta}_{\sigma,\nu}; \sigma, \nu) \lesssim B\sqrt{\frac{p \log\{(K+1)(D+1)\}}{n}} + \frac{B \vee \chi_0\big(|\log \sigma \wedge \nu| + \log \frac{\nu \vee \sigma}{\sigma \wedge \nu}\big)}{\sqrt{n}} \sqrt{\log \frac{1}{\epsilon}}, \tag{A.15}$$

with probability at least $1-\epsilon$.

Proof. Let $\{X, y\}$ denote the training data, i.e., $\{X, y\} := \big\{(X_i, y_i), 1 \le i \le n\big\}$ with $(X_i, y_i) \overset{i.i.d.}{\sim} F^*$. Given $\sigma, \nu > 0$, define a function class consisting of all $l_{\sigma,\nu}(\theta; \cdot)$ $\theta \in \Omega = \mathbb{R}^p \times \mathbb{R}$

$$\mathcal{L}_{\sigma,\nu}(\Omega) := \big\{ l_{\sigma,\nu}(\theta; \cdot) : \theta \in \Omega \big\}. \tag{A.16}$$

For simplicity, we often use the shorthand notations $R(\cdot)$, $R^{(n)}(\cdot)$, and $l(\theta; \cdot)$ to denote $R_{\sigma,\nu}(\cdot)$, $(1/n)\sum_{i=1}^n l_{\sigma,\nu}(\cdot; X_i, y_i)$, and $l_{\sigma,\nu}(\theta; \cdot)$, respectively, when there is no ambiguity.

First, the standard bound for excess risk through uniform laws yields

$$\begin{aligned} R(\hat{\theta}_{\sigma,\nu}) - R(\theta^*_{\sigma,\nu}) &\le R(\hat{\theta}_{\sigma,\nu}) - R^{(n)}(\hat{\theta}_{\sigma,\nu}) + R^{(n)}(\hat{\theta}_{\sigma,\nu}) - R^{(n)}(\theta^*_{\sigma,\nu}) + R^{(n)}(\theta^*_{\sigma,\nu}) - R(\theta^*_{\sigma,\nu}) \\ &\le 2 \sup_{\theta \in \Omega} \big| R(\theta) - R^{(n)}(\theta) \big| = 2 \sup_{l \in \mathcal{L}_{\sigma,\nu}(\Omega)} \big| \mathbb{P}_n l - \mathbb{P} l \big| = 2 \big\| \mathbb{P}_n - \mathbb{P} \big\|_{\mathcal{L}_{\sigma,\nu}(\Omega)}, \end{aligned} \tag{A.17}$$

where $\mathbb{P}$ is the distribution F^* and $\mathbb{P}_n$ is the empirical measure that places probability mass $1/n$ on each $(X_i, y_i), 1 \le i \le n$.

Let

$$g_{\sigma,\nu}(m) := \chi_0 \log\big[\sigma\Phi(m) + \nu\{1 - \Phi(m)\}\big]. \tag{A.18}$$

Without loss of generality, we assume that $\sigma \le \nu$, that is, $\sigma \wedge \nu = \sigma$, $\sigma \vee \nu = \nu$, then

$$g_{\sigma,\nu}(m) = \chi_0 \log \sigma + \chi_0 \log\big[1 + (\frac{\nu}{\sigma} - 1)\{1 - \Phi(m)\}\big]. \tag{A.19}$$

Because $\log(1 + (\frac{\nu}{\sigma} - 1)t)$ is an increasing function for $t \ge 0$, we know

$$\chi_0 \log \sigma \le g_{\sigma,\nu}(m) \le \chi_0 \log \sigma + \chi_0 \log \frac{\nu}{\sigma}, \tag{A.20}$$

from which it follows that

$$|g_{\sigma,\nu}(m)| \le \big|\chi_0 \log\sigma + \chi_0 \log\frac{\nu}{\sigma}\big| \vee |\chi_0 \log\sigma| \le |\chi_0 \log\sigma| + \chi_0 \log\frac{\nu}{\sigma}. \tag{A.21}$$

By Assumption $\mathcal{A}$ and (A.21),

$$|l_{\sigma,\nu}| \le |\rho| + |g_{\sigma,\nu}| \le B + \chi_0\big(|\log\sigma| + \log\frac{\nu}{\sigma}\big). \tag{A.22}$$

Due to (A.22) and $(X_i, y_i) \overset{i.i.d.}{\sim} F^*$, applying McDiarmid's inequality and symmetrization in empirical process theory yields a data-dependent bound with probability at least $1-\epsilon$,

$$\big\|\mathbb{P}_n - \mathbb{P}\big\|_{\mathcal{L}_{\sigma,\nu}(\Omega)} \le 2\mathcal{R}_{X,y}(\mathcal{L}_{\sigma,\nu}(\Omega)) + 6\Big\{B + \chi_0\big(|\log\sigma| + \log\frac{\nu}{\sigma}\big)\Big\}\sqrt{\frac{\log(3/\epsilon)}{2n}}, \tag{A.23}$$

where $\mathcal{R}_{X,y}(\mathcal{L}_{\sigma,\nu}(\Omega))$ is the empirical Rademacher complexity of $\mathcal{L}_{\sigma,\nu}(\Omega)$ with respect to the training data $\{X, y\}$:

$$\mathcal{R}_{X,y}(\mathcal{L}_{\sigma,\nu}(\Omega)) := \frac{1}{n}\mathbb{E}_\epsilon \sup_{l\in\mathcal{L}_{\sigma,\nu}(\Omega)} \sum_{i=1}^n \epsilon_i l(\theta; X_i, y_i), \tag{A.24}$$

and ϵ_i's are i.i.d. Rademacher random variables; see, e.g., Theorem 3.4.5 in Giné and Nickl (2015). Note that the expectation in (A.24) is taken with respect to ϵ only, and (A.24) depends on σ, ν through the function class $\mathcal{L}_{\sigma,\nu}(\Omega)$.

It remains to bound the empirical Rademacher complexity. Toward this, denote by A_σ, A_ν two augmented design matrices

$$A_\sigma = -\frac{1}{\sigma}[X, (1-\sigma)1_n], \quad A_\nu = -\frac{1}{\nu}[X, (1-\nu)1_n], \tag{A.25}$$

where 1_n is a column vector of n ones. Let

$$\alpha_\sigma = \frac{1}{\sigma}y, \quad \alpha_\nu = \frac{1}{\nu}y. \tag{A.26}$$

Suppose that $\mathrm{rank}(A_\sigma) \le r$ and $\mathrm{rank}(A_\nu) \le r$. By the singular value decomposition,

$$A_\sigma = U_\sigma D_\sigma V_\sigma^T, \quad A_\nu = U_\nu D_\nu V_\nu^T, \tag{A.27}$$

where U_σ, U_ν are orthogonal matrices with r columns: $U_\sigma^T U_\sigma = I_{r\times r}, U_\nu^T U_\nu = I_{r\times r}$. Define

$$\bar{U}_\sigma = [U_\sigma, \alpha_\sigma], \quad \bar{U}_\nu = [U_\nu, \alpha_\nu]. \tag{A.28}$$

Now, by the sub-additivity of sup and (A.28),

$$\begin{aligned}
&\mathcal{R}_{X,y}(\mathcal{L}_{\sigma,\nu}(\Omega)) \\
&\le \frac{1}{n}\mathbb{E}_\epsilon \sup_{\theta\in\Omega}\langle\epsilon, \rho(A_\sigma\theta + \alpha_\sigma)\rangle + \frac{1}{n}\mathbb{E}_\epsilon \sup_{\theta\in\Omega}\langle\epsilon, \rho(A_\nu\theta + \alpha_\nu)\rangle + \frac{1}{n}\mathbb{E}_\epsilon \sup_{m\in\mathbb{R}}\langle\epsilon, g_{\sigma,\nu}(1m)\rangle \\
&\le \frac{1}{n}\mathbb{E}_\epsilon \sup_{\xi\in\mathbb{R}^{r+1}} \langle\epsilon, \rho(\bar{U}_\sigma\xi)\rangle + \frac{1}{n}\mathbb{E}_\epsilon \sup_{\xi\in\mathbb{R}^{r+1}} \langle\epsilon, \rho(\bar{U}_\nu\xi)\rangle + \frac{1}{n}\mathbb{E}_\epsilon \sup_{m\in\mathbb{R}}\langle\epsilon, g_{\sigma,\nu}(1m)\rangle.
\end{aligned} \tag{A.29}$$

To bound the first term in (A.29), let

$$Z(X, y, \sigma) = \bar{U}_\sigma = [Z_1, \cdots, Z_n]^T, \tag{A.30}$$

and given ρ, define a class of functions

$$\mathcal{F} := \big\{\rho(\langle \xi, \cdot \rangle) : \xi \in \mathbb{R}^{r+1}\big\}, \tag{A.31}$$

which does not depend on the two scale parameters. Given $Z(X, y, \sigma)$, let $\mathbb{Q}_n$ be the empirical measure determined by Z_i, $1 \le i \le n$, and

$$\|f - \tilde{f}\|^2_{\mathbb{Q}_n} := \frac{1}{n}\sum_{i=1}^n \big\{f(Z_i) - \tilde{f}(Z_i)\big\}^2, \quad \forall f, \tilde{f} \in \mathcal{F}. \tag{A.32}$$

Then

$$\frac{1}{n}\mathbb{E}_\epsilon \sup_{\xi \in \mathbb{R}^{r+1}} \langle \epsilon, \rho(\bar{U}_\sigma \xi)\rangle = \frac{1}{\sqrt{n}}\mathbb{E}_\epsilon \sup_{f \in \mathcal{F}} \frac{1}{\sqrt{n}}\sum_{i=1}^n \epsilon_i f(Z_i) \le \frac{C}{\sqrt{n}}\int_0^B \sqrt{\log \mathcal{N}(\varepsilon, \mathcal{F}, \|\cdot\|_{\mathbb{Q}_n})}\, \mathrm{d}\varepsilon, \tag{A.33}$$

where the last inequality is due to Dudley's integral bound. Because $|f| \le B$, $\forall f \in \mathcal{F}$, by Theorem 2.6.7 in van der Vaart and Wellner (2013) we know

$$\mathcal{N}(\varepsilon, \mathcal{F}, \|\cdot\|_{\mathbb{Q}_n}) \le C\mathcal{V}(\mathcal{F})\big(\frac{cB}{\varepsilon}\big)^{2\mathcal{V}(\mathcal{F})}, \tag{A.34}$$

where $\mathcal{V}(\mathcal{F})$ denotes the VC-dimension of $\mathcal{F}$ defined through the notion of subgraph (cf. Definition 3.6.8 in Giné and Nickl (2015)). In more details, $\mathcal{V}(\mathcal{F})$ is the VC-dimension of the following $\{0, 1\}$-valued function class defined on $\mathbb{R}^{r+1} \times \mathbb{R}$:

$$\mathcal{H} := \big\{h_\xi(z, t) = 1_{\rho(\langle \xi, z\rangle) > t} = \mathrm{sign}(\rho(\langle \xi, z\rangle) - t) : \xi \in \mathbb{R}^{r+1}\big\}. \tag{A.35}$$

Here, the sign function is defined by $\mathrm{sign}(a) = 1$ if $a > 0$, and 0 otherwise, and recall that $1_{\rho(\langle \xi, z\rangle) > t}$ is the indicator function of the set $\big\{(z, t) : \rho(\langle \xi, z\rangle) > t, z \in \mathbb{R}^{r+1}, t \in \mathbb{R}\big\}$ or the subgraph of $\rho(\langle \xi, \cdot\rangle)$ for each given ξ.

To bound $\mathcal{V}(\mathcal{H})$, we introduce a more general function class $\tilde{\mathcal{H}}$

$$\tilde{\mathcal{H}} := \big\{\tilde{h}_{\xi, \omega_1, \omega_2}(z, t) = \mathrm{sign}\big(\omega_1 \rho(\langle \xi, z\rangle) + \omega_2 t\big) : \xi \in \mathbb{R}^{r+1}, \omega_1, \omega_2 \in \mathbb{R}\big\}, \tag{A.36}$$

where $\tilde{h}_{\xi, \omega_1, \omega_2}(z, t)$ is defined on $(z, t) \in \mathbb{R}^{r+1} \times \mathbb{R}$ and has two additional parameters ω_1 and ω_2 than $h_\xi(z, t)$ in the definition of $\mathcal{H}$. Since $\mathcal{H} \subset \tilde{\mathcal{H}}$ (by fixing $\omega_1 = 1, \omega_2 = -1$),

$$\mathcal{V}(\mathcal{F}) = \mathcal{V}(\mathcal{H}) \le \mathcal{V}(\tilde{\mathcal{H}}). \tag{A.37}$$

$\tilde{\mathcal{H}}$ corresponds to the set of functions computed by a neural network $\mathbf{N}$ shown in Figure A.1. It has two computation units, one in the hidden layer with activation function ρ, and the other in the output layer applying a sign operation on the combined inputs $\omega_1\rho(\langle \xi, z\rangle) + \omega_2 t$. Lemma A.6 is Theorem 10 in Bartlett et al. (2019) and can be proved based on Theorem 2.2 of Goldberg and Jerrum (1995).

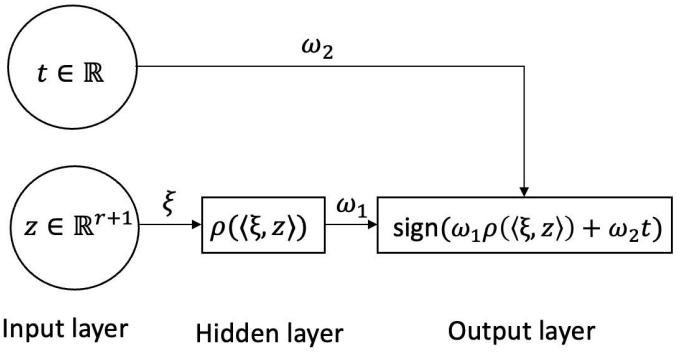


Figure A.1: Architecture of the network $\mathbf{N}$ that computes $\tilde{h}_{\xi,\omega_1,\omega_2}(z,t) \in \tilde{\mathcal{H}}$.

Lemma A.6. *Suppose that a neural network $\mathbf{N}_0$ satisfies (i) $\mathbf{N}_0$ has a directed acyclic graph, that is, the connections from input or computation units to computation units do not form any loops, (ii) the unique output unit is the only computation unit in the output layer (layer J), where $J \geq 2$ denotes the length of the longest path in the graph of $\mathbf{N}_0$, and the activation function of the output unit is a sign function that takes inputs from units in any layer $j < J$, including the input layer (layer 0), and (iii) within each computation unit except the output unit, Ψ is the activation function and is piecewise polynomial on $I \geq 1$ intervals $\Psi(t) = P_i(t)$, $\forall t \in [u_{i-1}, u_i), i = 1, \ldots, I$, where $u_0 = -\infty, u_I = \infty$, and each polynomial function P_i has degree at most $M \geq 0$. Let $W \geq 2$ be the number of parameters (weights and biases) and $U \geq 2$ be the number of computation units. Let $\mathcal{G}$ denote the set of $\{0,1\}$-valued functions computed by the network $\mathbf{N}_0$, then $\mathcal{V}(\mathcal{G}) \leq 2W \log_2 \left[16e\{M^{U-1} + \sum_{i=0}^{U-1} M^i\}(1+I)^U\right]$.*

Our network $\mathbf{N}$ as shown in Figure A.1 is a feed-forward neural network and has a directed acyclic graph structure. The output unit of $\mathbf{N}$ in the output layer takes the input $\rho(\langle \xi, z \rangle)$ from the computation unit in the hidden layer and the input unit $t \in \mathbb{R}$ in the input layer. Under Assumption $\mathcal{A}$, our network $\mathbf{N}$ satisfies the assumptions in Lemma A.6 and has $r+3$ parameters (no bias parameters) and two computation units. Using Lemma A.6, we get

$$\mathcal{V}(\tilde{\mathcal{H}}) \leq C(r+3) \log \left\{e(2D+1)(1+K)^2\right\} \lesssim r \log \left\{(K+1)(D+1)\right\}. \tag{A.38}$$

Now, based on (A.37) and (A.38),

$$\mathcal{V}(\mathcal{F}) \leq Cr \log \left\{(K+1)(D+1)\right\}, \tag{A.39}$$

and combining (A.33), (A.34), and (A.39) results in

$$\frac{1}{n}\mathbb{E}_\epsilon \sup_{\xi \in \mathbb{R}^{r+1}} \langle \epsilon, \rho(\bar{U}_\sigma \xi) \rangle \leq \frac{C}{\sqrt{n}} \int_0^B \sqrt{\mathcal{V}(\mathcal{F}) \log \frac{B}{\varepsilon}}\, \mathrm{d}\varepsilon \lesssim B\sqrt{\frac{r \log \left\{(K+1)(D+1)\right\}}{n}}. \tag{A.40}$$

Similarly, the second term in (A.29) satisfies

$$\frac{1}{n}\mathbb{E}_\epsilon \sup_{\xi \in \mathbb{R}^{r+1}} \langle \epsilon, \rho(\bar{U}_\nu \xi) \rangle \lesssim B\sqrt{\frac{r \log \left\{(K+1)(D+1)\right\}}{n}}. \tag{A.41}$$

Note that (A.40) and (A.41) hold for all $\sigma, \nu > 0$.

Finally, we bound the third term on the right-hand side of (A.29). Let

$$q_{\sigma,\nu}(m) := \chi_0 \log\big[1 + (\frac{\nu}{\sigma} - 1)\{1 - \Phi(m)\}\big], \tag{A.42}$$

and so $0 \le q_{\sigma,\nu}(m) \le \chi_0 \log \frac{\nu}{\sigma}$. Again, by the sub-additivity of sup and Dudley's integral bound,

$$\begin{aligned}
\frac{1}{n}\mathbb{E}_\epsilon \sup_{m\in\mathbb{R}} \langle \epsilon, g_{\sigma,\nu}(1m)\rangle &\le \frac{1}{n}\chi_0 \log\sigma \mathbb{E}_\epsilon \sup_{m\in\mathbb{R}} \langle \epsilon, 1\rangle + \frac{1}{n}\mathbb{E}_\epsilon \sup_{m\in\mathbb{R}} \langle \epsilon, q_{\sigma,\nu}(1m)\rangle \\
&\le \frac{C}{n}\int_0^{\frac{1}{2}\chi_0 \log\frac{\nu}{\sigma}\sqrt{n}} \sqrt{\log \mathcal{N}\big(\varepsilon, \big\{[q_{\sigma,\nu}(1m)] : m \in \mathbb{R}\big\}, \|\cdot\|_2\big)}\, \mathrm{d}\varepsilon \\
&\le \frac{C}{n}\int_0^{\frac{1}{2}\chi_0 \log\frac{\nu}{\sigma}\sqrt{n}} \sqrt{\log \mathcal{N}\Big(\frac{\varepsilon}{\sqrt{n}}, \big[0, \chi_0 \log\frac{\nu}{\sigma}\big], |\cdot|\Big)}\, \mathrm{d}\varepsilon \\
&\lesssim \frac{\chi_0 \log\frac{\nu}{\sigma}}{\sqrt{n}}.
\end{aligned} \tag{A.43}$$

Plugging (A.40), (A.41), and (A.43) into (A.29) yields

$$\mathcal{R}_{X,y}(\mathcal{L}_{\sigma,\nu}(\Omega)) \lesssim B\sqrt{\frac{r\log\big\{(K+1)(D+1)\big\}}{n}} + \frac{\chi_0 \log\frac{\nu}{\sigma}}{\sqrt{n}}. \tag{A.44}$$

Summarizing (A.17), (A.23), and (A.44), we have the following bound with probability at least $1-\epsilon$ for any $0 < \epsilon < 1$,

$$\begin{aligned}
&\mathcal{E}(\hat{\theta}_{\sigma,\nu}; \sigma, \nu) \\
&\le C\left[B\sqrt{\frac{r\log\big\{(K+1)(D+1)\big\}}{n}} + \frac{\chi_0 \log\frac{\nu}{\sigma}}{\sqrt{n}} + \frac{B + \chi_0\big(\log\frac{\nu}{\sigma} + |\log\sigma|\big)}{\sqrt{n}}\sqrt{\log\frac{1}{\epsilon}}\right] \\
&\lesssim B\sqrt{\frac{p\log\big\{(K+1)(D+1)\big\}}{n}} + \frac{B \vee \chi_0\big(\log\frac{\nu}{\sigma} + |\log\sigma|\big)}{\sqrt{n}}\sqrt{\log\frac{1}{\epsilon}},
\end{aligned} \tag{A.45}$$

where the last inequality is due to $r \le p + 1 \lesssim p$. The proof of Lemma A.5 is complete. □

Lemma A.5 shows the effect of skewness for fixed values of σ, ν. For example, when $\sigma = \sigma^*, \nu = \nu^*$, the excess risk $\mathcal{E}(\hat{\theta}_{\sigma^*,\nu^*}; \sigma^*, \nu^*)$ satisfies

$$\mathcal{E}(\hat{\theta}_{\sigma^*,\nu^*}; \sigma^*, \nu^*) \lesssim \frac{B\sqrt{p\log\big\{(K+1)(D+1)\big\}}}{\sqrt{n}} + \frac{B \vee \chi_0\big\{\log\frac{\sigma^* \vee \nu^*}{\sigma^* \wedge \nu^*} + |\log(\sigma^* \wedge \nu^*)|\big\}}{\sqrt{n}}\sqrt{\log\frac{1}{\epsilon}} \tag{A.46}$$

with probability at least $1-\epsilon$. Our objective is to establish a uniform law that applies to all scale parameters $\sigma, \nu > 0$. Toward this, let

$$\tau = |\log(\sigma \wedge \nu)| + \log\frac{\sigma \vee \nu}{\sigma \wedge \nu}, \tag{A.47}$$

which is nonnegative. By Lemma A.5, there exists a universal constant C such that the event $A(\tau,t)$

$$A(\tau,t):=\left\{\mathcal{E}(\hat{\theta}_{\sigma,\nu};\sigma,\nu)-C\left[B\sqrt{\frac{p\log\{(K+1)(D+1)\}}{n}}+(B+\chi_0\tau)t\right]\geq 0\right\} \tag{A.48}$$

occurs with probability

$$\mathbb{P}\big[A(\tau,t)\big]\leq\exp(-2nt^2), \tag{A.49}$$

for any $t>0$.

Let $[0,+\infty)=\bigcup_{l=0}^{\infty}[\tau_l,\tau_{l+1}]$ with $0=\tau_0<\tau_1<\ldots<+\infty$ to be determined. Let $Q(\tau)\geq 0$ $(\forall\tau\geq 0)$ be an increasing function. Then, for all $l\geq 0$ and $l\in\mathbb{Z}$, $\tau\in[\tau_l,\tau_{l+1}]$ implies $\tau\geq\tau_l$, $Q(\tau)\geq Q(\tau_l)$, and $\tau Q(\tau)\geq\tau Q(\tau_l)\geq\tau_l Q(\tau_l)$. Based on (A.49) and the union bound,

$$\begin{aligned}&\mathbb{P}\left(\sup_{\sigma,\nu>0}\mathcal{E}(\hat{\theta}_{\sigma,\nu};\sigma,\nu)-C\left[B\sqrt{\frac{p\log\{(K+1)(D+1)\}}{n}}+(B+\chi_0\tau)\{t+Q(\tau)\}\right]\geq 0\right)\\ \leq&\sum_{l=0}^{\infty}\mathbb{P}\Big[A\big(\tau_l,t+Q(\tau_l)\big)\Big].\end{aligned} \tag{A.50}$$

Take $Q(\tau)=\sqrt{\log(1+\tau)/n}$ and $\tau_l=l$ for all $l\geq 0$ and $l\in\mathbb{Z}$, then (A.50) is bounded by

$$\sum_{l=0}^{\infty}\mathbb{P}\left[A\left(l,t+\sqrt{\frac{\log(1+l)}{n}}\right)\right]\leq\sum_{l=0}^{\infty}\frac{1}{(l+1)^2}\exp(-2nt^2)\leq C\exp(-2nt^2). \tag{A.51}$$

Plugging (A.47) into (A.51) gives the following uniform law for all $\sigma,\nu>0$ with probability $1-\epsilon$

$$\begin{aligned}\mathcal{E}(\hat{\theta}_{\sigma,\nu};\sigma,\nu)\leq&C\left[\frac{B\sqrt{p\log\{(K+1)(D+1)\}}}{\sqrt{n}}+\frac{B\vee\chi_0\{\log\frac{\sigma\vee\nu}{\sigma\wedge\nu}+|\log(\sigma\wedge\nu)|\}}{\sqrt{n}}\right.\\ &\left.\times\left\{\sqrt{\log\frac{1}{\epsilon}}+\sqrt{\log\big(\log\frac{\sigma\vee\nu}{\sigma\wedge\nu}\vee|\log(\sigma\wedge\nu)|\vee 1\big)+\log 2}\right\}\right]\\ \lesssim&\frac{B\sqrt{p\log\{(K+1)(D+1)\}}}{\sqrt{n}}+\frac{B\vee\chi_0\{\log\frac{\sigma\vee\nu}{\sigma\wedge\nu}+|\log(\sigma\wedge\nu)|\}}{\sqrt{n}}\\ &\times\left\{\sqrt{\log\frac{1}{\epsilon}}+\sqrt{\log\big(\log\frac{\sigma\vee\nu}{\sigma\wedge\nu}\vee|\log(\sigma\wedge\nu)|\vee 1\big)}\right\},\quad\forall\sigma,\nu>0,\end{aligned} \tag{A.52}$$

The proof of Theorem A.1 is complete.

A.3 Proof of Theorem 1

By the optimality of $\hat{\mu}$: $l(\hat{\mu}) + \frac{\tau}{2}\|\hat{\gamma}\|_2^2 + \lambda\varrho\|\hat{\bar{\beta}}\|_{2,1} \le l(\mu^*) + \frac{\tau}{2}\|\gamma^*\|_2^2 + \lambda\varrho\|\bar{\beta}^*\|_{2,1}$, we obtain a basic inequality as follows

$$\begin{aligned}
&\boldsymbol{\Delta}_l(\hat{\mu}, \mu^*) + \frac{\tau}{2}\|\hat{\gamma} - \gamma^*\|_2^2 \\
\le& \left\langle \epsilon_{\bar{\eta}}, \bar{X}\hat{\bar{\beta}} - \bar{X}\bar{\beta}^* \right\rangle + \langle \epsilon_{\bar{m}}, \hat{\bar{m}} - \bar{m}\rangle + \langle \epsilon_{\varsigma}, \hat{\varsigma} - \varsigma\rangle + \langle -\tau\gamma^*, \hat{\gamma} - \gamma^*\rangle \\
&+ \lambda\varrho\|\bar{\beta}^*\|_{2,1} - \lambda\varrho\|\hat{\bar{\beta}}\|_{2,1} \\
\le& \left\langle \epsilon_{\bar{\eta}}, \bar{X}\hat{\bar{\beta}} - \bar{X}\bar{\beta}^* \right\rangle + \langle \epsilon_{\bar{m}}, \hat{\bar{m}} - \bar{m}\rangle + \langle \epsilon_{\varsigma}, \hat{\varsigma} - \varsigma\rangle + \frac{b}{2}\|\hat{\gamma} - \gamma^*\|_2^2 + \frac{\tau^2}{2b}\|\gamma^*\|_2^2 \\
&+ \lambda\varrho\|\bar{\beta}^*\|_{2,1} - \lambda\varrho\|\hat{\bar{\beta}}\|_{2,1}
\end{aligned} \tag{A.53}$$

for any $b > 0$.

First, we bound $\left\langle \epsilon_{\bar{\eta}}, \bar{X}\hat{\bar{\beta}} - \bar{X}\bar{\beta}^* \right\rangle$. Let $\bar{X}_k$ denote the kth column of $\bar{X}$ and $\Xi_k := [\bar{X}_k, \bar{X}_{p+k}]^T$ and recall $\bar{\beta}_k = [\beta_k, b_k]^T$, $k = 1, \ldots, p$. By Hölder's inequality,

$$\langle \epsilon_{\bar{\eta}}, \bar{X}\hat{\bar{\beta}} - \bar{X}\bar{\beta}^* \rangle = \sum_{k=1}^{p} \langle \Xi_k\epsilon_{\bar{\eta}}, \hat{\bar{\beta}}_k - \bar{\beta}_k^* \rangle \le \sum_{k=1}^{p} \|\Xi_k\epsilon_{\bar{\eta}}\|_2 \|\hat{\bar{\beta}}_k - \bar{\beta}_k^*\|_2 \le \sqrt{2}\left\|\bar{X}^T\epsilon_{\bar{\eta}}\right\|_\infty \|\hat{\bar{\beta}} - \bar{\beta}^*\|_{2,1}. \tag{A.54}$$

Let $\lambda = A\left\|\bar{X}^T\epsilon_{\bar{\eta}}\right\|_\infty/\varrho$ for $A > 1$. Then we have

$$\begin{aligned}
&\left\|\bar{X}^T\epsilon_{\bar{\eta}}\right\|_\infty \left\|\hat{\bar{\beta}} - \bar{\beta}^*\right\|_{2,1} + \lambda\varrho\left\|\bar{\beta}^*\right\|_{2,1} - \lambda\varrho\left\|\hat{\bar{\beta}}\right\|_{2,1} \\
\le& \left(\lambda\varrho + \left\|\bar{X}^T\epsilon_{\bar{\eta}}\right\|_\infty\right)\left\|\bar{\beta}^*\right\|_{2,1} + \left(\left\|\bar{X}^T\epsilon_{\bar{\eta}}\right\|_\infty - \lambda\varrho\right)\left\|\hat{\bar{\beta}}\right\|_{2,1} \\
=& (1 + A)\left\|\bar{X}^T\epsilon_{\bar{\eta}}\right\|_\infty \left\|\bar{\beta}^*\right\|_{2,1} + (1 - A)\left\|\bar{X}^T\epsilon_{\bar{\eta}}\right\|_\infty \left\|\hat{\bar{\beta}}\right\|_{2,1} \\
\le& (1 + A)\left\|\bar{X}^T\epsilon_{\bar{\eta}}\right\|_\infty \left\|\bar{\beta}^*\right\|_{2,1}.
\end{aligned} \tag{A.55}$$

By the assumption on $\epsilon_{\bar{\eta}}$, the ψ-norm of $|\bar{X}_k^T\epsilon_{\bar{\eta}}|$ is bounded by

$$\left\|\bar{X}_k^T\epsilon_{\bar{\eta}}\right\|_\psi = \|\langle \epsilon_{\bar{\eta}}, \bar{X}_k\rangle\|_\psi \le \|\epsilon_{\bar{\eta}}\|_\psi \|\bar{X}_k\|_2 \le \omega_{\bar{\eta}}\varrho. \tag{A.56}$$

The following inequality is essentially Massart's finite class lemma, adapted for our purpose

$$\left\| \max_{1\le k\le 2p} |\bar{X}_k^T\epsilon_{\bar{\eta}}| \right\|_\psi \le 2(1 \vee c_1 \vee 2\psi(c_2))^2 c_0 \psi^{-1}(p)\omega_{\bar{\eta}}\varrho. \tag{A.57}$$

It can be obtained by modifying the proof of Lemma 2.2.2 in van der Vaart and Wellner (2013) (details omitted). By Lemma A.4,

$$\mathbb{E}\left[\|\bar{X}^T\epsilon_{\bar{\eta}}\|_\infty\right] \le \left\|\|\bar{X}^T\epsilon_{\bar{\eta}}\|_\infty\right\|_{L_2} \le \psi^{-1}(1)\left\|\left\|\bar{X}^T\epsilon_{\bar{\eta}}\right\|_\infty\right\|_\psi \le Cc_\psi\psi^{-1}(p)\omega_{\bar{\eta}}\varrho, \tag{A.58}$$

where $c_\psi = (1 \vee c_1 \vee 2\psi(c_2))^2 c_0 \psi^{-1}(1)$. Therefore,

$$\begin{aligned}\mathbb{E}\Big\{\big\langle \epsilon_{\bar\eta}, \bar X\hat{\bar\beta} - \bar X\bar\beta^*\big\rangle + \lambda\varrho\big\|\bar\beta^*\big\|_{2,1} - \lambda\varrho\big\|\hat{\bar\beta}\big\|_{2,1}\Big\} &\le (1+A)\mathbb{E}\Big[\big\|\bar X^T\epsilon_{\bar\eta}\big\|_\infty\Big]\big\|\bar\beta^*\big\|_{2,1}\\ &\le C(1+A)c_\psi\psi^{-1}(p)\omega_{\bar\eta}\varrho\big\|\bar\beta^*\big\|_{2,1}. \end{aligned} \tag{A.59}$$

Next, we bound the stochastic term $\langle \epsilon_{\bar m}, \hat{\bar m} - \bar m^*\rangle$. For any $a > 0$,

$$\langle \epsilon_{\bar m}, \hat{\bar m} - \bar m^*\rangle = \langle P_{1_n}\epsilon_{\bar m}, \hat{\bar m} - \bar m^*\rangle \le |u^T\epsilon_{\bar m}| \cdot \|\hat{\bar m} - \bar m^*\|_2 \le a(u^T\epsilon_{\bar m})^2 + \frac{1}{2a}\mathbf{D}_2(\hat{\bar m}, \bar m^*), \tag{A.60}$$

where P_{1_n} is the orthogonal projection matrix onto the column space of 1_n, or equivalently, uu^T with $u = (1/\sqrt{n})1_n$. By Lemma A.4,

$$\mathbb{E}\big[(u^T\epsilon_{\bar m})^2\big] \le \{\psi^{-1}(1)\}^2\|u^T\epsilon_{\bar m}\|_\varphi^2 \le \{\psi^{-1}(1)\}^2\omega_{\bar m}^2. \tag{A.61}$$

Likewise, for the stochastic term $\langle \epsilon_\varsigma, \hat\varsigma - \varsigma^*\rangle$ with $\epsilon_\varsigma = [\epsilon_{\frac{1}{\sigma}1_n}^T, \epsilon_{\frac{1}{\nu}1_n}^T]^T$ and $\hat\varsigma - \varsigma^* = [(1/\hat\sigma - 1/\sigma^*)1_n^T, (1/\hat\nu - 1/\nu^*)1_n^T]^T$, we have

$$\begin{aligned}&\big\langle \epsilon_{\frac{1}{\sigma}1_n}, (\frac{1}{\hat\sigma} - \frac{1}{\sigma^*})1_n\big\rangle + \big\langle \epsilon_{\frac{1}{\nu}1_n}, (\frac{1}{\hat\nu} - \frac{1}{\nu^*})1_n\big\rangle\\ =&\big\langle P_{1_n}\epsilon_{\frac{1}{\sigma}1_n}, (\frac{1}{\hat\sigma} - \frac{1}{\sigma^*})1_n\big\rangle + \big\langle P_{1_n}\epsilon_{\frac{1}{\nu}1_n}, (\frac{1}{\hat\nu} - \frac{1}{\nu^*})1_n\big\rangle\\ \le& a(u^T\epsilon_{\frac{1}{\sigma}1_n})^2 + a(u^T\epsilon_{\frac{1}{\nu}1_n})^2 + \frac{1}{2a}\mathbf{D}_2(\hat\varsigma, \varsigma^*),\end{aligned}$$

and

$$\mathbb{E}\big[(u^T\epsilon_{\frac{1}{\sigma}1_n})^2\big] \le \{\varphi^{-1}(1)\}^2\|u^T\epsilon_{\frac{1}{\sigma}1_n}\|_\varphi^2 \le \{\varphi^{-1}(1)\}^2\omega_\varsigma^2,$$

and $\mathbb{E}[(u^T\epsilon_{\frac{1}{\nu}1_n})^2] \le \{\varphi^{-1}(1)\}^2\omega_\varsigma^2$. To sum up, for any $a > 0$,

$$\begin{aligned}&\mathbb{E}\big\{\langle \epsilon_{\bar m}, \hat{\bar m} - \bar m^*\rangle + \langle \epsilon_\varsigma, \hat\varsigma - \varsigma^*\rangle\big\}\\ \le& a\{\psi^{-1}(1)\}^2\omega_{\bar m}^2 + 2a\{\varphi^{-1}(1)\}^2\omega_\varsigma^2 + \frac{1}{2a}\{\mathbf{D}_2(\hat{\bar m}, \bar m^*) + \mathbf{D}_2(\hat\varsigma, \varsigma^*)\}.\end{aligned} \tag{A.62}$$

Plugging (A.59) and (A.62) into (A.53) to obtain

$$\begin{aligned}&\mathbb{E}\Big\{\boldsymbol{\Delta}_l(\hat\mu, \mu^*) + (\tau - b)\mathbf{D}_2(\hat\gamma, \gamma^*)\Big\} - \frac{1}{2a}\{\mathbf{D}_2(\hat{\bar m}, \bar m^*) + \mathbf{D}_2(\hat\varsigma, \varsigma^*)\}\\ \le& C(1+A)c_\psi\psi^{-1}(p)\omega_{\bar\eta}\varrho\big\|\bar\beta^*\big\|_{2,1} + a\{\psi^{-1}(1)\}^2\omega_{\bar m}^2 + 2a\{\varphi^{-1}(1)\}^2\omega_\varsigma^2 + \frac{\tau^2}{2b}\|\gamma^*\|_2^2.\end{aligned} \tag{A.63}$$

Choosing $b = \tau/4$, $a = 2/\tau$, we get

$$\begin{aligned}&\mathbb{E}\Big\{\boldsymbol{\Delta}_l(\hat\mu, \mu^*) + \frac{\tau}{2}\mathbf{D}_2(\hat\gamma, \gamma^*)\Big\}\\ \le& C(1+A)c_\psi\psi^{-1}(p)\omega_{\bar\eta}\varrho\big\|\bar\beta^*\big\|_{2,1} + \frac{2}{\tau}\big[\{\psi^{-1}(1)\}^2\omega_{\bar m}^2 + 2\{\varphi^{-1}\}^2\omega_\varsigma^2\big] + 2\tau\|\gamma^*\|_2^2\\ \lesssim& c_\psi\psi^{-1}(p)\omega_{\bar\eta}\varrho\big\|\bar\beta^*\big\|_{2,1} + \frac{1}{\tau}(\omega_{\bar m}^2 + \omega_\varsigma^2) + \tau\|\gamma^*\|_2^2.\end{aligned} \tag{A.64}$$

The proof is complete.

A.4 Proof of Theorem 2

We prove a more general result, which includes Theorem 2 as a special case.

Theorem A.2. *Assume that the effective noises $\epsilon_{\bar\eta}, \epsilon_{\bar m}$, and ϵ_ς satisfy $\|\epsilon_{\bar\eta}\|_\psi \le \omega_{\bar\eta}$, $\|\epsilon_{\bar m}\|_\psi \le \omega_{\bar m}$, and $\|\epsilon_\varsigma\|_\varphi \le \omega_\varsigma$. Let $\hat\zeta$ denote the optimal solution for (19) with $\varrho \ge \kappa_{2,\infty}$. Suppose that there exist some $\vartheta > 0$ and some large $K > 0$ such that the following condition holds for any $\bar\beta, \gamma$ (recall $\zeta^T = [\bar\beta^T, \gamma^T]$, $\mu^T = [\bar\eta^T, \gamma^T]$ defined based on (17))*

$$(1+\vartheta)\lambda\varrho\big\|(\bar\beta - \bar\beta^*)_{\mathcal{J}^*}\big\|_{2,1} \le \mathbf{\Delta}_l(\mu, \mu^*) + \lambda\varrho\big\|(\bar\beta - \bar\beta^*)_{\mathcal{J}^{*c}}\big\|_{2,1} + K\lambda^2 J^*, \tag{A.65}$$

where

$$\lambda = \big(1 \vee \frac{1}{\vartheta}\big)\omega_{\bar\eta}\psi^{-1}\big[p\psi\big\{A\psi^{-1}(p)\big\}\big] \tag{A.66}$$

for some large enough $A > 0$. Then for any $L, L' > 0$,

$$\begin{aligned}\mathbf{\Delta}_l(\hat\mu, \mu^*) + \frac{\tau}{2}\mathbf{D}_2(\gamma, \gamma^*) \lesssim\ & \big(\frac{1}{\vartheta} \vee \frac{1}{\vartheta^3}\big)K\omega_{\bar\eta}^2\Big(\psi^{-1}\Big[p\psi\big\{A\psi^{-1}(p)\big\}\Big]\Big)^2 J^* \\ & + \big(1 \vee \frac{1}{\vartheta}\big)\frac{1}{\tau}\big(L^2\omega_{\bar m}^2 + L'^2\omega_\varsigma^2\big) + \big(1 \vee \frac{1}{\vartheta}\big)\tau\|\gamma^*\|_2^2\end{aligned} \tag{A.67}$$

holds with probability at least $1 - C/\psi\{A\psi^{-1}(p)\} - 1/\psi(cL) - C/\varphi(cL')$, where C, c are positive constants.

Theorem A.2 implies Theorem 2. In fact, (A.65), assuming ϑ is a constant, is just (27), and letting $\psi = \psi_q$ for some $q > 0$, (A.66) yields $\lambda = A\omega_{\bar\eta}(\log p)^{\frac{1}{q}}$. Then, choosing $L = C\psi^{-1}(p^{A^q}), L' = C\varphi^{-1}(p^{A^q})$, we can obtain the result in Theorem 2:

$$\begin{aligned}&\mathbf{\Delta}_l(\hat\mu, \mu^*) + \frac{\tau}{2}\mathbf{D}_2(\hat\gamma, \gamma^*) \\ \lesssim\ & KA^2\omega_{\bar\eta}^2 J^*(\log p)^{\frac{2}{q}} + \frac{1}{\tau}\{\psi_q^{-1}(p^{A^q})\}^2\omega_{\bar m}^2 + \frac{1}{\tau}\{\varphi^{-1}(p^{A^q})\}^2\omega_\varsigma^2 + \tau\|\gamma^*\|_2^2\end{aligned}$$

with probability at least $1 - Cp^{-cA^q}$.

Proof. From $l(\hat\mu) + \tau\|\hat\gamma\|_2^2 + \lambda\varrho\|\hat{\bar\beta}\|_{2,1} \le l(\mu^*) + \tau\|\gamma^*\|_2^2 + \lambda\varrho\|\bar\beta^*\|_{2,1}$, we obtain

$$\begin{aligned}&\mathbf{\Delta}_l(\hat\mu, \mu^*) + \frac{\tau}{2}\|\hat\gamma - \gamma^*\|_2^2 \\ \le\ & \big\langle\epsilon_{\bar\eta}, \bar X\hat{\bar\beta} - \bar X\bar\beta^*\big\rangle + \langle\epsilon_{\bar m}, \hat{\bar m} - \bar m\rangle + \langle\epsilon_\varsigma, \hat\varsigma - \varsigma\rangle + \langle -\tau\gamma^*, \hat\gamma - \gamma^*\rangle \\ & + \lambda\varrho\|\bar\beta^*\|_{2,1} - \lambda\varrho\|\hat{\bar\beta}\|_{2,1} \\ \le\ & \big\langle\epsilon_{\bar\eta}, \bar X\hat{\bar\beta} - \bar X\bar\beta^*\big\rangle + \langle\epsilon_{\bar m}, \hat{\bar m} - \bar m\rangle + \langle\epsilon_\varsigma, \hat\varsigma - \varsigma\rangle + \frac{b}{2}\|\hat\gamma - \gamma^*\|_2^2 + \frac{\tau^2}{2b}\|\gamma^*\|_2^2 \\ & + \lambda\varrho\|\bar\beta^*\|_{2,1} - \lambda\varrho\|\hat{\bar\beta}\|_{2,1}\end{aligned} \tag{A.68}$$

for any $b > 0$.

To bound $\langle \epsilon_{\bar{\eta}}, \bar{X}\hat{\bar{\beta}} - \bar{X}\bar{\beta}^* \rangle$, we use the same notations in Appendix A.3 and recall

$$\langle \epsilon_{\bar{\eta}}, \bar{X}\hat{\bar{\beta}} - \bar{X}\bar{\beta}^* \rangle \leq c_0 \max_{1\leq k\leq 2p} |\bar{X}_k^T \epsilon_{\bar{\eta}}| \|\hat{\bar{\beta}} - \bar{\beta}^*\|_{2,1}, \tag{A.69}$$

where the constant $c_0 \geq \sqrt{2}$, and

$$\big\|\bar{X}_k^T \epsilon_{\bar{\eta}}\big\|_{\psi} \leq \omega_{\bar{\eta}} \varrho, \ \forall k. \tag{A.70}$$

Let

$$\lambda_0 = \omega_{\bar{\eta}} \psi^{-1}\Big[p\psi\big\{A\psi^{-1}(p)\big\}\Big] \tag{A.71}$$

for some large enough $A > 0$. By the union bound and Markov's inequality, the event $\max_{1\leq k\leq 2p} |\bar{X}_k^T \epsilon_{\bar{\eta}}| \geq \lambda_0 \varrho$ occurs with probability

$$\mathbb{P}(\max_{1\leq k\leq 2p} |\bar{X}_k^T \epsilon_{\bar{\eta}}| \geq \lambda_0 \varrho) \leq \frac{2p}{\psi\Big(\frac{\omega_{\bar{\eta}} \varrho \psi^{-1}\left[p\psi\{A\psi^{-1}(p)\}\right]}{\omega_{\bar{\eta}}\varrho}\Big)} \leq \frac{2p}{p\psi\{A\psi^{-1}(p)\}} \leq \frac{2}{\psi\{A\psi^{-1}(p)\}}. \tag{A.72}$$

Let $\lambda = \lambda_0/b'$ with $b' > 0$. By the subadditivity of the ℓ_1-penalty, we get

$$\begin{aligned}
&c_0\lambda_0\varrho\big\|\hat{\bar{\beta}} - \bar{\beta}^*\big\|_{2,1} + \lambda\varrho\big\|\bar{\beta}^*\big\|_{2,1} - \lambda\varrho\big\|\hat{\bar{\beta}}\big\|_{2,1} \\
\leq\ & c_0\lambda_0\varrho\big\|(\hat{\bar{\beta}} - \bar{\beta}^*)_{\mathcal{J}^*}\big\|_{2,1} + c_0\lambda_0\varrho\big\|(\hat{\bar{\beta}} - \bar{\beta}^*)_{\mathcal{J}^{*C}}\big\|_{2,1} + \lambda\varrho\big\|\bar{\beta}^*_{\mathcal{J}^*}\big\|_{2,1} \\
&- \lambda\varrho\big\|\hat{\bar{\beta}}_{\mathcal{J}^*}\big\|_{2,1} - \lambda\varrho\big\|(\hat{\bar{\beta}} - \bar{\beta}^*)_{\mathcal{J}^{*C}}\big\|_{2,1} \\
\leq\ & \big(c_0\lambda_0 + \lambda\big)\varrho\big\|(\hat{\bar{\beta}} - \bar{\beta}^*)_{\mathcal{J}^*}\big\|_{2,1} - \big(\lambda - c_0\lambda_0\big)\varrho\big\|(\hat{\bar{\beta}} - \bar{\beta}^*)_{\mathcal{J}^{*C}}\big\|_{2,1} \\
=\ & \big(1 + c_0b'\big)\lambda\varrho\big\|(\hat{\bar{\beta}} - \bar{\beta}^*)_{\mathcal{J}^*}\big\|_{2,1} - \big(1 - c_0b'\big)\lambda\varrho\big\|(\hat{\bar{\beta}} - \bar{\beta}^*)_{\mathcal{J}^{*C}}\big\|_{2,1},
\end{aligned} \tag{A.73}$$

where $\mathcal{J}^{*C}$ is the complement of $\mathcal{J}^*$.

For the stochastic term $\langle \epsilon_{\bar{m}}, \hat{\bar{m}} - \bar{m}^* \rangle$, for any $a > 0$,

$$\langle \epsilon_{\bar{m}}, \hat{\bar{m}} - \bar{m}^* \rangle = \langle P_{1_n}\epsilon_{\bar{m}}, \hat{\bar{m}} - \bar{m}^* \rangle \leq |u^T\epsilon_{\bar{m}}| \cdot \|\hat{\bar{m}} - \bar{m}^*\|_2 \leq a(u^T\epsilon_{\bar{m}})^2 + \frac{1}{2a}\mathbf{D}_2(\hat{\bar{m}}, \bar{m}^*), \tag{A.74}$$

where $P_{1_n} = uu^T$ with $u = (1/\sqrt{n})1_n$. By the assumption on $\epsilon_{\bar{m}}$, $\|u^T\epsilon_{\bar{m}}\|_{\psi} \leq \omega_{\bar{m}}$. By Markov's inequality,

$$\mathbb{P}\Big((u^T\epsilon_{\bar{m}})^2 \geq cL^2\omega_{\bar{m}}^2\Big) \leq \frac{1}{\psi(\frac{cL\omega_{\bar{m}}}{\omega_{\bar{m}}})} = \frac{1}{\psi(cL)}, \tag{A.75}$$

from which it follow that

$$\langle \epsilon_{\bar{m}}, \hat{\bar{m}} - \bar{m}^* \rangle \leq acL^2\omega_{\bar{m}}^2 + \frac{1}{2a}\mathbf{D}_2(\hat{\bar{m}}, \bar{m}^*) \tag{A.76}$$

occurs with probability at least $1-1/\psi(cL)$. For the stochastic term $\langle \epsilon_\varsigma, \hat{\varsigma}-\varsigma\rangle$ with $\epsilon_\varsigma = [\epsilon^T_{\frac{1}{\hat{\sigma}}1_n}, \epsilon^T_{\frac{1}{\hat{\nu}}1_n}]^T$ and $\hat{\varsigma}-\varsigma^* = [(1/\hat{\sigma}-1/\sigma^*)1_n^T, (1/\hat{\nu}-1/\nu^*)1_n^T]^T$, similarly,

$$\begin{aligned}
&\langle \epsilon_{\frac{1}{\hat{\sigma}}1_n}, (\frac{1}{\hat{\sigma}}-\frac{1}{\sigma^*})1_n\rangle + \langle \epsilon_{\frac{1}{\hat{\nu}}1_n}, (\frac{1}{\hat{\nu}}-\frac{1}{\nu^*})1_n\rangle \\
=&\langle P_{1_n}\epsilon_{\frac{1}{\hat{\sigma}}1_n}, (\frac{1}{\hat{\sigma}}-\frac{1}{\sigma^*})1_n\rangle + \langle P_{1_n}\epsilon_{\frac{1}{\hat{\nu}}1_n}, (\frac{1}{\hat{\nu}}-\frac{1}{\nu^*})1_n\rangle \\
\leq& a(u^T\epsilon_{\frac{1}{\hat{\sigma}}1_n})^2 + a(u^T\epsilon_{\frac{1}{\hat{\nu}}1_n})^2 + \frac{1}{2a}\mathbf{D}_2(\hat{\varsigma},\varsigma^*).
\end{aligned}$$

By the assumption on ϵ_ς, $\|u^T\epsilon_{\frac{1}{\hat{\sigma}}1_n}\|_\varphi \leq \omega_\varsigma$, $\|u^T\epsilon_{\frac{1}{\hat{\nu}}1_n}\|_\varphi \leq \omega_\varsigma$. It follows from

$$\mathbb{P}\Big((u^T\epsilon_{\frac{1}{\hat{\sigma}}1_n})^2 \geq cL'^2\omega_\varsigma^2\Big) \leq \frac{1}{\psi(\frac{cL\omega_\varsigma}{\omega_\varsigma})} = \frac{1}{\psi(cL')} \tag{A.77}$$

and $\mathbb{P}\Big((u^T\epsilon_{\frac{1}{\hat{\nu}}1_n})^2 \geq cL'^2\omega_\varsigma^2\Big) \leq 1/\psi(cL')$ that

$$\langle \epsilon_\varsigma, \hat{\varsigma}-\varsigma\rangle \leq 2acL'^2\omega_\varsigma^2 + \frac{1}{2a}\mathbf{D}_2(\hat{\varsigma},\varsigma^*), \tag{A.78}$$

with probability at least $1-2/\varphi(cL')$.

Plugging (A.72), (A.73), (A.76), and (A.78) into (A.68) results in

$$\begin{aligned}
&\boldsymbol{\Delta}_l(\hat{\mu},\mu^*) + (\tau-b)\mathbf{D}_2(\hat{\gamma},\gamma^*) - \frac{1}{2a}\mathbf{D}_2(\hat{\bar{m}},\bar{m}^*) - \frac{1}{2a}\mathbf{D}_2(\hat{\varsigma},\varsigma^*) \\
\leq& \big(1+c_0b'\big)\lambda\varrho\big\|(\hat{\bar{\beta}}-\bar{\beta}^*)_{\mathcal{J}^*}\big\|_{2,1} - \big(1-c_0b'\big)\lambda\varrho\big\|(\hat{\bar{\beta}}-\bar{\beta}^*)_{\mathcal{J}^{*C}}\big\|_{2,1} + \frac{\tau^2}{2b}\|\gamma^*\|_2^2 \\
&+ acL^2\omega_{\bar{m}}^2 + 2acL'^2\omega_\varsigma^2.
\end{aligned} \tag{A.79}$$

with probability at least $1-C/\psi\{A\psi^{-1}(p)\} - 1/\psi(cL) - C/\varphi(cL')$, where C, c are positive constants.

The regularity condition (A.65) implies

$$\begin{aligned}
&(1+\frac{\vartheta}{2+\vartheta})\lambda\varrho\big\|(\hat{\bar{\beta}}-\bar{\beta}^*)_{\mathcal{J}^*}\big\|_{2,1} \\
\leq& \frac{2}{2+\vartheta}\boldsymbol{\Delta}_l(\hat{\mu},\mu^*) + (1-\frac{\vartheta}{2+\vartheta})\lambda\varrho\big\|(\hat{\bar{\beta}}-\bar{\beta}^*)_{\mathcal{J}^{*C}}\big\|_{2,1} + \frac{2}{2+\vartheta}K\lambda^2J^*.
\end{aligned} \tag{A.80}$$

Set $b' = \vartheta/\{(2+\vartheta)c_0\}$ and add (A.79) and (A.80) to get

$$\begin{aligned}
&\frac{\vartheta}{2+\vartheta}\boldsymbol{\Delta}_l(\hat{\mu},\mu^*) + (\tau-b)\mathbf{D}_2(\hat{\gamma},\gamma^*) - \frac{1}{2a}\mathbf{D}_2(\hat{\bar{m}},\bar{m}^*) - \frac{1}{2a}\mathbf{D}_2(\hat{\varsigma},\varsigma^*) \\
\leq& \frac{2}{2+\vartheta}K\lambda^2J^* + acL^2\omega_{\bar{m}}^2 + 2acL'^2\omega_\varsigma^2 + \frac{\tau^2}{2b}\|\gamma^*\|_2^2.
\end{aligned} \tag{A.81}$$

Taking $b=\tau/4$ and $a=2/\tau$ leads to

$$\frac{\vartheta}{2+\vartheta}\boldsymbol{\Delta}_l(\hat{\mu},\mu^*) + \frac{\tau}{2}\mathbf{D}_2(\hat{\gamma},\gamma^*) \leq \frac{2}{2+\vartheta}K\lambda^2J^* + \frac{2}{\tau}cL^2\omega_{\bar{m}}^2 + \frac{4}{\tau}cL'^2\omega_\varsigma^2 + 2\tau\|\gamma^*\|_2^2, \tag{A.82}$$

or equivalently

$$\boldsymbol{\Delta}_l(\hat{\mu},\mu^*)+\frac{(2+\vartheta)\tau}{2\vartheta}\mathbf{D}_2(\hat{\gamma},\gamma^*)\leq\frac{2}{\vartheta}K\lambda^2J^*+\frac{2+\vartheta}{\vartheta}\frac{2}{\tau}\big(cL^2\omega_{\bar{m}}^2+2cL'^2\omega_{\varsigma}^2\big)+\frac{2+\vartheta}{\vartheta}2\tau\|\gamma^*\|_2^2. \tag{A.83}$$

Note that $\{\tau(2+\vartheta)\}/(2\vartheta)\geq\tau/2$. With $\lambda=\lambda_0/b'=\{(2+\vartheta)/\vartheta\}c_0\lambda_0$, we can derive from (A.83) that

$$\begin{aligned}
&\boldsymbol{\Delta}_l(\hat{\mu},\mu^*)+\frac{\tau}{2}\mathbf{D}_2(\hat{\gamma},\gamma^*)\\
&\leq Kc_0^2\frac{2(2+\vartheta)^2}{\vartheta^3}\lambda_0^2J^*+\frac{2+\vartheta}{\vartheta}\frac{2}{\tau}\big(cL^2\omega_{\bar{m}}^2+2cL'^2\omega_{\varsigma}^2\big)+\frac{2+\vartheta}{\vartheta}2\tau\|\gamma^*\|_2^2\\
&\lesssim(\frac{1}{\vartheta}\vee\frac{1}{\vartheta^3})K\omega_{\bar{\eta}}^2\bigg(\psi^{-1}\Big[p\psi\big\{A\psi^{-1}(p)\big\}\Big]\bigg)^2J^*+(1\vee\frac{1}{\vartheta})\frac{1}{\tau}\big(L^2\omega_{\bar{m}}^2+L'^2\omega_{\varsigma}^2\big)\\
&\quad+(1\vee\frac{1}{\vartheta})\tau\|\gamma^*\|_2^2,
\end{aligned} \tag{A.84}$$

with probability at least $1-C/\psi\{A\psi^{-1}(p)\}-1/\psi(cL)-C/\varphi(cL')$, where C,c are positive constants. The proof is complete.

□

A.5 An Elementwise Estimation Error Bound

Theorem A.3. *Assume that the effective noises $\epsilon_{\bar{\eta}},\epsilon_{\bar{m}}$, and ϵ_{ς} satisfy $\|\epsilon_{\bar{\eta}}\|_{\psi_q}\leq\omega_{\bar{\eta}}$, $\|\epsilon_{\bar{m}}\|_{\psi_q}\leq\omega_{\bar{m}}$, and $\|\epsilon_{\varsigma}\|_{\varphi}\leq\omega_{\varsigma}$ for some $q>0$. Consider $\hat{\zeta}$ as the optimal solution to (19) with $\tau=0$, $\varrho\geq\kappa_{2,\infty}$, and*

$$\lambda=A\omega_{\bar{\eta}}(\log p)^{\frac{1}{q}} \tag{A.85}$$

for some large enough $A>0$. Suppose that there exist some $\vartheta,\alpha>0$ and some large $K>0$ such that for any $\zeta^T=[\bar{\beta}^T,\gamma^T]$

$$\begin{aligned}
&\alpha nJ^*\|\bar{\beta}-\bar{\beta}^*\|_{2,\infty}^2+\alpha\mathbf{D}_2(\gamma,\gamma^*)+(1+\vartheta)\lambda\varrho\big\|(\bar{\beta}-\bar{\beta}^*)_{\mathcal{J}^*}\big\|_{2,1}\\
&\leq\boldsymbol{\Delta}_l(\mu,\mu^*)+(1-\vartheta)\lambda\varrho\big\|(\bar{\beta}-\bar{\beta}^*)_{\mathcal{J}^{*C}}\big\|_{2,1}+K\lambda^2J^*.
\end{aligned} \tag{A.86}$$

Then

$$\|\hat{\bar{\beta}}-\bar{\beta}^*\|_{2,\infty}\leq C\frac{\sqrt{K\alpha}\vee\vartheta}{\alpha\sqrt{n}}A\Big\{\omega_{\bar{\eta}}+\frac{1}{\sqrt{J^*}}(\omega_{\bar{m}}+\omega_{\varsigma})\Big\}(\log p)^{\frac{1}{q}} \tag{A.87}$$

with probability at least $1-Cp^{-c(A\vartheta)^q}-C/\varphi(cA\vartheta(\log p)^{\frac{1}{q}})$, where C,c are positive constants.

The element-wise error bound (A.87) together with a signal strength condition guarantees faithful variable selection with high probability; see Remark 4.

Proof. By the optimality of $\hat{\mu}$: $l(\hat{\mu})+\lambda\varrho\|\hat{\bar{\beta}}\|_{2,1}\leq l(\mu^*)+\lambda\varrho\|\bar{\beta}^*\|_{2,1}$, we obtain the basic inequality

$$\boldsymbol{\Delta}_l(\hat{\mu},\mu^*)\leq\big\langle\epsilon_{\bar{\eta}},\bar{X}\hat{\bar{\beta}}-\bar{X}\bar{\beta}^*\big\rangle+\langle\epsilon_{\bar{m}},\hat{\bar{m}}-\bar{m}\rangle+\langle\epsilon_{\varsigma},\hat{\varsigma}-\varsigma\rangle+\lambda\varrho\|\bar{\beta}^*\|_{2,1}-\lambda\varrho\|\hat{\bar{\beta}}\|_{2,1}. \tag{A.88}$$

We follow the same lines as in Section A.4 to bound the stochastic terms on the right-hand side of (A.88). Let

$$\lambda_0 = A\omega_{\bar{\eta}}(\log p)^{\frac{1}{q}} \tag{A.89}$$

for some large enough $A > 0$ and $\lambda = \lambda_0/b'$ for some $b' > 0$. Similar to the previous analysis, we have with probability at least $1 - Cp^{-A^q}$,

$$\begin{aligned}
&\langle \epsilon_{\bar{\eta}}, \bar{X}\hat{\bar{\beta}} - \bar{X}\bar{\beta}^* \rangle + \lambda\varrho\|\bar{\beta}^*\|_{2,1} - \lambda\varrho\|\hat{\bar{\beta}}\|_{2,1} \\
&\leq \big(1 + c_0 b'\big)\lambda\varrho\big\|(\hat{\bar{\beta}} - \bar{\beta}^*)_{\mathcal{J}^*}\big\|_{2,1} - \big(1 - c_0 b'\big)\lambda\varrho\big\|(\hat{\bar{\beta}} - \bar{\beta}^*)_{\mathcal{J}^{*C}}\big\|_{2,1},
\end{aligned} \tag{A.90}$$

where the constant $c_0 \geq \sqrt{2}$. Moreover, for any $a > 0$, we get

$$\langle \epsilon_{\bar{m}}, \hat{\bar{m}} - \bar{m}^* \rangle + \langle \epsilon_{\varsigma}, \hat{\varsigma} - \varsigma \rangle \leq acL^2\omega_{\bar{m}}^2 + 2acL'^2\omega_{\varsigma}^2 + \frac{1}{2a}\big\{\mathbf{D}_2(\hat{\bar{m}}, \bar{m}^*) + \mathbf{D}_2(\hat{\varsigma}, \varsigma^*)\big\}, \tag{A.91}$$

with probability at least $1 - 1/\psi_q(cL) - C/\varphi(cL')$, where $L, L' > 0$ can be customized.

Plugging the bounds of (A.90) and (A.91) into (A.88), we get

$$\begin{aligned}
&\boldsymbol{\Delta}_l(\hat{\mu}, \mu^*) - \frac{1}{2a}\big\{\mathbf{D}_2(\hat{\bar{m}}, \bar{m}^*) + \mathbf{D}_2(\hat{\varsigma}, \varsigma^*)\big\} \\
&\leq \big(1 + c_0 b'\big)\lambda\varrho\big\|(\hat{\bar{\beta}} - \bar{\beta}^*)_{\mathcal{J}^*}\big\|_{2,1} - \big(1 - c_0 b'\big)\lambda\varrho\big\|(\hat{\bar{\beta}} - \bar{\beta}^*)_{\mathcal{J}^{*C}}\big\|_{2,1} + acL^2\omega_{\bar{m}}^2 + 2acL'^2\omega_{\varsigma}^2,
\end{aligned}$$

with probability at least $1 - Cp^{-A^q} - 1/\psi_q(cL) - C/\varphi(cL')$, where C, c are positive constants.

With the regularity condition and and $b' = \vartheta/c_0$, we obtain

$$\begin{aligned}
&\alpha\big(nJ^*\|\bar{\beta} - \bar{\beta}^*\|_{2,\infty}^2 + \mathbf{D}_2(\gamma, \gamma^*)\big) - \frac{1}{2a}\big\{\mathbf{D}_2(\hat{\bar{m}}, \bar{m}^*) + \mathbf{D}_2(\hat{\varsigma}, \varsigma^*)\big\} \\
&\leq K\lambda^2 J^* + acL^2\omega_{\bar{m}}^2 + 2acL'^2\omega_{\varsigma}^2.
\end{aligned} \tag{A.92}$$

Setting $a = 1/\alpha$ and using $\lambda = \lambda_0/b' = (c_0/\vartheta)\lambda_0$, we obtain

$$\|\bar{\beta} - \bar{\beta}^*\|_{2,\infty}^2 \leq \frac{Kc_0^2}{n\alpha\vartheta^2}A^2\omega_{\bar{\eta}}^2(\log p)^{\frac{2}{q}} + \frac{cL^2\omega_{\bar{m}}^2}{n\alpha^2 J^*} + \frac{2cL'^2\omega_{\varsigma}^2}{n\alpha^2 J^*} \tag{A.93}$$

with probability at least $1 - Cp^{-A^q} - 1/\psi_q(cL) - C/\varphi(cL')$. With $L = L' = A(\log p)^{\frac{1}{q}}$,

$$\begin{aligned}
\|\bar{\beta} - \bar{\beta}^*\|_{2,\infty}^2 &\leq \frac{Kc_0^2}{n\alpha\vartheta^2}A^2\omega_{\bar{\eta}}^2(\log p)^{\frac{2}{q}} + \frac{c}{n\alpha^2 J^*}A^2(\log p)^{\frac{2}{q}}(\omega_{\bar{m}}^2 + \omega_{\varsigma}^2) \\
&\leq C\frac{A^2}{n\alpha^2\vartheta^2}(\log p)^{\frac{2}{q}}\big\{K\alpha\omega_{\bar{\eta}}^2 + \frac{\vartheta^2}{J^*}(\omega_{\bar{m}}^2 + \omega_{\varsigma}^2)\big\},
\end{aligned} \tag{A.94}$$

which implies

$$\|\bar{\beta} - \bar{\beta}^*\|_{2,\infty} \leq C\frac{\sqrt{K\alpha} \vee \vartheta}{\sqrt{n}\alpha\vartheta}A(\log p)^{\frac{1}{q}}\big\{\omega_{\bar{\eta}} + \frac{1}{\sqrt{J^*}}(\omega_{\bar{m}} + \omega_{\varsigma})\big\}, \tag{A.95}$$

with probability at least $1 - Cp^{-cA^q} - C/\varphi(cA(\log p)^{1/q})$. Hence by taking $A' = A/\vartheta$ and $\lambda = A'\omega_{\bar{\eta}}(\log p)^{\frac{1}{q}}$, we have

$$\|\bar{\beta} - \bar{\beta}^*\|_{2,\infty} \leq C\frac{\sqrt{K\alpha} \vee \vartheta}{\sqrt{n}\alpha}A'(\log p)^{\frac{1}{q}}\big\{\omega_{\bar{\eta}} + \frac{1}{\sqrt{J^*}}(\omega_{\bar{m}} + \omega_{\varsigma})\big\}, \tag{A.96}$$

with probability at least $1 - Cp^{-c(A'\vartheta)^q} - C/\varphi(cA'\vartheta(\log p)^{\frac{1}{q}})$. The proof is complete. □

A.6 Analysis of a General Penalty

Consider the following problem associated with a general sparsity-inducing penalty P

$$l(\mu) + \|\varrho\bar{\beta}\|_{2,P} + \frac{\tau}{2}\|\mu\|_2^2, \tag{A.97}$$

where $\|\cdot\|_{2,P}$ is short for $\|\cdot\|_{2,P(\cdot;\lambda)}$ and $\|\bar{\beta}\|_{2,P(\cdot;\lambda)} := \sum_{k=1}^{p} P(\|\bar{\beta}_k\|_2;\lambda)$.

Since a sparsity-inducing penalty necessarily possesses thresholding power, we assume, without loss of generality, that $P(\cdot;\lambda) \geq P_H(\cdot;\lambda)$ throughout the subsection, where the "hard penalty" $P_H(t;\lambda) := (-t^2/2 + \lambda|t|)1_{|t|<\lambda} + (\lambda^2/2)1_{|t|\geq\lambda}$ is induced by the hard-thresholding, and $P_{2,H}(\bar{\beta};\lambda) := \sum_{k=1}^{p} P_H(\|\bar{\beta}_k\|_2;\lambda)$ (see She (2016) for more details). Furthermore, a penalty is referred to as subadditive if it satisfies $P(t+s;\lambda) \leq P(t;\lambda)+P(s;\lambda)$. In fact, when $P(t)$ is concave on $\mathbb{R}_+$ and $P(0) = 0$, $P(|t|)$ is necessarily subadditive. Well-known examples include the widely used ℓ_1-penalty, ℓ_0-penalty, SCAD, MCP, and bridge ℓ_r $(0 < r < 1)$ in the literature.

Theorem A.4. *Assume that the effective noises $\epsilon_{\bar{\eta}}, \epsilon_{\bar{m}}$, and ϵ_ς satisfy $\|\epsilon_{\bar{\eta}}\|_{\psi_2} \leq \omega_{\bar{\eta}}$, $\|\epsilon_{\bar{m}}\|_{\psi_2} \leq \omega_{\bar{m}}$, and $\|\epsilon_\varsigma\|_\varphi \leq \omega_\varsigma$ where $\{\varphi^{-1}(\cdot)\}^2$ concave or $\{\varphi^{-1}(t)\}^2 \lesssim t$ on $\mathbb{R}_+$ (for example, φ can be an L_q-norm with $q \geq 2$ or a ψ_q-norm with $q > 0$). Let $\hat{\zeta}$ denote the optimal solution of (A.97) with $\varrho \geq \kappa_2 := \|\bar{X}\|_2$.*

(a) Let $\lambda = A\omega_{\bar{\eta}}\sqrt{\log(ep)}/\sqrt{\tau \wedge 1}$ with A a sufficiently large constant. Then the following bound always holds

$$\mathbb{E}\left\{\boldsymbol{\Delta}_l(\hat{\mu},\mu^*) + \frac{\tau}{2}\mathbf{D}_2(\hat{\mu},\mu^*)\right\} \lesssim \left\|\varrho\bar{\beta}^*\right\|_{2,P(\cdot;\lambda)} + \frac{1}{\tau\wedge 1}\omega_{\bar{\eta}}^2 + \frac{1}{\tau}\left(\omega_{\bar{m}}^2 + \omega_\varsigma^2\right) + \tau\|\mu^*\|_2^2. \tag{A.98}$$

(b) Let P be a sub-additive penalty. Assume that there exist some $\alpha \geq 0$, $\vartheta > 0$, and some large $K > 0$ such that

$$\alpha\mathbf{D}_2(\mu,\mu^*) + (1+\vartheta)\|\varrho(\bar{\beta}-\bar{\beta}^*)_{\mathcal{J}^*}\|_{2,P} \leq \boldsymbol{\Delta}_l(\mu,\mu^*) + (1-\vartheta)\|\varrho(\bar{\beta}-\bar{\beta}^*)_{\mathcal{J}^{*C}}\|_{2,P} + K\lambda^2 J^*, \tag{A.99}$$

for all $\zeta^T = [\bar{\beta}^T, \bar{m}^T, \varsigma^T]$, where $\lambda = A\omega_{\bar{\eta}}\sqrt{\log(ep)}/\sqrt{\{(\tau+\alpha)\wedge\vartheta\}\vartheta}$ with A a sufficiently large constant. Then

$$\begin{aligned}\mathbb{E}\,\mathbf{D}_2(\hat{\mu},\mu^*) \lesssim\ & \frac{\omega_{\bar{\eta}}^2}{(\tau+\alpha)\{(\tau+\alpha)\wedge\vartheta\}\vartheta}\left\{KA^2J^*\log(ep) + \vartheta\right\} \\ & + \frac{\omega_{\bar{m}}^2 + \{\varphi^{-1}(1)\}^2\omega_\varsigma^2}{(\tau+\alpha)^2} + \left(1\wedge\frac{\tau}{\alpha}\right)^2\|\mu^*\|_2^2.\end{aligned} \tag{A.100}$$

According to the proof below, (A.99) can be relaxed to $\vartheta P_{2,H}\left(\varrho(\bar{\beta}-\bar{\beta}^*)_{\mathcal{J}^*};\lambda\right)+\alpha\mathbf{D}_2(\mu,\mu^*)+\|\varrho(\bar{\beta}-\bar{\beta}^*)_{\mathcal{J}^*}\|_{2,P} \leq \boldsymbol{\Delta}_l(\mu,\mu^*) + \|\varrho(\bar{\beta}-\bar{\beta}^*)_{\mathcal{J}^{*C}}\|_{2,P} - \vartheta P_{2,H}\left(\varrho(\bar{\beta}-\bar{\beta}^*)_{\mathcal{J}^{*C}};\lambda\right) + K\lambda^2J^*$, or $\vartheta P_{2,H}\left(\varrho(\bar{\beta}-\bar{\beta}^*);\lambda\right) + \alpha\mathbf{D}_2(\mu,\mu^*) + \|\varrho\bar{\beta}^*\|_{2,P} \leq \boldsymbol{\Delta}_l(\mu,\mu^*) + \|\varrho\bar{\beta}\|_{2,P} + K\lambda^2J^*$ if P is not subadditive.

Proof. By definition, $l(\hat{\zeta}) + \frac{\tau}{2}\|\hat{\mu}\|_2^2 + \big\|\varrho\hat{\bar{\beta}}\big\|_{2,P} \le l(\zeta^*) + \frac{\tau}{2}\|\mu^*\|_2^2 + \big\|\varrho\bar{\beta}^*\big\|_{2,P}$, which means

$$\begin{aligned}
&\boldsymbol{\Delta}_l(\hat{\mu},\mu^*) + \frac{\tau}{2}\|\hat{\mu}-\mu^*\|_2^2 \\
\le& \big\langle \epsilon_{\bar{\eta}}, \bar{X}\hat{\bar{\beta}} - \bar{X}\bar{\beta}^*\big\rangle + \langle \epsilon_{\bar{m}}, \hat{\bar{m}} - \bar{m}^*\rangle + \langle \epsilon_{\varsigma}, \hat{\varsigma}-\varsigma^*\rangle + \langle -\tau\mu^*, \hat{\mu}-\mu^*\rangle \\
&+ \big\|\varrho\bar{\beta}^*\big\|_{2,P} - \big\|\varrho\hat{\bar{\beta}}\big\|_{2,P} \\
\le& \big\langle \epsilon_{\bar{\eta}}, \bar{X}\hat{\bar{\beta}} - \bar{X}\bar{\beta}^*\big\rangle + \langle \epsilon_{\bar{m}}, \hat{\bar{m}} - \bar{m}^*\rangle + \langle \epsilon_{\varsigma}, \hat{\varsigma}-\varsigma^*\rangle + \frac{b}{2}\|\hat{\mu}-\mu^*\|_2^2 + \frac{\tau^2}{2b}\|\mu^*\|_2^2 \\
&+ \big\|\varrho\bar{\beta}^*\big\|_{2,P} - \big\|\varrho\hat{\bar{\beta}}\big\|_{2,P},
\end{aligned} \tag{A.101}$$

where b can be any positive number.

To bound the stochastic term $\langle \epsilon_{\bar{\eta}}, \bar{X}(\hat{\bar{\beta}} - \bar{\beta}^*)\rangle$, define $\lambda_0 = \omega_{\bar{\eta}}\sqrt{\log(ep)}$ and

$$R = \sup_{\bar{\beta}\in\mathbb{R}^{2p}} \Big\{ \langle \epsilon_{\bar{\eta}}, \bar{X}(\bar{\beta}-\bar{\beta}^*)\rangle - \frac{1}{2a'}\big\|\bar{X}(\bar{\beta}-\bar{\beta}^*)\big\|_2^2 - \frac{1}{2b'}P_{2,H}\big(\varrho(\bar{\beta}-\bar{\beta}^*); \sqrt{a'b'}A\lambda_0\big)\Big\}. \tag{A.102}$$

By Lemma A.7, for any $a' \ge 2b' > 0$ and a sufficiently large constant A, $\mathbb{P}(R \ge a'\omega_{\bar{\eta}}^2 t) \le C\exp(-ct)p^{-cA^2}$. Therefore, $\mathbb{E}[R] \lesssim a'\omega_{\bar{\eta}}^2$.

Next, we bound $\langle \epsilon_{\bar{m}}, \hat{\bar{m}} - \bar{m}^*\rangle$ by

$$\langle \epsilon_{\bar{m}}, \hat{\bar{m}} - \bar{m}^*\rangle = \langle P_{1_n}\epsilon_{\bar{m}}, \hat{\bar{m}} - \bar{m}^*\rangle \le |u^T\epsilon_{\bar{m}}|\cdot\|\hat{\bar{m}} - \bar{m}^*\|_2 \le a(u^T\epsilon_{\bar{m}})^2 + \frac{1}{2a}\mathbf{D}_2(\hat{\bar{m}}, \bar{m}^*) \tag{A.103}$$

for any $a > 0$, where $P_{1_n} = uu^T$ with $u = (1/\sqrt{n})1_n$. By the assumption on $\epsilon_{\bar{m}}$, $\|(u^T\epsilon_{\bar{m}})^2\|_{\psi_1} \lesssim \|u^T\epsilon_{\bar{m}}\|_{\psi_2}^2 \le \omega_{\bar{m}}^2$, from which it follows that $\mathbb{E}[(u^T\epsilon_{\bar{m}})^2] \le C\omega_{\bar{m}}^2$. Similarly, for the stochastic term $\langle \epsilon_{\varsigma}, \hat{\varsigma}-\varsigma^*\rangle$ with $\epsilon_{\varsigma} = [\epsilon_{\frac{1}{\sigma}1_n}^T, \epsilon_{\frac{1}{\nu}1_n}^T]^T$ and $\hat{\varsigma}-\varsigma^* = [(1/\hat{\sigma}-1/\sigma^*)1_n^T, (1/\hat{\nu}-1/\nu^*)1_n^T]^T$, we have

$$\begin{aligned}
&\big\langle \epsilon_{\frac{1}{\sigma}1_n}, (\frac{1}{\hat{\sigma}} - \frac{1}{\sigma^*})1_n\big\rangle + \big\langle \epsilon_{\frac{1}{\nu}1_n}, (\frac{1}{\hat{\nu}} - \frac{1}{\nu^*})1_n\big\rangle \\
=& \big\langle P_{1_n}\epsilon_{\frac{1}{\sigma}1_n}, (\frac{1}{\hat{\sigma}} - \frac{1}{\sigma^*})1_n\big\rangle + \big\langle P_{1_n}\epsilon_{\frac{1}{\nu}1_n}, (\frac{1}{\hat{\nu}} - \frac{1}{\nu^*})1_n\big\rangle \\
\le& a(u^T\epsilon_{\frac{1}{\sigma}1_n})^2 + a(u^T\epsilon_{\frac{1}{\nu}1_n})^2 + \frac{1}{2a}\mathbf{D}_2(\hat{\varsigma}, \varsigma^*).
\end{aligned}$$

From Lemma A.4, we get a bound in L_2-norm:

$$\mathbb{E}\big[(u^T\epsilon_{\frac{1}{\sigma}1_n})^2\big] \le \{\varphi^{-1}(1)\}^2\|u^T\epsilon_{\frac{1}{\sigma}1_n}\|_{\varphi}^2 \le \{\varphi^{-1}(1)\}^2\omega_{\varsigma}^2,$$

and $\mathbb{E}[(u^T\epsilon_{\frac{1}{\nu}1_n})^2] \le \{\varphi^{-1}(1)\}^2\omega_{\varsigma}^2$. To sum up, for any $a > 0$,

$$\mathbb{E}\big\{\langle \epsilon_{\bar{m}}, \hat{\bar{m}} - \bar{m}^*\rangle + \langle \epsilon_{\varsigma}, \hat{\varsigma}-\varsigma^*\rangle\big\} \le aC\omega_{\bar{m}}^2 + 2a\{\varphi^{-1}(1)\}^2\omega_{\varsigma}^2 + \frac{1}{2a}\{\mathbf{D}_2(\hat{\bar{m}}, \bar{m}^*) + \mathbf{D}_2(\hat{\varsigma}, \varsigma^*)\}. \tag{A.104}$$

Now, plugging the bounds (A.102) and (A.104) into (A.101) yields

$$\begin{aligned}
&\mathbb{E}\Big\{\boldsymbol{\Delta}_l(\hat{\mu},\mu^*)+\frac{\tau-b}{2}\|\hat{\mu}-\mu^*\|_2^2\Big\}-\frac{1}{a'}\mathbf{D}_2(\hat{\bar{\eta}},\bar{\eta}^*)-\frac{1}{2a}\{\mathbf{D}_2(\hat{\bar{m}},\bar{m}^*)+\mathbf{D}_2(\hat{\varsigma},\varsigma^*)\}\\
&\le\frac{1}{2b'}P_{2,H}\big(\varrho(\hat{\bar{\beta}}-\bar{\beta}^*);\sqrt{a'b'}A\lambda_0\big)+\big\|\varrho\bar{\beta}^*\big\|_{2,P(\cdot;\lambda)}-\big\|\varrho\hat{\bar{\beta}}\big\|_{2,P(\cdot;\lambda)}\\
&\quad+aC\omega_{\bar{m}}^2+2a\{\varphi^{-1}(1)\}^2\omega_\varsigma^2+\frac{\tau^2}{2b}\|\mu^*\|_2^2+Ca'\omega_{\bar{\eta}}^2. \qquad \text{(A.105)}
\end{aligned}$$

To prove part (a), we use the subadditivity of P_H:

$$\begin{aligned}
&\mathbb{E}\Big\{\boldsymbol{\Delta}_l(\hat{\mu},\mu^*)+\frac{\tau-b}{2}\|\hat{\mu}-\mu^*\|_2^2\Big\}-\frac{1}{a'}\mathbf{D}_2(\hat{\bar{\eta}},\bar{\eta}^*)-\frac{1}{2a}\{\mathbf{D}_2(\hat{\bar{m}},\bar{m}^*)+\mathbf{D}_2(\hat{\varsigma},\varsigma^*)\}\\
&\le\frac{1}{2b'}P_{2,H}\big(\varrho\hat{\bar{\beta}};\sqrt{a'b'}A\lambda_0\big)+\frac{1}{2b'}P_{2,H}\big(\bar{\beta}^*;\sqrt{a'b'}A\lambda_0\big)+\big\|\varrho\bar{\beta}^*\big\|_{2,P(\cdot;\lambda)}-\big\|\varrho\hat{\bar{\beta}}\big\|_{2,P(\cdot;\lambda)}\\
&\quad+aC\omega_{\bar{m}}^2+2a\{\varphi^{-1}(1)\}^2\omega_\varsigma^2+\frac{\tau^2}{2b}\|\mu^*\|_2^2+Ca'\omega_{\bar{\eta}}^2. \qquad \text{(A.106)}
\end{aligned}$$

Because $P(\cdot;\lambda)\ge P_H(\cdot;\lambda)$, taking $b=\tau/4, b'=1/2, a=2/\tau, a'=4/(\tau\wedge 1)$ gives

$$\begin{aligned}
&\mathbb{E}\Big\{\boldsymbol{\Delta}_l(\hat{\mu},\mu^*)+\frac{\tau}{2}\mathbf{D}_2(\hat{\mu},\mu^*)\Big\}\\
&\le 2\big\|\varrho\bar{\beta}^*\big\|_{2,P(\cdot;\lambda)}+2\tau\|\mu^*\|_2^2+\frac{2C}{\tau}\omega_{\bar{m}}^2+\frac{4}{\tau}\{\varphi^{-1}(1)\}^2\omega_\varsigma^2+\frac{4C}{\tau\wedge 1}\omega_{\bar{\eta}}^2\\
&\lesssim\big\|\varrho\bar{\beta}^*\big\|_{2,P(\cdot;\lambda)}+\tau\|\mu^*\|_2^2+\frac{1}{\tau}(\omega_{\bar{m}}^2+\omega_\varsigma^2)+\frac{1}{\tau\wedge 1}\omega_{\bar{\eta}}^2. \qquad \text{(A.107)}
\end{aligned}$$

To prove part (b), we use $\big\|\varrho\bar{\beta}^*_{\mathcal{J}^{*c}}\big\|_{2,P(\cdot;\lambda)}=\big\|\varrho(\bar{\beta}-\bar{\beta}^*)_{\mathcal{J}^{*c}}\big\|_{2,P(\cdot;\lambda)}$ and $\big\|\varrho\bar{\beta}^*_{\mathcal{J}^*}\big\|_{2,P(\cdot;\lambda)}-\big\|\varrho\bar{\beta}_{\mathcal{J}^*}\big\|_{2,P(\cdot;\lambda)}\le\big\|\varrho(\bar{\beta}-\bar{\beta}^*)_{\mathcal{J}^*}\big\|_{2,P(\cdot;\lambda)}$ and rewrite (A.105) as

$$\begin{aligned}
&\mathbb{E}\Big\{\boldsymbol{\Delta}_l(\hat{\mu},\mu^*)+\frac{\tau-b}{2}\|\hat{\mu}-\mu^*\|_2^2\Big\}-\frac{1}{a'}\mathbf{D}_2(\hat{\bar{\eta}},\bar{\eta}^*)-\frac{1}{2a}\{\mathbf{D}_2(\hat{\bar{m}},\bar{m}^*)+\mathbf{D}_2(\hat{\varsigma},\varsigma^*)\}\\
&\le\frac{1}{2b'}P_{2,H}\big(\varrho(\hat{\bar{\beta}}-\bar{\beta}^*)_{\mathcal{J}^*};\sqrt{a'b'}A\lambda_0\big)+\frac{1}{2b'}P_{2,H}\big(\varrho(\hat{\bar{\beta}}-\bar{\beta}^*)_{\mathcal{J}^{*C}};\sqrt{a'b'}A\lambda_0\big)\\
&\quad+\big\|\varrho(\bar{\beta}-\bar{\beta}^*)_{\mathcal{J}^*}\big\|_{2,P(\cdot;\lambda)}-\big\|\varrho(\bar{\beta}-\bar{\beta}^*)_{\mathcal{J}^{*c}}\big\|_{2,P(\cdot;\lambda)}+aC\omega_{\bar{m}}^2+2a\{\varphi^{-1}(1)\}^2\omega_\varsigma^2\\
&\quad+\frac{\tau^2}{2b}\|\mu^*\|_2^2+Ca'\omega_{\bar{\eta}}^2. \qquad \text{(A.108)}
\end{aligned}$$

The condition (A.99) implies

$$\begin{aligned}
&\alpha\mathbf{D}_2(\mu,\mu^*)+\vartheta P_{2,H}\big(\varrho(\bar{\beta}-\bar{\beta}^*)_{\mathcal{J}^*};\lambda\big)+\|\varrho(\bar{\beta}-\bar{\beta}^*)_{\mathcal{J}^*}\|_{2,P(\cdot;\lambda)}\\
&\le\boldsymbol{\Delta}_l(\mu,\mu^*)+\|\varrho(\bar{\beta}-\bar{\beta}^*)_{\mathcal{J}^{*C}}\|_{2,P(\cdot;\lambda)}-\vartheta P_{2,H}\big(\varrho(\bar{\beta}-\bar{\beta}^*)_{\mathcal{J}^{*C}};\lambda\big)+K\lambda^2J^*. \qquad \text{(A.109)}
\end{aligned}$$

With $b'=1/(2\vartheta)$, we add (A.108) and (A.109) to get

$$\begin{aligned}
&\mathbb{E}\Big\{(\tau+\alpha-b)\mathbf{D}_2(\hat{\mu},\mu^*)\Big\}-\frac{1}{a'}\mathbf{D}_2(\hat{\bar{\eta}},\bar{\eta}^*)-\frac{1}{2a}\{\mathbf{D}_2(\hat{\bar{m}},\bar{m}^*)+\mathbf{D}_2(\hat{\varsigma},\varsigma^*)\}\\
&\le K\lambda^2J^*+aC\omega_{\bar{m}}^2+2a\{\varphi^{-1}(1)\}^2\omega_\varsigma^2+\frac{\tau^2}{2b}\|\mu^*\|_2^2+Ca'\omega_{\bar{\eta}}^2, \qquad \text{(A.110)}
\end{aligned}$$

where $\lambda = \sqrt{a'b'}A\lambda_0$ and $A \geq A_0$ with A_0 given in the Lemma A.7. Now, setting $b = (\tau+\alpha)/4, a' = 4/\{(\tau+\alpha)\wedge\vartheta\}, a = 2/(\tau+\alpha)$ gives

$$\begin{aligned}&\mathbb{E}\Big\{\frac{\tau+\alpha}{2}\mathbf{D}_2(\hat{\mu},\mu^*)\Big\}\\ &\leq \frac{2KA^2}{\{(\tau+\alpha)\wedge\vartheta\}\vartheta}\omega_{\bar{\eta}}{}^2 J^*\log(ep) + \frac{2C\omega_{\bar{m}}^2}{(\tau+\alpha)} + \frac{4\{\varphi^{-1}(1)\}^2\omega_\varsigma^2}{(\tau+\alpha)} + \frac{2\tau^2\|\mu^*\|_2^2}{\tau+\alpha} + \frac{4C\omega_{\bar{\eta}}^2}{(\tau+\alpha)\wedge\vartheta}.\end{aligned}$$

The proof is complete. □

Lemma A.7. *Let $\epsilon = [\epsilon_i]$ be an n-dimensional random vector (not necessarily mean centered or having independent components) satisfying $\|\epsilon\|_{\psi_2} \leq \omega$. Suppose that $\bar{X} \in \mathbb{R}^{n\times 2p}$ satisfies $\|\bar{X}\|_2 \leq \varrho$. Let $\lambda_0 = \omega\sqrt{\log(ep)}$. Then there exist universal constants $A_0, C, c > 0$ such that for any $a \geq 2b > 0$, $A_1 \geq A_0$, the following event*

$$\sup_{\bar{\beta}\in\mathbb{R}^{2p}} 2\langle\epsilon, \bar{X}\bar{\beta}\rangle - \frac{1}{a}\|\bar{X}\bar{\beta}\|_2^2 - \frac{1}{b}P_{2,H}(\varrho\bar{\beta}; \sqrt{ab}A_1\lambda_0) \geq a\omega^2 t \tag{A.111}$$

occurs with probability at most $C\exp(-ct)p^{-cA_1^2}$.

Proof. Let $P_0 = (\lambda^2/2)1_{t\neq 0}$. Define $l_H(\bar{\beta}) = 2\langle\epsilon, \bar{X}\bar{\beta}\rangle - \frac{1}{a}\|\bar{X}\bar{\beta}\|_2^2 - \frac{1}{b}P_{2,H}(\bar{\beta}; \sqrt{ab}A_0\lambda_0), l_0(\bar{\beta}) = 2\langle\epsilon, \bar{X}\bar{\beta}\rangle - \frac{1}{a}\|\bar{X}\bar{\beta}\|_2^2 - \frac{1}{b}P_{2,0}(\bar{\beta}; \sqrt{ab}A_0\lambda_0)$. Introduce two events $\varepsilon_H = \{\sup_{\bar{\beta}\in\Gamma} l_H(\bar{\beta}) \geq at\omega^2\}$, and $\varepsilon_0 = \{\sup_{\bar{\beta}\in\Gamma} l_0(\bar{\beta}) \geq at\omega^2\}$

First, we use an optimization technique to prove that $\varepsilon_H = \varepsilon_0$. Since $P_0 \geq P_H$, $P_{2,0} \geq P_{2,H}$ and thus $\varepsilon_0 \subset \varepsilon_H$. The occurrence of ε_H implies that $l_H(\bar{\beta}^o) \geq at\omega^2$ for any $\bar{\beta}^o$ defined by

$$\bar{\beta}^o \in \arg\min_{\bar{\beta}\in\mathbb{R}^{2p}} \frac{1}{a}\|\bar{X}\bar{\beta}\|_2^2 - 2\langle\epsilon, \bar{X}\bar{\beta}\rangle + \frac{1}{b}P_{2,H}(\bar{\beta}; \sqrt{ab}A_0\lambda_0). \tag{A.112}$$

From Lemma 5 in She (2016), under $\|\bar{X}\|_2 \leq 1$, there exists a globally optimal solution $\bar{\beta}^0$ to the problem

$$\min_{\bar{\beta}\in\mathbb{R}^{2p}} \frac{1}{2}\|y - \bar{X}\bar{\beta}\|_2^2 + P_{2,H}(\bar{\beta}; \lambda), \tag{A.113}$$

such that for any $j = 1, \ldots, p$, either $\bar{\beta}_j^0 = 0$ or $\|\bar{\beta}_j^0\|_2 \geq \lambda$. Therefore, with $a \geq 2b > 0$, there exists at least one global minimizer $\bar{\beta}^{oo}$ satisfying $P_{2,H}(\bar{\beta}^{oo}; \sqrt{ab}A_1\lambda_0) = P_{2,0}(\bar{\beta}^{oo}; \sqrt{ab}A_1\lambda_0)$ and thus $l_H(\bar{\beta}^{oo}) = l_0(\bar{\beta}^{oo})$. This means $\sup_{\bar{\beta}\in\mathbb{R}^{2p}} l_0(\bar{\beta}) \geq l_0(\bar{\beta}^{oo}) = l_H(\bar{\beta}^{oo}) \geq at\omega^2$, and so $\varepsilon_H \subset \varepsilon_0$. It suffices to prove $\mathbb{P}(\varepsilon_0) \leq C\exp(-ct)p^{-cA_1^2}$.

Next, we use Lemma A.8 to bound the tail probability of R defined by

$$R = \sup_{1\leq J\leq p}\sup_{\bar{\beta}\in\Gamma_J}\Big\{\langle\epsilon, \bar{X}\bar{\beta}\rangle - \frac{1}{2a}\|\bar{X}\bar{\beta}\|_2^2 - \frac{1}{2b}P_{2,0}(\bar{\beta}; \sqrt{ab}A_1\lambda_0)\Big\}, \tag{A.114}$$

where $P_{2,0}(\bar{\beta}; \lambda_0) = P_0(J; \lambda_0) = (1/2)J\lambda_0^2$ for $\bar{\beta} \in \Gamma_J$ and $\Gamma_J = \{\bar{\beta} \in \mathbb{R}^{2p} : J(\bar{\beta}) = J\}$ (in the trivial case of $J = 0$, the quantity inside the braces is 0).

By a scaling argument, for any $a > 0$,

$$\langle\epsilon, \bar{X}\bar{\beta}/\|\bar{X}\bar{\beta}\|_2\rangle\|\bar{X}\bar{\beta}\|_2 \leq \frac{a}{2}\sup_{\alpha\in\Gamma_J}(\langle\epsilon, \alpha\rangle)^2 + \frac{1}{2a}\|\bar{X}\bar{\beta}\|_2^2. \tag{A.115}$$

Applying Lemma A.8 with $G = 2, q = 2$ results in

$$\mathbb{P}\Big(\sup_{\bar{\beta}\in\Gamma_J} \langle \epsilon, \bar{X}\bar{\beta}\rangle - \frac{1}{2a}\|\bar{X}\bar{\beta}\|_2^2 - \frac{a}{2}LJ\log(ep)\omega^2 \geq \frac{a}{2}\omega^2 t\Big) \leq C\exp(-ct), \tag{A.116}$$

or

$$\mathbb{P}\Big(\sup_{\bar{\beta}\in\Gamma_J} \langle \epsilon, \bar{X}\bar{\beta}\rangle - \frac{1}{2a}\|\bar{X}\bar{\beta}\|_2^2 - aLP_{2,0}(\bar{\beta};\lambda_0) \geq a\omega^2 t\Big) \leq C\exp(-ct). \tag{A.117}$$

Set $A_1 \geq \sqrt{2L}$. Noticing that (i) $(A_1^2/2)P_0(J;\lambda_0) \geq LP_0(J;\lambda_0) + cA_1^2P_0(J;\lambda_0)$ for some $c > 0$, and (ii) $J\log(ep) \geq \log p + J$ for any $J \geq 1$, we get

$$\begin{aligned}
&\mathbb{P}\big(R \geq a\omega^2 t\big) \\
&\leq \sum_{J=1}^{p} \mathbb{P}\Big(\sup_{\bar{\beta}\in\Gamma_J} \langle \epsilon, \bar{X}\bar{\beta}\rangle - \frac{1}{2a}\|\bar{X}\bar{\beta}\|_2^2 - \frac{1}{2b}P_{2,0}(\bar{\beta};\sqrt{ab}A_1\lambda_0) \geq a\omega^2 t\Big) \\
&= \sum_{J=1}^{p} \mathbb{P}\Big(\sup_{\bar{\beta}\in\Gamma_J} \langle \epsilon, \bar{X}\bar{\beta}\rangle - \frac{1}{2a}\|\bar{X}\bar{\beta}\|_2^2 - \frac{1}{4}aA_1^2J\log(ep)\omega_{\bar{\eta}}^2 \geq a\omega^2 t\Big) \\
&\leq \sum_{J=1}^{p} \mathbb{P}\Big(\sup_{\bar{\beta}\in\Gamma_J} \langle \epsilon, \bar{X}\bar{\beta}\rangle - \frac{1}{2a}\|\bar{X}\bar{\beta}\|_2^2 - \frac{a}{2}LJ\log(ep)\omega^2 \geq a\omega^2 t + caA_1^2J\log(ep)\omega^2\Big) \\
&\leq \sum_{J=1}^{p} C\exp(-ct)\sum_{J=1}^{p}\exp\{-cA_1^2(J+\log p)\} \\
&\leq C\exp(-ct)p^{-cA_1^2},
\end{aligned} \tag{A.118}$$

where the last inequality is due to the sum of geometric sequence. □

Lemma A.8. *Given a matrix $X \in \mathbb{R}^{n\times pG}$ with a block form of $X = [X_1, \ldots, X_p]$ with $X_j \in \mathbb{R}^{n\times G}, 1 \leq j \leq p$, let $X_{\mathcal{J}}$ denote the submatrix formed by the column blocks of X indexed by $\mathcal{J}$. Define $\Gamma_{\mathcal{J}} = \{\alpha \in \mathbb{R}^{pG} : \|\alpha\|_2 \leq 1, \alpha \in \mathcal{R}(X_{\mathcal{J}})\}$ and $\Gamma_J = \bigcup_{|\mathcal{J}|=J}\Gamma_{\mathcal{J}}$, where $1 \leq J \leq p$. Let $\epsilon = [\epsilon_i]$ be an n-dimensional random vector. (i) Assume ϵ satisfies $\|\epsilon\|_{\psi_q} \leq \omega$ for some $q \geq 1$. Then for any $t > 0$,*

$$\mathbb{P}\Big(\sup_{\alpha\in\Gamma_{\mathcal{J}}} (\langle \epsilon, \alpha\rangle)^2 - LJG\omega^2 - L(JG)^{\frac{2}{q}}\omega^2 \geq \omega^2 t\Big) \leq C\exp\big(-ct^{\frac{q}{2}}\big), \tag{A.119}$$

$$\mathbb{P}\Big(\sup_{\alpha\in\Gamma_J} (\langle \epsilon, \alpha\rangle)^2 - L\big[JG + (JG)^{\frac{2}{q}} + (2J)^{\frac{2}{q}}\{\log(ep)\}^{\frac{2}{q}}\big]\omega^2 \geq \omega^2 t\Big) \leq C\exp\big(-ct^{\frac{q}{2}}\big), \tag{A.120}$$

where C is a universal constant and c, L are constants depending on q only. (ii) Assume that $\epsilon_1, \ldots, \epsilon_n$ are independent, centered, and $\|\epsilon_i\|_{\psi_q} \leq \omega$ for some $q \in (0, 2]$. Then (A.119) and (A.120) are replaced by

$$\mathbb{P}\Big(\sup_{\alpha\in\Gamma_{\mathcal{J}}} (\langle \epsilon, \alpha\rangle)^2 - LJG\omega^2 \geq \omega^2 t\Big) \leq C\exp\big(-ct^{\frac{q}{2}}\big), \tag{A.121}$$

$$\mathbb{P}\Big(\sup_{\alpha\in\Gamma_J} (\langle \epsilon, \alpha\rangle)^2 - L\big[JG + (2J)^{\frac{2}{q}}\{\log(ep)\}^{\frac{2}{q}}\big]\omega^2 \geq \omega^2 t\Big) \leq C\exp\big(-ct^{\frac{q}{2}}\big). \tag{A.122}$$

Proof. First, we show (A.119) under the assumption that $\|\epsilon\|_{\psi_q} \le \omega$ for some $q \ge 1$. Define a centered random vector $\epsilon_c = \epsilon - M$ with $M = \mathbb{E}[\epsilon]$, then

$$\sup_{\alpha\in\Gamma_{\mathcal{J}}} \langle \epsilon, \alpha\rangle \le \sup_{\alpha\in\Gamma_{\mathcal{J}}} \langle \epsilon_c, \alpha\rangle + \sup_{\alpha\in\Gamma_{\mathcal{J}}} \langle P_{X_{\mathcal{J}}} M, \alpha\rangle \le \sup_{\alpha\in\Gamma_{\mathcal{J}}} \langle \epsilon_c, \alpha\rangle + \|U^T M\|_2, \tag{A.123}$$

where $P_{X_{\mathcal{J}}} = UU^T$ is the orthogonal projection matrix onto $\mathcal{R}(X_{\mathcal{J}})$ with $\{U_1, \ldots, U_{JG}\}$ as an orthonormal basis and $\|P_{X_{\mathcal{J}}}\|_2 = 1$. We claim that $\|U^T M\|_2^2 = \sum_{i=1}^{JG} (\mathbb{E}[U_i^T \epsilon])^2 \le CJG\omega^2$. In fact, because $\|U_i^T \epsilon\|_{\psi_q} \le \omega$,

$$\mathbb{P}\big(|U_i^T \epsilon| > t\omega\big) \le \int_0^{+\infty} \frac{1}{\psi_q(\frac{t\omega}{\omega})}\, \mathrm{d}t = \int_0^{+\infty} \exp(-t^q)\, \mathrm{d}t < \infty, \tag{A.124}$$

and so $\mathbb{E}|U_i^T \epsilon| \le C\omega$, $1 \le i \le J$.

By definition, $\{\langle \epsilon_c, \alpha\rangle : \alpha \in \Gamma_{\mathcal{J}}\}$ is a (centered) ψ_q-process. The induced metric on $\Gamma_{\mathcal{J}}$ is: $d(\alpha, \alpha') = \omega\|\alpha - \alpha'\|_2$. To bound the metric entropy $\log(\mathcal{N}(\varepsilon, \Gamma_{\mathcal{J}}, d)))$, where the $\mathcal{N}(\varepsilon, \Gamma_{\mathcal{J}}, d))$ is the smallest cardinality of an ε-net that covers $\Gamma_{\mathcal{J}}$ under the metric d, we apply a standard volume argument to get

$$\log(\mathcal{N}(\varepsilon, \Gamma_{\mathcal{J}}, d))) \le \log(\frac{C\omega}{\varepsilon})^{JG} = JG \log(C\omega/\varepsilon). \tag{A.125}$$

By Theorem 5.36 in Wainwright (2019), we get

$$\mathbb{P}\Big(\sup_{\alpha,\alpha'\in\Gamma_{\mathcal{J}}} |\langle \epsilon_c, \alpha - \alpha'\rangle| - L\int_0^D \psi_q^{-1}(\mathcal{N}(\varepsilon, \Gamma_{\mathcal{J}}, d))\, \mathrm{d}\varepsilon \ge Lt\Big) \le 2\exp\big(-\frac{t^q}{D^q}\big), \tag{A.126}$$

for any $t > 0$, where $D = \sup_{\alpha,\alpha'\in\Gamma_{\mathcal{J}}} d(\alpha, \alpha') = 2\omega$. Equivalently,

$$\mathbb{P}\Big(\sup_{\alpha,\alpha'\in\Gamma_{\mathcal{J}}} |\langle \epsilon_c, \alpha - \alpha'\rangle| - L\int_0^{2\omega} \psi_q^{-1}(\mathcal{N}(\varepsilon, \Gamma_{\mathcal{J}}, d))\, \mathrm{d}\varepsilon \ge \omega t\Big) \le C\exp\big(-ct^q\big), \tag{A.127}$$

which implies

$$\mathbb{P}\Big(\sup_{\alpha\in\Gamma_{\mathcal{J}}} \langle \epsilon_c, \alpha\rangle - L\int_0^{2\omega} \psi_q^{-1}(\mathcal{N}(\varepsilon, \Gamma_{\mathcal{J}}, d))\, \mathrm{d}\varepsilon \ge \omega t\Big) \le C\exp\big(-ct^q\big). \tag{A.128}$$

Based on (A.124) and (A.128), we obtain

$$\mathbb{P}\Big(\sup_{\alpha\in\Gamma_{\mathcal{J}}} \langle \epsilon, \alpha\rangle - L\sqrt{JG}\omega - L\int_0^{2\omega} \psi_q^{-1}(\mathcal{N}(\varepsilon, \Gamma_{\mathcal{J}}, d))\, \mathrm{d}\varepsilon \ge \omega t\Big) \le C\exp\big(-ct^q\big), \tag{A.129}$$

or

$$\mathbb{P}\Big(\sup_{\alpha\in\Gamma_{\mathcal{J}}} \langle \epsilon, \alpha\rangle - L\sqrt{JG}\omega - L(JG)^{\frac{1}{q}}\omega \ge \omega t\Big) \le C\exp\big(-ct^q\big). \tag{A.130}$$

Therefore,

$$\mathbb{P}\Big(\sup_{\alpha\in\Gamma_{\mathcal{J}}} (\langle \epsilon, \alpha\rangle)^2 - LJG\omega^2 - L(JG)^{\frac{2}{q}}\omega^2 \ge \omega^2 t^2\Big) \le C\exp\big(-ct^q\big), \tag{A.131}$$

and letting $s = t^2$ gives (A.119). With a union bound, we also obtain

$$\mathbb{P}\Big(\sup_{\alpha \in \Gamma_J} (\langle \epsilon, \alpha \rangle)^2 - LJG\omega^2 - L(JG)^{\frac{2}{q}}\omega^2 \geq \omega^2 t \Big) \leq \binom{p}{J} C \exp\big(- ct^{\frac{q}{2}} \big) \\ \leq C \exp\big\{ - ct^{\frac{q}{2}} + J \log(ep) \big\}, \tag{A.132}$$

from which it follows

$$\mathbb{P}\Big(\sup_{\alpha \in \Gamma_J} (\langle \epsilon, \alpha \rangle)^2 - LJG\omega^2 - L(JG)^{\frac{2}{q}}\omega^2 - L2^{\frac{2}{q}-1}\{J \log(ep)\}^{\frac{2}{q}}\omega^2 \geq \omega^2 t \Big) \leq C \exp\big(- ct^{\frac{q}{2}} \big). \tag{A.133}$$

Next, we prove (A.121) assuming $\epsilon_1, \ldots, \epsilon_n$ are independent, centered, and $\|\epsilon_i\|_{\psi_q} \leq \omega$ for some $0 < q \leq 2$. By Hölder's inequality,

$$\sup_{\alpha \in \Gamma_{\mathcal{J}}} \langle \epsilon, \alpha \rangle = \sup_{\alpha \in \Gamma_{\mathcal{J}}} \langle P_{X_{\mathcal{J}}} \epsilon, \alpha \rangle \leq \|U^T \epsilon\|_2, \tag{A.134}$$

where $P_{X_{\mathcal{J}}} = UU^T$ is the orthogonal projection matrix onto $\mathcal{R}(X_{\mathcal{J}})$ with $\{U_1, \ldots, U_{JG}\}$ as an orthonormal basis and $\|P_{X_{\mathcal{J}}}\|_2 = 1$. It remains to get a tail bound for $\|U^T \epsilon\|_2^2$.

Noticing that (i) $\mathbb{E}[\epsilon] = 0$, (ii) $\|\epsilon_i\|_{\psi_q} \leq \omega$ implies $\mathbb{E}[\epsilon_i^2] \leq C\omega^2$, (iii) $\|UU^T\|_2 = 1$, and (iv) $Tr(UU^T) = JG$, we apply a generalized Hanson-Wright inequality (Sambale (2020), Theorem 2.1) to obtain

$$\mathbb{P}\big(\|U^T \epsilon\|_2^2 - LJG\omega^2 \geq \omega^2 t \big) \leq C \exp\big(- ct^{\frac{q}{2}} \big), \tag{A.135}$$

where C can be 2 and c, L are constants depending on q only. This results in

$$\mathbb{P}\Big(\sup_{\alpha \in \Gamma_{\mathcal{J}}} (\langle \epsilon, \alpha \rangle)^2 - LJG\omega^2 \geq \omega^2 t \Big) \leq C \exp\big(- ct^{\frac{q}{2}} \big). \tag{A.136}$$

Finally, we prove (A.122). Applying the union bound on (A.136) yields

$$\mathbb{P}\Big(\sup_{\alpha \in \Gamma_J} (\langle \epsilon, \alpha \rangle)^2 - LJG\omega^2 \geq \omega^2 t \Big) \leq \binom{p}{J} C \exp\big(- ct^{\frac{q}{2}} \big) \leq C \exp\big\{ - ct^{\frac{q}{2}} + J \log(ep) \big\}. \tag{A.137}$$

Let $t' = t^{\frac{q}{2}} - (1/c) J \log(ep)$. Then (A.137) can be rewritten as

$$\mathbb{P}\Big(\sup_{\alpha \in \Gamma_J} (\langle \epsilon, \alpha \rangle)^2 - LJG\omega^2 \geq \omega^2 \{ t' + \frac{1}{c} J \log(ep) \}^{\frac{2}{q}} \Big) \leq C \exp(-ct'). \tag{A.138}$$

Given $0 < q \leq 2$, by the convexity of $x^{\frac{2}{q}}$ on $\mathbb{R}_+$, we have $(a+b)^{\frac{2}{q}} \leq 2^{\frac{2}{q}-1}(a^{\frac{2}{q}} + b^{\frac{2}{q}})$ for any $a, b \geq 0$, and thus (A.138) becomes

$$\mathbb{P}\Big(\sup_{\alpha \in \Gamma_J} (\langle \epsilon, \alpha \rangle)^2 - LJG\omega^2 - L2^{\frac{2}{q}-1}\{J \log(ep)\}^{\frac{2}{q}}\omega^2 \geq 2^{\frac{2}{q}-1}\omega^2 (t')^{\frac{2}{q}} \Big) \leq C \exp(-ct'), \tag{A.139}$$

or equivalently,

$$\mathbb{P}\Big(\sup_{\alpha \in \Gamma_J} (\langle \epsilon, \alpha \rangle)^2 - LJG\omega^2 - L2^{\frac{2}{q}-1}\{J \log(ep)\}^{\frac{2}{q}}\omega^2 \geq \omega^2 s \Big) \leq C \exp\big(- cs^{\frac{q}{2}} \big). \tag{A.140}$$

The proof is now complete. □

B Further Extensions

We showcase two nonparametric statistical applications of skewed pivot-blend, one based on data ranks and the other based on kernels.

Skewed rank-based estimation. Recently, there has been a lot of interest in rank-based nonparametric estimation methods that minimize $\sqrt{12}\sum_{i=1}^{n} r_i\{\mathrm{R}(r_i)/(n+1)-\frac{1}{2}\}$ (Jaeckel, 1972), where $\mathrm{R}(r_i)$ denotes the rank of r_i among $r_1,\ldots,r_n$, or equivalently, the ℓ_1 loss on the spread of residuals: $\frac{1}{n(n-1)}\sum_{i\neq j}|r_i-r_j|$ (Hettmansperger and McKean, 1978, 2010; Wang et al., 2020). Applying the technique of U-statistics (with kernel $h(r_i,r_j)=|r_i-r_j|$) can show that the criterion results in estimators with desired asymptotic properties.

The adoption of the symmetric (and relatively robust) ℓ_1-loss function heavily depends on the assumption that r_i follows an i.i.d. distribution (when assessed against the statistical truth). In turn, it implies that the distribution of differences r_i-r_j is symmetrical, as in the case of a double-exponential.

Nonetheless, the assumption may not hold in real-world applications **beyond** the i.i.d. setting. Such deviations can result in extra skewness of $|r_i-r_j|$ that cannot be captured by an exponential (or half-normal) distribution. We introduce a skewed criterion that operates on the *absolute* differences $|r_i-r_j|$ and $m,\sigma,\nu>0$:

$$\sum_{i\neq j}\Big\{\Big(\frac{|r_i-r_j|-m}{\sigma}+m\Big)\,1_{m(1-\sigma)\leq|r_i-r_j|\leq m}+\Big(\frac{|r_i-r_j|-m}{\nu}+m\Big)\,1_{|r_i-r_j|>m}\Big\}$$
$$+\,n(n-1)\log\{\sigma[1-\exp(-m)]+\nu\exp(-m)\}. \tag{B.1}$$

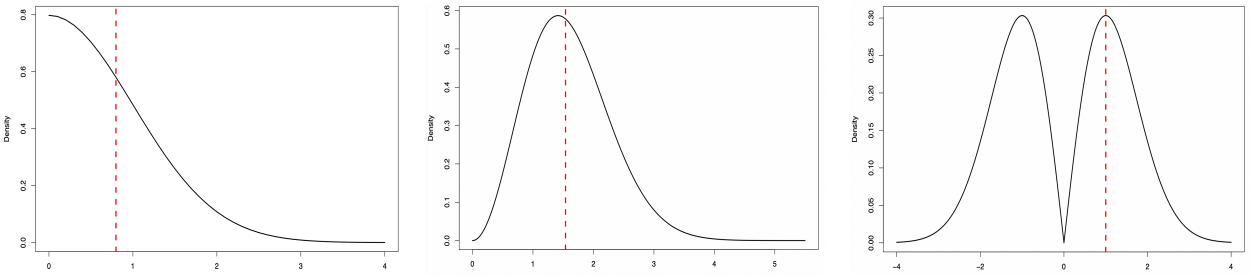
(B.1) results from applying skewed pivot-blend to the exponential density (instead of the double-exponential density). A regularization term (such as an ℓ_1 penalty) can be incorporated to capture structural parsimony.

Kernel-assisted nonparametric skew estimation. Assume $y-X\beta^*\sim \mathrm{SP}^{(\phi^*)}(\sigma^*,\nu^*,m^*)$, where the functional form of the density ϕ^* is also unknown. We can employ a backward-forward scheme for nonparametric estimation, along with *explicit* capture of skewness that aligns with the theme of this paper.

To estimate σ^*,ν^*,m^*, kernel may be employed to approximate ϕ^*, but the data follow $\mathrm{SP}^{(\phi^*)}$. We can address this by using backward pivot-blend. Holding β,σ,ν,m constant for now, and referring to the definitions of $\tilde{r}_i$ $(1\leq i\leq n)$, $L(m)$, and $R(m)$ in Remark 2, introduce a kernel density estimator with appropriate weights to estimate ϕ^* based on the transformed residuals $\tilde{r}_i$:

$$\phi_K(t;h)=\frac{1}{h}\sum_{i=1}^{n}\Big\{\frac{\nu}{L(m)\nu+R(m)\sigma}K_h\big(\frac{r_i-m}{\sigma}+m-t\big)1_{r_i\leq m}$$
$$+\frac{\sigma}{L(m)\nu+R(m)\sigma}K_h\big(\frac{r_i-m}{\nu}+m-t\big)1_{r_i>m}\Big\}. \tag{B.2}$$

where $K_h(t)=K\big(t/h\big)$, the kernel function $K(t)$ can be any continuous symmetric function with $\int_{-\infty}^{\infty}K(t),dt=1$, and $h>0$ represents the bandwidth parameter. The kernel approach in conjunction with (forward) skewed pivot-blend gives rise to a nonparametric skew-estimation problem, which warrants further investigation.



(a) ϕ: half-normal, m^*: mean (b) ϕ: Maxwell, m^*: median (c) ϕ: double Rayleigh, m^*: right mode

Figure C.1: Plots of ϕ for **Ex 3**–**Ex 5**. The choices of m^* are indicated in red dashed lines.

C More Experiments

In this part, we consider some densities which may not be unimodal or symmetric (as shown in Figure C.1). The experimental setup is the same as in Section 5.1.

Ex 3. (Half-normal with the mean as the pivotal point): Let ϕ be the half-normal density (with scale 1). We set $n = 700$, $\kappa = 0.1$, m^* as the mean, $\nu^* = 0.1$ and $\sigma^* = 0.2, 0.3, 0.4$.

Ex 4. (Maxwell-Boltzmann with the median as the pivotal point): Here, ϕ is the Maxwell-Boltzmann density which is widely used in statistical mechanics (Huang, 2008). We set $n = 500$, $\kappa = 0.2$, m^* as the median, $\nu^* = 0.1$ and $\sigma^* = 0.4, 0.5, 0.6$.

Ex 5. (Double Rayleigh with m^* as the right mode): Let ϕ be the double Rayleigh density (with scale 1), $n = 500$, $\kappa = 0.1$, m^* be the right mode, $\nu^* = 0.1$ and $\sigma^* = 0.2, 0.3, 0.4$.

Table C.1 illustrates significant issues with conventional methods in capturing skewness, even when asymmetry is centered around the median, mean, or mode. Specifically, QR* and BQR* failed to accurately recover the true β^* despite using the true quantile as input. Moreover, in all three cases, AME or AME provided at least one misleading estimate of the scales. In contrast, our proposed method demonstrated remarkable performance, with the associated β-error being at most 1/3 of that of the other methods.

We also conducted experiments on the air pollution data of Leeds from 1994 to 1998 (Heffernan and Tawn, 2004; Southworth et al., 2020). The dataset contain 578 measurements of the daily maximum levels of ozone (O3), nitrogen dioxide (NO2), nitrogen oxide (NO), sulfur dioxide (SO2) and particulate matter (PM10). We forecasted PM10 levels for the summer months (April to July) using other air pollutants. We considered the Gumbel model and its skewed pivot-blend enhancement SPEUS, in addition to ZQR, AME, and ESN. The p-values from the Kolmogorov-Smirnov tests for these models were 8e-2, 0.71, 9e-2, 0.20, and 8e-8, respectively. The results may appear surprising, considering the popularity of the skewed Gumbel distribution for modeling such extreme values (Boldi and Davison, 2007). Figure C.2 presents the residual Q-Q plots, demonstrating that our skew-reinforced method significantly improves data fit compared to the classical Gumbel model.

In real-world applications, data scientists frequently confront challenges related to skewness. However, fitting a common distribution that inadequately addresses these distortions can lead to subpar model fits and misleading inferences. The skewed pivot-blend technique refines skewness management in these distributions, thus enhancing the accuracy and relia-

	Skewed half-normal								
	$\sigma^*/\nu^* = 2$			$\sigma^*/\nu^* = 3$			$\sigma^*/\nu^* = 4$		
	Err(β)	Err(σ)	Err(ν)	Err(β)	Err(σ)	Err(ν)	Err(β)	Err(σ)	Err(ν)
QR*	0.80	—	—	0.80	—	—	0.80	—	—
BQR*	0.80	0.49	0.62	0.80	0.51	0.64	0.80	0.53	0.65
AME	0.64	0.99	1.22	0.56	1.0	2.02	0.48	1.0	2.82
ZQR	0.65	0.97	0.09	0.59	0.95	0.37	0.52	0.95	0.67
SPEUS	0.14	0.11	0.1	0.08	0.08	0.13	0.06	0.09	0.17
	Skewed Maxwell-Boltzmann								
	$\sigma^*/\nu^* = 4$			$\sigma^*/\nu^* = 5$			$\sigma^*/\nu^* = 6$		
	Err(β)	Err(σ)	Err(ν)	Err(β)	Err(σ)	Err(ν)	Err(β)	Err(σ)	Err(ν)
QR*	1.59	—	—	1.59	—	—	1.59	—	—
BQR*	1.59	0.41	0.74	1.59	0.43	0.75	1.59	0.44	0.75
AME	1.47	0.22	0.78	1.46	0.19	0.84	1.46	0.17	0.91
ZQR	1.49	0.58	0.30	1.48	0.58	0.32	1.48	0.57	0.34
SPEUS	0.18	0.14	0.20	0.17	0.13	0.21	0.16	0.12	0.30
	Skewed double Rayleigh								
	$\sigma^*/\nu^* = 2$			$\sigma^*/\nu^* = 3$			$\sigma^*/\nu^* = 4$		
	Err(β)	Err(σ)	Err(ν)	Err(β)	Err(σ)	Err(ν)	Err(β)	Err(σ)	Err(ν)
QR*	1.0	—	—	1.0	—	—	1.0	—	—
BQR*	1.0	0.68	0.27	1.0	0.70	0.26	1.0	0.71	0.26
AME	0.87	1.78	1.76	0.88	2.02	1.9	0.89	2.15	1.92
ZQR	0.67	0.50	1.67	0.50	0.48	2.69	0.32	0.45	3.78
SPEUS	0.22	0.03	0.18	0.15	0.03	0.24	0.10	0.03	0.26

Table C.1: Performance comparison for skewed half-normal with pivotal point at the mean, skewed Maxwell with pivotal point at the median, and skewed double Rayleigh with pivotal point at the right mode (Ex 3–Ex 5)

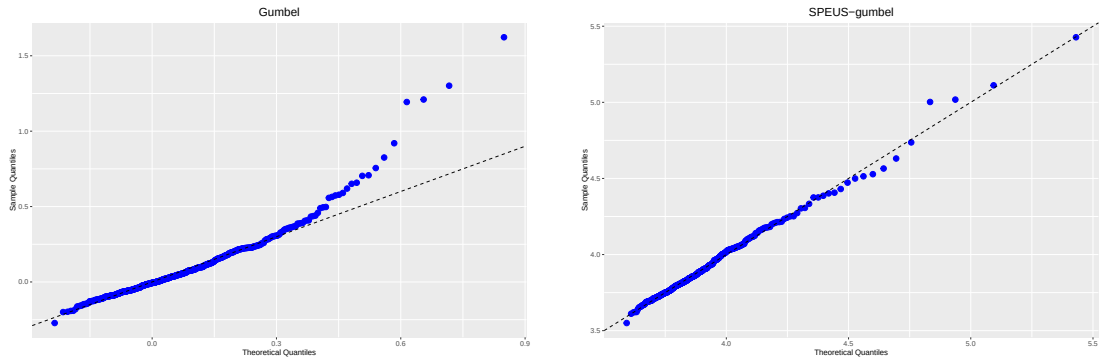


Figure C.2: Q-Q plots of residuals for Gumbel (left) and SPEUS (right) on air pollution data.

bility of the models.

References

Bartlett, P. L., Harvey, N., Liaw, C., and Mehrabian, A. (2019). Nearly-tight VC-dimension and pseudo-dimension bounds for piecewise linear neural networks. *The Journal of Machine Learning Research*, 20(1):2285–2301.

Boldi, M. O. and Davison, A. C. (2007). A mixture model for multivariate extremes. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69(2):217–229.

Devroye, L., Györfi, L., and Lugosi, G. (1996). *A Probabilistic Theory of Pattern Recognition*. New York: Springer.

Giné, E. and Nickl, R. (2015). *Mathematical Foundations of Infinite-Dimensional Statistical Models*. Cambridge: Cambridge University Press.

Goldberg, P. W. and Jerrum, M. R. (1995). Bounding the Vapnik-Chervonenkis dimension of concept classes parameterized by real numbers. *Machine Learning*, 18(2-3):131–148.

Götze, F., Sambale, H., and Sinulis, A. (2021). Concentration inequalities for polynomials in α-sub-exponential random variables. *Electronic Journal of Probability*, 26:1–22.

Hampel, F. R., Ronchetti, E. M., Rousseeuw, P. J., and Stahel, W. A. (2011). *Robust Statistics: The Approach Based on Influence Functions*. New Jersey: John Wiley & Sons.

Heffernan, J. E. and Tawn, J. A. (2004). A conditional approach for multivariate extreme values. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 66(3):497–546.

Hettmansperger, T. P. and McKean, J. W. (1978). Statistical inference based on ranks. *Psychometrika*, 43(1):69–79.

Hettmansperger, T. P. and McKean, J. W. (2010). *Robust Nonparametric Statistical Methods*. Florida: CRC Press.

Huang, K. (2008). *Statistical Mechanics*. New Jersey: John Wiley & Sons.

Jaeckel, L. A. (1972). Estimating regression coefficients by minimizing the dispersion of the residuals. *The Annals of Mathematical Statistics*, 43(5):1449–1458.

Kuchibhotla, A. K. and Chakrabortty, A. (2022). Moving beyond sub-Gaussianity in high-dimensional statistics: Applications in covariance estimation and linear regression. *Information and Inference: A Journal of the IMA*, 11:1389–1456.

Sambale, H. (2020). Some notes on concentration for α-subexponential random variables. *arXiv preprint:2002.10761*.

She, Y. (2016). On the finite-sample analysis of Θ-estimators. *Electronic Journal of Statistics*, 10:1874–1895.

Southworth, H., Heffernan, J. E., and Metcalfe, P. (2020). Package 'texmex'. https://cran.r-project.org/package=texmex.

van der Vaart, A. W. and Wellner, J. A. (2013). *Weak Convergence and Empirical Processes: With Applications to Statistics*. Berlin: Springer Science & Business Media.

Wainwright, M. J. (2019). *High-dimensional Statistics: A Non-Asymptotic Viewpoint*. Cambridge: Cambridge University Press.

Wang, L., Peng, B., Bradic, J., Li, R., and Wu, Y. (2020). A tuning-free robust and efficient approach to high-dimensional regression. *Journal of the American Statistical Association*, 115(532):1700–1714.