Appendix

The appendix is organized as follows. In Appendix A, we provide an introduction to some fundamental probability tools that are utilized in our proofs. Specifically, we discuss sub-Gaussian and sub-exponential distributions in Appendix A.1, and present Bernstein-type inequalities in Appendix A.2. In Appendix B, we summarize the properties of the multi-class logistic regression model that are needed in our proofs. Specifically, in Appendix B.2, we present the generalized linear model formulation of the multi-class logistic model and in Appendix B.1, we discuss the gradient and Hessian of the loss function. In Appendix B.3, we introduce pseudo self-concordant functions. In Appendix C, we present a thorough proof of one of our fundamental results, specifically Theorem 3. In Appendix D, we delve into the properties of some essential constants utilized in constructing the results of Theorem 3. In Appendix E, we provide the excess risk bounds for the case of $p(x)$ having bounded support. The proofs of the main results of Section 4 are provided in Appendix F. Finally, in Appendix G, we provide more details of our numerical experiments.

A Probability tools

A.1 Sub-Gaussian and sub-exponential distributions

Definition 11 (Sub-Gaussian random variable)**.** *A random variable x is sub-Gaussian if there exists $c_1 > 0$ such that $\mathbb{P}(|x| > t) \leq \exp(1 - t^2/c_1^2)$ for all $t \geq 0$.*

Lemma 12 (Proposition [22] in [23])**.** *Let x be a sub-Gaussian random variable. Then the following properties are equivalent, with parameters $c_i > 0$:*

(1) $\mathbb{P}(|x| > t) \leq \exp(1 - t^2/c_1^2)$, *for all* $t \geq 0$.

(2) $(\mathbb{E}\,|x|^p)^{1/p} \leq c_2\sqrt{p}$, *for all* $p \geq 1$.

(3) $\mathbb{E}\exp(x^2/c_3^2) \leq 2$.

Definition 13 (Sub-Gaussian norm)**.** *Let x a sub-Gaussian random variable. The sub-Gaussian norm of x, denoted $\|x\|_{\psi_2}$, is defined as follows:*

$$\|x\|_{\psi_2} \triangleq \inf\{t > 0 : \mathbb{E}\exp(x^2/t^2) \leq 2\}.$$

Lemma 14 (Sub-exponential random variable)**.** *Let x be a random variable. We say that x is sub-exponential if there exists $c_i > 0$ for which one of following properties is true. Furthermore, these properties are equivalent.*

(1) $\mathbb{P}(|x| > t) \leq \exp(1 - t/c_1)$ *for all* $t \geq 0$.

(2) $(\mathbb{E}\,|x|^p)^{1/p} \leq c_2 p$ *for all* $p \geq 1$.

(3) $\mathbb{E}\exp(|x|/c_3) \leq 2$.

Definition 15 (Sub-exponential norm)**.** *The sub-exponential norm of x, denoted $\|x\|_{\psi_1}$, is defined as follows:*

$$\|x\|_{\psi_1} \triangleq \inf\{t > 0 : \mathbb{E}\exp(|x|/t) \leq 2\}.$$

Lemma 16 (Sub-exponential is sub-Gaussian squared, Lemma 2.7.6 in [23])**.** *A random variable x is sub-Gaussian if and only if x^2 is sub-exponential. Moreover,*

$$\|x^2\|_{\psi_1} = \|x\|_{\psi_2}^2.$$

Definition 17 (Sub-Gaussian random vectors)**.** *A random vector $Z \in \mathbb{R}^d$ is sub-Gaussian if $\langle Z, u\rangle$ is sub-Gaussian for all $u \in \mathbb{R}^d$, with $\|u\|_2 = 1$. The sub-Gaussian norm of Z is defined as*

$$\|Z\|_{\psi_2} \triangleq \sup_{u\in\mathcal{S}^{d-1}} \|\langle Z, u\rangle\|_{\psi_2}.$$

Lemma 18. *Let $Z_1, \cdots, Z_n$ be independent centered sub-Gaussian random vectors, then $\|\sum_{i=1}^n Z_i\|_{\psi_2}^2 \lesssim \sum_{i=1}^n \|Z_i\|_{\psi_2}^2$.*

Lemma 19 (Affine transformation of sub-Gaussian vectors, Lemma A.5 in [24])**.** *Let* $X \in \mathbb{R}^d$ *such that* $\mathbb{E}[X] = 0$*,* $\mathbf{\Sigma} := \mathbb{E}[XX^\top]$ *and* $\|\mathbf{\Sigma}^{-1/2}X\|_{\psi_2} \le K$*. Then for any* $\mathbf{A} \in \mathbb{R}^{d\times d}$ *and* $b \in \mathbb{R}^d$*,* $\widehat{X} = \mathbf{A}X + b$ *satisfies*

$$\|\widehat{\mathbf{\Sigma}}^{-1/2}\widehat{X}\|_{\psi_2} \lesssim K, \quad \textit{where} \quad \widehat{\mathbf{\Sigma}} = \mathbb{E}[\widehat{X}\widehat{X}^\top].$$

The following lemma gives a high probability bound for the quadratic form $\|x\|^2_{\mathbf{\Sigma}^{-1}}$ of a non-centered sub-Gaussian vector x, where $\mathbf{\Sigma}$ is the covariance of x. The result can be viewed as a corollary of Theorem 2.1 in [25].

Lemma 20 (Tail inequalities for quadratic form of sub-Gaussian vectors)**.** *Let* $\mathbf{J} \in \mathbb{R}^{d\times d}$ *be a symmetric, positive semi-definite matrix. For any* $\delta \in (0,1)$ *the following is true:*

(1) If $x \in \mathbb{R}^d$ *is a zero-centered sub-Gaussian random vector, i.e.* $\mathbb{E}[x] = 0$ *and there exits* $K > 0$ *such that* $\|x\|_{\psi_2} \le K$*. Then we have with probability at least* $1-\delta$*,*

$$\|x\|^2_{\mathbf{J}} \lesssim K^2\big(\operatorname{Trace}(\mathbf{J}) + \sqrt{d}\|\mathbf{J}\|\log(e/\delta)\big). \tag{25}$$

(2) If $x \in \mathbb{R}^d$ *is a sub-Gaussian random vector with* $\|\mathbf{\Sigma}^{-1/2}x\|_{\psi_2} \le K$*, where* $\mathbf{\Sigma} = \mathbb{E}[xx^T]$*. Then with probability at least* $1-\delta$*,*

$$\|x\|^2_{\mathbf{\Sigma}^{-1}} \lesssim K^2\big(d + \sqrt{d}\log(e/\delta)\big). \tag{26}$$

Proof.

(1) By Theorem 2.1 in [25], we have for all $t > 0$,

$$\mathbb{P}\left[\|x\|^2_{\mathbf{J}} > K^2\big(\operatorname{Trace}(\mathbf{J}) + 2\sqrt{\operatorname{Trace}(\mathbf{J}^2)t} + 2\|\mathbf{J}\|t\big)\right] \le \exp(-t). \tag{27}$$

Let $t = \log(1/\delta)$ in Eq. (27), since $\sqrt{\operatorname{Trace}(\mathbf{J}^2)} = \|\mathbf{J}\|_F \le \sqrt{d}\|\mathbf{J}\|$, we can get Eq. (25).

(2) Note that we can not directly derive Eq. (26) from Eq. (25) since x is not zero-mean. But we can shift x to an isotropic sub-Gaussian random vector. Indeed, let $\mu = \mathbb{E}[x]$ and $\mathbf{\Sigma}_0 = \mathbb{E}[(x-\mu)(x-\mu)^\top]$. Then $\mathbf{\Sigma}_0^{-1/2}(x-\mu)$ is centered isotropic random vector. By Lemma 19, affine transformation of sub-Gaussian random vectors are also sub-Gaussian, i.e. $\mathbf{\Sigma}_0^{-1/2}(x-\mu)$ is also sub-Gaussian and

$$\|\mathbf{\Sigma}_0^{-1/2}(x-\mu)\|_{\psi_2} \lesssim K. \tag{28}$$

Denote $\mathbf{J} = \mathbf{\Sigma}_0^{1/2}\mathbf{\Sigma}^{-1}\mathbf{\Sigma}_0^{1/2}$. By Sherman–Morrison formula, we have

$$\mathbf{\Sigma}^{-1} = (\mathbf{\Sigma}_0 + \mu\mu^\top)^{-1} = \mathbf{\Sigma}_0^{-1} - \frac{\mathbf{\Sigma}_0^{-1}\mu\mu^\top\mathbf{\Sigma}_0^{-1}}{1+\mu^\top\mathbf{\Sigma}_0^{-1}\mu}, \tag{29}$$

and thus

$$\|\mathbf{J}\|_\infty \le 1, \tag{30}$$

$$\|\mathbf{J}\|_2 = \left\|\mathbf{I}_d - \frac{(\mathbf{\Sigma}_0^{-1/2}\mu)(\mathbf{\Sigma}_0^{-1/2}\mu)^\top}{1+\|\mathbf{\Sigma}_0^{-1/2}\mu\|_2^2}\right\|_2 \le \|\mathbf{I}_d\|_2 + \frac{\|\mathbf{\Sigma}_0^{-1/2}\mu\|_2^2}{1+\|\mathbf{\Sigma}_0^{-1/2}\mu\|_2^2} \le 2, \tag{31}$$

$$\operatorname{Trace}(\mathbf{J}) = \langle\mathbf{\Sigma}^{-1}, \mathbf{\Sigma}_0\rangle = \operatorname{Trace}(\mathbf{I}_d) - \frac{\mu^\top\mathbf{\Sigma}_0^{-1}\mu}{1+\mu^\top\mathbf{\Sigma}_0^{-1}\mu} \le d. \tag{32}$$

By Eq. (25), we have with probability at least $1-\delta$,

$$\begin{aligned}\|x-\mu\|^2_{\mathbf{\Sigma}^{-1}} = \|\mathbf{\Sigma}_0^{-1/2}(x-\mu)\|^2_{\mathbf{J}} &\lesssim \operatorname{Trace}(\mathbf{J}) + K^2(\|\mathbf{J}\|_2\sqrt{\log(1/\delta)} + \|\mathbf{J}\|_\infty\log(1/\delta)) \\ &\lesssim K^2\Big(d + \sqrt{d}\log(e/\delta)\Big).\end{aligned} \tag{33}$$

In addition, by Eq. (29),

$$\|\mu\|^2_{\mathbf{\Sigma}^{-1}} = \mu^\top\mathbf{\Sigma}^{-1}\mu = \mu^\top\mathbf{\Sigma}_0^{-1}\mu - \frac{(\mu^\top\mathbf{\Sigma}_0^{-1}\mu)^2}{1+\mu^\top\mathbf{\Sigma}_0^{-1}\mu} = \frac{\mu^\top\mathbf{\Sigma}_0^{-1}\mu}{1+\mu^\top\mathbf{\Sigma}_0^{-1}\mu} \le 1. \tag{34}$$

Combining Eqs. (33) and (34), we obtain

$$\|x\|^2_{\mathbf{\Sigma}^{-1}} \le (\|x-\mu\|_{\mathbf{\Sigma}^{-1}} + \|\mu\|_{\mathbf{\Sigma}^{-1}})^2 \lesssim K^2\Big(d + \sqrt{d}\log(e/\delta)\Big). \tag{35}$$

□

A.2 Bernstein-type inequalities

We give Bernstein-type inequalities for vectors and matrices in the following lemmas. These properties are used in the proof of excess risk bounds in the bounded domain case (Appendix E).

Lemma 21 (Vector Bernstein inequality; see Theorem 18 in [26])**.** *Let $x_1, x_2, \cdots, x_n$ be independent random vectors such that*

$$\mathbb{E}[x_i] = 0, \quad \|x_i\|_2 \le \mu \quad \textit{and } \mathbb{E}[\|x_i\|_2^2] \le \nu, \qquad \forall i \in [n].$$

Let $S = \frac{1}{n}\sum_{i=1}^n x_i$. Then if $0 < \epsilon < \nu/\mu$,

$$\mathbb{P}[\|S\|_2 \ge \epsilon] \le \exp\big(-\frac{n\epsilon^2}{8\nu} + \frac{1}{4}\big). \tag{36}$$

Lemma 22 (Matrix Bernstein inequality; see Theorem 19 in [26])**.** *Let $\mathbf{X}_1, \mathbf{X}_2, \cdots, \mathbf{X}_n$ be independent random Hermitian matrices with common dimension $d \times d$ such that*

$$\mathbb{E}[\mathbf{X}_i] = 0, \quad \|\mathbf{X}_i\|_2 \le \mu \quad \textit{and } \mathbb{E}[\|\mathbf{X}_i\|_2^2] \le \nu, \qquad \forall i \in [n].$$

Let $\mathbf{S} = \frac{1}{n}\sum_{i=1}^n \mathbf{X}_i$. Then if $0 < \epsilon < 2\nu/\mu$,

$$\mathbb{P}[\|\mathbf{S}\|_2 \ge \epsilon] \le 2d \cdot \exp\big(-\frac{n\epsilon^2}{4\nu}\big). \tag{37}$$

B Multi-class logistic regression and pseudo self-concordance

In Appendix B.1, we present some properties of the gradient and Hessian of $\ell_{(x,y)}(\theta)$ with respect to θ. In Appendix B.2, we show that the multi-class logistic regression model is a Generalized Linear Model. Then we present some properties related with the pseudo-concordance in Appendix B.3.

Notation. Given $y \in [c]$ and $\eta \in \mathbb{R}^{c-1}$, we define the loss function $\ell(y, \eta)$ by

$$\ell(y,\eta) \triangleq \begin{cases} -\log\big(\frac{\exp(\eta_y)}{1+\sum_{l\in[c-1]}\exp(\eta_l)}\big), & y \in [c-1] \\ -\log\big(\frac{1}{1+\sum_{l\in[c-1]}\exp(\eta_l)}\big), & y = c. \end{cases} \tag{38}$$

where η_y is the y-th component of η. Note that given $x \in \mathbb{R}^d$, $y \in [c]$ and $\theta \in \mathbb{R}^{(c-1)\times d}$, if we let $\eta = \theta x$, then

$$\ell(y,\eta) = \ell_{(x,y)}(\theta),$$

where $\ell_{(x,y)} \triangleq -\log p(y|x,\theta)$ (Eq. (1)).

To differentiate the derivatives with respect to η and θ, we use $\ell'(y,\eta)$ to represent the gradient of the loss with respect to η, and $\nabla \ell_{(x,y)}(\theta)$ to represent the gradient of the loss with respect to θ. Similar notations hold for higher order derivatives.

B.1 Properties of multi-class logistic regression

We present the expressions of gradient and Hessian of the loss function $\ell_{(x,y)}(\theta)$ with respect to θ in the following proposition.

Proposition 23. *Given a sample point $x \in \mathbb{R}^d$, its label $y \in [c]$, and parameter $\theta \in \mathbb{R}^{(c-1)\times d}$ in the multiclass logistic regression model. We consider the negative log-likelihood loss $\ell_{(x,y)}(\theta) = -\log p(y|x,\theta)$, where $p(y|x,\theta)$ is defined in Eq. (1). Let $\widetilde{c} \triangleq c-1$, $\widetilde{d} \triangleq d(c-1)$, θ_i be the i-th row of θ. Define vector $\mathbf{h}(x,\theta) \in \mathbb{R}^{\widetilde{c}}$ by*

$$\mathbf{h}_i(x,\theta) = p(y = i|x,\theta) = \frac{\exp(x^\top\theta_i)}{1+\sum_{s\in[\widetilde{c}]}\exp(x^\top\theta_s)}, \qquad \forall i \in [\widetilde{c}].$$

Then the gradient and Hessian of $\ell_{(x,y)}(\theta)$ w.r.t θ can be expressed in the following ways:

(1) Gradient $\nabla \ell_{x,y}(\theta) \in \mathbb{R}^{\widetilde{c} \times d}$ *is given by*

$$\nabla \ell_{(x,y)}(\theta) = \begin{bmatrix} \beta_1(y,x,\theta)x^\top \\ \cdots \\ \beta_{\widetilde{c}}(y,x,\theta)x^\top \end{bmatrix}, \tag{39}$$

where $\beta_i(x,y,\theta) = -1_{\{y=i\}} + \mathbf{h}_i(x,\theta)$.

(2) Hessian $\nabla^2 \ell_{(x,y)}(\theta) \in \mathbb{R}^{\widetilde{d} \times \widetilde{d}}$ *is given by*

$$\begin{aligned} \nabla^2 \ell_{(x,y)}(\theta) &= \Big(\mathrm{diag}(\mathbf{h}(x,\theta)) - \mathbf{h}(x,\theta)\mathbf{h}(x,\theta)^\top\Big) \otimes (xx^\top) \\ &= \begin{bmatrix} \alpha_{11}(x,\theta)xx^\top & \cdots & \alpha_{1\widetilde{c}}(x,\theta)xx^\top \\ \vdots & \ddots & \vdots \\ \alpha_{\widetilde{c}1}(x,\theta)xx^\top & \cdots & \alpha_{\widetilde{c}\widetilde{c}}(x,\theta)xx^\top \end{bmatrix}, \end{aligned} \tag{40}$$

where $\alpha_{i,j}(\theta) = 1_{\{i=j\}}\mathbf{h}_i(x,\theta) - \mathbf{h}_i(x,\theta)\mathbf{h}_j(x,\theta)$.

Lemma 24. *Given a point* $x \in \mathbb{R}^d$, $\mathbb{E}_{y\sim p(y|x,\theta_*)}[\nabla \ell_{(x,y)}(\theta_*)] = 0$. *In addition, let* $p(x)$ *be a point distribution and* $L_p(\theta)$ *be the expected loss at* θ*, then*

$$\nabla L_p(\theta_*) = 0. \tag{41}$$

Proof. Since $\nabla \ell_{(x,y)}(\theta) = -\nabla_\theta \log p(y|x,\theta)$, we have

$$\begin{aligned} \mathbb{E}_{y\sim p(y|x,\theta_*)}[\nabla \ell_{(x,y)}(\theta_*)] &= -\sum_{k\in[c]} p(y=k|x,\theta_*)\nabla_\theta \log p(y=k|x,\theta_*) \\ &= -\sum_{k\in[c]} p(y=k|x,\theta_*)\frac{\nabla_\theta p(y=k|x,\theta_*)}{p(y=k|x,\theta_*)} \\ &= -\nabla_\theta\Big(\sum_{k\in[c]} p(y=k|x,\theta_*)\Big) = -\nabla_\theta 1 = 0. \end{aligned} \tag{42}$$

Thus,

$$\nabla_\theta\big(\mathbb{E}_{y\sim p(y|x,\theta_*)}[\ell_{(x,y)}(\theta)]\big)\big|_{\theta=\theta_*} = \mathbb{E}_{y\sim p(y|x,\theta_*)}[\nabla \ell_{(x,y)}(\theta_*)] = 0. \tag{43}$$

Since $\nabla L_p(\theta) = \nabla_\theta \int p(x)\,\mathbb{E}_{y\sim p(y|x,\theta_*)}[\ell_(x,y)(\theta)]dx = \int p(x)\nabla_\theta\,\mathbb{E}_{y\sim p(y|x,\theta_*)}[\ell_(x,y)(\theta)]dx$, by Eq. (43), we have

$$\nabla L_p(\theta_*) = \int p(x)\nabla_\theta\big(\mathbb{E}_{y\sim p(y|x,\theta_*)}[\ell_{(x,y)}(\theta)]\big)\big|_{\theta=\theta_*}dx = 0. \tag{44}$$

□

The following lemma is a basic property for Fisher information matrix.

Lemma 25. *The Fisher information matrix for a point* x *at parameter* θ *is defined by* $\mathbb{E}_{y\sim p(y|x,\theta_*)}[\nabla \ell_{(x,y)}(\theta)(\nabla \ell_{(x,y)}(\theta))^\top]$*, then*

$$\mathbb{E}_{y\sim p(y|x,\theta_*)}[\nabla \ell_{(x,y)}(\theta_*)(\nabla \ell_{(x,y)}(\theta_*))^\top] = \mathbb{E}_{y\sim p(y|x,\theta_*)}[\nabla^2 \ell_{(x,y)}(\theta_*)]. \tag{45}$$

Proof.

$$\begin{aligned} \nabla^2 \ell_{(x,y)}(\theta_*) &= -\frac{\nabla^2 p(y|x,\theta_*))}{p(y|x,\theta_*)} + \frac{\nabla p(y|x,\theta_*)\nabla p(y|x,\theta_*)^\top}{p(y|x,\theta_*)^2} \\ &= -\frac{\nabla^2 p(y|x,\theta_*))}{p(y|x,\theta_*)} + \nabla \ell_{(x,y)}(\theta_*)(\nabla \ell_{(x,y)}(\theta_*))^\top \end{aligned}$$

Thus,

$$\mathbb{E}_{y\sim p(y|x,\theta_*)}[\nabla \ell_{(x,y)}(\theta_*)(\nabla \ell_{(x,y)}(\theta_*))^\top]$$

$$
\begin{aligned}
&= \mathbb{E}_{y\sim p(y|x,\theta_*)}[\nabla^2 \ell_{(x,y)}(\theta_*)] + \mathbb{E}_{y\sim p(y|x,\theta_*)}\left[\frac{\nabla^2 p(y|x,\theta_*))}{p(y|x,\theta_*)}\right] \\
&= \mathbb{E}_{y\sim p(y|x,\theta_*)}[\nabla^2 \ell_{(x,y)}(\theta_*)] + \int p(y|x,\theta_*)\frac{\nabla^2 p(y|x,\theta_*))}{p(y|x,\theta_*)} d\sigma \\
&= \mathbb{E}_{y\sim p(y|x,\theta_*)}[\nabla^2 \ell_{(x,y)}(\theta_*)] + \nabla^2 \int p(y|x,\theta_*) d\sigma = \mathbb{E}_{y\sim p(y|x,\theta_*)}[\nabla^2 \ell_{(x,y)}(\theta_*)].
\end{aligned}
$$

□

B.2 Multi-class logistic regression as a Generalized Linear Model (GLM)

Definition 26 (Exponential family model)**.** *Suppose μ is a base measure on space $\mathcal{Y}$ and there exists a sufficient statistic $T : \mathcal{Y} \to \mathbb{R}^c$. Then the exponential family associated with the function $T(y)$ and measure μ is defined as the set of distributions with densities $p(y|\eta)$ w.r.t μ, where*

$$p(y|\eta) = \exp(\langle \eta, T(y)\rangle - A(\eta)) \tag{46}$$

and $a(\eta)$ is the cumulant function defined by

$$A(\eta) \triangleq \log \int_{\mathcal{Y}} \exp(\langle \eta, T(y)\rangle) d\mu(y) \tag{47}$$

whenever a is finite.

Definition 27 (Generalized linear model with canonical response function)**.** *Generalized linear model with canonical response function is a model assuming that:*

1. *the input $x \in \mathbb{R}^d$ enter into the model via a linear combination $\eta = \theta x$,*
2. *the output y is characterized by an exponential family distribution (Definition 26).*

In the following lemma, we remark that the multi-class logistic regression model is a generalized linear model. The proof is trivial.

Lemma 28. *Multi-class logistic regression is a generalized linear model with canonical response function with η, $A(\eta)$ and $T(y)$ defined as the followings:*

$$\eta = [\log(\mathbf{h}_1/\mathbf{h}_c), \log(\mathbf{h}_2/\mathbf{h}_c), \cdots, \log(\mathbf{h}_{c-1}/\mathbf{h}_c)]^\top \tag{48}$$

$$A(\eta) = -\log \mathbf{h}_c \tag{49}$$

$$T(1) = [1, 0, \cdots, 0]^\top, \quad \cdots, \quad T(c-1) = [0, \cdots, 1]^\top, \quad T(c) = [0, \cdots, 0]^\top, \tag{50}$$

where $\mathbf{h}_i = p(y = i|x, \theta)$ ($p(y|x, \theta)$ is defined in Eq. (1)).

B.3 Pseudo self-concordance

Lemma 29 (pseudo self-concordance of multi-class logistic regression model)**.** *$\ell(y, \eta)$ is pseudo self-concordant, i.e.*

$$\forall h \in \mathbb{R}^{c-1}, \qquad |\ell'''(y,\eta)[h,h,h]| \le 2\|h\|_\infty \ell''(y,\eta)[h,h]. \tag{51}$$

Proof. By Lemma 28 and Equation (46),

$$\ell(y,\eta) = -\log p(y,\eta) = -\langle \eta, T(y)\rangle + A(\eta).$$

From theory of the exponential family distributions, we have

$$A'(\eta) = \mathbb{E}_\eta[T(y)], \quad A''(\eta) = \mathbb{E}_\eta[(T(y) - \mathbb{E}_\eta[T(y)])^{\otimes 2}], \quad A'''(\eta) = \mathbb{E}_\eta[(T(y) - \mathbb{E}_\eta[T(y)])^{\otimes 3}]. \tag{52}$$

where we denote the pth order tensor for a vector x as

$$x^{\otimes p} = \underbrace{x \otimes x \otimes \cdots \otimes x}_{p \text{ times}}.$$

Note that $\ell^{(p)}(y,\eta) = A^{(p)}(\eta)$ whenever $p \geq 2$, then we have

$$\begin{aligned}
\left|\ell'''(y,\eta)[h,h,h]\right| &= \left|\,\mathbb{E}\left[(T(y)-\mathbb{E}_\eta[T(y)])^{\otimes 3}[h,h,h]\right]\right| \\
&= \left|\,\mathbb{E}\left[(T(y)-\mathbb{E}_\eta[T(y)])^{\otimes 2}[h,h]\langle T(y)-\mathbb{E}_\eta[T(y)],h\rangle\right]\right| \\
&\leq \sup_{y\in\mathcal{Y}}\left|\langle T(y)-\mathbb{E}_\eta[T(y)],h\rangle\right|\ell''(y,\eta)[h,h] \\
&\overset{(a)}{\leq} 2\sup_{y\in\mathcal{Y}}\|T(y)\|_1\|h\|_\infty\ell''(y,\eta)[h,h] \\
&\overset{(b)}{\leq} 2\|h\|_\infty\ell''(y,\eta)[h,h], \qquad (53)
\end{aligned}$$

where (a) follows by Cauchy-Schwarz inequality, triangle inequality, and $\|E_\eta[T(y)]\|_2 \leq E_\eta\|T(y)\|_2 \leq \sup_{y\in\mathcal{Y}}\|T(y)\|_2$, (b) follows by the fact that $\|T(y)\|_2 = 1$ for $y \neq c$ and $\|T(y)\|_2 = 0$ for $y = c$ (Lemma 28). ☐

The previous lemma states the pseudo self-concordance of $\ell(y,\eta)$ w.r.t η. The following proposition states that the empirical loss function is pseudo self-concordant w.r.t θ, which is a corollary of the previous lemma via chain rule.

Proposition 30. *For multi-class regression model, we fix $\theta_0,\theta_1 \in \mathbb{R}^{(c-1)\times d}$. Let $\theta_t = \theta_0 + t(\theta_1-\theta_0)$, we define $\phi_n(t)$ by*

$$\phi_n(t) == \frac{1}{n}\sum_{i=1}^n \ell_{(x_i,y_i)}(\theta_t). \qquad (54)$$

Then we have

$$|\phi_n'''(t)| \leq 2\phi_n''(t)\max_{i\in[n]}\|(\theta_1-\theta_0)x_i\|_\infty \qquad (55)$$

Proof. Denote $\Delta = \theta_1-\theta_0$, then $\theta_t = \theta_0 = t\Delta$. Following chain rule and the smoothness of ℓ, we obtain that the derivatives of $\phi(t)$ and $\phi_n(t)$ are given by

$$\phi_n^{(p)}(t) = \frac{1}{n}\sum_{i=1}^n \ell^{(p)}(y,\theta_t x)[\underbrace{\Delta x,\cdots,\Delta x}_{p\text{ times}}].$$

Applying Lemma 29, we can get

$$\begin{aligned}
|\phi'''(t)| &\leq \frac{1}{n}\sum_{i=1}^n\left|\ell'''(y_i,\theta_t x_i)[\Delta x_i,\Delta x_i,\Delta x_i]\right| \\
&\leq \frac{1}{n}\sum_{i=1}^n 2\|\Delta x\|_\infty\ell''(y_i,\theta_t x_i)[\Delta x_i,\Delta x_i] \\
&\leq 2\phi_n''(t)\max_{i\in[n]}\|(\theta_1-\theta_0)x_i\|_\infty.
\end{aligned}$$

☐

The following proposition forms the foundation of our proof of Theorem 3. It gives lower and upper bounds to perturbations of pseudo self-concordant function.

Proposition 31 (Proposition 1 in [27])**.** *Let $F:\Theta\to\mathbb{R}$ be a convex C^3-mapping. Fix $\theta_0,\theta_1\in\Theta$, let $\Delta = \theta_1-\theta_0$ and $\theta_t = \theta_0+t\Delta$ for $t\in\mathbb{R}$. Define function $\phi_F(t) = F(\theta_t)$. Assume that $\mathbf{H}_0 \triangleq \nabla^2F(\theta_0)\succ 0$, $|\phi_F'''(t)| \leq R\|\Delta\|_2\cdot\phi_F''(t)$ for some $R\geq 0$. Denote $S = R\|\Delta\|_2$, we have*

$$\frac{e^{-S}+S-1}{S^2}\|\Delta\|_{\mathbf{H}_0}^2 \leq F(\theta_1)-F(\theta_0)-\big(\nabla F(\theta_0)\big)^\top\Delta \leq \frac{e^S-S-1}{S^2}\|\Delta\|_{\mathbf{H}_0}^2, \qquad (56)$$

$$e^{-S}\mathbf{H}_0 \preceq \nabla^2F(\theta_1)\preceq e^S\mathbf{H}_0. \qquad (57)$$

C Proof of Theorem 3

We first give the detailed version of Theorem 3 in Appendix C.1. In Appendix C.2, we present a sketch of the proof for the excess risk bounds in Eq. (7). In Appendix C.3, we provide and prove a tail bound for a certain type of random matrices, which is useful in our full proof. Finally, we give the full proof of Theorem 3 (Theorem 32) in Appendix C.4.

Notation. For the ease of notation, we define the empirical risk over finite samples $Q_n(\theta)$ and its Hessian $\mathbf{H}_n(\theta)$ by

$$\theta_n \in \arg\min_{\theta} Q_n(\theta) \triangleq \frac{1}{n}\sum_{i\in[n]} \ell_{(x_i,y_i)}(\theta), \qquad (x_i,y_i) \overset{\text{i.i.d.}}{\sim} \pi_q(x,y), \tag{58}$$

$$\mathbf{H}_n(\theta) \triangleq \nabla^2 Q_n(\theta). \tag{59}$$

In addition, let $\vec{\mathbf{A}} \in \mathbb{R}^{mn}$ be the vectorization of a matrix $\mathbf{A} \in \mathbb{R}^{m\times n}$ by stacking all rows together, i.e. $\vec{\mathbf{A}} = (\mathbf{A}_1^\top, \cdots, \mathbf{A}_m^\top)^\top$ where $\mathbf{A}_i$ is i-th row of $\mathbf{A}$.

C.1 Detailed version of Theorem 3

Theorem 32. *Suppose Assumption 1 holds for both $p(x)$ and $q(x)$.Let σ, ρ and $\nu > 0$ be constants such that $\mathbf{H}_p(\theta_*) \preceq \sigma \mathbf{H}_q$, $\mathbf{I}_{c-1} \otimes \mathbf{V}_p \preceq \rho \mathbf{H}_p(\theta_*)$ and $\mathbf{V}_q \preceq \nu \mathbf{V}_p$ hold. Whenever*

$$n \gtrsim \max\left\{K_{2,q}^2(r)\widetilde{d}\log(ed/\delta),\ \sigma\rho\nu K_{0,q}^2 K_{1,q}^2 K_{2,q}^2(r)\left(\widetilde{d} + \sqrt{\widetilde{d}}\log(e/\delta)\right)\right\}, \tag{60}$$

where $\widetilde{d} \triangleq d(c-1)$, we have with probability at least $1-\delta$,

$$L_q(\theta_n) - L_q(\theta_*) \lesssim K_{1,q}^2 \frac{\widetilde{d} + \sqrt{\widetilde{d}}\log(e/\delta)}{n}, \tag{61}$$

$$\frac{e^{-\alpha}+\alpha-1}{\alpha^2}\,\frac{\mathbf{H}_q{}^{-1}\cdot\mathbf{H}_p}{n} \lesssim \mathbb{E}[L_p(\theta_n)] - L_p \lesssim \frac{e^{\alpha}-\alpha-1}{\alpha^2}\,\frac{\mathbf{H}_q{}^{-1}\cdot\mathbf{H}_p}{n}. \tag{62}$$

Here $\mathbf{H}_p = \mathbf{H}_p(\theta_*)$ and $\mathbf{H}_q = \mathbf{H}_q(\theta_*)$; and $\mathbb{E}$ is the expectation over $\{y_i \sim p(y_i|x_i,\theta_*)\}_{i=1}^n$. Furthermore,

$$\alpha = \mathcal{O}\left(\sqrt{\sigma\rho}K_{0,p}K_{1,q}K_{2,p}(r)\sqrt{\left(\widetilde{d}+\sqrt{\widetilde{d}}\log(e/\delta)\right)/n}\right). \tag{63}$$

C.2 Proof sketch of Eq.(7)

Here we present the basics of step 6 in the full proof of Theorem 3 (see Appendix C.4). Some details of this step are established in the steps 1-5 of the full proof.

Let $\theta_0 = \theta_*$, $\theta_1 = \theta_n$ and $\Delta \triangleq \theta_n - \theta_*$. Define $\phi_p(t) = L_p(\theta_* + t\Delta)$, we first prove that there exits $\alpha > 0$ s.t. $|\phi_p'''(t)| \le \alpha\phi_p''(t)$. Thus the premise of Proposition 31 is satisfied. By Eq. (56) and the fact that $\nabla L_p(\theta_*) = 0$ (Lemma 24), we have

$$\frac{e^{-\alpha}+\alpha-1}{\alpha^2}\|\vec{\Delta}\|_{\mathbf{H}_p}^2 \le L_p(\theta_n) - L_p(\theta_0) \le \frac{e^{\alpha}-\alpha-1}{\alpha^2}\|\vec{\Delta}\|_{\mathbf{H}_p}^2 \tag{64}$$

By Taylor theorem, there exists $\tilde{\theta}$ between θ_n and θ_* such that

$$\vec{\nabla}Q_n(\theta_*) = \vec{\nabla}Q_n(\theta_n) + \mathbf{H}_n(\tilde{\theta})\vec{\Delta} = \mathbf{H}_n(\tilde{\theta})\vec{\Delta}, \tag{65}$$

where the last equality follows by $\vec{\nabla}Q_n(\theta_n) = 0$ because the empirical loss Q_n is convex and θ_n is its solution. We can prove that if the sample bound Eq. (6) holds,

$$\mathbf{H}_n(\tilde{\theta}) \approx \mathbf{H}_q, \tag{66}$$

where "≈" means that there exits $a_1, a_2 > 0$ such that $a_1\mathbf{H}_q \preceq \mathbf{H}_n(\tilde{\theta}) \preceq a_2\mathbf{H}_q$. Thus we have

$$\|\vec{\Delta}\|_{\mathbf{H}_p}^2 = \vec{\Delta}^\top\mathbf{H}_p\vec{\Delta} \approx \vec{\nabla}Q_n(\theta_*)^\top\left(\mathbf{H}_q{}^{-1}\mathbf{H}_p\mathbf{H}_q{}^{-1}\right)\vec{\nabla}Q_n(\theta_*)$$

$$= \big\langle \mathbf{H}_q{}^{-1}\mathbf{H}_p\mathbf{H}_q{}^{-1}, \vec{\nabla}Q_n(\theta_*)\vec{\nabla}Q_n(\theta_*)^\top \big\rangle. \tag{67}$$

Then we prove that

$$\mathbb{E}_{\{y_i\sim p(y_i|x_i,\theta_*)\}_{i=1}^n}\left[\vec{\nabla}Q_n(\theta_*)\vec{\nabla}Q_n(\theta_*)^\top\right] = \frac{1}{n}\mathbf{H}_n(\theta_*) \approx \frac{1}{n}\mathbf{H}_q. \tag{68}$$

Substitute this into Eq. (67), we have

$$\mathbb{E}_{\{y_i\sim p(y_i|x_i,\theta_*)\}_{i=1}^n}[\|\vec{\Delta}\|^2_{\mathbf{H}_p}] \approx \frac{1}{n}\langle \mathbf{H}_q{}^{-1}, \mathbf{H}_p\rangle. \tag{69}$$

By taking expectation over Eq. (64) and using Eq. (69), we can get Eq. (7).

C.3 Supporting tools

In the following proposition, we present and prove a tail bound for the average sum of independent random matrices $\{\mathbf{A}_i\}_{i\in[n]}$ satisfying $\mathbb{E}[\mathbf{A}_i] = \mathbf{I}$ and Eq. (70).

Proposition 33. *Let $\mathbf{A}_1, \cdots, \mathbf{A}_n$ be $\widetilde{d}\times\widetilde{d}$ be independent symmetric matrices such that $\mathbb{E}[\mathbf{A}_i] = \mathbf{I}_{\widetilde{d}}$. There is constant $K > 0$ such that for any $i \in [n]$,*

$$\sup_{u\in\mathcal{S}^{\widetilde{d}-1}} \|u^\top \mathbf{A}_i u\|_{\psi_1} \le K, \tag{70}$$

where $\mathcal{S}^{\widetilde{d}-1}$ is the unit sphere in $\mathbb{R}^{\widetilde{d}}$, $\|\cdot\|_{\psi_1}$ is the norm for sub-exponential random variable (Definition 15). Define matrix $\mathbf{S}_n = \frac{1}{n}\sum_{i=1}^n \mathbf{A}_i$. Then for every $t \ge 0$, with probability at least $1 - 2\exp(-c_K t^2)$ we have

$$\|\mathbf{S}_n - \mathbf{I}_{\widetilde{d}}\| \le \max\{a, a^2\}, \qquad \textit{where } a = \frac{C_K\sqrt{\widetilde{d}} + t}{\sqrt{n}}. \tag{71}$$

Here c_K, C_k are constants that depend on K.

Proof. The proof follows a covering argument. We consider $1/4-$net $\mathcal{N}$ of the unit sphere $\mathcal{S}^{\widetilde{d}-1}$. By Lemma 5.2 in [22], $|\mathcal{N}| \le 9^{\widetilde{d}}$. Since $\mathbf{S}_n$ is symmetric, we can use Lemma 5.4 in [22] to bound matrix operator norm using points in $1/4-$net $\mathcal{N}$:

$$\|\mathbf{S}_n - \mathbf{I}_{\widetilde{d}}\| \le 2\max_{x\in\mathcal{N}}\left|\Big\langle(\mathbf{S}_n - \mathbf{I}_{\widetilde{d}})x, x\Big\rangle\right| = 2\max_{x\in\mathcal{N}}\left|x^\top\mathbf{S}_n x - 1\right|, \tag{72}$$

where the last equality follows by $\|x\|_2 = 1$ on $\mathcal{N}$. Thus it is sufficient to prove with the given probability,

$$2\max_{x\in\mathcal{N}}\left|x^\top\mathbf{S}_n x - 1\right| \le \max\{a, a^2\} \triangleq \epsilon. \tag{73}$$

Pick an arbitrary $x \in \mathcal{N}$, then

$$nx^\top\mathbf{S}_n x = \sum_{i=1}^n x^\top\mathbf{A}_i x \triangleq \sum_{i=1}^n Z_i^2, \tag{74}$$

where we define random variable $Z_i \triangleq x^\top\mathbf{A}_i x$. We have the following properties for Z_i:

$$\mathbb{E}[Z_i] = \mathbb{E}[x^\top\mathbf{A}_i x] = \langle x^\top, \mathbb{E}[\mathbf{A}_i]x\rangle = 1,$$
$$\|Z_i\|_{\psi_1} = \|x^\top\mathbf{A}_i x\|_{\psi_1} \overset{(a)}{\le} K,$$
$$\|Z_i - 1\|_{\psi_1} = \|Z_i - \mathbb{E}[Z_i]\|_{\psi_1} \overset{(b)}{\le} 2\|Z_i\|_{\psi_1} \le 2K,$$

where inequality (a) follows by Eq. (70), inequality (b) follows by Jensen's inequality.

Thus $Z_1 - 1, Z_2 - 1, \cdots, Z_n - 1$ are independent centered sub-exponential random variables. Using Corollary 5.17 in [22], we can get

$$\mathbb{P}\left(\left|x^\top\mathbf{S}_n x - 1\right| \ge \frac{\epsilon}{2}\right) = \mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^n(Z_i - 1)\right| \ge \frac{\epsilon}{2}\right) \le 2\exp[-\frac{c_1}{K^2}\min(\epsilon, \epsilon^2)n]$$

$$\leq 2\exp[-\frac{c_1}{K^2}a^2 n] \leq 2\exp[-\frac{c_1}{K^2}(C_K^2\widetilde{d}+t^2)]. \tag{75}$$

Take the union bound of all $x \in \mathcal{N}$, let

$$c_K = \frac{c_1}{K^2}, \qquad C_K = K\sqrt{\log 9/c_1}, \tag{76}$$

we have

$$\begin{aligned}\mathbb{P}\left(\max_{x\in\mathcal{N}}\left|x^\top \mathbf{S}_n x - 1\right| \geq \frac{\epsilon}{2}\right) &\leq 9^n \cdot 2\exp[-\frac{c_1}{K^2}(C_K^2\widetilde{d}+t^2)] \\ &\leq 2\exp\left[p\log 9 - d_1\log 9 - \frac{c_1 t^2}{K^2}\right] \\ &= 2\exp(-\frac{c_1 t^2}{K^2}) = 2\exp(-c_K t^2). \end{aligned} \tag{77}$$

As we noted in Eq. (73), this completes the proof. ☐

Corollary 34. *Under the premise of Proposition 33, whenever*

$$n \gtrsim K^2(\widetilde{d}+\log(1/\delta)), \tag{78}$$

with probability at least $1-\delta$*,*

$$1/2\mathbf{I}_{\widetilde{d}} \preceq \mathbf{S}_n \preceq 3/2\mathbf{I}_{\widetilde{d}}. \tag{79}$$

Proof. Let $t = 2K\sqrt{\log(1/\delta)/c_1}$, by Eq. (76) we have

$$2\exp(-c_K t^2) \leq 2\exp\big(-\frac{c_1}{K^4}\frac{K^2\log(1/2\delta)}{c_1}\big) = \delta. \tag{80}$$

Let $n = \frac{32}{c_1}K^2(\widetilde{d}+\log(1/\delta)$, then

$$a = \frac{C_K\sqrt{\widetilde{d}}+t}{\sqrt{n}} = \frac{\frac{2}{\sqrt{c_1}}K^2(\sqrt{\widetilde{d}}+\sqrt{\log(1/\delta)})}{\frac{4\sqrt{2}}{\sqrt{c_1}}K^2\sqrt{\widetilde{d}+\log(1/\delta)}} \leq \frac{1}{2}, \tag{81}$$

and thus $\max\{a, a^2\} \leq 1/2$. Therefore, with probability at least $1-\delta$, we have

$$\|\mathbf{S}_n - \mathbf{I}_{\widetilde{d}}\| \leq \frac{1}{2}, \tag{82}$$

and thus $1/2\mathbf{I}_{\widetilde{d}} \preceq \mathbf{S}_n \preceq 3/2\mathbf{I}_{\widetilde{d}}$. ☐

C.4 Proof of Theorem 3 (Theorem 32)

We present the full proof of Theorem 3 as the following. Some of the techniques used in the proof are inspired by [24].

Proof. By the definitions of σ, ρ and ν in Theorem 3, we have the following basic inequalities. Given vectors $v \in \mathbb{R}^d$ and $u \in \mathbb{R}^{\widetilde{d}}$, we have the following norm relations:

$$\|v\|_{\mathbf{V}_q} \leq \sqrt{\nu}\|v\|_{\mathbf{V}_p}, \qquad \|v\|_{\mathbf{V}_p^{-1}} \leq \sqrt{\nu}\|v\|_{\mathbf{V}_q^{-1}}, \tag{83}$$

$$\|u\|_{\mathbf{H}_p} \leq \sqrt{\sigma}\|u\|_{\mathbf{H}_q}, \tag{84}$$

$$\|u\|_{\widetilde{\mathbf{V}}_p} \leq \sqrt{\rho}\|u\|_{\mathbf{H}_p}, \tag{85}$$

where $\widetilde{\mathbf{V}}_p \triangleq \mathbf{I}_{c-1}\otimes\mathbf{V}_p$.

step 1. Let $V_n = \sqrt{n}\mathbf{H}_p{}^{-1/2}\vec{\nabla}Q_n(\theta_*)$, then V_n is a centered, isotropic sub-Gaussian random vector. Indeed, since $\nabla Q_n(\theta_*) = \frac{1}{n}\sum_{i\in[n]}\vec{\nabla}\ell_{z_i}(\theta_*)$, we have

$$\mathop{\mathbb{E}}_{\{z_i\sim\pi_q\}_{i=1}^n}[V_n] = \frac{1}{\sqrt{n}}\mathbf{H}_q{}^{-1/2}\sum_{i\in[n]}\mathbb{E}_{z_i\sim\pi_q}[\vec{\nabla}\ell_{z_i}(\theta_*)] = 0$$

$$
\begin{aligned}
\mathop{\mathbb{E}}_{\{z_i\sim\pi_q\}_{i=1}^n}[V_nV_n^\top] &= \mathbf{H}_q{}^{-1/2}\left(\frac{1}{n}\sum_{i\in[n]}\mathbb{E}_{z_i\sim\mathcal{P}}[\vec{\nabla}\ell_{z_i}(\theta_*)\vec{\nabla}\ell_{z_i}(\theta_*)^\top]\right)\mathbf{H}_q{}^{-1/2} \\
&= \mathbf{H}_q{}^{-1/2}\mathbf{H}_q\mathbf{H}_q{}^{-1/2} = \mathbf{I}_{\widetilde{d}}. \qquad (86)
\end{aligned}
$$

By Lemma 18,

$$
\|V_n\|_{\psi_2}^2 \lesssim \sum_{i\in[n]}\|\frac{1}{\sqrt{n}}\mathbf{H}_q{}^{-1/2}\vec{\nabla}\ell_{z_i}(\theta_*)\|_{\psi_2}^2 = K_{1,q}^2. \qquad (87)
$$

Now we apply the upper bound for quadratic form of sub-Gaussian random vector derived in Eq. (25) from Lemma 20, we can get

$$
\|\vec{\nabla}Q_n(\theta_*)\|_{\mathbf{H}_q{}^{-1}}^2 = \frac{1}{n}\|V_n\|_2^2 \lesssim \frac{K_{1,q}^2\left(\widetilde{d}+\sqrt{\widetilde{d}}\log(e/\delta)\right)}{n}. \qquad (88)
$$

step 2. W.l.o.g we assume that Assumption 1-(3) holds with $r=\mathcal{O}(1)$ and denote $\overline{K}_{2,q}\triangleq K_{2,q}(r)$ $\overline{K}_{2,p}\triangleq K_{2,p}(r)$ for ease of discussion. Now we show that the Hessian $\mathbf{H}_q(\theta)$ is a good approximation to $\mathbf{H}_q$ for any $\theta\in\mathcal{B}_{q,\widehat{r}}(\theta_*)=\{\theta:\|\theta-\theta_*\|_{\mathbf{V}_q,\infty}\le\widehat{r}\}$, where $\widehat{r}=1/c$ for some constant c depending on $K_{0,q}$ and $\overline{K}_{2,q}$.

Fix $\theta_0=\theta_*$ and pick arbitrary $\theta_1\in\Theta$, let $\theta_t=\theta_0+t\Delta$, where $\Delta\triangleq\theta_1-\theta_0$. Define function

$$
\phi_q(t)\triangleq L_q(\theta_t)=\mathbb{E}_{z\sim\pi_q}[\ell_z(\theta_t)] \qquad (89)
$$

Our goal is to show that $\phi_q(t)$ is pseudo self-concordant, i.e. we intend to get some constant $C>0$ s.t. $|\phi_q'''(t)|\le C\phi_q''(t)$. First we observe that

$$
\begin{aligned}
\phi_q''(t) &= \mathbb{E}_{(x,y)\sim\pi_q}[\ell''(y,\theta_tx)[\Delta x,\Delta x]] = \mathbb{E}_{(x,y)\sim\pi_q}[\vec{\Delta}^\top\left(\nabla^2\ell_{(x,y)}(\theta_tx)\right)\vec{\Delta}] \\
&= \vec{\Delta}^\top\,\mathbb{E}_{(x,y)\sim\pi_q}[\nabla^2\ell_{(x,y)}(\theta_tx)]\vec{\Delta} = \|\vec{\Delta}\|_{\mathbf{H}_q(\theta_t)}^2. \qquad (90)
\end{aligned}
$$

Note that $\ell(y,\eta)$ is the loss function defined in Eq. (38) and $\ell''(y,\eta)$ is the Hessian w.r.t η.

On the other hand, by Lemma 29 we have

$$
\begin{aligned}
|\phi_q'''(t)| &\le \mathbb{E}_{(x,y)\sim\pi_q}\left[\left|\ell'''(y,\theta_tx)[\Delta x,\Delta x,\Delta x]\right|\right] \\
&\le 2\,\mathbb{E}_{(x,y)\sim\pi_q}\left[\ell''(y,\theta_tx)[\Delta x,\Delta x]\|\Delta x\|_\infty\right] \\
&\le 2\sqrt{\mathbb{E}_{(x,y)\sim\pi_q}\left[\left(\ell''(y,\theta_tx)[\Delta x,\Delta x]\right)^2\right]}\sqrt{\mathbb{E}_{(x,y)\sim\pi_q}\left[\|\Delta x\|_\infty^2\right]}, \qquad (91)
\end{aligned}
$$

where the last inequality follows by Cauchy-Schwartz inequality.

Now we bound both of the square root terms in Eq. (91). For the first square root term, let $\widehat{\Delta}\triangleq\mathbf{H}_q(\theta_t)^{1/2}\vec{\Delta}/\|\vec{\Delta}\|_{\mathbf{H}_q(\theta_t)}$, then $\vec{\Delta}=\|\vec{\Delta}\|_{\mathbf{H}_q(\theta_t)}\mathbf{H}_q(\theta_t)^{-1/2}\widehat{\Delta}$ and $\|\widehat{\Delta}\|_2=1$. We have

$$
\begin{aligned}
\ell''(y,\theta_tx)[\Delta x,\Delta x] &= \vec{\Delta}^\top\nabla^2\ell_{(x,y)}(\theta_tx)\vec{\Delta} \\
&= \|\vec{\Delta}\|_{\mathbf{H}_q(\theta_t)}^2\widehat{\Delta}^\top\mathbf{H}_q(\theta_t)^{-1/2}\nabla^2\ell_{(x,y)}(\theta_tx)\mathbf{H}_q(\theta_t)^{-1/2}\widehat{\Delta}. \qquad (92)
\end{aligned}
$$

We claim that $\ell''(y,\theta_tx)[\Delta x,\Delta x]$ is a sub-exponential random variable. Indeed,

$$
\begin{aligned}
\left\|\ell''(y,\theta_tx)[\Delta x,\Delta x]\right\|_{\psi_1} &\overset{(a)}{\le} \|\vec{\Delta}\|_{\mathbf{H}_q(\theta_t)}^2\|\widehat{\Delta}^\top\mathbf{H}_q(\theta_t)^{-1/2}\nabla^2\ell_{(x,y)}(\theta_tx)\mathbf{H}_q(\theta_t)^{-1/2}\widehat{\Delta}\|_{\psi_1} \\
&\overset{(b)}{\le} \|\vec{\Delta}\|_{\mathbf{H}_q(\theta_t)}^2\sup_{u\in\mathcal{S}^{\widetilde{d}-1}}\|u^\top\mathbf{H}_q(\theta_t)^{-1/2}\nabla^2\ell_{(x,y)}(\theta_tx)\mathbf{H}_q(\theta_t)^{-1/2}u\|_{\psi_1} \\
&\overset{(c)}{\le} \|\vec{\Delta}\|_{\mathbf{H}_q(\theta_t)}^2\overline{K}_{2,q}, \qquad (93)
\end{aligned}
$$

where (a) follows by Eq. (92), (b) follows by the fact that $\|\widehat{\Delta}\|_2=1$, (c) follows by Assumption 1-(3). By the property of sub-exponential random variable in Lemma 14-(1), we can obtain that

$$
\mathbb{E}_{(x,y)\sim\pi_q}\left[\left(\ell''(y,\theta_tx)[\Delta x,\Delta x]\right)^2\right] \lesssim \overline{K}_{2,q}^2\|\vec{\Delta}\|_{\mathbf{H}_q(\theta_t)}^4 \overset{Eq.\ (90)}{=} \overline{K}_{2,q}^2\phi_q''(t)^2. \qquad (94)
$$

On the other hand, let $\Delta_i^\top$ be the ith row of $\Delta \in \mathbb{R}^{(c-1)\times d}$. For $x \sim q(x)$, define random variable $\xi(x) \triangleq \|\Delta x\|_\infty$, we claim that $\xi(x)$ is sub-Gaussian. Indeed,

$$\begin{aligned}\xi(x) = \|\Delta x\|_\infty &= \max_{i\in[c-1]} |\langle x, \Delta_i\rangle| = \max_{i\in[c-1]} |\langle \mathbf{V}_q^{-1/2}x, \mathbf{V}_q^{1/2}\Delta_i\rangle| \\ &= \max_{i\in[c-1]} \|\mathbf{V}_q^{1/2}\Delta_i\|_2 \left|\left\langle \mathbf{V}_q^{-1/2}x, \frac{\mathbf{V}_q^{1/2}\Delta_i}{\|\mathbf{V}_q^{1/2}\Delta_i\|_2}\right\rangle\right| \\ &\le \|\Delta\|_{\mathbf{V}_q,\infty} \max_{i\in[c-1]} \left|\left\langle \mathbf{V}_q^{-1/2}x, \frac{\mathbf{V}_q^{1/2}\Delta_i}{\|\Delta_i\|_{\mathbf{V}_q}}\right\rangle\right| \triangleq \|\Delta\|_{\mathbf{V}_q,\infty}\left|\left\langle \mathbf{V}_q^{-1/2}x, \frac{\mathbf{V}_q^{1/2}\Delta_{i(x)}}{\|\Delta_{i(x)}\|_{\mathbf{V}_q}}\right\rangle\right| \end{aligned} \tag{95}$$

where we define $i(x)$ for each x as the index such that the maximum is attained. Now we have

$$\begin{aligned}\|\xi(x)\|_{\psi_2} &\le \|\Delta\|_{\mathbf{V}_q,\infty}\left\|\left\langle \mathbf{V}_q^{-1/2}x, \frac{\mathbf{V}_q^{1/2}\Delta_{i(x)}}{\|\Delta_{i(x)}\|_{\mathbf{V}_q}}\right\rangle\right\|_{\psi_2} \\ &\le \|\Delta\|_{\mathbf{V}_q,\infty} \sup_{u\in\mathcal{S}^{d-1}} \|\langle \mathbf{V}_p^{-1/2}x, u\rangle\|_{\psi_2} = \|\Delta\|_{\mathbf{V}_q,\infty}\|\mathbf{V}_q^{-1/2}x\|_{\psi_2} \\ &\le \|\Delta\|_{\mathbf{V}_q,\infty} K_{0,q},\end{aligned} \tag{96}$$

where the last inequality follows by Assumption 1-(1). Applying Lemma 12-(2), we have

$$\mathbb{E}_{(x,y)\sim\pi_q}[\|\Delta x\|_\infty^2] = \mathbb{E}_{x\sim q}[|\xi(x)|^2] \lesssim \|\Delta\|^2_{\mathbf{V}_q,\infty} K_{0,q}^2. \tag{97}$$

Now substitute Eqs. (94) and (97) into Eq. (91), we can prove that $\phi_p(t)$ is pseudo self-concordant:

$$|\phi_q'''(t)| \le C\|\Delta\|_{\mathbf{V}_q,\infty} K_{0,q}\overline{K}_{2,q}\|\vec{\Delta}\|^2_{\mathbf{H}_q(\theta_t)} = C\|\Delta\|_{\mathbf{V}_q,\infty} K_{0,q}\overline{K}_{2,q}\phi_q''(t), \tag{98}$$

where the last equality follows by Eq. (90). We consider the ball $\mathcal{B}_{q,\widehat{r}}(\theta_*) = \{\theta \in \Theta : \|\theta - \theta_*\|_{\mathbf{V}_q,\infty} \le \widehat{r}\}$, where $\widehat{r}$ is defined by

$$\widehat{r} \triangleq \frac{1}{C\log\sqrt{2}\cdot K_{0,q}\overline{K}_{2,q}}. \tag{99}$$

Thus for any $\theta \in \mathcal{B}_{q,\widehat{r}}(\theta_*)$, by Eq. (98)

$$|\phi_q'''(t)| \le \log\sqrt{2}\cdot\phi_q''(t). \tag{100}$$

Now we satisfy the premise of Proposition 31 by setting $S = \log\sqrt{2}$. With Eq. (57) we can conclude that for any $\theta \in \mathcal{B}_{q,\widehat{r}}(\theta_*)$,

$$1/\sqrt{2}\mathbf{H}_q \preceq \mathbf{H}_q(\theta) \preceq \sqrt{2}\mathbf{H}_q. \tag{101}$$

step 3. In this step, we consider an ϵ-net $\mathcal{N}_\epsilon$ on ball $\mathcal{B}_{q,\widehat{r}}(\theta_*)$ under metric $\|\cdot\|_{\mathbf{V}_q,\infty}$($\widehat{r}$ is defined in Eq. (99)). We intend to approximate empirical Hessian $\mathbf{H}_n(\theta)$ using $\mathbf{H}_n(\theta')$, where $\theta' \in \mathcal{N}_\epsilon$.

Since $\{x_i\}_{i=1}^n$ are drawn independently from $q(x)$, by (26) in Lemma 20 it holds with probability at least $1-\delta$ that

$$\|x_i\|^2_{\mathbf{V}_q^{-1}} \lesssim K_{0,q}^2\Big(d + \sqrt{d}\log(e/\delta)\Big). \tag{102}$$

By union bound and Eq. (83), with probability at least $1-\delta$ we have

$$\max_{i\in[n]} \|x_i\|^2_{\mathbf{V}_q^{-1}} \lesssim K_{0,q}^2\Big(d + \sqrt{d}\log(en/\delta)\Big) \triangleq R^2. \tag{103}$$

Let $\mathcal{N}_\epsilon$ be an ϵ-net on ball $\mathcal{B}_{q,\widehat{r}}(\theta_*)$ with ϵ defined as

$$\epsilon \triangleq \frac{\log\sqrt{2}}{2\cdot R}. \tag{104}$$

Denote $\mathcal{P} : \mathcal{B}_{q,\widehat{r}}(\theta_*) \to \mathcal{N}_\epsilon$ as the projection of $\theta \in \mathcal{B}_{q,\widehat{r}}(\theta_*)$ onto the $\epsilon-$net, i.e. $\mathcal{P}(\theta)$ is the closest point in $\mathcal{N}_\epsilon$ to θ under norm $\|\cdot\|_{\mathbf{V}_q,\infty}$:

$$\mathcal{P}(\theta) \in \arg\min_{\theta' \in \mathcal{N}_\epsilon} \|\theta - \theta'\|_{\mathbf{V}_q,\infty}. \tag{105}$$

We remark that the choice of $\mathcal{P}(\theta)$ does not effect our results. Now pick arbitrary $\theta_1 \in \Theta_{\overline{r}}(\theta_*)$, $\theta_0 = \mathcal{P}(\theta)$, $\theta_t = \theta_0 + t(\theta_1 - \theta_0)$, and $\phi_n(t) = Q_n(\theta_t)$. Using Proposition 30, we have

$$\begin{aligned}
\phi_n'''(t)| &\le 2\phi_n''(t) \max_{i\in[n]} \|(\theta_1 - \theta_0)x_i\|_\infty \\
&\le 2\phi_n''(t)\|\theta_1 - \theta_0\|_{\mathbf{V}_q,\infty} \max_{i\in[n]} \|x_i\|_{\mathbf{V}_q^{-1}} \\
&\le 2R\epsilon\phi_n''(t) = \log\sqrt{2}\cdot\phi_n''(t),
\end{aligned} \tag{106}$$

where the last inequality follows by Eqs. (103) and (105). Thus $\phi_n(t)$ is pseudo self-concordant, and we can apply Proposition 31 with $S = \log\sqrt{2}$. By Eq. (57) we have

$$1/\sqrt{2}\mathbf{H}_n(\mathcal{P}(\theta)) \preceq \mathbf{H}_n(\theta) \preceq \sqrt{2}\mathbf{H}_n(\mathcal{P}(\theta)), \qquad \forall \theta \in \mathcal{B}_{q,\widehat{r}}(\theta_*). \tag{107}$$

step 4. In this step we approximate empirical Hessian $\mathbf{H}_n(\theta)$ using $\mathbf{H}_q(\theta)$, for all $\theta \in \mathcal{N}_\epsilon$. Note that $\mathbf{H}_n(\theta) = \nabla^2 Q_n(\theta) = \frac{1}{n}\sum_{i=1}^n \nabla^2\ell_{z_i}(\theta x_i)$. For an arbitrary $\theta \in \mathcal{N}_\epsilon$, let $\mathbf{A}_i = \mathbf{H}_q(\theta)^{-1/2}\nabla^2\ell_{z_i}(\theta)\mathbf{H}_q(\theta)^{-1/2}$, then $\mathbb{E}[\mathbf{A}_i] = \mathbf{I}_{\widetilde{d}}$ and

$$\frac{1}{n}\sum_{i\in[n]} \mathbf{A}_i = \mathbf{H}_q(\theta)^{-1/2}\mathbf{H}_n(\theta)\mathbf{H}_q(\theta)^{-1/2}. \tag{108}$$

By Assumption 1-(3), $\{\mathbf{A}_i\}_{i=1}^n$ satisfy the premise of Proposition 33. Applying Corollary 34 and then using union bound over all $\theta \in \mathcal{N}_\epsilon$, we obtain that whenever

$$n \gtrsim \overline{K}_{2,q}^2(\widetilde{d} + \log(|\mathcal{N}_\epsilon|/\delta), \tag{109}$$

where $|\mathcal{N}_\epsilon|$ is the number of points contained in $\mathcal{N}_\epsilon$, then with probability at least $1-\delta$,

$$1/2\mathbf{I}_{\widetilde{d}} \preceq \frac{1}{n}\sum_{i\in[n]} \mathbf{A}_i \preceq 3/2\mathbf{I}_{\widetilde{d}}, \qquad \forall \theta \in \mathcal{N}_\epsilon. \tag{110}$$

By Eq. (108), Eq. (110) is equivalent to

$$1/2\mathbf{H}_q(\theta) \preceq \mathbf{H}_n(\theta) \preceq 3/2\mathbf{H}_q(\theta), \qquad \forall \theta \in \mathcal{N}_\epsilon. \tag{111}$$

Now we intend to derive a bound for n to satisfy Eq. (109). First we need to estimate an upper bound for $|\mathcal{N}_\epsilon|$. By Proposition 4.2.12 in [23], we have $|\mathcal{N}_\epsilon| \le (\frac{3\widehat{r}}{\epsilon})^{\widetilde{d}}$. Thus a sufficient condition for (109) is

$$n \gtrsim \overline{K}_{2,p}^2\Big(\widetilde{d} + \widetilde{d}\log\big(\frac{e\widehat{r}}{\epsilon\delta}\big)\Big). \tag{112}$$

Recall that $\widehat{r} = O\Big(1/(K_{0,q}\overline{K}_{2,q})\Big)$, $\epsilon = O\Big(1/\Big(K_{0,q}\sqrt{d + \sqrt{d}\log(en/\delta)}\Big)\Big)$, then

$$\log\big(\frac{e\overline{r}}{\epsilon\delta}\big) = \log\Bigg(\frac{eK_{0,q}\sqrt{d+\sqrt{d}\log(en/\delta)}}{K_{0,q}\overline{K}_{2,q}}\Bigg). \tag{113}$$

Thus it is sufficient to let

$$n \gtrsim \overline{K}_{2,q}^2\widetilde{d}\log(ed/\delta), \tag{114}$$

which is the first bound at Eq. (6).

step 5. Next we prove that if n is larger than the second bound of Eq. (6), then $\theta_n \in \mathcal{B}_{q,\widehat{r}}(\theta_*)$ and Eq. (61) holds. First, combining Eqs. (101), (107) and (111), we have with probability at least $1-\delta$,

$$\frac{1}{4}\mathbf{H}_q \preceq \mathbf{H}_n(\theta) \preceq 3\mathbf{H}_q, \qquad \forall \theta \in \mathcal{B}_{q,\widehat{r}}(\theta_*). \tag{115}$$

Let $\theta_0 = \theta_*$, pick arbitrary $\theta_1 \in \mathcal{B}_{q,\widehat{r}}(\theta_*)$, $\theta_t = \theta_0 + t\Delta$, where $\Delta \triangleq \theta_1 - \theta_0$. By Eq. (90), we already have $\phi_q''(0) = \|\vec{\Delta}\|_{\mathbf{H}_q}$. On the other hand, we can show that

$$\phi_n''(t) = \frac{1}{n}\sum_{i=1}^{n} \ell''(y_i, \theta x_i)[\Delta x, \Delta x] = \|\vec{\Delta}\|_{\mathbf{H}_n(\theta_t)}, \tag{116}$$

Thus Eq. (115) reduces to

$$\frac{1}{4}\phi_q''(0) \le \phi_n''(t) \le 3\phi_q''(0), \qquad t \in [0,1]. \tag{117}$$

Integrating this twice, we have $\frac{1}{4}\phi_q''(0)t^2 \le \phi_n(t) - \phi_n(0) - \phi_n'(0)t \le 3\phi_q''(0)t^2$. Let $t = 1$, we can get with probability at least $1-\delta$,

$$\frac{1}{4}\|\vec{\Delta}\|_{\mathbf{H}_q}^2 \le Q_n(\theta) - Q_n(\theta_*) - \langle \vec{\nabla} Q_n(\theta_*), \vec{\Delta}\rangle \le 3\|\vec{\Delta}\|_{\mathbf{H}_q}^2. \tag{118}$$

Using Cauchy-Schwartz inequality, we can obtain

$$\begin{aligned} Q_n(\theta) - Q_n(\theta_*) &\ge \frac{1}{4}\|\vec{\Delta}\|_{\mathbf{H}_q}^2 + \langle \vec{\nabla} Q_n(\theta_*), \vec{\Delta}\rangle \\ &\ge \frac{1}{4}\|\vec{\Delta}\|_{\mathbf{H}_q}\Big(\|\vec{\Delta}\|_{\mathbf{H}_q} - 4\|\vec{\nabla} Q_n(\theta_*)\|_{\mathbf{H}_q{}^{-1}}\Big). \end{aligned} \tag{119}$$

Our goal is to prove that given n lower bounded by the second bound in Eq. (6), $\theta_n \in \mathcal{B}_{q,\widehat{r}}$. Since $Q_n(\theta)$ is a convex function and $\Theta_{\overline{r}}(\theta_*)$ is a convex set, it suffices to show that the right hand side of Eq. (119) is non-negative for all $\theta \in \partial\mathcal{B}_{q,\widehat{r}}$, i.e. $\|\Delta\|_{\mathbf{V}_q,\infty} = \widehat{r}$. First note that

$$\begin{aligned} \|\vec{\Delta}\|_{\mathbf{H}_q} &\overset{Eq.\ (84)}{\ge} \frac{1}{\sqrt{\sigma}}\|\vec{\Delta}\|_{\mathbf{H}_p} \overset{Eq.\ (85)}{\ge} \sqrt{\frac{1}{\sigma\rho\nu}}\|\vec{\Delta}\|_{\widetilde{\mathbf{V}}_p} \ge \sqrt{\frac{1}{\sigma\rho}}\|\Delta\|_{\mathbf{V}_p} \\ &\overset{Eq.\ (83)}{\ge} \sqrt{\frac{1}{\sigma\rho\nu}}\|\Delta\|_{\mathbf{V}_q} = \sqrt{\frac{1}{\sigma\rho\nu}}\cdot\widehat{r} \ge \frac{1}{C\sqrt{\sigma\rho\nu}K_{0,q}\overline{K}_{2,q}}. \end{aligned} \tag{120}$$

Since we have proved that $\|\vec{\nabla} Q_n(\theta_*)\|_{\mathbf{H}_p{}^{-1}} \lesssim \sqrt{\frac{K_{1,q}^2\left(\widetilde{d}+\sqrt{\widetilde{d}}\log(e/\delta)\right)}{n}}$ in step 1, connecting this with Eqs. (120) and (119), we have $\theta_n \in \mathcal{B}_{q,\widehat{r}}(\theta_*)$ if

$$n \gtrsim \sigma\rho\nu K_{0,q}^2 K_{1,q}^2 \overline{K}_{2,q}^2\Big(\widetilde{d} + \sqrt{\widetilde{d}}\log(e/\delta)\Big). \tag{121}$$

Now let $\theta_1 = \theta_n$, then $\vec{\Delta} = \mathrm{vec}(\theta_n - \theta_*)$. Since $Q_n(\theta_n) \le Q_n(\theta_*)$, from Eq. (119) we can get

$$\|\mathrm{vec}(\theta_n - \theta_*)\|_{\mathbf{H}_q}^2 \le \|\vec{\nabla} Q_n(\theta_*)\|_{\mathbf{H}_q{}^{-1}}. \tag{122}$$

We have proved that $1/\sqrt{2}\mathbf{H}_q \preceq \mathbf{H}_q(\theta) \preceq \sqrt{2}\mathbf{H}_q$ in Eq. (101), it can be reduced to

$$\frac{1}{\sqrt{2}}\phi_q''(0) \le \phi_q''(t) \le \sqrt{2}\phi_q''(0), \quad 0 \le t \le 1. \tag{123}$$

Integrating twice on $[0,1]$, we have $\frac{1}{2\sqrt{2}}\phi_q''(0)t^2 \le \phi_q(t) - \phi_q(0) \le \frac{\sqrt{2}}{2}\phi_q''(0)t^2$. Since $\theta_n \in \mathcal{B}_{q,\widehat{r}}(\theta_*)$, we can assume $\theta_1 = \theta_n$. Let $t = 1$, we can get

$$\begin{aligned} L_q(\theta_n) - L_q(\theta_*) &\overset{Eq.\ (89)}{=} \phi_q(\theta_n) - \phi_q(\theta_*) \overset{Eq.\ (90)}{\le} \frac{\sqrt{2}}{2}\|\mathrm{vec}(\theta_n - \theta_*)\|_{\mathbf{H}_q}^2 \\ &\overset{Eq.\ (122)}{\le} \frac{\sqrt{2}}{2}\|\vec{\nabla} Q_n(\theta_*)\|_{\mathbf{H}_q{}^{-1}} \overset{Eq.\ (88)}{\lesssim} \sqrt{\frac{K_{1,q}^2\Big(\widetilde{d} + \sqrt{\widetilde{d}}\log(e/\delta)\Big)}{n}}. \end{aligned} \tag{124}$$

step 6. Now we bound the excess risk with respect to $p(x)$, i.e. $L_p(\theta_n) - L_p(\theta_*)$.

Our goal is to use the Taylor expansion property in Proposition 31. First we have to show that $L_p(\theta)$ is pseudo self-concordant. Let $\theta_0 = \theta_*$, $\theta_1 = \theta_n$, and $\theta_t = \theta_0 + t\Delta$, where $\Delta = \theta_1 - \theta_0$. Define

$$\phi_p(t) \triangleq L_p(\theta_t) = \mathbb{E}_{z\sim\pi_p}[\ell_z(\theta_t)]. \tag{125}$$

We can follow the argument from step 2 and obtain that

$$|\phi_p'''(t)| \leq C\|\Delta\|_{\mathbf{V}_p,\infty} K_{0,p}\overline{K}_{2,p}\phi_p''(t). \tag{126}$$

Note that

$$\begin{aligned} \|\Delta\|_{\mathbf{V}_p,\infty} \leq \|\vec{\Delta}\|_{\widetilde{\mathbf{V}}_p} &\overset{Eq.\ (85)}{\leq} \sqrt{\rho}\|\vec{\Delta}\|_{\mathbf{H}_p} \overset{Eq.\ (84)}{\leq} \sqrt{\sigma\rho}\|\vec{\Delta}\|_{\mathbf{H}_q} \\ &\overset{Eq.\ (124)}{\lesssim} \sqrt{\sigma\rho}K_{1,q}\sqrt{\frac{\widetilde{d} + \sqrt{\widetilde{d}}\log(e/\delta)}{n}}. \end{aligned} \tag{127}$$

Substitute this into Eq. (126), we have $|\phi_p'''(t)| \leq \alpha\phi_p''(t)$, where

$$\alpha = \mathcal{O}\left(\sqrt{\sigma\rho}K_{0,p}K_{1,q}\overline{K}_{2,p}\sqrt{\frac{\widetilde{d} + \sqrt{\widetilde{d}}\log(e/\delta)}{n}}\right). \tag{128}$$

Now we can use Proposition 31 and let $S = \alpha$. Note that $\nabla L_p(\theta_*) = 0$, by Eq. (56) we have

$$\frac{e^{-\alpha} + \alpha - 1}{\alpha^2}\|\vec{\Delta}\|_{\mathbf{H}_p}^2 \leq L_p(\theta_n) - L_p(\theta_*) \leq \frac{e^{\alpha} - \alpha - 1}{\alpha^2}\|\vec{\Delta}\|_{\mathbf{H}_p}^2. \tag{129}$$

By Taylor theorem, there exits $\widetilde{\theta} \in \mathcal{B}_{q,\widehat{r}}(\theta_*)$ between θ_* and θ_n such that

$$\vec{\nabla} Q_n(\theta_*) = \vec{\nabla} Q_n(\theta_n) + \mathbf{H}_n(\widetilde{\theta})\vec{\Delta}. \tag{130}$$

Since $\vec{\nabla} Q_n(\theta_n) = 0$, we have

$$\vec{\nabla} Q_n(\theta_*) = \mathbf{H}_n(\widetilde{\theta})\vec{\Delta}. \tag{131}$$

By Eq. (115), we have $\frac{1}{4}\mathbf{H}_q \preceq \mathbf{H}_n(\widetilde{\theta}) \preceq 3\mathbf{H}_q$. Define $\mathbf{M}_{q,n} \triangleq {\mathbf{H}_q}^{1/2}(\mathbf{H}_n(\widetilde{\theta}))^{-1}{\mathbf{H}_q}^{1/2}$, then

$$\frac{1}{3}\mathbf{I}_{\widetilde{d}} \preceq \mathbf{M}_{q,n} \preceq 4\mathbf{I}_{\widetilde{d}}. \tag{132}$$

For the lower bound in Eq. (129), we have with probability at least $1 - \delta$,

$$\begin{aligned} L_p(\theta_n) - L_p(\theta_*) &\geq \frac{e^{-\alpha} + \alpha - 1}{\alpha^2}\vec{\Delta}^\top\mathbf{H}_p\vec{\Delta} \\ &= \frac{e^{-\alpha} + \alpha - 1}{\alpha^2}\big(\vec{\Delta}^\top\mathbf{H}_n(\widetilde{\theta})\big)\big(\mathbf{H}_n(\widetilde{\theta})^{-1}\mathbf{H}_p\mathbf{H}_n(\widetilde{\theta})^{-1}\big)\big(\mathbf{H}_n(\widetilde{\theta})\vec{\Delta}\big) \\ &\overset{Eq.\ (131)}{=} \frac{e^{-\alpha} + \alpha - 1}{\alpha^2}\vec{\nabla} Q_n(\theta_*)^\top{\mathbf{H}_q}^{-1/2}\mathbf{M}_{q,n}\big({\mathbf{H}_q}^{-1/2}\mathbf{H}_p{\mathbf{H}_q}^{-1/2}\big)\mathbf{M}_{q,n}{\mathbf{H}_q}^{-1/2}\vec{\nabla} Q_n(\theta_*) \\ &\overset{Eq.\ (136)}{\geq} \frac{e^{-\alpha} + \alpha - 1}{9\alpha^2}\big\langle{\mathbf{H}_q}^{-1}\mathbf{H}_p{\mathbf{H}_q}^{-1}, \vec{\nabla} Q_n(\theta_*)\vec{\nabla} Q_n(\theta_*)^\top\big\rangle. \end{aligned} \tag{133}$$

Similarly, we can derive the upper bound:

$$\begin{aligned} L_p(\theta_n) - L_p(\theta_*) &\leq \frac{e^{\alpha} - \alpha - 1}{\alpha^2}\vec{\Delta}^\top\mathbf{H}_p\vec{\Delta} \\ &= \frac{e^{\alpha} - \alpha - 1}{\alpha^2}\vec{\nabla} Q_n(\theta_*)^\top{\mathbf{H}_q}^{-1/2}\mathbf{M}_{q,n}\big({\mathbf{H}_q}^{-1/2}\mathbf{H}_p{\mathbf{H}_q}^{-1/2}\big)\mathbf{M}_{q,n}{\mathbf{H}_q}^{-1/2}\vec{\nabla} Q_n(\theta_*) \\ &\leq 16\frac{e^{\alpha} - \alpha - 1}{\alpha^2}\big\langle{\mathbf{H}_q}^{-1}\mathbf{H}_p{\mathbf{H}_q}^{-1}, \vec{\nabla} Q_n(\theta_*)\vec{\nabla} Q_n(\theta_*)^\top\big\rangle. \end{aligned} \tag{134}$$

Given $\{x_i\}_{i=1}^n \overset{i.i.d}{\sim} q(x)$, we have

$$\mathbb{E}_{\{y_i\sim p(y_i|x_i,\theta_*)\}_{i=1}^n}[\vec{\nabla} Q_n(\theta_*)\vec{\nabla} Q_n(\theta_*)^\top]$$

$$
\begin{aligned}
&=\frac{1}{n^2}\mathbb{E}_{\{y_i\sim p(y_i|x_i,\theta_*)\}_{i=1}^n}\left[\sum_{i=1}^n\vec{\nabla}\ell_{z_i}(\theta_*)\sum_{j=1}^n(\vec{\nabla}\ell_{z_i}(\theta_*))^\top\right]\\
&=\frac{1}{n^2}\sum_{i=1}^n\mathbb{E}_{y_i\sim p(y_i|x_i,\theta_*)}[\vec{\nabla}\ell_{z_i}(\theta_*)\vec{\nabla}\ell_{z_i}(\theta_*)^\top]+\frac{2}{n^2}\sum_{i\neq j}\mathbb{E}_{\substack{y_i\sim p(y_i|x_i,\theta_*)\\ y_j\sim p(y_j|x_j,\theta_*)}}[\vec{\nabla}\ell_{z_i}(\theta_*)\vec{\nabla}\ell_{z_j}(\theta_*)^\top]\\
&\overset{(a)}{=}\frac{1}{n^2}\sum_{i=1}^n\mathbb{E}_{y_i\sim p(y_i|x_i,\theta_*)}[\vec{\nabla}\ell_{z_i}(\theta_*)\vec{\nabla}\ell_{z_i}(\theta_*)^\top]\overset{(b)}{=}\frac{1}{n^2}\sum_{i=1}^n\mathbb{E}_{y_i\sim p(y_i|x_i,\theta_*)}[\nabla^2\ell_{z_i}(\theta_*)]\\
&=\frac{1}{n}\mathbf{H}_n(\theta_*)
\end{aligned}
\tag{135}
$$

where (a) follows by the independence between y_i and y_j and the fact that $\mathbb{E}_{y_i\sim p(y_i|x_i,\theta_*)}[\nabla\ell_{(x_i,y_i)}(\theta_*)]=0$ from Lemma 24, (b) follows by Lemma 25.

Similar to the argument in step 4, using Corollary 34 we have with probability at least $1-\delta$,

$$
\frac{1}{2}\mathbf{H}_q\preceq\mathbf{H}_n(\theta_*)\preceq\frac{3}{2}\mathbf{H}_q,
\tag{136}
$$

where the requirement for n is already satisfied due to the second bound for n in Eq. (6). Since ${\mathbf{H}_q}^{-1/2}\mathbf{H}_p{\mathbf{H}_q}^{-1/2}$ is symmetric positive definite, we can assume it has eigen-decomposition ${\mathbf{H}_q}^{-1/2}\mathbf{H}_p{\mathbf{H}_q}^{-1/2}=\sum_{i=1}^{\tilde{d}}\lambda_i v_i v_i^\top$. Then

$$
\begin{aligned}
\left\langle{\mathbf{H}_q}^{-1}\mathbf{H}_p{\mathbf{H}_q}^{-1},\mathbf{H}_n(\theta_*)\right\rangle&=\left\langle{\mathbf{H}_q}^{-1/2}\mathbf{H}_p{\mathbf{H}_q}^{-1/2},{\mathbf{H}_q}^{-1/2}\mathbf{H}_n(\theta_*){\mathbf{H}_q}^{-1/2}\right\rangle\\
&=\sum_{i=1}^{d'}\lambda_i v_i^\top\left({\mathbf{H}_p}^{-1/2}\mathbf{H}_n(\theta_*){\mathbf{H}_p}^{-1/2}\right)v_i.
\end{aligned}
\tag{137}
$$

Using Eq. (136), we can get upper bound and lower bound of Eq. (137):

$$
\frac{1}{2}\langle{\mathbf{H}_q}^{-1},\mathbf{H}_p\rangle\leq\left\langle{\mathbf{H}_q}^{-1}\mathbf{H}_p{\mathbf{H}_q}^{-1},\mathbf{H}_n(\theta_*)\right\rangle\leq\frac{3}{2}\langle{\mathbf{H}_q}^{-1},\mathbf{H}_p\rangle.
\tag{138}
$$

Combining Eqs. (138) and (135), we have

$$
\begin{aligned}
\frac{\langle{\mathbf{H}_q}^{-1},\mathbf{H}_p\rangle}{2n}&\leq\mathbb{E}_{\{y_i\sim p(y_i|x_i,\theta_*)\}_{i=1}^n}\left\langle{\mathbf{H}_q}^{-1}\mathbf{H}_p{\mathbf{H}_q}^{-1},\vec{\nabla}Q_n(\theta_*)\vec{\nabla}Q_n(\theta_*)^\top\right\rangle\\
&=\frac{1}{n}\left\langle{\mathbf{H}_q}^{-1}\mathbf{H}_p{\mathbf{H}_q}^{-1},\mathbf{H}_n(\theta_*)\right\rangle\leq\frac{3\langle{\mathbf{H}_q}^{-1},\mathbf{H}_p\rangle}{2n}.
\end{aligned}
\tag{139}
$$

Combining this with the upper bound Eq. (134) and lower bound Eq. (133), we can obtain with probability at least $1-\delta$,

$$
\frac{e^{-\alpha}+\alpha-1}{18\alpha^2}\frac{\langle{\mathbf{H}_q}^{-1},\mathbf{H}_p\rangle}{n}\leq\mathbb{E}[L_p(\theta_n)]-L_p(\theta_*)\leq\frac{24(e^{\alpha}-\alpha-1)}{\alpha^2}\frac{\langle{\mathbf{H}_q}^{-1},\mathbf{H}_p\rangle}{n}.
\tag{140}
$$

where the expectation $\mathbb{E}$ is w.r.t $\{y_i\sim p(y_i|x_i,\theta_*)\}_{i=1}^n$.

□

D Parameter discussion

In this section, we discuss the constants introduced in Lemma 2. In Proposition 35, we derive upper bounds for $K_{1,p}$ and $K_{2,p}(r)$ when Assumption 1 holds. If we additionally assume that $p(x)\sim\mathcal{N}(\mathbf{0},\mathbf{V}_p)$, then we can derive bounds for ρ, $K_{0,p}$, $K_{1,p}$ and $K_{2,p}(r)$ in Proposition 37. Note that we discuss constants for $p(x)$ here as example, but the results can be similarly extended to $q(x)$ if the same assumption holds for $q(x)$.

Proposition 35. *Suppose Assumption 1 holds for $p(x)$. ρ is the minimum constant defined in Theorem 3 such that $\mathbf{I}_{c-1}\otimes\mathbf{V}_p\preceq\rho\mathbf{H}_p$. Then*

(1) For $K_{1,p}$ defined in Lemma 2-(2), we have

$$K_{1,p} < 2\sqrt{\rho}K_{0,p}. \tag{141}$$

(2) For $K_{2,p}(r)$ defined in Lemma 2-(3), let $\rho(\theta) > 0$ be constant s.t. $\mathbf{I}_{c-1} \otimes \mathbf{V}_p \preceq \rho(\theta)\mathbf{H}_p(\theta)$ for $\theta \in \mathcal{B}_r(\theta_*)$, we have

$$K_{2,p}(r) < 2 \sup_{\theta \in \mathcal{B}_r(\theta_*)} \rho(\theta) K_{0,p}^2. \tag{142}$$

Proof. For the ease of notation, we use $\widetilde{c} = c - 1$ and $\widetilde{d} = d(c-1)$. We define $\mathbf{h}(x,\theta)\mathbb{R}^{\widetilde{c}}$ for a given $x \in \mathbb{R}^d$ and $\theta \in \mathbb{R}^{\widetilde{c} \times d}$ by

$$\mathbf{h}_i(x,\theta) = \frac{\exp(x^\top \theta_i)}{1 + \sum_{s \in [\widetilde{c}]} \exp(x^\top \theta_s)}, \qquad \forall i \in [\widetilde{c}] \tag{143}$$

where θ_i is the i-th row of θ.

(1) Denote $\widetilde{\mathbf{V}}_p \triangleq \mathbf{I}_{\widetilde{c}} \otimes \mathbf{V}_p$, then $\widetilde{\mathbf{V}}_p \preceq \rho \mathbf{H}_p$ and $\mathbf{H}_p{}^{-1/2} \preceq \sqrt{\rho}\widetilde{\mathbf{V}}_p^{-1/2}$. Thus

$$\|\mathbf{H}_p{}^{-1/2} \vec{\nabla} \ell_{(x,y)}(\theta_*)\|_{\psi_2} \leq \sqrt{\rho} \|\widetilde{\mathbf{V}}_p^{-1/2} \vec{\nabla} \ell_{(x,y)}(\theta_*)\|_{\psi_2}. \tag{144}$$

By Proposition 23, the i-th row ($i \in [\widetilde{c}]$) of matrix $\nabla \ell_{(x,y)}(\theta_*)$ is

$$[\nabla \ell_{(x,y)}(\theta_*)]_i = \frac{\partial \ell_{(x,y)}(\theta_*)}{\partial \theta_{*,i}} = \beta_i(x,y)x,$$

where $\beta_i(x,y) \triangleq -1_{\{y=i\}} + \mathbf{h}_i(x,\theta_*)$.

Therefore $\left(\vec{\nabla} \ell_{(x,y)}(\theta_*)\right)^\top = [\beta_1(x,y)x^\top, \beta_2(x,y)x^\top, \cdots, \beta_{\widetilde{c}}(x,y)x^\top]$ and thus

$$\begin{aligned}&\left(\widetilde{\mathbf{V}}_p^{-1/2} \vec{\nabla} \ell_{(x,y)}(\theta_*)\right)^\top \\ =&\left[\beta_1(x,y)(\mathbf{V}_p^{-1/2}x)^\top, \beta_2(x,y)(\mathbf{V}_p^{-1/2}x)^\top, \cdots, \beta_{\widetilde{c}}(x,y)(\mathbf{V}_p^{-1/2}x)^\top\right].\end{aligned} \tag{145}$$

We also observe that for any (x,y),

$$\sum_{i \in [\widetilde{c}]} |\beta_i(x,y)| \leq 1 + \frac{\sum_{j \in [\widetilde{c}]} \exp(x^\top \theta_j^*)}{1 + \sum_{j \in [\widetilde{c}]} \exp(x^\top \theta_j^*)} < 2. \tag{146}$$

By definition of the sub-Gaussian vector norm we have

$$\|\widetilde{\mathbf{V}}_p^{-1/2} \vec{\nabla} \ell_{(x,y)}(\theta_*)\|_{\psi_2} \triangleq \sup_{u \in \mathcal{S}^{d\widetilde{c}-1}} \|\langle \widetilde{\mathbf{V}}_p^{-1/2} \vec{\nabla} \ell_{(x,y)}(\theta_*), u\rangle\|_{\psi_2} \tag{147}$$

where $\mathcal{S}^{\widetilde{d}-1}$ is the unit sphere in $\mathbb{R}^{\widetilde{d}}$. For any $u \in \mathcal{S}^{d\widetilde{c}-1}$, we represent $u^\top = [u_1^\top, u_2^\top, \cdots, u_{\widetilde{c}}^\top]$, where $u_i \in \mathbb{R}^d$ for each $i \in [\widetilde{c}]$. Then for any $y \in [c]$, by Eq. (145) we have

$$\|\langle \widetilde{\mathbf{V}}_p^{-1/2} \vec{\nabla} \ell_{(x,y)}(\theta_*), u\rangle\|_{\psi_2} = \left\| \sum_{i \in [\widetilde{c}]} \beta_i(x,y) u_i^\top \mathbf{V}_p^{-1/2} x \right\|_{\psi_2}. \tag{148}$$

For a given x and $u \in \mathcal{S}^{\widetilde{d}-1}$, define

$$u(x) \in \operatorname*{arg\,max}_{u_i, i \in [\widetilde{c}]} |u_i^\top \mathbf{V}_p^{-1/2} x|, \tag{149}$$

where the choice of $u(x)$ does not effect our result. By Eq. (146),

$$\|\langle \widetilde{\mathbf{V}}_p^{-1/2} \vec{\nabla} \ell_{(x,y)}(\theta_*), u\rangle\|_{\psi_2} < 2\|(u(x))^\top \mathbf{V}_p^{-1/2} x\|_{\psi_2}. \tag{150}$$

Since $\|u(x)\| \leq 1$, by combining Eqs. (150) and (147) we can get

$$\|\widetilde{\mathbf{V}}_p^{-1/2} \vec{\nabla} \ell_{(x,y)}(\theta_*)\|_{\psi_2} < 2 \sup_{v \in \mathcal{S}^{d-1}} \|v^\top \mathbf{V}_p^{-1/2} x\|_{\psi_2} = 2\|\mathbf{V}_p^{-1/2} x\|_{\psi_2} \leq 2K_{0,p}. \tag{151}$$

(2) Let $\mathbf{W}_p(\theta) \triangleq \widetilde{\mathbf{V}}_p^{1/2}\mathbf{H}_p(\theta)^{-1/2}$, then $\mathbf{W}_p(\theta) \preceq \sqrt{\rho(\theta)}\mathbf{I}_{\widetilde{d}}$. First, we observe that

$$\begin{aligned}
&\sup_{u\in\mathcal{S}^{\widetilde{d}-1}} \|u^\top \mathbf{H}_p(\theta)^{-1/2}\nabla^2\ell_{(x,y)}(\theta)\mathbf{H}_p(\theta)^{-1/2}u\|_{\psi_1} \\
=&\sup_{\substack{v\triangleq\mathbf{W}_p(\theta)u\\ \|u\|_2\le 1}} \|v^\top \widetilde{\mathbf{V}}_p^{-1/2}\nabla^2\ell_{(x,y)}(\theta)\widetilde{\mathbf{V}}_p^{-1/2}v\|_{\psi_1} \\
\overset{(a)}{\le}&\sup_{\|u\|_2\le 1} \|(\sqrt{\rho(\theta)}u)^\top \widetilde{\mathbf{V}}_p^{-1/2}\nabla^2\ell_{(x,y)}(\theta)\widetilde{\mathbf{V}}_p^{-1/2}(\sqrt{\rho(\theta)}u)\|_{\psi_1} \\
\le&\rho(\theta)\sup_{u\in\mathcal{S}^{\widetilde{d}-1}} \|u^\top \widetilde{\mathbf{V}}_p^{-1/2}\nabla^2\ell_{(x,y)}(\theta)\widetilde{\mathbf{V}}_p^{-1/2}u\|_{\psi_1},
\end{aligned} \tag{152}$$

where (a) follows by the fact that $\lambda_{\max}(\mathbf{W}_p(\theta)) \le \sqrt{\rho(\theta)})$ and thus $\{v = \mathbf{W}_p(\theta))u : \|u\|_2 \le 1\} \subset \{\sqrt{\rho(\theta)}u : \|u\|_2 \le 1\}$.

By Proposition 23, we have the Hessian $\nabla^2\ell_{(x,y)}(\theta) \in \mathbb{R}^{\widetilde{d}\times\widetilde{d}}$ with the following form:

$$\nabla^2\ell_{(x,y)}(\theta) = \begin{bmatrix} \alpha_{11}(x,\theta)xx^\top & \cdots & \alpha_{1\widetilde{c}}(x,\theta)xx^\top \\ \vdots & \ddots & \vdots \\ \alpha_{\widetilde{c}1}(x,\theta)xx^\top & \cdots & \alpha_{\widetilde{c}\widetilde{c}}(x,\theta)xx^\top \end{bmatrix} \tag{153}$$

where

$$\alpha_{i,j}(\theta) = 1_{\{i=j\}}\mathbf{h}_i(x,\theta) - \mathbf{h}_i(x,\theta)\mathbf{h}_j(x,\theta). \tag{154}$$

For any $u \in \mathcal{S}^{\widetilde{d}-1}$, we decompose it into $\widetilde{c}$ chunks with dimension d, i.e. $u^\top = [u_1^\top, \cdots, u_{\widetilde{c}}^\top]$ and $u_i \in \mathbb{R}^d$. Since $\widetilde{\mathbf{V}}_p = \mathbf{I}_{\widetilde{c}} \otimes \mathbf{V}_p$, we have $\widetilde{\mathbf{V}}_p^{-1/2} = \mathbf{I}_{\widetilde{c}} \otimes \mathbf{V}_p^{-1/2}$. Define $\widetilde{u}_i \triangleq \mathbf{V}_p^{-1/2}u_i$, $\widetilde{u} \triangleq \widetilde{\mathbf{V}}_p^{-1/2}u$, then $\widetilde{u}^\top = [\widetilde{u}_1^\top, \cdots, \widetilde{u}_{\widetilde{c}}^\top]$. For the "sup" term in Eq. (152), we have

$$\begin{aligned}
&\sup_{u\in\mathcal{S}^{\widetilde{d}-1}} \|u^\top \widetilde{\mathbf{V}}_p^{-1/2}\nabla^2\ell_{(x,y)}(\theta)\widetilde{\mathbf{V}}_p^{-1/2}u\|_{\psi_1} = \sup_{u\in\mathcal{S}^{\widetilde{d}-1}} \|\widetilde{u}^\top\nabla^2\ell_{(x,y)}(\theta)\widetilde{u}\|_{\psi_1} \\
\overset{(a)}{=}&\sup_{u\in\mathcal{S}^{\widetilde{d}-1}} \Big\|\sum_{i\in[\widetilde{c}]}\sum_{j\in[\widetilde{c}]}\alpha_{ij}(x,\theta)\widetilde{u}_i^\top xx^\top\widetilde{u}_j\Big\|_{\psi_1} \\
\overset{(b)}{=}&\sup_{u\in\mathcal{S}^{\widetilde{d}-1}} \Big\|\sum_{i\in[\widetilde{c}]}\sum_{j\in[\widetilde{c}]}\alpha_{ij}(x,\theta)u_i^\top(\mathbf{V}_p^{-1/2}x)(\mathbf{V}_p^{-1/2}x)^\top u_j\Big\|_{\psi_1},
\end{aligned} \tag{155}$$

where (a) follows by Eq. (154), (b) follows by $\widetilde{u}_i = \mathbf{V}_p^{-1/2}u_i$.

Now we intend to upper bound Eq. (155) by using $\|\mathbf{V}_p^{-1/2}x\|_{\psi_2} \le K_{0,p}$. First for any $x \in \mathbb{R}$ and $u \in \mathcal{S}^{\widetilde{d}-1}$, we define

$$u(x) \in \arg\max_{u_i, i\in[\widetilde{c}]} \left|u_i^\top(\mathbf{V}_p^{-1/2}x)(\mathbf{V}_p^{-1/2}x)^\top u_i\right|,$$

where the choice of $u(x)$ does not effect our result. Since for any $a, b \in \mathbb{R}$, we have inequality $|ab| \le \frac{a^2+b^2}{2} \le \max\{a^2, b^2\}$, then

$$\left|u_i^\top(\mathbf{V}_p^{-1/2}x)(\mathbf{V}_p^{-1/2}x)^\top u_j\right| \le \left|u(x)^\top(\mathbf{V}_p^{-1/2}x)(\mathbf{V}_p^{-1/2}x)^\top u(x)\right|, \qquad \forall i,j \in [\widetilde{c}]. \tag{156}$$

On the other hand, by Eq. (154) we have

$$|\alpha_{ij}(x,\theta)| = \begin{cases} \mathbf{h}_i(x,\theta) - \mathbf{h}_i^2(x,\theta) & \text{if } i = j, \\ \mathbf{h}_i(x,\theta)\mathbf{h}_j(x,\theta) & \text{otherwise.} \end{cases} \tag{157}$$

Thus

$$\sum_{i\in[\widetilde{c}]}\sum_{j\in[\widetilde{c}]}|\alpha_{ij}(x,\theta)| = \sum_{i\in[\widetilde{c}]}\Big[\mathbf{h}_i(x,\theta) - \mathbf{h}_i^2(x,\theta) + \mathbf{h}_i(x,\theta)[\|\mathbf{h}(x,\theta)\|_1 - \mathbf{h}_i(x,\theta)]\Big]$$

$$
\begin{aligned}
&= \sum_{i\in[\widetilde{c}]} \Big[(1+\|\mathbf{h}(x,\theta)\|_1)\mathbf{h}_i(x,\theta) - 2\mathbf{h}_i^2(x,\theta)\Big] \\
&= (1+\|\mathbf{h}(x,\theta)\|_1)\|\mathbf{h}(x,\theta)\|_1 - 2\sum_{i\in[\widetilde{c}]} \mathbf{h}_i^2(x,\theta) \\
&< 2, \qquad (158)
\end{aligned}
$$

where the last inequality follows by the fact that $\|\mathbf{h}(x,\theta)\|_1 = 1 - \frac{1}{1+\sum_{s\in[\widetilde{c}]}\exp(x^\top\theta_s)} < 1$.

Now substitute Eq. (155) into Eq. (152), we can obtain that

$$
\begin{aligned}
&\sup_{u\in\mathcal{S}^{\widetilde{d}-1}} \|u^\top \mathbf{H}_p(\theta)^{-1/2}\nabla^2\ell_{(x,y)}(\theta)\mathbf{H}_p(\theta)^{-1/2}u\|_{\psi_1} \\
\leq&\rho(\theta)\sup_{u\in\mathcal{S}^{\widetilde{d}-1}} \Big\|\sum_{i\in[\widetilde{c}]}\sum_{j\in[\widetilde{c}]}\alpha_{ij}(x,\theta)u_i^\top(\mathbf{V}_p^{-1/2}x)(\mathbf{V}_p^{-1/2}x)^\top u_j\Big\|_{\psi_1} \\
\overset{(a)}{\leq}&\rho(\theta)\sup_{u\in\mathcal{S}^{\widetilde{d}-1}} \Big\|\Big(\sum_{i\in[\widetilde{c}]}\sum_{j\in[\widetilde{c}]}|\alpha_{ij}(x,\theta)|\Big)\Big(u(x)^\top(\mathbf{V}_p^{-1/2}x)(\mathbf{V}_p^{-1/2}x)^\top u(x)\Big)\Big\|_{\psi_1} \\
\overset{(b)}{<}&2\rho(\theta)\sup_{v\in\mathcal{S}^{d-1}}\|(v^\top\mathbf{V}_p^{-1/2}x)^2\|_{\psi_1} \\
\overset{(c)}{=}&2\rho(\theta)\sup_{v\in\mathcal{S}^{d-1}}\|(v^\top\mathbf{V}_p^{-1/2}x)\|_{\psi_2}^2 \\
=&2\rho(\theta)\|\mathbf{V}_p^{-1/2}x\|_{\psi_2}^2 \overset{(d)}{\leq} 2\rho(\theta)K_{0,p}^2, \qquad (159)
\end{aligned}
$$

where (a) follows by Eq. (156), (b) follows by Eq. (158) and the fact that $u(x)\in\mathbb{R}^d$ and $\|u(x)\|_2\leq 1$, (c) follows by Lemma 16, (d) follows by Lemma 2-(1). Comparing Eq. (159) to Eq. (5) (in Lemma 2-(3)), we can get

$$
K_{2,p}(r) < 2\sup_{\theta\in\mathcal{B}_r(\theta_*)}\sqrt{\rho(\theta)}K_{0,p}. \qquad (160)
$$

□

Before establishing the result for Gaussian design, we provide a form of Hessian expression of the loss function with respect to θ in the following lemma.

Lemma 36. *For any (x,y) and parameter θ, $\nabla^2\ell_{(x,y)}(\theta) = \widetilde{x}(\theta)\widetilde{x}(\theta)^\top$, where $\widetilde{x}(\theta) = (\ell''(y,\theta x))^{1/2}\otimes x$.*

Proof. The proof is trivial. By chain rule, $\nabla^2\ell_{(x,y)}(\theta) = \ell''(y,\theta x)\otimes xx^\top$. □

In the following proposition, we consider the case for a Gaussian design, i.e. $p(x)\sim\mathcal{N}(\mathbf{0},\mathbf{V}_p)$. In particular, we present the bounds for constants ρ, $K_{0,p}$, $K_{1,p}$ and $K_{2,p}(r)$ used in Theorem 3 by using θ_*, $\mathbf{V}_p$ and r. Our bound for ρ is inspired Proposition D.1 in [24], where the binary logistic regression on Gaussian design is considered.

Proposition 37 (Gaussian design)**.** *Suppose $p(x)\sim\mathcal{N}(\mathbf{0},\mathbf{V}_p)$, Assumption 1 holds for $p(x)$. Suppose that $\rho>0$ is the minimum constant such that $\widetilde{\mathbf{V}}_p \triangleq \mathbf{I}_{\widetilde{c}}\otimes\mathbf{V}_p \preceq \rho\mathbf{H}_p$, then for ρ and constant defined in Lemma 2, we have*

$$
\begin{aligned}
\rho &\lesssim \big(2+\max_{i\in[\widetilde{c}]}\|\theta_{*,i}\|_{\mathbf{V}_p}^2\big)^{3/2}, \qquad (161) \\
K_{0,p} &\lesssim 1, \qquad (162) \\
K_{1,p} &\lesssim \big(2+\max_{i\in[\widetilde{c}]}\|\theta_{*,i}\|_{\mathbf{V}_p}^2\big)^{3/4}, \qquad (163) \\
K_{2,p}(r) &\lesssim \big(2+r^2+\max_{i\in[\widetilde{c}]}\|\theta_{*,i}\|_{\mathbf{V}_p}^2\big)^{3/4}, \qquad (164)
\end{aligned}
$$

where $\theta_{*,i}$ is the i-th row of $\theta_*\in\mathbb{R}^{(c-1)\times d}$.

Proof.

(1) Proof of Eq. (161).

First, we consider the decorrelated design $z \triangleq \widetilde{\mathbf{V}}_p^{-1/2} x$, thus $z \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_{\widetilde{c}})$. Define parameter $\xi \triangleq \theta \widetilde{\mathbf{V}}_p^{1/2}$, and denote $\xi_* = \theta_* \widetilde{\mathbf{V}}_p^{1/2}$. Then we have $\theta x = \xi z$. By Lemma 36, we have

$$\mathbf{H}_p = \mathbf{H}_p(\theta_*) = \mathbb{E}_x[\widetilde{x}(\theta_*)\widetilde{x}(\theta_*)^\top], \tag{165}$$

where $\widetilde{x}(\theta) = [\ell''(y, \theta x)]^{1/2} \otimes x$, note that Hessian $\ell''(y, \theta x) \in \mathbb{R}^{\widetilde{c} \times \widetilde{c}}$ has no dependence on label y.

Now we define $\widetilde{z}(\xi) \triangleq \widetilde{\mathbf{V}}_p^{-1/2} \widetilde{x}(\theta)$, then

$$\begin{aligned} \widetilde{z}(\xi) &= (\mathbf{I}_{\widetilde{c}} \otimes \widetilde{\mathbf{V}}_p^{-1/2})([\ell''(y, \theta x)]^{1/2} \otimes x) = ([\ell''(y, \theta x)]^{1/2}) \otimes (\widetilde{\mathbf{V}}_p^{-1/2} x) \\ &= [\ell''(y, \xi z)]^{1/2} \otimes z. \end{aligned} \tag{166}$$

Then the covariance matrix of $\widetilde{z}(\xi_*)$ has the following form:

$$\begin{aligned} \mathbf{\Psi}(\xi_*) &\triangleq \mathbb{E}_z[\widetilde{z}(\xi_*)\widetilde{z}(\xi_*)^\top] \\ &= \mathbb{E}_z[\ell''(y, \xi_* z) \otimes (zz^\top)] \\ &= \widetilde{\mathbf{V}}_p^{-1/2} \mathbf{H}_p \widetilde{\mathbf{V}}_p^{-1/2}, \end{aligned} \tag{167}$$

where the last equality follows by definition of $\widetilde{z}(\xi_*)$ and Eq. (165). Thus, we can upper bound ρ by finding lower bound of $\lambda_{\min}(\mathbf{\Psi}(\xi_*))$ since by the definition of ρ, we have

$$\rho \le \frac{1}{\lambda_{\min}(\mathbf{\Psi}(\xi_*))}. \tag{168}$$

For any $z \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_{\widetilde{c}})$, we have

$$\ell''(y, \xi_* z) = \mathbf{\Gamma}(z) - \mathbf{h}(z)\mathbf{h}(z)^\top, \tag{169}$$

where $\mathbf{h}(z) \in \mathbb{R}^{\widetilde{c}}$ and

$$\mathbf{h}_i(z) = \frac{\exp(z^\top \xi_{*,i})}{1 + \sum_{j \in [\widetilde{c}]} \exp(z^\top \xi_{*,j})}, \tag{170}$$

and $\mathbf{\Gamma}(z) = \mathrm{diag}(\mathbf{h}_1(z), \mathbf{h}_2(z), \cdots, \mathbf{h}_{\widetilde{c}}(z))$. Thus for any $z \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_{\widetilde{c}})$,

$$\begin{aligned} \ell''(y, \xi_* z) &= \mathbf{\Gamma}(z)^{1/2} \Big[\mathbf{I}_{\widetilde{c}} - \big(\mathbf{\Gamma}(z)^{-1/2}\mathbf{h}(z)\big)\big(\mathbf{\Gamma}(z)^{-1/2}\mathbf{h}(z)\big)^\top\Big] \mathbf{\Gamma}(z)^{1/2} \\ &\succeq (1 - \|\mathbf{\Gamma}(z)^{-1/2}\mathbf{h}(z)\|_2^2)\mathbf{\Gamma}(z) \\ &= (1 - \|\mathbf{h}(z)\|_1)\mathbf{\Gamma}(z), \end{aligned} \tag{171}$$

where the last equality follows by the fact that the i-th component of $\mathbf{\Gamma}(z)^{-1/2}\mathbf{h}(z)$ is $\sqrt{\mathbf{h}_i(z)}$. Substitute this into Eq. (167), we can get

$$\mathbf{\Psi}(\xi_*) \succeq \mathbb{E}_z\Big[(1 - \|\mathbf{h}(z)\|_1)\mathbf{\Gamma}(z) \otimes (zz^\top)\Big]. \tag{172}$$

Note that $\mathbf{\Gamma}(z)$ is a diagonal matrix, we additionally have

$$\begin{aligned} \lambda_{\min}[\mathbf{\Psi}(\xi_*)] &= \lambda_{\min}\Big(\mathbb{E}_z\Big[(1 - \|\mathbf{h}(z)\|_1)\mathbf{\Gamma}(z) \otimes (zz^\top)\Big]\Big) \\ &= \min_{i \in [\widetilde{c}]} \lambda_{\min}\Big(\mathbb{E}_z\Big[\mathbf{h}_i(z)(1 - \|\mathbf{h}(z)\|_1)zz^\top\Big]\Big). \end{aligned} \tag{173}$$

For any arbitrary $i \in [\widetilde{c}]$, we have

$$\mathbf{h}_i(z)(1 - \|\mathbf{h}(z)\|_1) = \frac{\exp(z^\top \xi_{*,i})}{\Big(1 + \sum_{j \in [\widetilde{c}]} \exp(z^\top \xi_{*,j})\Big)^2}. \tag{174}$$

By the symmetry of $\mathcal{N}(\mathbf{0}, \mathbf{I}_{\widetilde{c}})$, w.l.o.g. we can assume that $\xi_{*,i}$ is parallel to e_1, where e_1 is the unit vector of the first coordinate. Thus we have $z^\top \xi_{*,i} = \|\xi_{*,i}\|_2 z_1$ and

$$\mathbf{h}_i(z)(1 - \|\mathbf{h}(z)\|_1) = \frac{\exp(t_i z_1)}{\Big(1 + \beta + \exp(t_i z_1)\Big)^2} \approx \exp(-|t_i z_1|), \tag{175}$$

where we use $\approx$ to represent the intersection of $\lesssim$ and $\gtrsim$, $\beta = \sum_{j \neq i} \exp(z^\top \xi_{*,j})$ and we define t_i by

$$t_i \triangleq \|\xi_{*,i}\|_2 = \|\theta_* \mathbf{V}_p^{1/2}\|_2 = \|\theta_*\|_{\mathbf{V}_p}. \tag{176}$$

Now by Eq. (175) we have

$$\begin{aligned} \mathbb{E}_z \Big[\mathbf{h}_i(z)(1 - \|\mathbf{h}(z)\|_1) z z^\top\Big] &\approx \mathbb{E}_{\{z_i \sim \mathcal{N}(0,1)\}_{i=1}^d} [\exp(-|t_i z_1|) z z^\top] \\ &= \begin{bmatrix} \kappa & \mathbf{0}_{d-1}^\top \\ \mathbf{0}_{d-1} & \kappa_\perp \mathbf{I}_{d-1}, \end{bmatrix} \end{aligned} \tag{177}$$

where κ and $\kappa_\perp$ have the following forms if we denote the standard one dimensional Gaussian density function as $\phi(\cdot)$:

$$\kappa = \int_{-\infty}^{\infty} \exp(-|t_i u|) u^2 \phi(u) du, \tag{178}$$

$$\kappa_\perp = \int_{-\infty}^{\infty} \exp(-|t_i u|) \phi(u) du. \tag{179}$$

By Eqs. (168), (173) and (177), we can upper bound ρ by finding the lower bounds for κ and $\kappa_\perp$. First we denote the Gaussian integral as $G(t) \triangleq \int_t^\infty e^{-u^2/2} du$, which has sharp bounds as

$$\frac{2e^{-t^2/2}}{t + \sqrt{t^2 + 4}} \leq G(t) \leq \frac{2e^{-t^2/2}}{t + \sqrt{t^2 + 8\pi}}, \qquad t \geq 0. \tag{180}$$

For κ, we have

$$\begin{aligned} \kappa &= \sqrt{\frac{2}{\pi}} \cdot \int_0^\infty e^{-t_i u - u^2} u^2 du = \sqrt{\frac{2}{\pi}} e^{t_i^2/2} \int_0^\infty e^{-(u+t_i)^2/2} u^2 du \\ &= \sqrt{\frac{2}{\pi}} \cdot e^{t_i^2/2} \int_{t_i}^\infty e^{-v^2/2} (v - t)^2 dv \\ &= \sqrt{\frac{2}{\pi}} \cdot e^{t_i^2/2} \big[(1 + t_i^2) G(t_i) - t_i e^{-t_i^2/2}\big]. \\ &\overset{\text{(a)}}{\gtrsim} \frac{2(t_i^2 + 1)}{t_i + \sqrt{t_i^2 + 4}} - t_i = \frac{t_i(t_i - \sqrt{t_i^2 + 4}) + 2}{t_i + \sqrt{t_i^2 + 4}} \\ &= \frac{2(\sqrt{t_i^2 + 4} - t_i)}{(\sqrt{t_i^2 + 4} + t_i)^2} = \frac{8}{(\sqrt{t_i^2 + 4} + t_i)^3} \geq \frac{1}{(t_i^2 + 2)^{3/2}}, \end{aligned} \tag{181}$$

where (a) follows by the lower bound of $G(t_i)$ from (180). Similarly for $\kappa_\perp$,

$$\begin{aligned} \kappa_\perp &= \sqrt{\frac{2}{\pi}} \cdot \int_0^\infty e^{-t_i u - u^2/2} du \\ &= \sqrt{\frac{2}{\pi}} e^{t_i^2/2} \cdot \int_{t_i}^\infty e^{-v^2/2} dv = \sqrt{\frac{2}{\pi}} e^{t_i^2/2} G(t_i) \\ &\gtrsim \frac{1}{(t_i^2 + 2)^{1/2}}. \end{aligned} \tag{182}$$

Combining (177), (181) and (182), we can get for each $i \in [\widetilde{c}]$,

$$\lambda_{\min}\Big(\mathbb{E}_z \Big[\mathbf{h}_i(z)(1 - \|\mathbf{h}(z)\|_1) z z^\top\Big]\Big) \gtrsim \min\{\kappa, \kappa_\perp\} \gtrsim \frac{1}{(t_i^2 + 2)^{3/2}}. \tag{183}$$

Substitute this into (173), we have

$$\lambda_{\min}[\mathbf{\Psi}(\xi_*)] \gtrsim \min_{i\in[\widetilde{c}]} \frac{1}{(t_i^2+2)^{3/2}}. \tag{184}$$

Combining this with the bound of ρ in (168) and the definition of t_i in (176), we can obtain that

$$\rho \leq \frac{1}{\lambda_{\min}[\mathbf{\Psi}(\xi_*)]} \lesssim \max_{i\in[\widetilde{c}]}(2+\|\theta_{*,i}\|^2_{\mathbf{V}_p})^{3/2} = \left(2+\max_{i\in[\widetilde{c}]}\|\theta_{*,i}\|^2_{\mathbf{V}_p}\right)^{3/2}. \tag{185}$$

(2) Since $x \sim \mathcal{N}(\mathbf{0}, \mathbf{V}_p)$, $\mathbf{V}_p^{-1/2}x \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_d)$. For any $u \in \mathcal{S}^{d-1}$, $u^\top \mathbf{V}_p^{-1/2}x \sim \mathcal{N}(0,1)$. Thus

$$\|\mathbf{V}_p^{-1/2}x\|_{\psi_2} = \sup_{u\in\mathcal{S}^{d-1}} \|u^\top \mathbf{V}_p^{-1/2}x\|_{\psi_2} \lesssim 1 \tag{186}$$

and $K_{0,p} \lesssim 1$.

(3) Substitute Eqs. (161) and (162) into Eq. (141), we have

$$K_{1,p} < 2\sqrt{\rho}K_{0,p} \lesssim \left(2+\max_{i\in[\widetilde{c}]}\|\theta_{*,i}|^2_{\mathbf{V}_p}\right)^{3/4}. \tag{187}$$

(4) Substitute Eqs. (161) and (162) into Eq. (142), we have

$$\begin{aligned} K_{2,p}(r) &< 2\sup_{\theta\in\mathcal{B}_r(\theta_*)} \rho(\theta)K_{0,p}^2 \\ &\lesssim \sup_{\max_{i\in[\widetilde{c}]}\|\theta_i-\theta_{*,i}\|_{\mathbf{V}_p}\leq r} (2+\max_{i\in[\widetilde{c}]}\|\theta_i\|^2_{\mathbf{V}_p})^{3/4} \\ &\lesssim \left(2+r^2+\max_{i\in[\widetilde{c}]}\|\theta_{*,i}\|^2_{\mathbf{V}_p}\right)^{3/4}, \end{aligned} \tag{188}$$

where the last inequality follows by the triangle inequality $\|\theta_i\|_{\mathbf{V}_p} \leq \|\theta_i - \theta_{*,i}\|_{\mathbf{V}_p} + \|\theta_{*,i}\|_{\mathbf{V}_p}$.

□

E Bounded domain

For the case of bounded domain, we present the assumptions in Assumption 38, which are similar to the regularity assumptions used in [11]. Then we present the excess risk $L_p(\theta_n) - L_p(\theta_*)$ bounds in Theorem 40. Our proof is inspired by the proof of Theorem 5.1 in [28].

Assumption 38. *There exist constants L_1, L_2 and $L_3 > 0$, for any sample (x, y) randomly drawn from distribution $\pi_p(x, y)$ or $\pi_q(x, y)$, the following conditions are satisfied:*

(1) $\mathbf{H}_p$ and $\mathbf{H}_q$ are positive definite.

(2) gradient and Hessian of loss function with respect to θ at θ_* are bounded:

$$\|\mathrm{vec}(\nabla\ell_{(x,y)}(\theta_*))\|_{\mathbf{H}_p{}^{-1}} \leq L_1, \qquad \|\mathbf{H}_p{}^{-1/2}\nabla^2\ell_{(x,y)}(\theta_*)\mathbf{H}_p{}^{-1/2}\| \leq L_2, \tag{189}$$

(3) Lipschitz continuity of Hessian: there exits a neighborhood around θ_* denoted by $\mathcal{B}(\theta_*)$ such that $\forall\theta' \in \mathcal{B}(\theta_*)$,

$$\left\|\mathbf{H}_p{}^{-1/2}\left(\nabla^2\ell_{(x,y)}(\theta_*) - \nabla^2\ell_{(x,y)}(\theta')\right)\mathbf{H}_p{}^{-1/2}\right\| \leq L_3\|\mathrm{vec}(\theta_* - \theta')\|_{\mathbf{H}_p}. \tag{190}$$

Remark 39. *We did not explicitly assume that $x \in \mathbb{R}^d$ is bounded. However, by Proposition 23, each row of gradient $\nabla_{(}x, y)(\theta_*)$ is the scaling of x. Thus Assumption 38-(2) assumes that x is bounded implicitly.*

Theorem 40. *Suppose Assumption 38 holds. Let $\sigma > 0$ be the constant such that $\mathbf{H}_p \preceq \sigma\mathbf{H}_q$. For any $\delta \in (0, 1)$, whenever*

$$n \geq 256\max\left\{L_2^2\sigma^2\log(2d(c-1)/\delta), \log(1/\delta)\sigma^4L_1^2L_3^2\right\}, \tag{191}$$

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

$$\frac{3}{8}\frac{(1-\epsilon_p)}{(1+\epsilon_q)^2}\frac{\mathrm{Trace}(\mathbf{H}_q{}^{-1}\mathbf{H}_p)}{n} \le \mathbb{E}[L_p(\theta_n)] - L_p(\theta_*) \le \frac{5}{8}\frac{(1+\epsilon_p)}{(1-\epsilon_q)^2}\frac{\mathrm{Trace}(\mathbf{H}_q{}^{-1}\mathbf{H}_p)}{n}, \tag{192}$$

where $\mathbb{E}$ *is the expectation over* $\{y_i \sim p(y_i|x_i,\theta_*)\}_{i=1}^n$*,* ϵ_p *and* ϵ_q *are given by*

$$\epsilon_p = 2\sigma^2 L_1 L_3 \sqrt{\frac{2+8\log(1/\delta)}{n}} \qquad \epsilon_q = 4\sigma L_2 \sqrt{\frac{\log(2d(c-1)/\delta)}{n}} + 2\sigma^2 L_1 L_3 \sqrt{\frac{2+8\log(1/\delta)}{n}}. \tag{193}$$

Remark 41. *For Theorem 40, if Eq. (191) holds, we can upper bound* ϵ_p *and* ϵ_q*. This results in a simpler upper bound for the excess risk with respect to* $p(x)$*:*

$$\mathbb{E}[L_p(\theta_n)] - L_p(\theta_*) \le \frac{9}{5}\frac{\mathrm{Trace}(\mathbf{H}_q{}^{-1}\mathbf{H}_p)}{n}. \tag{194}$$

We show this at the end of the proof of Theorem 40.

proof of Theorem 40. We deploy the notation of $Q_n(\theta)$ and $\mathbf{H}_n(\theta)$ defined in Eqs. (58) and (59) for the ease of notation. *Throughout the whole proof, we treat parameter as vector*, i.e. $\theta \in \mathbb{R}^{\widetilde{d}}$. Denote the samples drawn from $\pi_q(x,y)$ by $\{z_i = (x_i,y_i) \overset{\text{i.i.d}}{\sim} \pi_q(x,y)\}_{i=1}^n$. Since $\mathbf{H}_p \preceq \sigma\mathbf{H}_q$, for a vector $v \in \mathbb{R}^{\widetilde{d}}$ we have

$$\|v\|_{\mathbf{H}_q{}^{-1}} \le \sqrt{\sigma}\|v\|_{\mathbf{H}_p{}^{-1}}, \qquad \|v\|_{\mathbf{H}_p} \le \sqrt{\sigma}\|v\|_{\mathbf{H}_q}. \tag{195}$$

For the ease of notation, we define norms for a matrix $\mathbf{A} \in \mathbb{R}^{\widetilde{d}\times\widetilde{d}}$ by

$$\|\mathbf{A}\|_P \triangleq \|\mathbf{H}_p{}^{-1/2}\mathbf{A}\mathbf{H}_p{}^{-1/2}\|, \qquad \|\mathbf{A}\|_Q \triangleq \|\mathbf{H}_q{}^{-1/2}\mathbf{A}\mathbf{H}_q{}^{-1/2}\|. \tag{196}$$

Note that for a matrix symmetric semi-positive definite matrix $\mathbf{A} \in \mathbb{S}_+^{\widetilde{d}}$,

$$\begin{aligned}\mathbf{H}_q{}^{-1/2}\mathbf{A}\mathbf{H}_q{}^{-1/2} &= (\mathbf{H}_q{}^{-1/2}\mathbf{H}_p{}^{1/2})(\mathbf{H}_p{}^{-1/2}\mathbf{A}\mathbf{H}_p{}^{-1/2})(\mathbf{H}_p{}^{1/2}\mathbf{H}_q{}^{-1/2})\\ &\preceq \sigma\mathbf{H}_p{}^{-1/2}\mathbf{A}\mathbf{H}_p{}^{-1/2}\end{aligned} \tag{197}$$

where the last inequality follows by the fact $\lambda_{\max}(\mathbf{H}_q{}^{-1/2}\mathbf{H}_p{}^{1/2}) = \sqrt{\sigma}$. Thus we have the following relation between these two norms:

$$\|\mathbf{A}\|_Q \le \sigma\|\mathbf{A}\|_P. \tag{198}$$

step 1. We aim to choose a ball $\mathcal{B}_1(\theta_*)$ centered at θ_* and n sufficiently large such that for any $\theta \in \mathcal{B}_1(\theta_*)$, $\mathbf{H}_n(\theta)$ approximates $\mathbf{H}_q$ in the spectral sense with high probability.

First, we have by triangle inequality that

$$\|\mathbf{H}_n(\theta) - \mathbf{H}_q\|_Q \le \|\mathbf{H}_n(\theta) - \mathbf{H}_n(\theta_*)\|_Q + \|\mathbf{H}_n(\theta_*) - \mathbf{H}_q\|_Q. \tag{199}$$

To bound the first term in Eq. (199), we can use Assumption 38-(3), i.e. if $\theta \in \mathcal{B}(\theta_*)$, then

$$\|\mathbf{H}_n(\theta) - \mathbf{H}_n(\theta_*)\|_Q \overset{Eq.\ (198)}{\le} \sigma\|\mathbf{H}_n(\theta) - \mathbf{H}_n(\theta_*)\|_P \le \sigma L_3\|\theta - \theta_*\|_{\mathbf{H}_p}. \tag{200}$$

Now we consider the second term on the right hand side of Eq. (199). Let $\mathbf{X}_i = \mathbf{H}_p{}^{-1/2}\big(\nabla^2\ell_{z_i}(\theta_*) - \mathbf{H}_q\big)\mathbf{H}_p{}^{-1/2}$ for each $i \in [n]$ and $\mathbf{S} = \frac{1}{n}\sum_{i=1}^n \mathbf{X}_i$. Since $\mathbb{E}[\nabla^2\ell_{z_i}(\theta_*)] = \nabla^2 L_q(\theta_*) = \mathbf{H}_q$, then $\mathbb{E}[\mathbf{X}_i] = 0$. By Eq. (189), we have $\|\nabla^2\ell_{z_i}(\theta_*)\|_P \le L_2$. Thus for any $i \in [n]$:

$$\begin{aligned}\|\mathbf{X}_i\| = \|\nabla^2\ell_{z_i}(\theta_*) - \mathbf{H}_q\|_P \le 2L_2,\\ \|\,\mathbb{E}(\mathbf{X}_i^2)\| \le \mathbb{E}\,\|\mathbf{X}_i^2\| \le \mathbb{E}\,\|\mathbf{X}_i\|^2 \le 4L_2^2.\end{aligned} \tag{201}$$

Let $\mu = 2L_2$ and $\nu = 4L_2^2$ in the matrix Bernstein inequality (i.e. **??**), we have with probability at least $1-\delta$,

$$\|\mathbf{S}\| \le 4L_2\sqrt{\frac{\log(2\widetilde{d}/\delta)}{n}} \triangleq \epsilon_1. \tag{202}$$

Note that $\|\mathbf{H}_n(\theta_*) - \mathbf{H}_q\|_P = \|\mathbf{S}\|$. Then with probability at least $1-\delta$,

$$\|\mathbf{H}_n(\theta_*) - \mathbf{H}_q\|_Q \leq \sigma \|\mathbf{H}_n(\theta_*) - \mathbf{H}_p\|_P \leq \sigma\epsilon_1. \tag{203}$$

Substitute Eqs. (200) and (203) into Eq. (199), we can get

$$\|\mathbf{H}_n(\theta) - \mathbf{H}_q\|_Q \leq \sigma L_3 \|\theta - \theta_*\|_{\mathbf{H}_p} + \sigma\epsilon_1. \tag{204}$$

Now consider a ball centered at θ_*:

$$\mathcal{B}_1(\theta_*) \triangleq \{\theta : \|\theta - \theta_*\|_{\mathbf{H}_p} \leq \frac{1}{4\sigma L_3}\},$$

then $\sigma L_3 \|\theta - \theta_*\|_{\mathbf{H}_q} \leq 1/4$ for any $\theta \in \mathcal{B}_1(\theta_*)$. Besides, if we choose n such that

$$n \geq 256 L_2^2 \sigma^2 \log(2\widetilde{d}/\delta), \tag{205}$$

we have

$$\epsilon_1 \leq \frac{1}{4\sigma}. \tag{206}$$

Substitute Eq. (206) into Eq. (204), we have $\|\mathbf{H}_n(\theta) - \mathbf{H}_q\|_Q \leq 1/2$ and thus with probability at least $1-\delta$,

$$\frac{1}{2}\mathbf{H}_q \preceq \mathbf{H}_n(\theta) \preceq \frac{3}{2}\mathbf{H}_q. \tag{207}$$

step 2. Next we show that when n is large enough, $\theta_n \in \mathcal{B}_1(\theta_*)$ with high probability. Given θ, by Taylor's expansion there exits $\tilde{\theta}$ between θ and θ_* such that

$$Q_n(\theta) = Q_n(\theta_*) + \nabla Q_n(\theta_*)^\top (\theta - \theta_*) + \frac{1}{2}(\theta - \theta_*)^\top \nabla^2 Q_n(\tilde{\theta})(\theta - \theta_*).$$

Then for all $\theta \in \mathcal{B}_1(\theta_*)$,

$$\begin{aligned}
Q_n(\theta) - Q_n(\theta_*) &= \nabla Q_n(\theta_*)^\top (\theta - \theta_*) + \frac{1}{2}\|\theta - \theta_*\|^2_{\mathbf{H}_n(\tilde{\theta})} \\
&\overset{(a)}{\geq} \nabla Q_n(\theta_*)^\top (\theta - \theta_*) + \frac{1}{4}\|\theta - \theta_*\|^2_{\mathbf{H}_q} \\
&\overset{(b)}{\geq} \|\theta - \theta_*\|_{\mathbf{H}_q} \left(\frac{1}{4}\|\theta - \theta_*\|_{\mathbf{H}_q} - \|\nabla Q_n(\theta_*)\|_{\mathbf{H}_q{}^{-1}}\right) \\
&\overset{(c)}{\geq} \|\theta - \theta_*\|_{\mathbf{H}_q} \left(\frac{1}{4\sqrt{\sigma}}\|\theta - \theta_*\|_{\mathbf{H}_p} - \sqrt{\sigma}\|\nabla Q_n(\theta_*)\|_{\mathbf{H}_p{}^{-1}}\right)
\end{aligned} \tag{208}$$

where (a) follows by Eq. (207), (b) follows by Cauchy-Schwartz inequality, and (c) follows by Eq. (195).

Now if we can show for all $\theta \in \partial\mathcal{B}_1\theta_*)$, the right hand side of Eq. (208) is non negative, then $\theta_n \in \mathcal{B}_1(\theta_*)$ because $Q_n(\theta)$ is a convex function. Let $\xi_i = \mathbf{H}_p{}^{-1/2}\nabla \ell_{z_i}(\theta_*)$ and $S = \frac{1}{n}\sum_{i=1}^n \xi_i$. Then $\mathbb{E}[\xi_i] = \mathbf{H}_p{}^{-1/2}\nabla L_p(\theta_*) = 0$ by Lemma 24. By Assumption 38-(2), for any $i \in [n]$ we have

$$\begin{aligned}
\|\xi_i\| = \|\nabla \ell_{z_i}(\theta_*)\|_{\mathbf{H}_p{}^{-1}} &\leq L_1, \\
\mathbb{E}[\|\xi_i\|^2] &\leq L_1^2.
\end{aligned} \tag{209}$$

Let $\mu = L_1$ and $\nu = L_1^2$ in the vector Bernstein inequality (i.e. **??**), with probability at least $1-\delta$ we have

$$\|\nabla Q_n(\theta_*)\|_{\mathbf{H}_p{}^{-1}} = \|S\| \leq L_1\sqrt{\frac{2 + 8\log(1/\delta)}{n}} \triangleq \epsilon_2. \tag{210}$$

Now if we choose n such that

$$n \geq 256(2 + 8\log(1/\delta))\sigma^4 L_1^2 L_3^2,$$

then

$$\epsilon_2 \leq \frac{1}{16 L_3 \sigma^2}. \tag{211}$$

Thus for all $\theta \in \partial\mathcal{B}_1(\theta_*)$, combining Eqs. (208), (210) and (211) we have

$$\begin{aligned} Q_n(\theta) - Q_n(\theta_*) &\geq \|\theta - \theta_*\|_{\mathbf{H}_q} \left(\frac{1}{4\sqrt{\sigma}} \|\theta - \theta_*\|_{\mathbf{H}_p} - \sqrt{\sigma} \|\nabla Q_n(\theta_*)\|_{{\mathbf{H}_p}^{-1}} \right) \\ &\geq \|\theta - \theta_*\|_{\mathbf{H}_q} \left(\frac{1}{4\sqrt{\sigma}} \frac{1}{4\sigma L_3} - \sqrt{\sigma} \frac{1}{16\sigma^2 L_3} \right) = 0. \end{aligned} \tag{212}$$

Then with probability at least $1-\delta$, $\theta_n \in B_1(\theta_*)$.

step 3. We denote $\Delta \triangleq \theta_n - \theta_*$, then by Taylor's theorem, there exits $\widetilde{\theta}_n$ between θ_n and θ_* such that

$$0 = \nabla Q_n(\theta_n) = \nabla Q_n(\theta_*) + \mathbf{H}_n(\widetilde{\theta}_n)\Delta. \tag{213}$$

In this step, we get a spectral relation between $\mathbf{H}_n(\widetilde{\theta}_n)$ and $\mathbf{H}_q$.

We have ensured that $\mathbf{H}_n(\widetilde{\theta}_n)$ is positive definite in step 1 (by Eq. (207)), thus

$$\Delta = -\big(\mathbf{H}_n(\widetilde{\theta}_n)\big)^{-1} \nabla Q_n(\theta_*), \tag{214}$$

and with probability at least $1-\delta$ we have

$$\begin{aligned} \|\Delta\|_{\mathbf{H}_q} &= (\Delta^\top \mathbf{H}_q \Delta)^{1/2} = [\nabla Q_n(\theta_*)^\top \big(\mathbf{H}_n(\widetilde{\theta}_n)\big)^{-1} \mathbf{H}_q \big(\mathbf{H}_n(\widetilde{\theta}_n)\big)^{-1} \nabla Q_n(\theta_*)]^{1/2} \\ &= \left[\left(\nabla Q_n(\theta_*)^\top {\mathbf{H}_q}^{-1/2} \right) \left({\mathbf{H}_q}^{1/2} \big(\mathbf{H}_n(\widetilde{\theta}_n)\big)^{-1} \mathbf{H}_q \big(\mathbf{H}_n(\widetilde{\theta}_n)\big)^{-1} {\mathbf{H}_q}^{1/2} \right) \left({\mathbf{H}_q}^{-1/2} \mathbf{H}_n(\theta_*) \right) \right]^{1/2} \\ &\leq \|{\mathbf{H}_q}^{1/2} \big(\mathbf{H}_n(\widetilde{\theta}_n)\big)^{-1} \mathbf{H}_q \big(\mathbf{H}_n(\widetilde{\theta}_n)\big)^{-1} {\mathbf{H}_q}^{1/2}\|^{1/2} \|{\mathbf{H}_q}^{-1/2} \nabla Q_n(\theta_*)\| \\ &\leq \|{\mathbf{H}_q}^{1/2} \big(\mathbf{H}_n(\widetilde{\theta}_n)\big)^{-1} {\mathbf{H}_q}^{1/2}\| \|\nabla Q_n(\theta_*)\|_{{\mathbf{H}_q}^{-1}} \\ &\overset{(a)}{\leq} 2\sqrt{\sigma} \|\nabla Q_n(\theta_*)\|_{{\mathbf{H}_p}^{-1}} \\ &\overset{(b)}{\leq} 2\sqrt{\sigma}\epsilon_2, \end{aligned} \tag{215}$$

where (a) follows by Eq. (195) and $1/2\mathbf{H}_p \preceq \mathbf{H}_n(\widetilde{\theta}_n)$ from Eq. (207) since $\widetilde{\theta}_n \in \mathcal{B}(\theta_*)$, (b) follows by Eq. (210).

Denote $\widetilde{\Delta} \triangleq \widetilde{\theta}_n - \theta_*$, since $\widetilde{\theta}_n$ lies between θ_n and θ_*, we have

$$\|\widetilde{\Delta}\|_{\mathbf{H}_q} \leq \|\Delta\|_{\mathbf{H}_q} \leq 2\sqrt{\sigma}\epsilon_2. \tag{216}$$

Following a similar argument as step 1, we can obtain that

$$\begin{aligned} \|\mathbf{H}_n(\widetilde{\theta}_n) - \mathbf{H}_q\|_Q &\leq \|\mathbf{H}_n(\widetilde{\theta}_n) - \mathbf{H}_n(\theta_*)\|_Q + \|\mathbf{H}_n(\theta_*) - \mathbf{H}_q\|_Q \\ &\leq \sigma \|\mathbf{H}_n(\widetilde{\theta}_n) - \mathbf{H}_n(\theta_*)\|_P + \sigma\epsilon_1 \\ &\leq \sigma L_3 \|\widetilde{\Delta}\|_{\mathbf{H}_p} + \sigma\epsilon_1 \\ &\overset{(a)}{\leq} 2\sigma^2 L_3 \epsilon_2 + \sigma\epsilon_1 \triangleq \epsilon_q, \end{aligned} \tag{217}$$

where (a) follows by Eq. (216) and the fact that $\|\widetilde{\Delta}\|_{\mathbf{H}_p} \leq \sqrt{\sigma}\|\widetilde{\Delta}\|_{\mathbf{H}_q}$. Note that we can upper bound ϵ_q by using Eqs. (206) and (211):

$$\epsilon_q = 2\sigma^2 L_3 \epsilon_2 + \sigma\epsilon_1 \leq \frac{3}{8}. \tag{218}$$

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

$$(1-\epsilon_q)\mathbf{H}_q \preceq \mathbf{H}_n(\widetilde{\theta}_n) \leq (1+\epsilon_q)\mathbf{H}_q. \tag{219}$$

step 4. Now we use Taylor's expansion to get bounds for $L_p(\theta_n) - L_p(\theta_*)$. By Taylor's theorem, there exits $\widetilde{z}_n$ between θ_n and θ_* such that

$$L_p(\theta_n) - L_p(\theta_*) = \frac{1}{2}\|\Delta\|^2_{\mathbf{H}_p(\widetilde{z}_n)}, \tag{220}$$

where the first order term vanishes because $\nabla L_p(\theta_*) = 0$ by Lemma 24.

From the Lipschitz condition Assumption 38-(3), we have

$$\|\mathbf{H}_p(\widetilde{z}_n) - \mathbf{H}_p\|_P \le L_3\|\widetilde{z}_n - \theta_*\|_{\mathbf{H}_p} \overset{(a)}{\le} 2\sigma^2 L_3 \epsilon_2 \triangleq \epsilon_p,$$

where inequality (a) follows by

$$\|\widetilde{z}_n - \theta_*\|_{\mathbf{H}_p} \le \|\Delta\|_{\mathbf{H}_p} \overset{Eq.\ (195)}{\le} \sqrt{\sigma}\|\Delta\|_{\mathbf{H}_q} \overset{Eq.\ (215)}{\le} 2\sigma^2\epsilon_2.$$

Note that we can upper bound ϵ_p by using Eq. (211):

$$\epsilon_p = 2\sigma^2 L_3 \epsilon_2 \le \frac{1}{8}. \tag{221}$$

Thus,

$$(1-\epsilon_p)\mathbf{H}_p \preceq \mathbf{H}_p(\widetilde{z}_n) \le (1+\epsilon_p)\mathbf{H}_p. \tag{222}$$

Define matrices $\mathbf{M}_{q,n}$ and $\mathbf{M}_{p,n}$ as follows:

$$\begin{aligned}\mathbf{M}_{q,n} &\triangleq {\mathbf{H}_q}^{1/2}\big(\mathbf{H}_n(\widetilde{\theta}_n)\big)^{-1}{\mathbf{H}_q}^{1/2},\\ \mathbf{M}_{p,n} &\triangleq {\mathbf{H}_p}^{-1/2}\mathbf{H}_p(\widetilde{z}_n){\mathbf{H}_p}^{-1/2}.\end{aligned}$$

By Eqs. (219) and (222), we have

$$\lambda_{\max}(\mathbf{M}_{q,n}) \le \frac{1}{1-\epsilon_q}, \qquad \lambda_{\min}(\mathbf{M}_{q,n}) \ge \frac{1}{1+\epsilon_q}, \tag{223}$$

$$\lambda_{\max}(\mathbf{M}_{p,n}) \le (1+\epsilon_p), \qquad \lambda_{\min}(\mathbf{M}_{p,n}) \ge (1-\epsilon_p). \tag{224}$$

Now we can bound the excess risk $L_p(\theta_n) - L_p(\theta_*)$ by using the Taylor expansion in Eq. (220):

$$\begin{aligned}L_p(\theta_n) - L_p(\theta_*) &= \frac{1}{2}\Delta^\top \mathbf{H}_p(\widetilde{z}_n)\Delta\\ &= \frac{1}{2}\Delta^\top {\mathbf{H}_p}^{1/2}\Big({\mathbf{H}_p}^{-1/2}\mathbf{H}_p(\widetilde{z}_n){\mathbf{H}_p}^{-1/2}\Big){\mathbf{H}_p}^{1/2}\Delta\\ &= \frac{1}{2}\Delta^\top {\mathbf{H}_p}^{1/2}\mathbf{M}_{p,n}{\mathbf{H}_p}^{1/2}\Delta.\end{aligned} \tag{225}$$

Observe that,

$$\begin{aligned}&\Delta^\top \mathbf{H}_p \Delta\\ =&\Delta^\top \mathbf{H}_n(\widetilde{\theta}_n){\mathbf{H}_q}^{-1/2}\underbrace{\big({\mathbf{H}_q}^{1/2}\big(\mathbf{H}_n(\widetilde{\theta}_n)\big)^{-1}\mathbf{H}_p\big(\mathbf{H}_n(\widetilde{\theta}_n)\big)^{-1}{\mathbf{H}_q}^{1/2}\big)}_{\triangleq \mathbf{M}}{\mathbf{H}_q}^{-1/2}\mathbf{H}_n(\widetilde{\theta}_n)\Delta,\end{aligned} \tag{226}$$

and

$$\begin{aligned}\mathbf{M} &= \big({\mathbf{H}_q}^{1/2}\big(\mathbf{H}_n(\widetilde{\theta}_n)\big)^{-1}{\mathbf{H}_q}^{1/2}\big)\big({\mathbf{H}_q}^{-1/2}\mathbf{H}_p{\mathbf{H}_q}^{-1/2}\big)\big({\mathbf{H}_q}^{1/2}\mathbf{H}_n(\widetilde{\theta}_n)\big)^{-1}{\mathbf{H}_q}^{1/2}\big)\\ &= \mathbf{M}_{q,n}\big({\mathbf{H}_q}^{-1/2}\mathbf{H}_p{\mathbf{H}_q}^{-1/2}\big)\mathbf{M}_{q,n}.\end{aligned} \tag{227}$$

Substitute Eq. (227) into Eq. (226), we have

$$\Delta^\top \mathbf{H}_p \Delta = \big(\Delta^\top \mathbf{H}_n(\widetilde{\theta}_n){\mathbf{H}_q}^{-1/2}\big)\mathbf{M}_{q,n}\big({\mathbf{H}_q}^{-1/2}\mathbf{H}_p{\mathbf{H}_q}^{-1/2}\big)\mathbf{M}_{q,n}\big({\mathbf{H}_q}^{-1/2}\mathbf{H}_n(\widetilde{\theta}_n)\Delta\big). \tag{228}$$

Based on the previous steps, with probability at least $1-\delta$, we have a lower bound for $L_p(\theta_n) - L_p(\theta_*)$ by Eq. (225):

$$L_p(\theta_n) - L_p(\theta_*)$$

$$
\begin{aligned}
&= \frac{1}{2} \Delta^\top {\mathbf{H}_p}^{1/2} \mathbf{M}_{p,n} {\mathbf{H}_p}^{1/2} \Delta \\
&\geq \frac{1}{2} \lambda_{\min}(\mathbf{M}_{p,n}) \Delta^\top \mathbf{H}_p \Delta \\
&\overset{(228)}{\geq} \frac{1}{2} \lambda_{\min}(\mathbf{M}_{p,n}) \big(\Delta^\top \mathbf{H}_n(\widetilde{\theta}_n) {\mathbf{H}_q}^{-1/2}\big) \mathbf{M}_{q,n} \big({\mathbf{H}_q}^{-1/2} \mathbf{H}_p {\mathbf{H}_q}^{-1/2}\big) \mathbf{M}_{q,n} \big({\mathbf{H}_q}^{-1/2} \mathbf{H}_n(\widetilde{\theta}_n) \Delta\big) \\
&\geq \frac{1}{2} \lambda_{\min}(\mathbf{M}_{p,n}) \lambda^2_{\min}(\mathbf{M}_{q,n}) \big(\Delta^\top \mathbf{H}_n(\widetilde{\theta}_n) {\mathbf{H}_q}^{-1} \mathbf{H}_p {\mathbf{H}_q}^{-1} \mathbf{H}_n(\widetilde{\theta}_n) \Delta\big) \\
&\geq \frac{1}{2} \frac{(1-\epsilon_p)}{(1+\epsilon_q)^2} \big\langle {\mathbf{H}_q}^{-1} \mathbf{H}_p {\mathbf{H}_q}^{-1}, \nabla Q_n(\theta_*) \nabla Q_n(\theta_*)^\top \big\rangle, \qquad (229)
\end{aligned}
$$

where the last inequality follows by Eqs. (223) and (224), and the fact that $\mathbf{H}_n(\widetilde{\theta}_n)\Delta = -\nabla Q_n(\theta_*)$ from Eq. (214).

By similar argument, we can get an upper bound:

$$
\begin{aligned}
L_p(\theta_n) - L_p(\theta_*) &\leq \frac{1}{2} \lambda_{\max}(\mathbf{M}_{p,n}) \lambda^2_{\max}(\mathbf{M}_{q,n}) \big(\Delta^\top \mathbf{H}_n(\widetilde{\theta}_n) {\mathbf{H}_q}^{-1} \mathbf{H}_p {\mathbf{H}_q}^{-1} \mathbf{H}_n(\widetilde{\theta}_n) \Delta\big) \\
&\leq \frac{1}{2} \frac{(1+\epsilon_p)}{(1-\epsilon_q)^2} \big\langle {\mathbf{H}_q}^{-1} \mathbf{H}_p {\mathbf{H}_q}^{-1}, \nabla Q_n(\theta_*) \nabla Q_n(\theta_*)^\top \big\rangle. \qquad (230)
\end{aligned}
$$

Following the same argument as we derive Eq. (135) in Appendix C.4, given $\{x_i\}_{i=1}^n$, we have

$$
\mathbb{E}_{\{y_i \sim p(y_i|x_i,\theta_*)\}_{i=1}^n} [\nabla Q_n(\theta_*) \nabla Q_n(\theta_*)^\top] = \frac{1}{n} \mathbf{H}_n(\theta_*). \qquad (231)
$$

Now if we take conditional expectation on both sides of Eqs. (229) and (230), we can obtain that

$$
\begin{aligned}
\frac{1}{2} \frac{(1-\epsilon_p)}{(1+\epsilon_q)^2} \frac{\big\langle {\mathbf{H}_q}^{-1} \mathbf{H}_p {\mathbf{H}_q}^{-1}, \mathbf{H}_n(\theta_*) \big\rangle}{n} &\leq \mathbb{E}_{\{y_i \sim p(y_i|x_i,\theta_*)\}_{i=1}^n} [L_p(\theta_n) - L_p(\theta_*)] \\
&\leq \frac{1}{2} \frac{(1+\epsilon_p)}{(1-\epsilon_q)^2} \frac{\big\langle {\mathbf{H}_q}^{-1} \mathbf{H}_p {\mathbf{H}_q}^{-1}, \mathbf{H}_n(\theta_*) \big\rangle}{n}. \qquad (232)
\end{aligned}
$$

From the analysis in step 1, we have with probability at least $1-\delta$,

$$
\|\mathbf{H}_n(\theta_*) - \mathbf{H}_q\|_Q \leq \sigma \epsilon_1 \leq \frac{1}{4}, \qquad (233)
$$

where the last inequality follows by Eq. (206). Thus

$$
\frac{3}{4} \mathbf{H}_q \preceq \mathbf{H}_n(\theta_*) \preceq \frac{5}{4} \mathbf{H}_q, \qquad (234)
$$

and

$$
\frac{3}{4} \operatorname{Trace}({\mathbf{H}_q}^{-1} \mathbf{H}_p) \leq \big\langle {\mathbf{H}_q}^{-1} \mathbf{H}_p {\mathbf{H}_q}^{-1}, \mathbf{H}_n(\theta_*) \big\rangle \leq \frac{5}{4} \operatorname{Trace}({\mathbf{H}_q}^{-1} \mathbf{H}_p). \qquad (235)
$$

Substitute Eq. (235) into Eq. (232), we have with probability at least $1-\delta$,

$$
\frac{3}{8} \frac{(1-\epsilon_p)}{(1+\epsilon_q)^2} \frac{\operatorname{Trace}({\mathbf{H}_q}^{-1} \mathbf{H}_p)}{n} \leq \mathbb{E}[L_p(\theta_n)] - L_p(\theta_*) \leq \frac{5}{8} \frac{(1+\epsilon_p)}{(1-\epsilon_q)^2} \frac{\operatorname{Trace}({\mathbf{H}_q}^{-1} \mathbf{H}_p)}{n}, \qquad (236)
$$

where $\mathbb{E}$ is the expectation over $\{y_i \sim p(y_i|x_i,\theta_*)\}_{i=1}^n$.

Note that, with the upper bounds given in Eqs. (218) and (221), we can additionally bound the upper bound of Eq. (236):

$$
\begin{aligned}
\mathbb{E}_{[}\, L_p(\theta_n)] - L_p(\theta_*) &\leq \frac{5}{8} \frac{(1+\epsilon_p)}{(1-\epsilon_q)^2} \frac{\operatorname{Trace}({\mathbf{H}_q}^{-1} \mathbf{H}_p)}{n} \\
&\leq \frac{5}{8} \frac{1+1/8}{(1-3/8)^2} \frac{\operatorname{Trace}({\mathbf{H}_q}^{-1} \mathbf{H}_p)}{n} \\
&= \frac{9}{5} \frac{\operatorname{Trace}({\mathbf{H}_q}^{-1} \mathbf{H}_p)}{n}. \qquad (237)
\end{aligned}
$$

□

F Proofs of Section 4

Notation. For a positive integer k, let $\mathbb{S}^k$ be the cone of symmetric matrices with dimension $k \times k$, $\mathbb{S}^k_+$ be the cone of symmetric semi-positive definite matrices with dimension $k \times k$, and $\mathbb{S}^k_{++}$ be the cone of symmetric positive definite matrices with dimension $k \times k$.

F.1 Proof of Lemma 5

Proof. 1. We can verify convexity by considering an arbitrary line, given by $\mathbf{Z}+t\mathbf{V}$, where $\mathbf{Z} \in \mathbb{S}^{\widetilde{d}}_{++}$ and $\mathbf{V} \in \mathbb{S}^{\widetilde{d}}$. We define $g(t) = f(\mathbf{Z}+t\mathbf{V})$, where t is restricted to the interval such that $\mathbf{Z}+t\mathbf{V} \in \mathbb{S}^{\widetilde{d}}_{++}$. From covex analysis, it is sufficient for us to prove the convexity of function g. We have

$$g(t) = \langle (\mathbf{Z}+t\mathbf{V})^{-1}, \mathbf{H}_p(\theta_0)\rangle$$
$$= \operatorname{Trace}\left(\mathbf{Z}^{1/2}\mathbf{H}_p(\theta_0)\mathbf{Z}^{1/2}\left(\mathbf{I}+t\mathbf{Z}^{-1/2}\mathbf{V}\mathbf{Z}^{-1/2}\right)^{-1}\right). \tag{238}$$

We can write $\mathbf{Z}^{-1/2}\mathbf{V}\mathbf{Z}^{-1/2}$ in its eigendecomposition form, i.e. $\mathbf{Z}^{-1/2}V\mathbf{Z}^{-1/2} = \mathbf{Q}\mathbf{\Sigma}\mathbf{Q}^\top$, where $\mathbf{\Sigma} = \operatorname{diag}\{\lambda_1, \cdots, \lambda_{\widetilde{d}}\}$. Then we have

$$g(t) = \operatorname{Trace}\left(\mathbf{Z}^{1/2}\mathbf{H}_p(\theta_0)\mathbf{Z}^{1/2}\mathbf{Q}\left(\mathbf{I}+t\mathbf{\Sigma}\right)^{-1}\mathbf{Q}^\top\right)$$
$$= \operatorname{Trace}\left(\left(\mathbf{Q}^\top\mathbf{Z}^{1/2}\mathbf{H}_p(\theta_0)\mathbf{Z}^{1/2}\mathbf{Q}\right)\left(\mathbf{I}+t\mathbf{\Sigma}\right)^{-1}\right)$$
$$= \sum_{i=1}^{\widetilde{d}} \frac{1}{1+t\lambda_i}\left[\mathbf{Q}^\top\mathbf{Z}^{1/2}\mathbf{H}_p(\theta_0)\mathbf{Z}^{1/2}\mathbf{Q}\right]_{ii}, \tag{239}$$

and thus

$$g''(t) = \sum_{i=1}^{\widetilde{d}} \frac{2\lambda_i^2}{(1+t\lambda_i)^3}\left[\mathbf{Q}^\top\mathbf{Z}^{1/2}\mathbf{H}_p(\theta_0)\mathbf{Z}^{1/2}\mathbf{Q}\right]_{ii} \tag{240}$$

Since $\mathbf{Z}+t\mathbf{V}$ is positive definite, so is $\mathbf{I}+t\mathbf{Z}^{-1/2}\mathbf{V}\mathbf{Z}^{-1/2}$. Thus $1+t\lambda_i > 0$ for all $i \in [\widetilde{d}]$. In addition, $\mathbf{Q}^\top\mathbf{Z}^{1/2}\mathbf{H}_p(\theta_0)\mathbf{Z}^{1/2}\mathbf{Q}$ is also positive definite, then its diagonals are all positive. Thus $g(t)'' \geq 0$ by Eq. (240), we conclude that g is convex, and thus f is convex.

2. If $\mathbf{A} \preceq \mathbf{B}$, then $\mathbf{B}^{-1} - \mathbf{A}^{-1} \preceq \mathbf{0}$. Thus $\langle \mathbf{B}^{-1} - \mathbf{A}^{-1}, \mathbf{H}_p(\theta_0)\rangle \leq 0$ since $\mathbf{H}_p(\theta_0)$ is positive definite, i.e.

$$f(\mathbf{A}) \geq f(\mathbf{B}). \tag{241}$$

3. Property 3 is trivial to prove.

□

F.2 Solving relaxed problem by entropic mirror descent

We present the algorithm for solving relaxed problem Eq. (14) using entropic mirror descent in Algorithm 2. Let $z = b\kappa$, then Eq. (14) is equivalent to:

$$\kappa_\diamond = \mathop{\arg\min}_{\substack{\kappa\in\mathbb{R}^m_+ \\ \|\kappa\|_1=1}} f(\kappa) \triangleq \Big\langle \Big(\sum_{i\in[m]} \kappa_i \mathbf{H}(x_i)\Big)^{-1}, \mathbf{H}_p(\theta_0)\Big\rangle. \tag{242}$$

Line 5 of the algorithm computes the gradient of $f(\kappa)$:

$$g_i \triangleq \frac{\partial f(\kappa)}{\partial \kappa_i} = -\big\langle \mathbf{H}(x_i), \mathbf{\Sigma}^{-1}\mathbf{H}_p(\theta_0)\mathbf{\Sigma}^{-1}\big\rangle, \tag{243}$$

where $\mathbf{\Sigma} = \sum_{i\in[m]} \kappa_i \mathbf{H}(x_i)$. We present the convergence rate of the algorithm in Theorem 42, which is adopted from Theorem 5.1 in [29].

Algorithm 2 RELAXSOLVE(b, $\mathbf{H}_p(\theta_0)$, $\{\mathbf{H}(x_i)\}_{i\in[m]}$)

```
Output: $z_\diamond$
$\kappa = (1/m, 1/m, \cdots, 1/m) \in \mathbb{R}^m$
for $t = 1$ to $T$ do                                    // $T$ is iteration number
    $\beta_t \leftarrow \mathcal{O}(\sqrt{\frac{\log m}{t}})$
    $\mathbf{\Sigma} \leftarrow \sum_{i\in[m]} \kappa_i \mathbf{H}(x_i)$
    $g_i \leftarrow -\langle \mathbf{H}(x_i), \mathbf{\Sigma}^{-1}\mathbf{H}_p(\theta_0)\mathbf{\Sigma}^{-1} \rangle, \forall i \in [m]$
    $\kappa_i \leftarrow \kappa_i \exp(-\beta_t g_i)$
    $\kappa_i \leftarrow \frac{\kappa_i}{\sum_{j\in[m]} \kappa_j}$
end for
$z_\diamond \leftarrow b\kappa$
```

Theorem 42. *Suppose $f : \mathbb{R}^n \supseteq \mathcal{X} \to \mathbb{R}$ is convex Lipschitz continuous function w.r.t $\|\cdot\|_1$, i.e. $|f(x) - f(y)| \leq L_f\|x - y\|_1$. Consider using entropic mirror descent algorithm with T steps and step size $\eta_t = \frac{1}{L_f}\sqrt{\frac{2\log n}{T}}$, denote solution at step t as x_t. Then we have*

$$\min_{1\leq t\leq T} f(x_t) - \min_{x\in\mathcal{X}} f(x) \leq L_f\sqrt{\frac{2\log n}{T}}. \tag{244}$$

F.3 Proof of Proposition 8

We first introduce the background of the regret minimization problem in Appendix F.3.1. Note that in this section, we consider that the loss matrix $\mathbf{F}_t$ at each step t can be any symmetric, semi-positive definite matrix (i.e. $\mathbf{F}_t \in \mathbb{S}_+^{\widetilde{d}}$). This is more general than the case of $\mathbf{F}_t \in \{\widetilde{\mathbf{H}}(x_i)\}_{i=1}^m$ in § 4.3. Then we give the proof of Proposition 8 in Appendix F.3.2.

F.3.1 Background of regret minimization

We introduce a regret minimization problem in the adversarial linear bandits setting with full information. Consider a game of b rounds. At each round $t \in [b]$:

- the player chooses an action $\mathbf{A}_t \in \Delta_{\widetilde{d}}$, where $\Delta_{\widetilde{d}} = \{\mathbf{A} \in \mathbb{R}^{\widetilde{d}\times\widetilde{d}} : \mathbf{A} \succeq \mathbf{0}, \mathrm{Trace}(\mathbf{A}) = 1\}$
- afterwards, the environment reveals a loss matrix $\mathbf{F}_t \in \mathbb{S}_+^{\widetilde{d}}$
- the loss $\langle \mathbf{A}_t, \mathbf{F}_t \rangle$ is incurred

The goal of the player is to minimize the *regret* over all rounds, which is defined by

$$\mathrm{Regret}(\{\mathbf{A}_t\}_{t=1}^b) \triangleq \sum_{t=1}^b \langle \mathbf{A}_t, \mathbf{F}_t \rangle - \inf_{\mathbf{U}\in\Delta_{\widetilde{d}}} \langle \mathbf{U}, \sum_{t=1}^b \mathbf{F}_t \rangle. \tag{245}$$

The regret represents the excess loss compared to the loss incurred by a single optimal action $\mathbf{U} \in \Delta_{\widetilde{d}}$ in hindsight. In our setting, the loss incurred by a single optimal action is actually the minimum eigenvalue of the summed matrix of the loss matrices. We remark this property in Lemma 43.

Lemma 43. *For any $\mathbf{A} \in \mathbb{S}_+^{\widetilde{d}}$, $\lambda_{\min}(\mathbf{A}) = \inf_{\mathbf{U}\in\Delta_{\widetilde{d}}} \langle \mathbf{U}, \mathbf{A} \rangle$.*

Proof. Since $\mathbf{A} \in \mathbb{S}_+^{\widetilde{d}}$, we have eigendecomposition $\mathbf{A} = \mathbf{V}\mathbf{\Lambda}\mathbf{V}^\top$, where $\mathbf{\Lambda} = \mathrm{diag}\{\lambda_1, \cdots, \lambda_{\widetilde{d}}\}$. Assume that $\lambda_1 \geq \cdots \geq \lambda_{\widetilde{d}} \geq 0$ and $\mathbf{v}_i$ is the eigenvector asscoiated with eigenvalue λ_i for $i \in [\widetilde{d}]$.

We first show $\lambda_{\min}(\mathbf{A}) \geq \inf_{\mathbf{U}\in\Delta_{\widetilde{d}}} \langle \mathbf{U}, \mathbf{A} \rangle$. Let $\mathbf{B} = \mathbf{v}_{\widetilde{d}}\mathbf{v}_{\widetilde{d}}^\top$, then $\mathbf{B} \succeq \mathbf{0}$ and $\mathrm{Trace}(\mathbf{B}) = 1$, i.e. $\mathbf{B} \in \Delta_{\widetilde{d}}$. Thus

$$\inf_{\mathbf{U}\in\Delta_{\widetilde{d}}} \langle \mathbf{U}, \mathbf{A} \rangle \leq \langle \mathbf{B}, \mathbf{A} \rangle = \mathbf{v}_{\widetilde{d}}^\top \mathbf{V}\mathbf{\Lambda}\mathbf{V}^\top \mathbf{v}_{\widetilde{d}} = \lambda_{\widetilde{d}} = \lambda_{\min}(\mathbf{A}). \tag{246}$$

On the other hand, for any $\mathbf{U} \in \Delta_{\widetilde{d}}$, we have

$$\begin{aligned} \langle \mathbf{U}, \mathbf{A} \rangle &= \langle \mathbf{U}, \sum_{i \in [\widetilde{d}]} \lambda_i \mathbf{v}_i \mathbf{v}_i^\top \rangle = \sum_{i \in [\widetilde{d}]} \lambda_i \mathbf{v}_i^\top \mathbf{U} \mathbf{v}_i \\ &\geq \lambda_{\widetilde{d}} \sum_{i \in [\widetilde{d}]} \mathbf{v}_i^\top \mathbf{U} \mathbf{v}_i = \lambda_{\widetilde{d}} \langle \mathbf{U}, \mathbf{V}\mathbf{V}^\top \rangle = \lambda_{\widetilde{d}} \operatorname{Trace}(\mathbf{U}) = \lambda_{\widetilde{d}}. \end{aligned} \tag{247}$$

Since Eq. (247) holds for any $\mathbf{U} \in \Delta_{\widetilde{d}}$, then

$$\lambda_{\min}(\mathbf{A}) \leq \inf_{\mathbf{U} \in \Delta_{\widetilde{d}}} \langle \mathbf{U}, \mathbf{A} \rangle. \tag{248}$$

Combining Eq. (246) and Eq. (248), we can get $\lambda_{\min}(\mathbf{A}) = \inf_{\mathbf{U} \in \Delta_{\widetilde{d}}} \langle \mathbf{U}, \mathbf{A} \rangle$.

□

Follow-The-Regularized-Leader (FTRL). FTRL algorithm chooses action $\mathbf{A}_t$ at the beginning of each round based on the previous loss matrices $\{\mathbf{F}_l\}_{l=1}^{t-1}$. In particular, for a given regularizer $w(\cdot)$ and learning rate $\eta > 0$.

$$\mathbf{A}_1 = \operatorname*{arg\,min}_{\mathbf{A} \in \Delta_{\widetilde{d}}} w(\mathbf{A}), \qquad \mathbf{A}_t = \operatorname*{arg\,min}_{\mathbf{A} \in \Delta_{\widetilde{d}}} \left\{ \eta \sum_{l=1}^{t-1} \langle \mathbf{A}, \mathbf{F}_l \rangle + w(\mathbf{A}) \right\} \quad (t \geq 2). \tag{249}$$

We deploy the $\ell_{1/2}$-regularizer introduced by [14]:$w(\mathbf{A}) = -2 \operatorname{Trace}(\mathbf{A}^{1/2})$. Under such a regularizer, we can derive the closed form for $\mathbf{A}_t$, i.e. Eq. (17).

F.3.2 Proof of Proposition 8

Proof. By Theorem 28.4 in [17], we have an upper bound for regret as following:

$$\operatorname{Regret}(\{\mathbf{A}_t\}_{t=1}^b) \triangleq \sum_{t=1}^b \langle \mathbf{A}_t, \mathbf{F}_t \rangle - \inf_{\mathbf{U} \in \Delta_{\widetilde{d}}} \langle \mathbf{U}, \sum_{t=1}^b \mathbf{F}_t \rangle \leq \frac{\operatorname{diam}_w(\Delta_{\widetilde{d}})}{\eta} + \frac{1}{\eta} \sum_{t=1}^b D_w(\mathbf{A}_t, \tilde{\mathbf{A}}_{t+1}), \tag{250}$$

where $\operatorname{diam}_w(\Delta_{\widetilde{d}}) \triangleq \max_{\mathbf{A},\mathbf{B} \in \Delta_{\widetilde{d}}} w(\mathbf{A}) - w(\mathbf{B})$ is the diameter of $\Delta_{\widetilde{d}}$ with respect to w , D_w is w-induced Bregman divergence, and $\tilde{\mathbf{A}}_{t+1}$ is defined by

$$\tilde{\mathbf{A}}_{t+1} = \operatorname*{arg\,min}_{\mathbf{A} \succeq \mathbf{0}} \big\{ \eta \langle \mathbf{A}, \mathbf{F}_t \rangle + D_w(\mathbf{A}, \mathbf{A}_t) \big\}. \tag{251}$$

Since the regularizer $w(\mathbf{A}) = -2 \operatorname{Trace}(\mathbf{A}^{1/2})$ for any $\mathbf{A} \succeq \mathbf{0}$, $w(\mathbf{A})$ is differentiable and it has gradient $\nabla w(\mathbf{A}) = -\mathbf{A}^{-1/2}$. By definition of Bregman divergence, we have for any $\mathbf{A}, \mathbf{B} \succeq \mathbf{0}$:

$$\begin{aligned} D_w(\mathbf{A}, \mathbf{B}) &= w(\mathbf{A}) - w(\mathbf{B}) - \langle \mathbf{A} - \mathbf{B}, \nabla w(\mathbf{B}) \rangle \\ &= -2 \operatorname{Trace}(\mathbf{A}^{1/2} + 2 \operatorname{Trace}(\mathbf{B}^{1/2}) + \langle \mathbf{A} - \mathbf{B}, \mathbf{B}^{-1/2} \rangle \\ &= \langle \mathbf{A}, \mathbf{B}^{-1/2} \rangle + \operatorname{Trace}(\mathbf{B}^{1/2}) - 2 \operatorname{Trace}(\mathbf{A}^{1/2}). \end{aligned} \tag{252}$$

Substitute Eq. (252) into (251), we can get

$$\tilde{\mathbf{A}}_{t+1} = \operatorname*{arg\,min}_{\mathbf{A} \succeq \mathbf{0}} \big\{ \eta \langle \mathbf{A}, \mathbf{F}_t \rangle + \langle \mathbf{A}, \mathbf{A}_t^{-1/2} \rangle + \operatorname{Trace}(\mathbf{A}_t^{1/2}) - 2 \operatorname{Trace}(\mathbf{A}^{1/2}) \big\} \triangleq g(\mathbf{A}).$$

By the first order optimality condition of convex optimization, we have

$$\eta \mathbf{F}_t + \mathbf{A}_t^{-1/2} - \tilde{\mathbf{A}}_{t+1}^{-1/2} = 0,$$

and thus $\tilde{\mathbf{A}}_{t+1} = (\mathbf{A}_t^{-1/2} + \eta \mathbf{F}_t)^{-2}$. Therefore, by Eq. (252)

$$\begin{aligned} D_w(\mathbf{A}_t, \tilde{\mathbf{A}}_{t+1}) &= \langle \mathbf{A}_t, \tilde{\mathbf{A}}_{t+1}^{-1/2} \rangle + \operatorname{Trace}(\tilde{\mathbf{A}}_{t+1}^{1/2}) - 2 \operatorname{Trace}(\mathbf{A}_t^{1/2}) \\ &= \langle \mathbf{A}_t, \mathbf{A}_t^{-1/2} + \eta \mathbf{F}_t \rangle + \operatorname{Trace}[(\mathbf{A}_t^{-1/2} + \eta \mathbf{F}_t)^{-1}] - 2 \operatorname{Trace}(\mathbf{A}_t^{1/2}) \end{aligned}$$

$$= \langle \mathbf{A}_t, \eta \mathbf{F}_t \rangle + \mathrm{Trace}[(\mathbf{A}_t^{-1/2} + \eta \mathbf{F}_t)^{-1} - \mathbf{A}_t^{1/2}]. \tag{253}$$

Substitute Eq. (253) into Eq. (250), we can get

$$\lambda_{\min}(\sum_{t=1}^{b} \mathbf{F}_t) \overset{(a)}{=} \inf_{\mathbf{U} \in \Delta_{\widetilde{d}}} \langle \mathbf{U}, \sum_{t=1}^{b} \mathbf{F}_t \rangle \geq -\frac{\mathrm{diam}_w(\Delta_{\widetilde{d}})}{\eta} + \frac{1}{\eta} \sum_{t=1}^{b} \mathrm{Trace}[\mathbf{A}_t^{1/2} - (\mathbf{A}_t^{-1/2} + \eta \mathbf{F}_t)^{-1}]$$

$$\overset{(b)}{\geq} -\frac{2\sqrt{\widetilde{d}}}{\eta} + \frac{1}{\eta} \sum_{t=1}^{b} \mathrm{Trace}[\mathbf{A}_t^{1/2} - (\mathbf{A}_t^{-1/2} + \eta \mathbf{F}_t)^{-1}], \tag{254}$$

where equality (a) follows by Lemma 43 and inequality (b) follows by the fact that $\mathrm{diam}_w(\Delta_{\widetilde{d}}) \leq 2\sqrt{\widetilde{d}}$.

Since Eq. (254) holds for any $\mathbf{F}_t \in \mathbb{S}_+^{\widetilde{d}}$, then let $\mathbf{F}_t \in \{\widetilde{\mathbf{H}}(x_i)\}_{i \in [m]}$ and Eq. (18) is proved. □

F.4 Proof of Proposition 9

In Appendix F.4.1, we present some key inequalities that we need for the proof. In Appendix F.4.2, we present the full proof of Proposition 9. It is worth noting that a similar property to Proposition 9 is proven in [14]. However, in their setting, the loss matrices are rank-1 matrices, specifically of the form $\widetilde{x}_i \widetilde{x}_i^\top$, where $\widetilde{x}_i$ is a vector. On the other hand, in our setting, the loss matrices are transformed Fisher information matrices (i.e. $\widetilde{\mathbf{H}}(x_i)$, as defined in Equation 15). This distinction significantly complicates the derivation of a general result such as Eq. (24) in Proposition 9. The proof is by no means trivial. We remark that we do *not* assume special structure on points from unlabeled pool $U = \{x_i\}_{i \in [m]}$ and the ground truth parameter θ_* in our proof to Proposition 9.

F.4.1 Supporting Lemmas

Lemma 44. *For any* $i \in [m]$, $a_i > 0$, $b_i > 0$, $\pi_i \geq 0$, *then* $\max_{i \in [m]} \frac{a_i}{b_i} \geq \frac{\sum_{i \in [m]} \pi_i a_i}{\sum_{i \in [m]} \pi_i b_i}$.

Proof. We can use induction to prove the inequality. If $n = 2$, without loss of generality, we can assume $a_1/b_1 \geq a_2/b_2$, then

$$a_1 b_2 \geq a_2 b_1$$
$$\pi_1 a_1 b_1 + \pi_2 a_1 b_2 \geq \pi_1 a_1 b_1 + \pi_2 a_2 b_1$$

and

$$\max\{\frac{a_1}{b_1}, \frac{a_2}{b_2}\} = \frac{a_1}{b_1} \geq \frac{\pi_1 a_1 + \pi_2 a_2}{\pi_1 b_1 + \pi_2 b_2}.$$

Suppose the inequality is satisfied when $n = m - 1$, i.e.

$$\max_{i \in [m-1]} \frac{a_i}{b_i} \geq \frac{\sum_{i \in [m-1]} \pi_i a_i}{\sum_{i \in [m-1]} \pi_i b_i}. \tag{255}$$

When $n = m$,

$$\max_{i \in [m]} \frac{a_i}{b_i} = \max\Big\{ \max_{i \in [m-1]} \frac{a_i}{b_i}, \frac{a_m}{b_m} \Big\} \geq \max\left\{ \frac{\sum_{i \in [m-1]} \pi_i a_i}{\sum_{i \in [m-1]} \pi_i b_i}, \frac{a_m}{b_m} \right\}$$
$$\geq \frac{\sum_{i \in [m]} \pi_i a_i}{\sum_{i \in [m]} \pi_i b_i}.$$

The last inequality follows by the previous derivation when $n = 2$. Thus by induction, the inequality is proved for any positive integer n. □

Lemma 45. *For any* $i \in [m]$, $a_i \geq 0$, $b_i \geq 0$, *then* $\sum_{i \in [m]} \frac{a_i}{1+b_i} \geq \frac{\sum_{i \in [m]} a_i}{1+\sum_{i \in [m]} b_i}$.

Proof. We can use induction to prove this inequality. When $n = 2$,

$$\begin{aligned}
&[a_1(1+b_2) + a_2(1+b_1)](1+b_1+b_2) \\
&= a_1(1+b_2)(1+b_1) + a_1 b_2(1+b_2) + a_2(1+b_1)(1+b_2) + a_2 b_1(1+b_1) \\
&= (a_1+a_2)(1+b_1)(1+b_2) + a_1 b_2(1+b_2) + a_2 b_1(1+b_1) \\
&\geq (a_1+a_2)(1+b_1)(1+b_2).
\end{aligned} \tag{256}$$

Divide $(1+b_1)(1+b_2)(1+b_1+b_2)$ on both sides of Eq. (256), we can get

$$\begin{aligned}
\frac{a_1}{1+b_1} + \frac{a_2}{1+b_2} &= \frac{[a_1(1+b_2) + a_2(1+b_1)](1+b_1+b_2)}{(1+b_1)(1+b_2)(1+b_1+b_2)} \\
&\overset{Eq.\ (256)}{\geq} \frac{(a_1+a_2)(1+b_1)(1+b_2)}{(1+b_1)(1+b_2)(1+b_1+b_2)} = \frac{a_1+a_2}{1+b_1+b_2}.
\end{aligned} \tag{257}$$

Suppose the inequality is satisfied when $n = m-1$, i.e.

$$\sum_{i \in m-1} \frac{a_i}{1+b_i} \geq \frac{\sum_{i\in[m-1]} a_i}{1 + \sum_{i\in[m-1]} b_i}. \tag{258}$$

When $n = m$,

$$\begin{aligned}
\sum_{i\in[m]} \frac{a_i}{1+b_i} = \sum_{i\in[m-1]} \frac{a_i}{1+b_i} + \frac{a_m}{1+b_m} &\overset{Eq.\ (258)}{\geq} \frac{\sum_{i\in[m-1]} a_i}{1+\sum_{i\in[m-1]} b_i} + \frac{a_m}{1+b_m} \\
&\overset{Eq.\ (257)}{\geq} \frac{\sum_{i\in[m]} a_i}{1+\sum_{i\in[m]} b_i}.
\end{aligned} \tag{259}$$

□

Lemma 46. *For any matrices* $\mathbf{A}, \mathbf{B} \in \mathbb{S}_+^p$, *we have*

$$\langle (\mathbf{I}+\mathbf{B})^{-1}, \mathbf{A} \rangle \geq \frac{\mathrm{Trace}(\mathbf{A})}{1+\mathrm{Trace}(\mathbf{B})}. \tag{260}$$

Proof. Denote eigenvalues of matrix $\mathbf{A}$ as $\alpha_1 \geq \alpha_2 \geq \cdots \geq \alpha_p \geq 0$ and eigenvalues of matrix $\mathbf{B}$ as $\beta_1 \geq \beta_2 \geq \cdots \geq \beta_p \geq 0$. Then eigenvalues of $(\mathbf{I}+\mathbf{B})^{-1}$ are $0 \leq 1+\beta_1)^{-1} \leq (1+\beta_2)^{-1} \leq \cdots \leq (1+\beta_p)^{-1}$. Thus we have

$$\begin{aligned}
\langle (\mathbf{I}+\mathbf{B})^{-1}, \mathbf{A} \rangle &\overset{(a)}{\geq} \sum_{i=1}^{p} \frac{\alpha_i}{1+\beta_i} \\
&\overset{(b)}{\geq} \frac{\sum_{i=1}^p \alpha_i}{1+\sum_{i=1}^p \beta_i} = \frac{\mathrm{Trace}(\mathbf{A})}{1+\mathrm{Trace}(\mathbf{B})},
\end{aligned} \tag{261}$$

where inequality (a) follows by the lower bound of Von Neumann's trace inequality [30], inequality (b) follows by Lemma 45.

□

F.4.2 Proof of Proposition 9

Proof. Recall that in § 4.3, we define $\mathbf{B}_t$ by

$$\mathbf{B}_t^{-1/2} = \mathbf{A}_t^{-1/2} + \alpha\widetilde{\mathbf{D}}, \tag{262}$$

where $\widetilde{\mathbf{D}} = (\mathbf{\Sigma}_\diamond)^{-1/2}\mathbf{D}(\mathbf{\Sigma}_\diamond)^{-1/2}$. In addition, we have

$$\mathbf{I}_{\widetilde{d}} \overset{Eq.\ (15)}{=} \sum_{i\in[m]} z_{\diamond,i}\widetilde{\mathbf{H}}(x_i) \overset{Eq.\ (21)}{=} \sum_{i\in[m]} z_{\diamond,i}\widetilde{\mathbf{D}} + \sum_{i\in[m]} z_{\diamond,i}\widetilde{\mathbf{P}}_i\widetilde{\mathbf{P}}_i^\top = b\widetilde{\mathbf{D}} + \sum_{i\in[m]} z_{\diamond,i}\widetilde{\mathbf{P}}_i\widetilde{\mathbf{P}}_i^\top. \tag{263}$$

step 1. We first decompose $\frac{1}{\eta}\,\mathrm{Trace}[\mathbf{A}_t^{1/2} - (\mathbf{A}_t^{-1/2} + \eta\widetilde{\mathbf{H}}(x_i))^{-1}]$ for any $i \in [m]$ into the sum of two inner products between matrices. By Woodbury's matrix identity, we have

$$\begin{aligned}(\mathbf{A}_t^{-1/2} + \eta\widetilde{\mathbf{H}}(x_i))^{-1} &= (\mathbf{B}_t^{-1/2} + \eta\widetilde{\mathbf{P}}_i\widetilde{\mathbf{P}}_i^\top)^{-1} \\ &= \mathbf{B}_t^{1/2} - \eta\mathbf{B}_t^{1/2}\widetilde{\mathbf{P}}_i(\mathbf{I} + \eta\widetilde{\mathbf{P}}_i^\top\mathbf{B}_t^{1/2}\widetilde{\mathbf{P}}_i)^{-1}\widetilde{\mathbf{P}}_i^\top\mathbf{B}_t^{1/2}. \end{aligned} \tag{264}$$

Thus

$$\begin{aligned}&\frac{1}{\eta}\,\mathrm{Trace}[\mathbf{A}_t^{1/2} - (\mathbf{A}_t^{-1/2} + \eta\widetilde{\mathbf{H}}(x_i))^{-1}] \\ =&\frac{1}{\eta}\,\mathrm{Trace}(\mathbf{A}_t^{1/2} - \mathbf{B}_t^{1/2}) + \Big\langle(\mathbf{I} + \eta\widetilde{\mathbf{P}}_i^\top\mathbf{B}_t^{1/2}\widetilde{\mathbf{P}}_i)^{-1}, \widetilde{\mathbf{P}}_i^\top\mathbf{B}_t\widetilde{\mathbf{P}}_i\Big\rangle. \end{aligned} \tag{265}$$

We apply Woodbury's matrix identity to $\mathbf{B}_t^{1/2}$ in Eq. (262), then

$$\begin{aligned}\mathbf{B}_t^{1/2} &= (\mathbf{A}_t^{-1/2} + \eta(\boldsymbol{\Sigma}_\diamond)^{-1/2}\mathbf{D}(\boldsymbol{\Sigma}_\diamond)^{-1/2})^{-1} \\ &= \mathbf{A}_t^{1/2} - \eta\mathbf{A}_t^{1/2}\underbrace{(\boldsymbol{\Sigma}_\diamond)^{-1/2}\Big[\mathbf{D}^{-1} + \eta(\boldsymbol{\Sigma}_\diamond)^{-1/2}\mathbf{A}_t^{1/2}(\boldsymbol{\Sigma}_\diamond)^{-1/2}\Big]^{-1}(\boldsymbol{\Sigma}_\diamond)^{-1/2}}_{\triangleq\mathbf{E}}\mathbf{A}_t^{1/2}. \end{aligned} \tag{266}$$

Thus

$$\begin{aligned}&\frac{1}{\eta}\,\mathrm{Trace}(\mathbf{A}_t^{1/2} - \mathbf{B}_t^{1/2}) \\ =&\Big\langle\Big(\mathbf{D}^{-1} + \eta(\boldsymbol{\Sigma}_\diamond)^{-1/2}\mathbf{A}_t^{1/2}(\boldsymbol{\Sigma}_\diamond)^{-1/2}\Big)^{-1}, (\boldsymbol{\Sigma}_\diamond)^{-1/2}\mathbf{A}_t(\boldsymbol{\Sigma}_\diamond)^{-1/2}\Big\rangle \\ =&\Big\langle\mathbf{D}^{1/2}\Big(\mathbf{I} + \eta\mathbf{D}^{1/2}(\boldsymbol{\Sigma}_\diamond)^{-1/2}\mathbf{A}_t^{1/2}(\boldsymbol{\Sigma}_\diamond)^{-1/2}\mathbf{D}^{1/2}\Big)^{-1}\mathbf{D}^{1/2}, (\boldsymbol{\Sigma}_\diamond)^{-1/2}\mathbf{A}_t(\boldsymbol{\Sigma}_\diamond)^{-1/2}\Big\rangle \\ =&\Big\langle\Big(\mathbf{I} + \eta\mathbf{D}^{1/2}(\boldsymbol{\Sigma}_\diamond)^{-1/2}\mathbf{A}_t^{1/2}(\boldsymbol{\Sigma}_\diamond)^{-1/2}\mathbf{D}^{1/2}\Big)^{-1}, \mathbf{D}^{1/2}(\boldsymbol{\Sigma}_\diamond)^{-1/2}\mathbf{A}_t(\boldsymbol{\Sigma}_\diamond)^{-1/2}\mathbf{D}^{1/2}\Big\rangle. \end{aligned} \tag{267}$$

Substitute Eq. (267) into Eq. (265), we can get

$$\begin{aligned}&\frac{1}{\eta}\,\mathrm{Trace}[\mathbf{A}_t^{1/2} - (\mathbf{A}_t^{-1/2} + \eta\widetilde{\mathbf{H}}(x_i))^{-1}] \\ =&\Big\langle\Big(\mathbf{I} + \eta\mathbf{D}^{1/2}(\boldsymbol{\Sigma}_\diamond)^{-1/2}\mathbf{A}_t^{1/2}(\boldsymbol{\Sigma}_\diamond)^{-1/2}\mathbf{D}^{1/2}\Big)^{-1}, \mathbf{D}^{1/2}(\boldsymbol{\Sigma}_\diamond)^{-1/2}\mathbf{A}_t(\boldsymbol{\Sigma}_\diamond)^{-1/2}\mathbf{D}^{1/2}\Big\rangle \\ &+\Big\langle(\mathbf{I} + \eta\widetilde{\mathbf{P}}_i^\top\mathbf{B}_t^{1/2}\widetilde{\mathbf{P}}_i)^{-1}, \widetilde{\mathbf{P}}_i^\top\mathbf{B}_t\widetilde{\mathbf{P}}_i\Big\rangle. \end{aligned} \tag{268}$$

step 2. Now we intend to find a lower bound for $\max_{i\in[m]}\frac{1}{\eta}\,\mathrm{Trace}[\mathbf{A}_t^{1/2} - (\mathbf{A}_t^{-1/2} + \eta\widetilde{\mathbf{H}}(x_i))^{-1}]$ using Eq. (268). For the first inner product on the right hand side of Eq. (268), we can apply Lemma 46:

$$\begin{aligned}&\Big\langle\Big(\mathbf{I} + \eta\mathbf{D}^{1/2}(\boldsymbol{\Sigma}_\diamond)^{-1/2}\mathbf{A}_t^{1/2}(\boldsymbol{\Sigma}_\diamond)^{-1/2}\Big)^{-1}, \mathbf{D}^{1/2}(\boldsymbol{\Sigma}_\diamond)^{-1/2}\mathbf{A}_t(\boldsymbol{\Sigma}_\diamond)^{-1/2}\mathbf{D}^{1/2}\Big\rangle \\ \geq&\frac{\mathrm{Trace}(\mathbf{D}^{1/2}(\boldsymbol{\Sigma}_\diamond)^{-1/2}\mathbf{A}_t(\boldsymbol{\Sigma}_\diamond)^{-1/2}\mathbf{D}^{1/2})}{1 + \eta\,\mathrm{Trace}(\mathbf{D}^{1/2}(\boldsymbol{\Sigma}_\diamond)^{-1/2}\mathbf{A}_t^{1/2}(\boldsymbol{\Sigma}_\diamond)^{-1/2}\mathbf{D}^{1/2})} \\ =&\frac{\langle\mathbf{A}_t, \widetilde{\mathbf{D}}\rangle}{1 + \eta\langle\mathbf{A}_t^{1/2}, \widetilde{\mathbf{D}}\rangle}. \end{aligned} \tag{269}$$

Similarly, applying Lemma 46 to the second term on the right hand side of (268), we can get

$$\Big\langle(\mathbf{I} + \eta\widetilde{\mathbf{P}}_i^\top\mathbf{B}_t^{1/2}\widetilde{\mathbf{P}}_i)^{-1}, \widetilde{\mathbf{P}}_i^\top\mathbf{B}_t\widetilde{\mathbf{P}}_i\Big\rangle \geq \frac{\mathrm{Trace}(\widetilde{\mathbf{P}}_i^\top\mathbf{B}_t\widetilde{\mathbf{P}}_i)}{1 + \eta\,\mathrm{Trace}(\widetilde{\mathbf{P}}_i^\top\mathbf{B}_t^{1/2}\widetilde{\mathbf{P}}_i)} = \frac{\langle\mathbf{B}_t, \widetilde{\mathbf{P}}_i\widetilde{\mathbf{P}}_i^\top\rangle}{1 + \eta\langle\mathbf{B}_t^{1/2}, \widetilde{\mathbf{P}}_i\widetilde{\mathbf{P}}_i^\top\rangle}. \tag{270}$$

Substitute Eq. (269) and Eq. (270) into Eq. (268) and apply Lemma 45, we can get

$$\frac{1}{\eta}\,\mathrm{Trace}[\mathbf{A}_t^{1/2} - (\mathbf{A}_t^{-1/2} + \eta\widetilde{\mathbf{H}}(x_i))^{-1}] \geq \frac{\langle\mathbf{A}_t, \widetilde{\mathbf{D}}\rangle}{1 + \eta\langle\mathbf{A}_t^{1/2}, \widetilde{\mathbf{D}}\rangle} + \frac{\langle\mathbf{B}_t, \widetilde{\mathbf{P}}_i\widetilde{\mathbf{P}}_i^\top\rangle}{1 + \eta\langle\mathbf{B}_t^{1/2}, \widetilde{\mathbf{P}}_i\widetilde{\mathbf{P}}_i^\top\rangle}$$

$$\geq \frac{\langle \mathbf{A}_t, \widetilde{\mathbf{D}}\rangle + \langle \mathbf{B}_t, \widetilde{\mathbf{P}}_i \widetilde{\mathbf{P}}_i^\top\rangle}{1 + \eta[\langle \mathbf{A}_t^{1/2}, \widetilde{\mathbf{D}}\rangle + \langle \mathbf{B}_t^{1/2}, \widetilde{\mathbf{P}}_i \widetilde{\mathbf{P}}_i^\top\rangle]}. \tag{271}$$

Now by Lemma 44 and Eq. (271):

$$\begin{aligned}
&\max_{i\in[m]} \frac{1}{\eta} \operatorname{Trace}[\mathbf{A}_t^{1/2} - (\mathbf{A}_t^{-1/2} + \eta \widetilde{\mathbf{H}}(x_i))^{-1}] \geq \max_{i\in[m]} \frac{\langle \mathbf{A}_t, \widetilde{\mathbf{D}}\rangle + \langle \mathbf{B}_t, \widetilde{\mathbf{P}}_i \widetilde{\mathbf{P}}_i^\top\rangle}{1 + \eta[\langle \mathbf{A}_t^{1/2}, \widetilde{\mathbf{D}}\rangle + \langle \mathbf{B}_t^{1/2}, \widetilde{\mathbf{P}}_i \widetilde{\mathbf{P}}_i^\top\rangle]} \\
\geq &\frac{\sum_{i\in[m]} z_{\diamond,i}\langle \mathbf{A}_t, \widetilde{\mathbf{D}}\rangle + \sum_{i\in[m]} z_{\diamond,i}\langle \mathbf{B}_t, \widetilde{\mathbf{P}}_i \widetilde{\mathbf{P}}_i^\top\rangle}{\sum_{i\in[m]} z_{\diamond,i} + \eta[\sum_{i\in[m]} z_{\diamond,i}\langle \mathbf{A}_t^{1/2}, \widetilde{\mathbf{D}}\rangle + \sum_{i\in[m]} z_{\diamond,i}\langle \mathbf{B}_t^{1/2}, \widetilde{\mathbf{P}}_i \widetilde{\mathbf{P}}_i^\top\rangle]} \\
= &\frac{\langle \mathbf{A}_t, b\widetilde{\mathbf{D}}\rangle + \langle \mathbf{B}_t, \mathbf{I} - b\widetilde{\mathbf{D}}\rangle}{b + \eta[\langle \mathbf{A}_t^{1/2}, b\widetilde{\mathbf{D}}\rangle + \langle \mathbf{B}_t^{1/2}, \mathbf{I} - b\widetilde{\mathbf{D}}\rangle]},
\end{aligned} \tag{272}$$

where the last equality follows by Eq. (263) and the fact that $\sum_{i\in[m]} z_{\diamond,i} = b$.

step 3. In this step, we will show that the numerator of Eq. (272) is lower bounded by $1 - \eta/2b$. First note that we have derived that $\mathbf{B}_t^{1/2} = \mathbf{A}_t^{1/2} - \eta \mathbf{A}_t^{1/2}\mathbf{E}\mathbf{A}_t^{1/2}$ in Eq. (266). Then

$$\begin{aligned}
\mathbf{B}_t &= (\mathbf{A}_t^{1/2} - \eta \mathbf{A}_t^{1/2}\mathbf{E}\mathbf{A}_t^{1/2})^2 \\
&= \mathbf{A}_t - \underbrace{(\eta \mathbf{A}_t \mathbf{E}\mathbf{A}_t^{1/2} + \eta \mathbf{A}_t^{1/2}\mathbf{E}\mathbf{A}_t - \eta^2 \mathbf{A}_t^{1/2}\mathbf{E}\mathbf{A}_t\mathbf{E}\mathbf{A}_t^{1/2})}_{\triangleq \mathbf{G}} = \mathbf{A}_t - \mathbf{G}.
\end{aligned} \tag{273}$$

Substitute this into the numerator of (272), we have

$$\begin{aligned}
\langle \mathbf{A}_t, b\widetilde{\mathbf{D}}\rangle + \langle \mathbf{B}_t, \mathbf{I} - b\widetilde{\mathbf{D}}\rangle &= \langle \mathbf{A}_t, b\widetilde{\mathbf{D}}\rangle + \langle \mathbf{A}_t - \mathbf{G}, \mathbf{I} - b\widetilde{\mathbf{D}}\rangle \\
&= \operatorname{Trace}(\mathbf{A}_t) - \langle \mathbf{G}, \mathbf{I} - b\widetilde{\mathbf{D}}\rangle \\
&= 1 - \langle \mathbf{G}, \mathbf{I} - b\widetilde{\mathbf{D}}\rangle,
\end{aligned} \tag{274}$$

where the last equality follows by $\operatorname{Trace}(\mathbf{A}_t) = 1$. Now we intend to find an upper bound for $\langle \mathbf{G}, \mathbf{I} - b\widetilde{\mathbf{D}}\rangle$. First note that since $\mathbf{A}_t^{1/2}\mathbf{E}\mathbf{A}_t\mathbf{E}\mathbf{A}_t^{1/2} \succeq \mathbf{0}$, by the definition of $\mathbf{G}$ in Eq. (273) we have

$$\mathbf{G} \preceq \eta \mathbf{A}_t \mathbf{E}\mathbf{A}_t^{1/2} + \eta \mathbf{A}_t^{1/2}\mathbf{E}\mathbf{A}_t. \tag{275}$$

Recall the definition of $\mathbf{E}$ in Eq. (266), we claim that $\mathbf{E} \preceq \widetilde{\mathbf{D}}$. Indeed, since $(\mathbf{\Sigma}_\diamond)^{-1/2}\mathbf{A}_t^{1/2}(\mathbf{\Sigma}_\diamond)^{-1/2}$ is positive definite, we have

$$\mathbf{D}^{-1} + \eta(\mathbf{\Sigma}_\diamond)^{-1/2}\mathbf{A}_t^{1/2}(\mathbf{\Sigma}_\diamond)^{-1/2} \succeq \mathbf{D}^{-1},$$

Thus $\Big[\mathbf{D}^{-1} + \eta(\mathbf{\Sigma}_\diamond)^{-1/2}\mathbf{A}_t^{1/2}(\mathbf{\Sigma}_\diamond)^{-1/2}\Big]^{-1} \preceq \mathbf{D}$ and therefore,

$$\mathbf{E} \triangleq (\mathbf{\Sigma}_\diamond)^{-1/2}\Big[\mathbf{D}^{-1} + \eta(\mathbf{\Sigma}_\diamond)^{-1/2}\mathbf{A}_t^{1/2}(\mathbf{\Sigma}_\diamond)^{-1/2}\Big]^{-1}(\mathbf{\Sigma}_\diamond)^{-1/2} \preceq (\mathbf{\Sigma}_\diamond)^{-1/2}\mathbf{D}(\mathbf{\Sigma}_\diamond)^{-1/2} = \widetilde{\mathbf{D}}. \tag{276}$$

Now we have

$$\begin{aligned}
\langle \mathbf{G}, \mathbf{I} - b\widetilde{\mathbf{D}}\rangle &\overset{Eq.\ (275)}{\leq} \eta\langle \mathbf{A}_t\mathbf{E}\mathbf{A}_t^{1/2} + \mathbf{A}_t^{1/2}\mathbf{E}\mathbf{A}_t, \mathbf{I} - b\widetilde{\mathbf{D}}\rangle \\
&= \eta\langle \mathbf{E}, \mathbf{A}_t^{1/2}(\mathbf{I} - b\widetilde{\mathbf{D}})\mathbf{A}_t\rangle + \eta\langle \mathbf{E}, \mathbf{A}_t(\mathbf{I} - b\widetilde{\mathbf{D}})\mathbf{A}_t^{1/2}\rangle \\
&\overset{Eq.\ (276)}{\leq} \eta\langle \widetilde{\mathbf{D}}, \mathbf{A}_t^{1/2}(\mathbf{I} - b\widetilde{\mathbf{D}})\mathbf{A}_t + \mathbf{A}_t(\mathbf{I} - b\widetilde{\mathbf{D}})\mathbf{A}_t^{1/2}\rangle \\
&= 2\eta \operatorname{Trace}(\mathbf{A}_t^{3/2}\widetilde{\mathbf{D}}) - 2\eta b \operatorname{Trace}(\mathbf{A}_t^{1/2}\widetilde{\mathbf{D}}\mathbf{A}_t\widetilde{\mathbf{D}}) \triangleq h(\widetilde{\mathbf{D}}),
\end{aligned} \tag{277}$$

where we define function $h : \mathbb{S}_+^{\tilde{d}} \to \mathbb{R}$. By Eq. (263), $b\widetilde{\mathbf{D}} \preceq \mathbf{I}$ and thus the domain of function h is $\operatorname{dom} h = \{\widetilde{\mathbf{D}} \in \mathbb{S}_+^{\tilde{d}} : \widetilde{\mathbf{D}} \preceq \frac{1}{b}\mathbf{I}\}$.

We intend to find an upper bound for $h(\widetilde{\mathbf{D}})$. First we prove that $h(\widetilde{\mathbf{D}})$ is a concave function. We can verify its concavity by considering an arbitrary line, given by $\mathbf{Z} + t\mathbf{V}$, where $\mathbf{Z}, \mathbf{V} \in \mathbb{S}_+^{\tilde{d}}$. Define

$g(t) := h(\mathbf{Z} + t\mathbf{V})$, where t is restricted to the interval such that $\mathbf{Z} + t\mathbf{V} \in \mathrm{dom}h$. By convex analysis theory, it is sufficient to prove the concavity of function g. Note that

$$\begin{aligned} g(t) &= 2\eta \operatorname{Trace}[\mathbf{A}_t^{3/2}(\mathbf{Z} + t\mathbf{V})] - 2\eta b \operatorname{Trace}[\mathbf{A}_t^{1/2}(\mathbf{Z} + t\mathbf{V})\mathbf{A}_t(\mathbf{Z} + t\mathbf{V})] \\ &= -2\eta b t^2 \operatorname{Trace}(\mathbf{A}_t^{1/2}\mathbf{V}\mathbf{A}_t\mathbf{V}) + 2\eta t \operatorname{Trace}(\mathbf{A}_t^{3/2}\mathbf{V}) \\ &\quad - 2\eta b t \operatorname{Trace}(\mathbf{A}_t^{1/2}\mathbf{V}\mathbf{A}_t\mathbf{Z} + \mathbf{A}_t^{1/2}\mathbf{Z}\mathbf{A}_t\mathbf{V}) + 2\eta \operatorname{Trace}(\mathbf{Z}\mathbf{A}_t^{3/2}) - 2\eta b \operatorname{Trace}(\mathbf{A}_t^{1/2}\mathbf{Z}\mathbf{A}_t\mathbf{Z}). \end{aligned} \tag{278}$$

Thus $g''(t) = -4\eta b \operatorname{Trace}(\mathbf{A}_t^{1/2}\mathbf{V}\mathbf{A}_t\mathbf{V})$ and $g''(t) \le 0$ because $\mathbf{A}_t^{1/2}\mathbf{V}\mathbf{A}_t\mathbf{V} \succeq \mathbf{0}$. Therefore $g(t)$ is concave and so is $h(\widetilde{\mathbf{D}})$. Now consider the gradient of $h(\widetilde{\mathbf{D}})$:

$$\nabla h(\widetilde{\mathbf{D}}) = 2\eta \mathbf{A}_t^{3/2} - 4\eta b \mathbf{A}_t^{1/2}\widetilde{\mathbf{D}}\mathbf{A}_t. \tag{279}$$

Let $\nabla h(\widetilde{\mathbf{D}}) = 0$, we can get $\widetilde{\mathbf{D}} = \frac{1}{2b}\mathbf{I} \in \mathrm{dom}h$. Thus

$$\sup_{\widetilde{\mathbf{D}} \in \mathrm{dom}h} h(\widetilde{\mathbf{D}}) = h\Big(\frac{1}{2b}\mathbf{I}\Big) = \frac{\eta}{b}\operatorname{Trace}(\mathbf{A}_t^{3/2}) - \frac{\eta}{2b}\operatorname{Trace}(\mathbf{A}_t^{3/2}) = \frac{\eta}{2b}\operatorname{Trace}(\mathbf{A}_t^{3/2}) \le \frac{\eta}{2b}, \tag{280}$$

where the last inequality follows by the fact that all eigenvalues of $\mathbf{A}_t$ lie in $[0, 1]$ and $\operatorname{Trace}(\mathbf{A}_t) = 1$.

Combining Eq. (274) , Eq. (277) and Eq. (280), we can conclude that

$$\langle \mathbf{A}_t, b\widetilde{\mathbf{D}} \rangle + \langle \mathbf{B}_t, \mathbf{I} - b\widetilde{\mathbf{D}} \rangle \ge 1 - \frac{\eta}{2b}. \tag{281}$$

step 4. Now we derive an upper bound for the denominator of the right hand side of Eq. (272). By Eq. (266), we have

$$\begin{aligned} \langle \mathbf{A}_t^{1/2}, b\widetilde{\mathbf{D}} \rangle + \langle \mathbf{B}_t^{1/2}, \mathbf{I} - b\widetilde{\mathbf{D}} \rangle &= \langle \mathbf{A}_t^{1/2}, b\widetilde{\mathbf{D}} \rangle + \langle \mathbf{A}_t^{1/2} - \eta \mathbf{A}_t^{1/2}\mathbf{E}\mathbf{A}_t^{1/2}, \mathbf{I} - b\widetilde{\mathbf{D}} \rangle \\ &= \operatorname{Trace}(\mathbf{A}_t^{1/2}) - \eta \langle \mathbf{A}_t^{1/2}\mathbf{E}\mathbf{A}_t^{1/2}, \mathbf{I} - b\widetilde{\mathbf{D}} \rangle \\ &\overset{(a)}{\le} \operatorname{Trace}(\mathbf{A}_t^{1/2}) \overset{(b)}{\le} \sqrt{\tilde{d}}, \end{aligned} \tag{282}$$

where (a) follows by the fact that both $\mathbf{A}_t^{1/2}\mathbf{E}\mathbf{A}_t^{1/2}$ and $\mathbf{I} - b\widetilde{\mathbf{D}}$ are positive semidefinite, (b) follows by the following property:

$$\operatorname{Trace}(\mathbf{A}_t^{1/2}) = \sum_{i \in [\tilde{d}]} \lambda_i(\mathbf{A}_t^{1/2}) \le \sqrt{\tilde{d}} \sqrt{\sum_{i \in [\tilde{d}]} \lambda_i^2(\mathbf{A}_t^{1/2})} = \sqrt{\tilde{d}} \sqrt{\sum_{i \in [\tilde{d}]} \lambda_i(\mathbf{A}_t)} = \sqrt{\tilde{d}}. \tag{283}$$

where $\lambda_i(\mathbf{A}_t)$ is the i-th eigenvalue of $\mathbf{A}_t$, the inequality follows by the Cauchy-Schwarz inequality, the last equality follows by $\operatorname{Trace}(\mathbf{A}_t) = 1$.

step 5. Now substitute Eq. (281) and Eq. (282) into Eq. (272), we have

$$\max_{i \in [m]} \frac{1}{\eta} \operatorname{Trace}[\mathbf{A}_t^{1/2} - (\mathbf{A}_t^{-1/2} + \eta \widetilde{\mathbf{H}}(x_i))^{-1}] \ge \frac{\langle \mathbf{A}_t, b\widetilde{\mathbf{D}} \rangle + \langle \mathbf{B}_t, \mathbf{I} - b\widetilde{\mathbf{D}} \rangle}{b + \eta[\langle \mathbf{A}_t^{1/2}, b\widetilde{\mathbf{D}} \rangle + \langle \mathbf{B}_t^{1/2}, \mathbf{I} - b\widetilde{\mathbf{D}} \rangle]} \ge \frac{1 - \frac{\eta}{2b}}{b + \eta\sqrt{\tilde{d}}}. \tag{284}$$

□

F.5 Proof of Theorem 10

Proof. Let $b = 32\tilde{d}/\epsilon^2 + 16\sqrt{\tilde{d}}/\epsilon^2$, $\eta = 8\sqrt{\tilde{d}}/\epsilon$, by Proposition 9, we have

$$\begin{aligned} &\sum_{t=1}^{b} \operatorname{Trace}[\mathbf{A}_t^{1/2} - (\mathbf{A}_t^{-1/2} + \eta \mathbf{F}_t)^{-1}] \\ &\ge \sum_{t=1}^{b} \frac{1 - \frac{\eta}{2b}}{b + \eta\sqrt{\tilde{d}}} = \frac{b - \frac{\eta}{2}}{b + \eta\sqrt{\tilde{d}}} \ge \frac{32\tilde{d}/\epsilon^2 + 16\sqrt{\tilde{d}}/\epsilon^2 - 4\sqrt{\tilde{d}}/\epsilon}{32\tilde{d}/\epsilon^2 + 16\sqrt{\tilde{d}}/\epsilon^2 + 8\tilde{d}/\epsilon} \end{aligned}$$

$$\begin{aligned}&\geq \frac{32\widetilde{d}/\epsilon^2 + 16\sqrt{\widetilde{d}}/\epsilon^2 + 8\widetilde{d}/\epsilon - (8\widetilde{d}/\epsilon + 4\sqrt{\widetilde{d}}/\epsilon)}{32\widetilde{d}/\epsilon^2 + 16\sqrt{\widetilde{d}}/\epsilon^2 + 8\widetilde{d}/\epsilon} = 1 - \frac{8\widetilde{d}/\epsilon + 4\sqrt{\widetilde{d}}/\epsilon}{\frac{4}{\epsilon}(8\widetilde{d}/\epsilon + 4\sqrt{\widetilde{d}}/\epsilon) + 8\sqrt{\widetilde{d}}/\epsilon} \\ &\geq 1 - \frac{\epsilon}{4}. \end{aligned} \tag{285}$$

Substitute Eq. (285) into Eq. (18) in Proposition 8, we have

$$\begin{aligned}\lambda_{\min}\Big(\sum_{t=1}^{b} \mathbf{F}_t\Big) &\geq -\frac{2\sqrt{\widetilde{d}}}{\eta} + \frac{1}{\eta}\sum_{t=1}^{b} \mathrm{Trace}[\mathbf{A}_t^{1/2} - (\mathbf{A}_t^{-1/2} + \eta \mathbf{F}_t)^{-1}] \\ &\geq -\frac{2\sqrt{\widetilde{d}}}{8\sqrt{\widetilde{d}}/\epsilon} + 1 - \frac{\epsilon}{4} = 1 - \frac{\epsilon}{2} \geq \frac{1}{1+\epsilon}. \end{aligned} \tag{286}$$

By Proposition 7, we can get

$$f\Big(\sum_{t=1}^{b} \mathbf{F}_t\Big) \leq (1+\epsilon) f^*. \tag{287}$$

□

F.6 Proof of Theorem 4

In this section, we intend to prove Theorem 4. Our main approach is combining Theorem 3 and Theorem 10. In order to account for the effect of using ERM θ_0 as surrogate for θ_*, we first define optimal sampling over θ_* (Definition 47) and optimal sampling over θ_0 (Definition 48). Corollary 49 is a direct result from Proposition 9. At the end of this section, we give the proof for Theorem 4.

Definition 47. *[optimal sampling in hindsight] Suppose we know θ_*, we select points X_* defined by*

$$X_* \in \mathop{\arg\min}_{\substack{X \subset U \\ |X| = b}} \big\langle \mathbf{H}_q(\theta_*)^{-1}, \mathbf{H}_p(\theta_*) \big\rangle, \quad \text{where} \quad q(x) \triangleq \frac{1}{n_0 + b} \sum_{x' \in X_0 \cup X} \delta(x' - x). \tag{288}$$

Denote the empirical distribution on points $X_0 \cup X_*$ by $q_*(x)$.

Definition 48. *[optimal sampling over ERM] The optimal sampling over ERM θ_0 is defined by*

$$\widehat{X}_* \in \mathop{\arg\min}_{\substack{X \subset U \\ |X| = b}} \big\langle \mathbf{H}_q(\theta_0)^{-1}, \mathbf{H}_p(\theta_0) \big\rangle, \quad \text{where} \quad q(x) \triangleq \frac{1}{n_0 + b} \sum_{x' \in X_0 \cup X} \delta(x' - x). \tag{289}$$

Denote the empirical distribution on points $X_0 \cup \widehat{X}_*$ by $\widehat{q}_*(x)$.

Corollary 49. *Given $\epsilon \in (0,1)$, consider $\eta = 8\sqrt{\widetilde{d}}/\epsilon$, $b \geq 32\widetilde{d}/\epsilon^2 + 16\sqrt{\widetilde{d}}/\epsilon^2$ in Algorithm 1. Then we have*

$$\big\langle \big(\mathbf{H}_q(\theta_0)\big)^{-1}, \mathbf{H}_p(\theta_0) \big\rangle \leq (1+\epsilon) \big\langle \big(\mathbf{H}_{\widehat{q}_*}(\theta_0)\big)^{-1}, \mathbf{H}_p(\theta_0) \big\rangle. \tag{290}$$

Proof. Let X be the set of points selected by Algorithm 1, by Eq. (11) we have:

$$\mathbf{H}_q(\theta_0) = \frac{1}{n} \sum_{x \in X} \mathbf{H}(x), \quad \mathbf{H}_{\widehat{q}_*}(\theta_0) = \frac{1}{n} \sum_{x \in \widehat{X}_*} \mathbf{H}(x),$$

where $n = n_0 + b$, and thus

$$\big\langle \big(\mathbf{H}_q(\theta_0)\big)^{-1}, \mathbf{H}_p(\theta_0) \big\rangle = n f\Big(\sum_{x \in X} \mathbf{H}(x)\Big). \tag{291}$$

By Definition 48, we know that $\widehat{X}_*$ is the optimal solution to optimization problem Eq. (13). Since f_* is the optimal value of the objective function in (13), we have

$$\big\langle \big(\mathbf{H}_{\widehat{q}_*}(\theta_0)\big)^{-1}, \mathbf{H}_p(\theta_0) \big\rangle = n \Big\langle \Big(\sum_{x \in \widehat{X}_*} \mathbf{H}(x)\Big)^{-1}, \mathbf{H}_p(\theta_0) \Big\rangle = n f_*. \tag{292}$$

By Theorem 10, we have $f\big(\sum_{x \in X} \mathbf{H}(x)\big) \leq (1+\epsilon) f_*$. Combining this with Eqs. (291) and (292), we can obtain Eq. (290). □

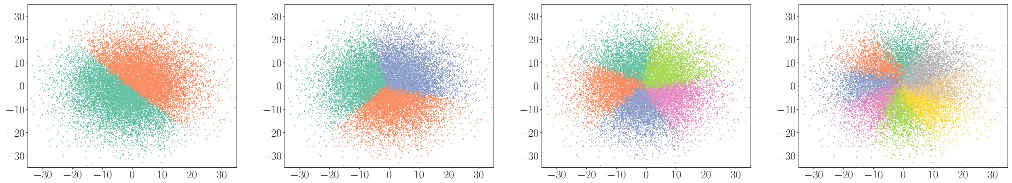

Figure 4: Plots of first two coordinates of points draw from the joint distribution $pi_p(x, y)$.

proof of Theorem 4. By Eq. (7) we have

$$\mathbb{E}[L_p(\theta_0)] - L_p(\theta_*) \lesssim \frac{e^{\alpha_1} - \alpha_1 - 1}{\alpha_1^2} \cdot \frac{\left\langle (\mathbf{H}_q(\theta_*))^{-1}, \mathbf{H}_p(\theta_*) \right\rangle}{n_0 + b}, \tag{293}$$

where

$$\alpha_1 = C_3 \sqrt{\sigma_1 \rho} \sqrt{\left(\widetilde{d} + \sqrt{\widetilde{d}} \log(e/\delta)\right)/(n_0 + b)}, \tag{294}$$

where $\sigma_1 = \lambda_{\max}(\mathbf{H}_q{}^{-1}\mathbf{H}_p)$. From the step 2 of the proof of Theorem 3, we have with probability at least $1 - \delta$,

$$\frac{1}{\sqrt{2}} \mathbf{H}_q(\theta_*) \preceq \mathbf{H}_q(\theta_{r-1}) \preceq \sqrt{2} \mathbf{H}_q(\theta_*). \tag{295}$$

Combining results from step 6 in the proof of Theorem 3 with Eq. (57) in Proposition 31, we can obtain that with probability at least $1 - \delta$,

$$e^{-\alpha_0} \mathbf{H}_p(\theta_*) \preceq \mathbf{H}_p(\theta_0) \preceq e^{\alpha_0} \mathbf{H}_p(\theta_*), \tag{296}$$

where

$$\alpha_0 = C_3' \sqrt{\sigma_0 \rho} \sqrt{\left(\widetilde{d} + \sqrt{\widetilde{d}} \log(e/\delta)\right)/n_0}, \tag{297}$$

where $\sigma_0 = \lambda_{\max}(\mathbf{H}_{q_0}^{-1}\mathbf{H}_p)$, $q_0(x)$ is the empirical distribution over the inital labeled points, i.e. $q_0(x) \triangleq \sum_{x' \in X_0} \delta(x - x')$.

Therefor we have

$$\begin{aligned}
\Big\langle \big(\mathbf{H}_q(\theta_*)\big)^{-1}, \mathbf{H}_q(\theta_*) \Big\rangle &\overset{(a)}{\leq} \sqrt{2} e^{\alpha_0} \Big\langle \big(\mathbf{H}_q(\theta_0)\big)^{-1}, \mathbf{H}_p(\theta_0) \Big\rangle \\
&\overset{(b)}{\leq} \sqrt{2} e^{\alpha_0} (1 + \epsilon) \Big\langle \big(\mathbf{H}_{\widehat{q}_*}(\theta_0)\big)^{-1}, \mathbf{H}_p(\theta_0) \Big\rangle \\
&\overset{(c)}{\leq} \sqrt{2} e^{\alpha_0} (1 + \epsilon) \Big\langle \big(\mathbf{H}_{q_*}(\theta_0)\big)^{-1}, \mathbf{H}_p(\theta_0) \Big\rangle \\
&\overset{(d)}{\leq} 2 e^{2\alpha_0} (1 + \epsilon) \Big\langle \big(\mathbf{H}_{q_*}(\theta_*)\big)^{-1}, \mathbf{H}_p(\theta_*) \Big\rangle \\
&= 2 e^{2\alpha_0} (1 + \epsilon) OPT,
\end{aligned} \tag{298}$$

where (a) and (d) follow by Eqs. (295) and (296), (b) follows by Corollary 49, (c) follows by the fact that $\widehat{q}_*$ is the optimal sampling distribution to minimize $\langle (\mathbf{H}_q(\theta_0))^{-1}, \mathbf{H}_p(\theta_0) \rangle$ (see the definition of optimal sampling over ERM in Definition 48).

By Eqs. (293) and (298), we can obtain Eq. (9).

□

G Additional experimental details

G.1 Synthetic experiments

We use numerical tests on synthetic datasets to demonstrate the two excess risk bounds derived in Theorem 32 (detailed version of Theorem 3): Eq. (61) and Eq. (62).

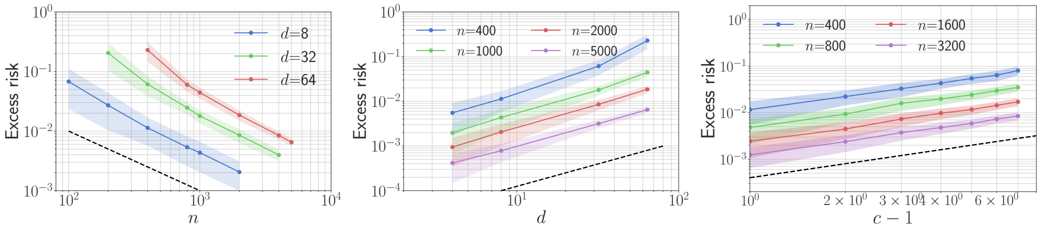

Figure 5: Excess risk of $q(x)$ as a function of n, d and $c-1$. The dashed black line in the left plot indicates inversely linear relation. The dashed black lines in the center and right plots indicate linear relations.

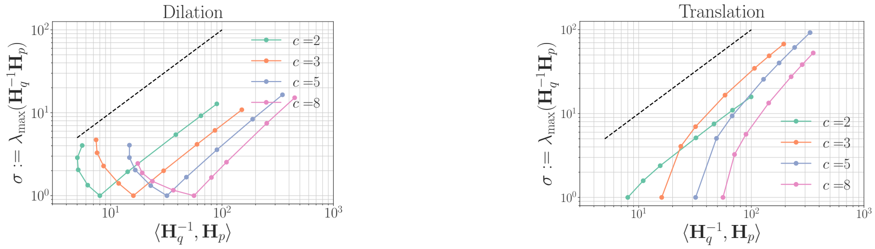

Figure 6: $\lambda_{\max}(\mathbf{H}_q{}^{-1}\mathbf{H}_p)$ vs $\langle \mathbf{H}_q{}^{-1}, \mathbf{H}_p \rangle$ in dilation tests (left plot) and translation tests (right plot).

Gaussian Setup. For a given dimension d, we choose $p(x) \sim \mathcal{N}(\mathbf{0}, \mathbf{V}_p)$, where $\mathbf{V}_p = 100\mathbf{I}_d$. For a given class number c, we define $\theta_* \in \mathbb{R}^{(c-1)\times d}$ such that points generated by $p(x)$ are almost equally distributed across the c classes. Besides, we normalize the row of θ_*, i.e. $\|\theta_{*,i}\|_2 = 1$. In Fig. 4, we plot the first two coordinates of the points draw from the joint distribution $pi_p(x, y)$, where each point is colored by its class id.

We use Monte Carlo method to approximate the risk of $p(x)$ at a given parameter θ, i.e. $L_p(\theta) = \mathbb{E}_{(x,y)\sim\pi_p(x,y)}[\ell_{(x,y)}(\theta)]$. In specific, we draw $N = 50,000$ i.i.d. points $\{x_i\}_{i\in[N]}$ from $p(x)$, for each x_i, we draw $M = 100$ i.i.d. labels $\{y_{ij}\}_{j\in[M]}$ from $p(y|x_i, \theta_*)$, then we can estimate the risk by

$$\begin{aligned} L_p(\theta) &\triangleq \mathbb{E}_{(x,y)\sim\pi_p(x,y)}[\ell_{(x,y)}(\theta)] = \mathbb{E}_{x\sim p(x)}\,\mathbb{E}_{y\sim p(y|x,\theta_*}[\ell_{x,y}(\theta)] \\ &\approx \frac{1}{N}\frac{1}{M}\sum_{i\in[N]}\sum_{j\in[M]}\ell_{(x_i,y_{ij})}(\theta). \end{aligned} \tag{299}$$

Demonstration of excess risk bound for $q(x)$ (Eq. (61)). We use $q(x) \sim \mathcal{N}(\mathbf{0}, 100\mathbf{I}_d)$ to demonstrate Eq. (61). Let $\{(x_i, y_i)\}_{i\in[n]}$ be samples i.i.d draw from $\pi_q(x, y)$. Denote the ERM estimate as θ_n defined by Eq. (4). In Fig. 5, we plot the excess risk with respect to $q(x)$ (i.e. $L_q(\theta_n) - L_q(\theta_*)$) against n, d and $c-1$. From theses plots, we can observe that the excess risk almost linearly depends on $\frac{1}{n}$, d and $c-1$ respectively. This observation is consistent to our upper bound derived in Eq. (61).

Demonstration of excess risk bounds for $p(x)$ (Eq. (62)). In § 5, we have introduced the different types of $q(x)$ used in dilation and translation tests. In Fig. 6, we plot the relations of $\lambda_{\max}(\mathbf{H}_q{}^{-1}\mathbf{H}_p)$ (which is σ in Theorem 32) and FIR ($\langle \mathbf{H}_q{}^{-1}, \mathbf{H}_p \rangle$. For the dilation tests, we present the plots of excess risk of $p(x)$ vs FIR, n, and FIR$/n$ respectively in Fig. 7. We plot the results for translation tests in Fig. 8. As mentioned in Section 5, these results are consistent to the bounds we derived in Eq. (62). One interesting finding is that from the lower rows of Figs. 7 and 8, the excess risk is upper bounded by $\frac{9}{5}\frac{\text{FIR}}{n}$ when n is large. This observation is consistent with the upper bound we derived in the bounded domain case (Eq. (194) in Appendix E).

Non-sub-Gaussian distributions. We consider two non-sub-Gaussian distributions: multivariate Laplace distribution and t-distribution. For $q(x)$, we only consider the translation case. We fix $c = 2$ and vary d, n and $q(x)$. In Fig. 9, we plot $\lambda_{\max}(\mathbf{H}_q{}^{-1}\mathbf{H}_p)$ vs FIR in different distributions. For multivariate Laplace distribution tests, we plot excess risk of $p(x)$ vs FIR, n and FIR$/n$ respectively

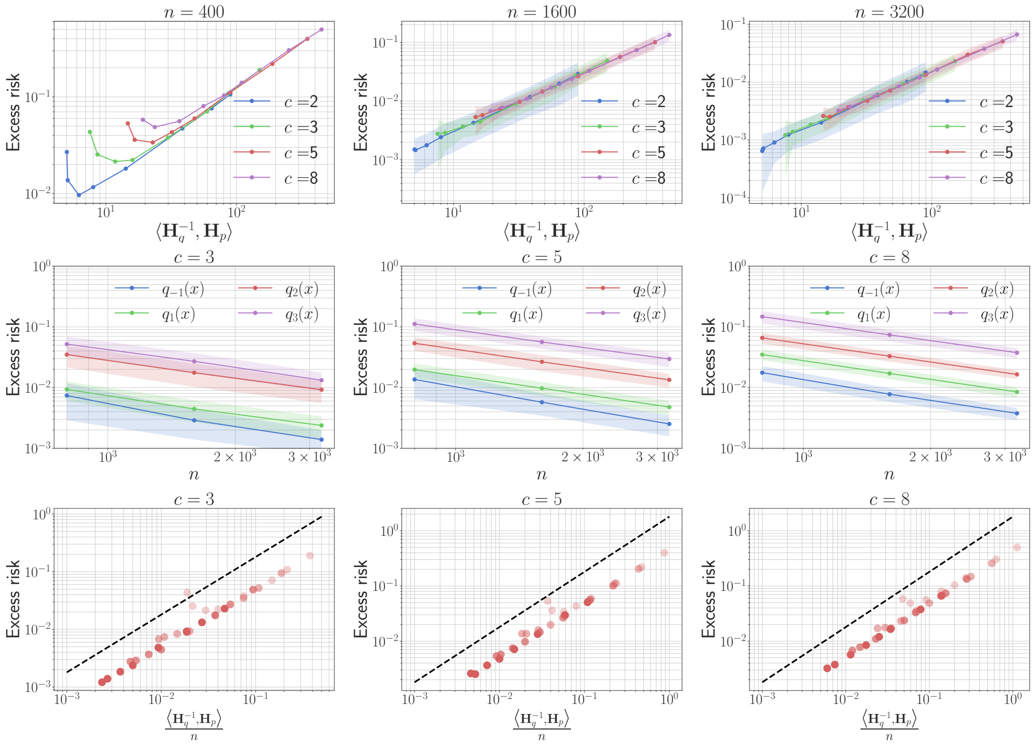

Figure 7: Gaussian *dilation* tests: excess risk of $p(x)$ vs FIR (upper row), n (middle row) and FIR$/n$ (lower row). For all plots in the lower row, the less transparent dots represent the larger sample size n, the black dashed lines represent linear relation $y = \frac{9}{5}x$.

in Fig. 10. We plot results for the multivariate t-distribution in Fig. 11. We can observe that the results are consistent to the excess risk bound derived in Eq. (7), even though we have sub-Gaussian distribution assumption in Theorem 3.

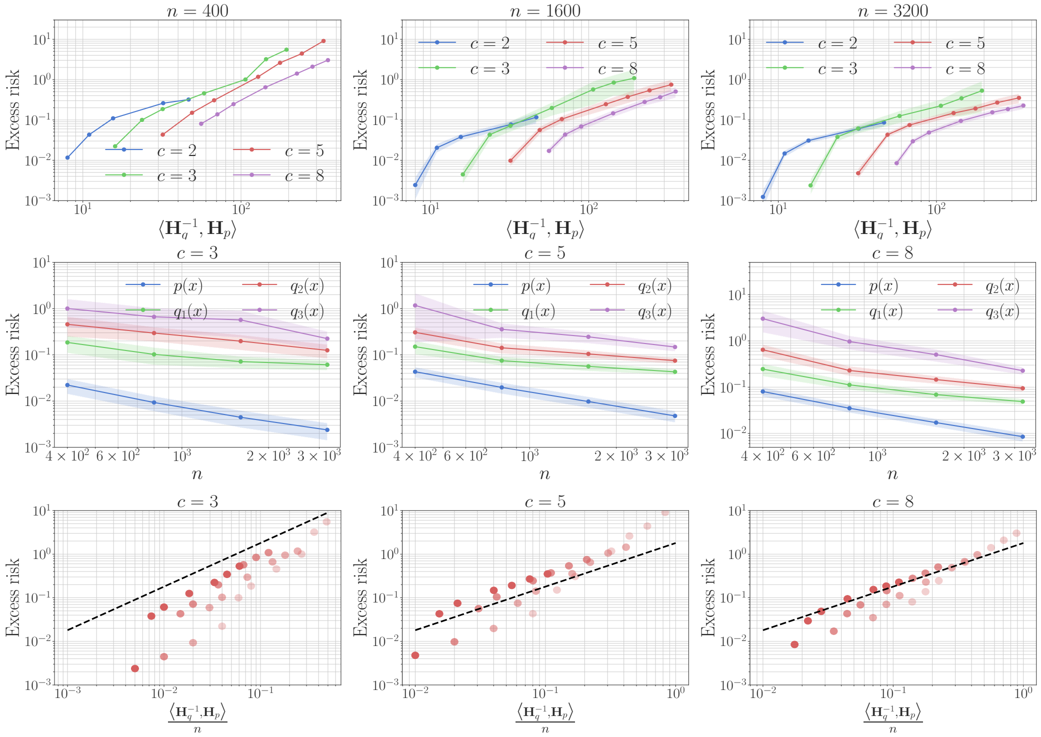

Figure 8: Gaussian *translation* tests: excess risk of $p(x)$ vs FIR (upper row), n (middle row) and FIR$/n$ (lower row). For all plots in the lower row, the less transparent dots represent the larger sample size n, the black dashed lines represent linear relation $y = \frac{9}{5}x$.

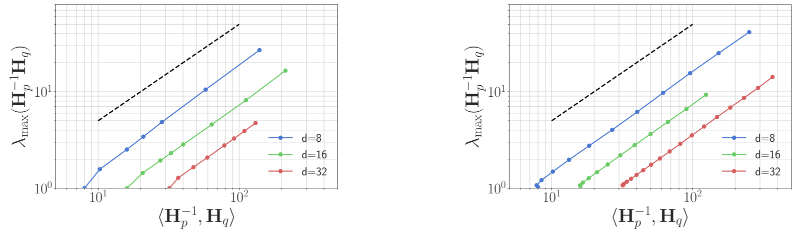

Figure 9: $\lambda_{\max}(\mathbf{H}_q{}^{-1}\mathbf{H}_p)$ vs Trace$(\mathbf{H}_q{}^{-1}\mathbf{H}_p)$ in Multivariate Laplace tests and t-distribution tests.

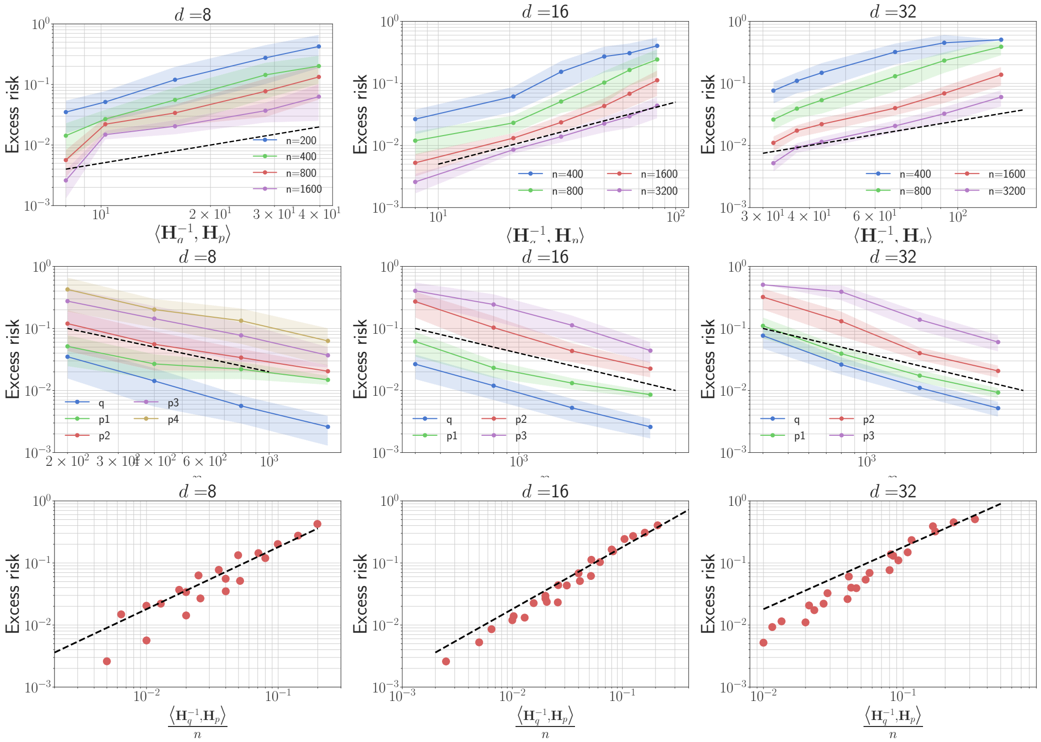

Figure 10: Multivariate Laplace distribution test: excess risk of $p(x)$ vs FIR (upper), n (middle), and $\frac{\text{FIR}}{n}$ (lower), the black dashed lines have slope 1 in upper and lower rows , and slope -1 in the middle row.

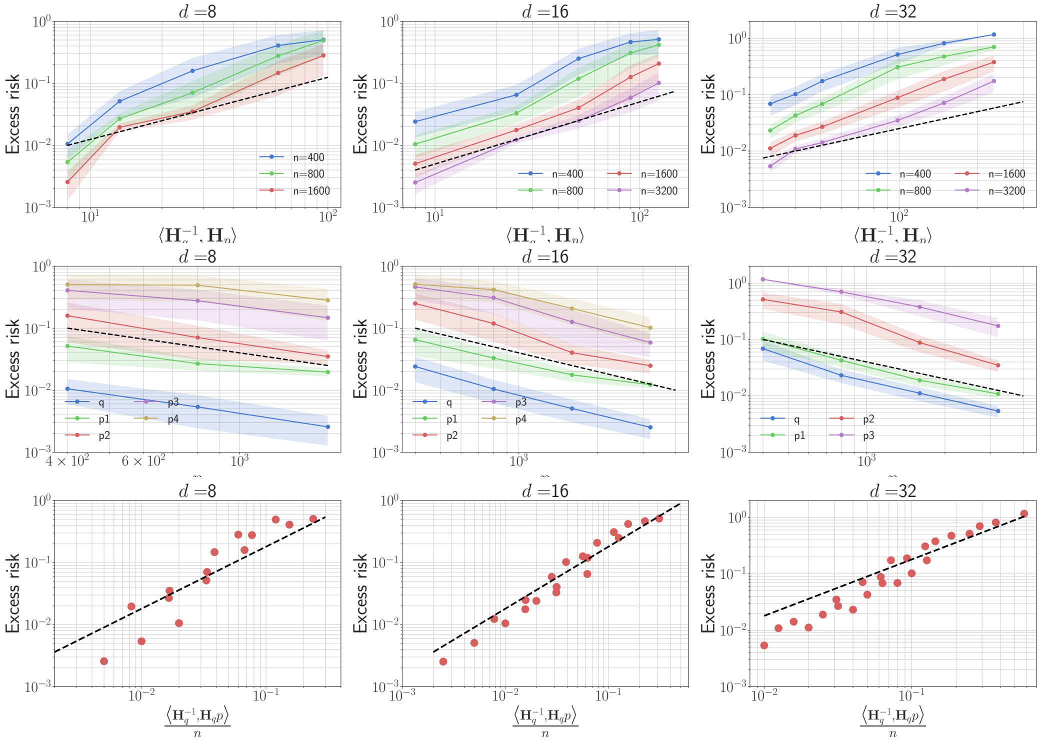

Figure 11: Multivariate t-distribution test: excess risk of $p(x)$ vs FIR (upper), n (middle), and $\frac{\text{FIR}}{n}$ (lower), the black dashed lines have slope 1 in upper and lower rows , and slope -1 in the middle row.

Algorithm 3 Spectral embedding via normalized graph Laplacian

```
Input: data points $\mathbf{X} \in \mathbb{R}^{N \times D}$, nearest neighbor number $k$, target out put dimension $d$
Output: $\widehat{\mathbf{X}} \in \mathbb{R}^{N \times d}$
Obtain $k$-nearest neighbor graph $\mathcal{G}$ on $\mathbf{X}$.
Obtain adjacency matrix $\mathbf{A}$ and its degree matrix $\mathbf{D}$ from $\mathcal{G}$ (using ones as weights).
Calculate normalized Laplacian $\mathbf{L} \leftarrow \mathbf{I} - \mathbf{D}^{-1/2}\mathbf{A}\mathbf{D}^{-1/2}$.
Calculate the first $d$ eigenvectors of $\mathbf{L}$ (corresponding to the $d$ smallest eigenvalues of $\mathbf{L}$): $\{v_i\}_{i \in [d]}$.
Form matrix $\widehat{\mathbf{X}}$ by stacking $\{v_i\}_{i \in [d]}$ column-wise.
```

G.2 Real-world Datasets

Data pre-processing. We use unsupervised learning to find an appropriate feature space that we can then use for multi-class logistic regression. SimCLR [21] is a framework for contrastive learning of visual representations. It learns representations by maximizing agreement between differently augmented views of the same data example via a contrastive loss in the latent space. We also employ a spectral embedding using the normalized nearest-neighbor graph Laplacian to extract features. We present the algorithm in Algorithm 3, where we use $k = 256$ as the number of nearest neighbor for all three datasets. Below, we provide a more detailed description of the preprocessing steps performed for each dataset.

- MNIST. We use the normalized Laplacian to reduce the dimension of the input data to dimension of 20. In Algorithm 3, $N = 60,000$, $D = 784$, and $d = 20$. For the active learning runs, we randomly select $m = 3,000$ points (with 300 points in each class id) to form the unlabeled data set U.
- CIFAR-10. First, we use pre-trained SimCLR model on the whole training data and extract the feature maps from the last layer (with dimension 512). Second, we use the normalized Laplacian to reduce the dimension of the training data to dimension of 20. In Algorithm 3, $N = 50,000$, $D = 512$, and $d = 20$. For the active learning tests, we randomly select $m = 3,000$ points (with 300 points in each class id) to form the unlabeled data set U.
- ImageNet-50. We first randomly select 50 classes from the training set of ImageNet. We use pre-trained SimCLR model and extract the features with dimension 2048. Then we use the normalized Laplacian to reduce the dimension of the training data to dimension of 40. In Algorithm 3, $D = 2048$, and $d = 40$. or the active learning tests, we randomly select $m = 5,000$ points (with 100 points in each class id) to form the unlabeled data set U.

Tuning hyperparameter η. In Algorithm 1, we have to set the learning rate η. We try different η and select the one that maximizes $\lambda_{\min}(\sum_{t=1}^{b} \widetilde{\mathbf{H}}(x_{i_t}))$ since this is our goal of the sparsification step (lines 3-11 in Algorithm 1). Note that for each round of active learning, we only need to solve the relaxed problem Eq. (14) once. Furthermore, tuning η does not require labeling information.

Additional results. We have presented the classification accuracy on unlabeled set in Fig. 3. In Fig. 12, we plot the normalized weights $z_\diamond$ (i.e. the solution of the relaxed problem Eq. (14)) at each round of active learning tests. We present the images selected by different active learning methods for MNIST (Fig. 13), CIFAR-10 (Fig. 14), and ImageNet-50 (Figs. 15 and 16).

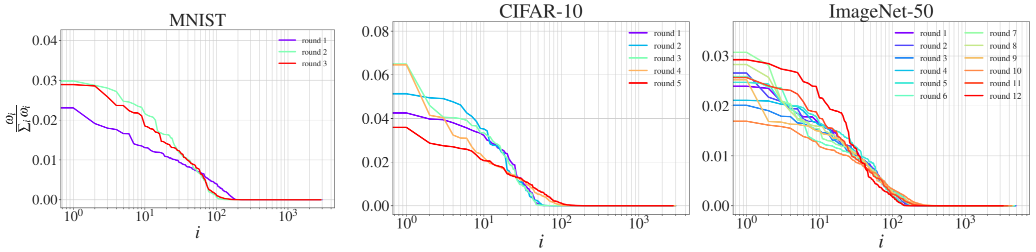

Figure 12: Normalized weights $z_\diamond$ (solution of Eq. (14)) at each round of active learning tests.

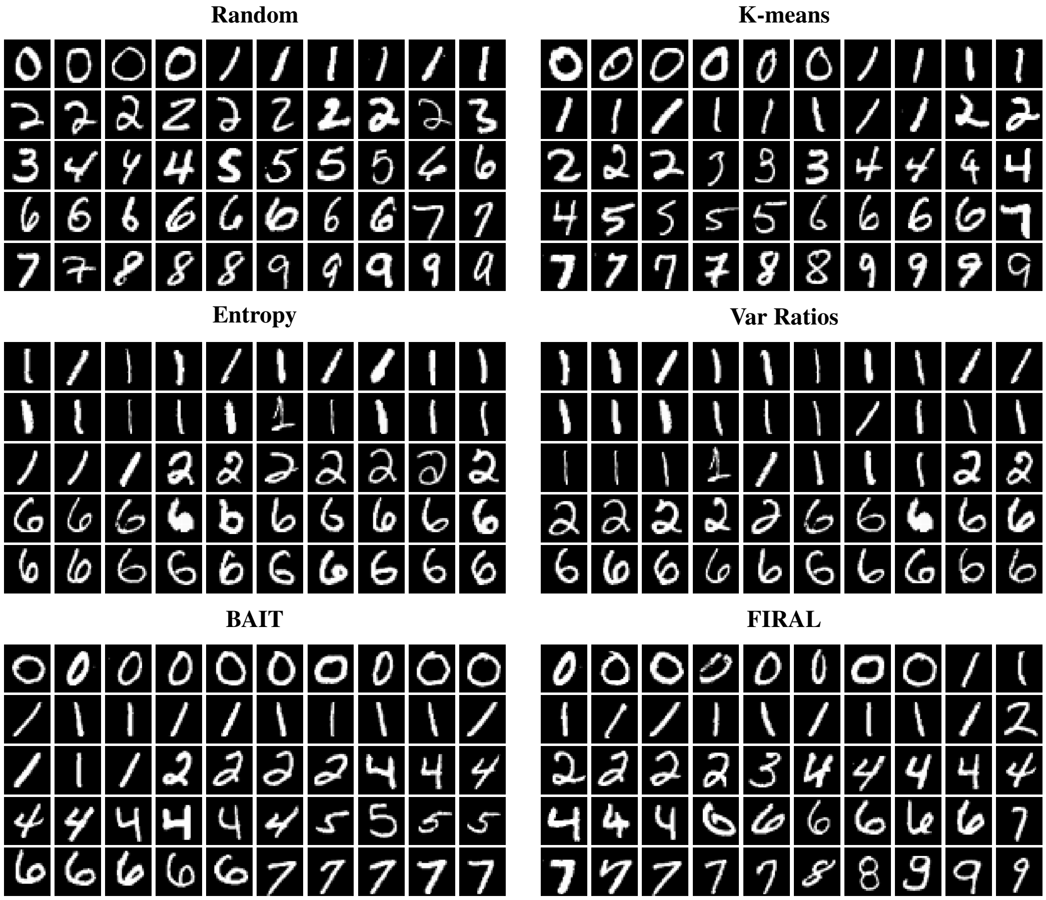

Figure 13: Selected samples for MNIST at the first round of active learning test.

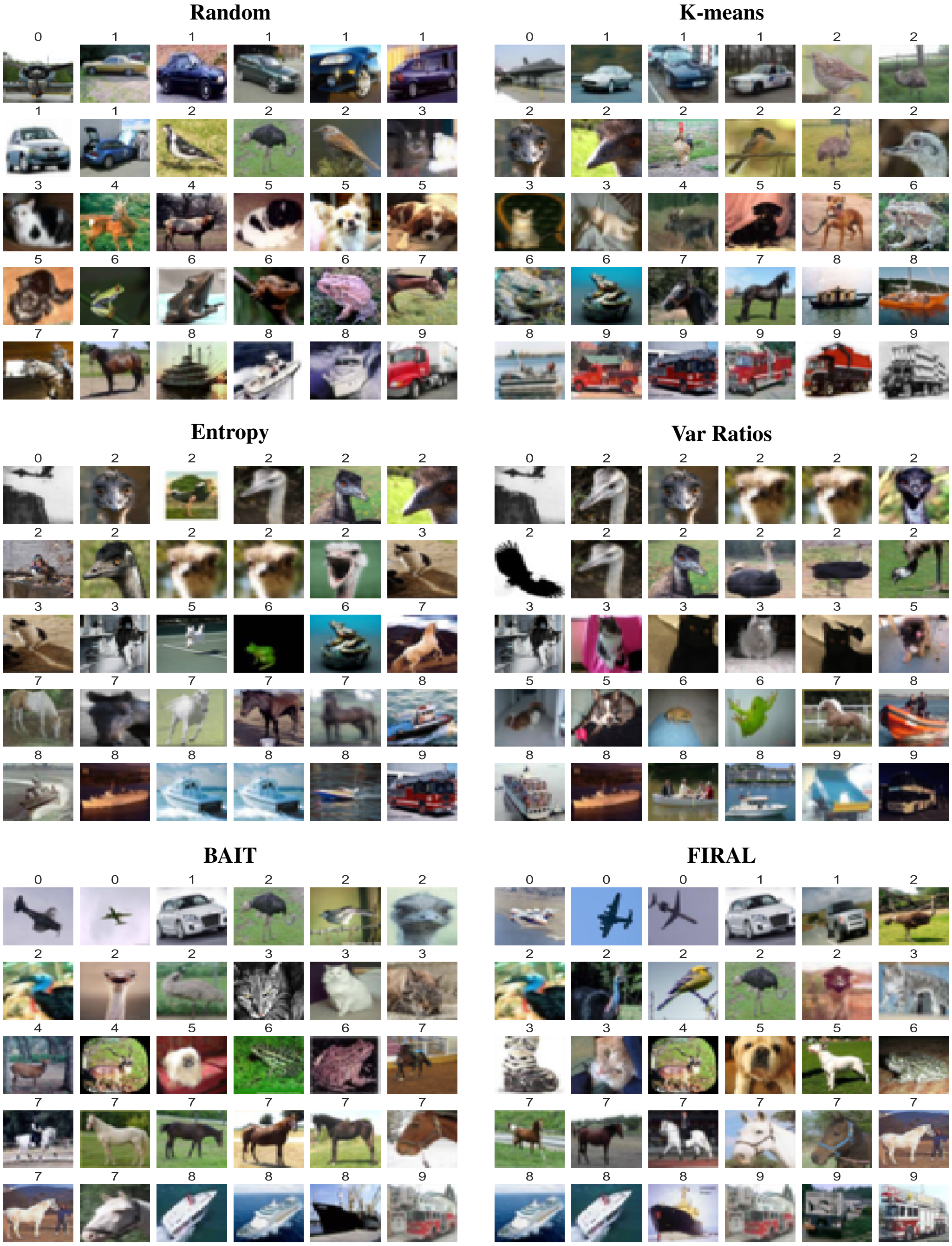

Figure 14: Selected samples for CIFAR10 at the first three rounds of active learning test.

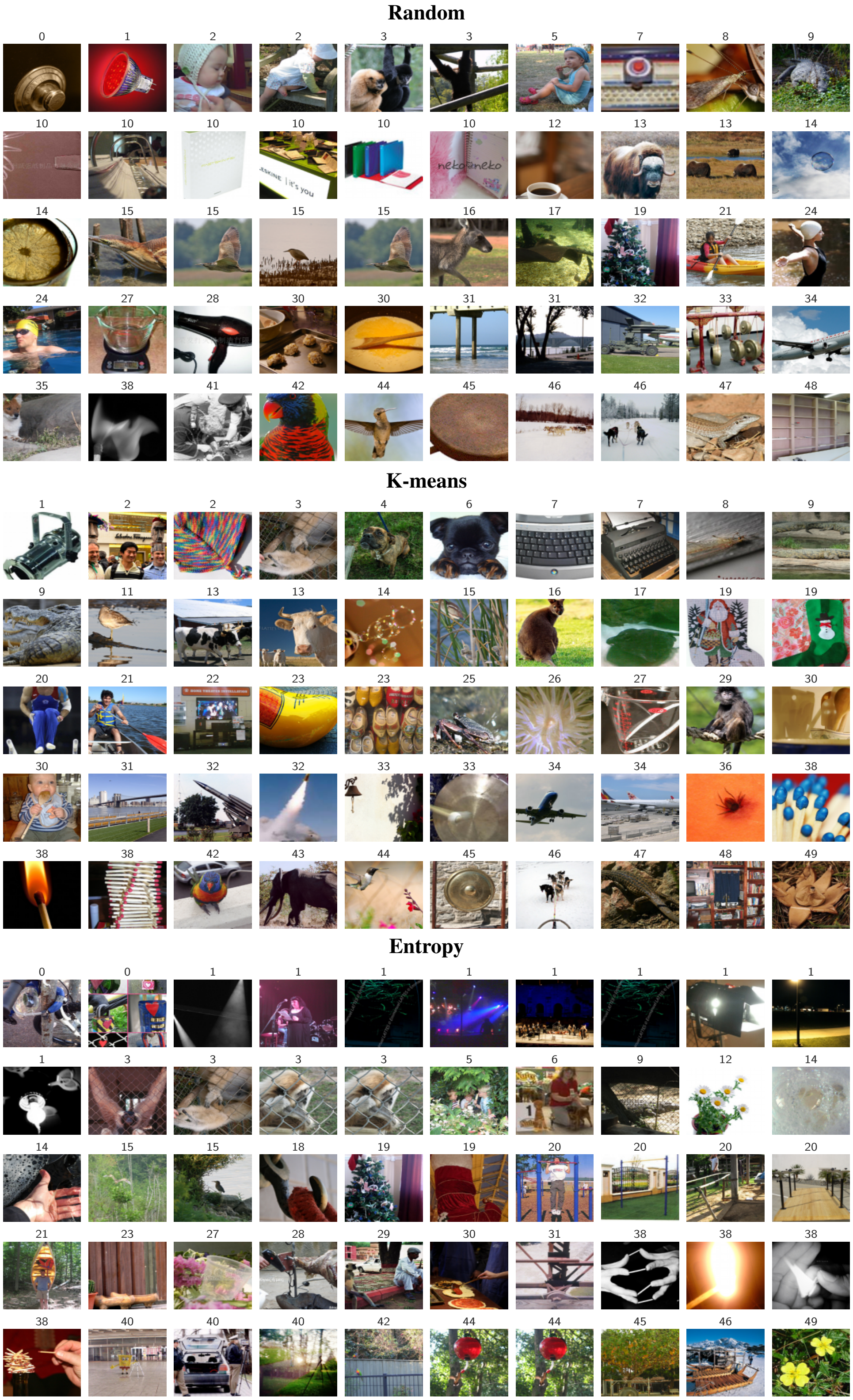

Figure 15: Selected samples for ImageNet-50 at the first round of active learning test.

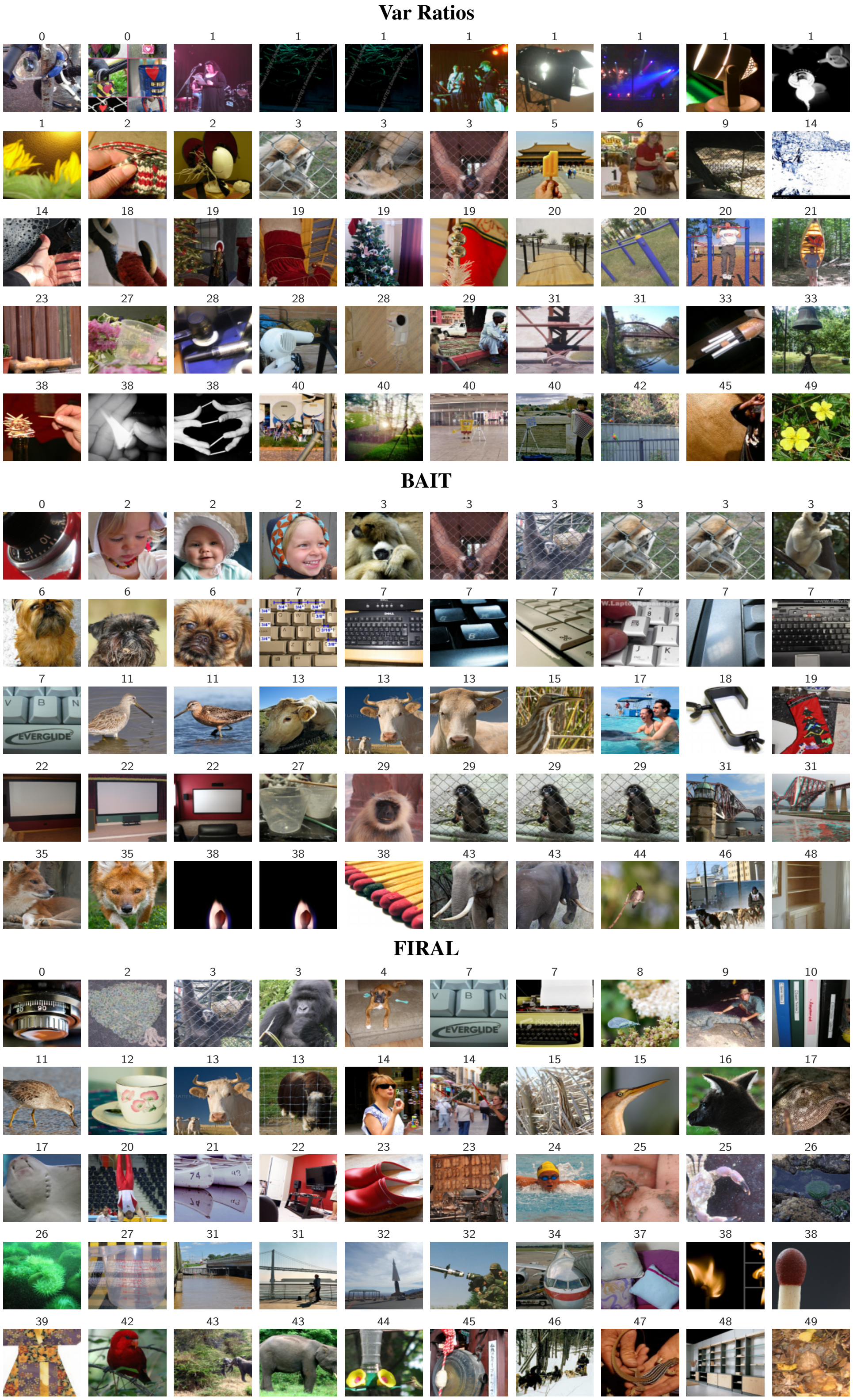

Figure 16: Selected samples for ImageNet-50 at the first round of active learning test.