Appendix

We provide proofs of main results and technical lemmas.

A. Proof of Theorem 1

First we invoke Lemma 5.1 with $\epsilon = \Delta/(10\rho)$ which outputs an orthonormal matrix $\mathbf{U}$ such that

$$\left\|\left(\mathbf{U}\mathbf{U}^\top - \mathbf{I}\right)\mathbf{w}_i\right\|_2 \;\; \leq \;\; \Delta/20 \tag{13}$$

with probability $1-\delta$. This step requires a dataset with

$$n_{L1} = \Omega\left(\frac{d}{t_{L1}} \cdot \min\left\{\Delta^{-6} p_{\min}^{-2}, \Delta^{-2}\lambda_{\min}^{-2}\right\} \cdot \log^3\left(\frac{d}{p_{\min}\Delta\delta}\right)\right)$$

i.i.d. tasks each with t_{L1} number of examples.

Second we invoke Lemma 5.2 with the matrix $\mathbf{U}$ estimated in Lemma 5.1 and $\widetilde{\epsilon} = \min\left\{\frac{\Delta}{20}, \frac{\Delta^2\sqrt{k}}{100}\right\}$ which outputs parameters satisfying

$$\begin{aligned} \left\|\mathbf{U}^\top(\widetilde{\mathbf{w}}_i - \mathbf{w}_i)\right\|_2 &\;\; \leq \;\; \Delta/20 \\ \left|\widetilde{r}_i^2 - r_i^2\right| &\;\; \leq \;\; \frac{\Delta^2}{100} r_i^2 \; . \end{aligned}$$

This step requires a dataset with

$$n_H \;\; = \;\; \Omega\left(\frac{\log(k/\delta)}{t_H \; p_{\min}\Delta^2}\left(k + \Delta^{-2}\right)\right)$$

i.i.d. tasks each with $t_H = \Omega\left(\Delta^{-2}\sqrt{k}\log\left(\frac{k}{p_{\min}\Delta\delta}\right)\right)$ number of examples.

Finally we invoke Lemma 5.3. Notice that in the last step we have estimated each $\mathbf{w}_i$ with error $\|\widetilde{\mathbf{w}}_i - \mathbf{w}_i\|_2 \leq \left\|\mathbf{U}\mathbf{U}^\top\widetilde{\mathbf{w}}_i - \mathbf{U}\mathbf{U}^\top\mathbf{w}_i\right\|_2 + \left\|\mathbf{U}\mathbf{U}^\top\mathbf{w}_i - \mathbf{w}_i\right\|_2 \leq \Delta/10$. Hence the input for Lemma 5.3 satisfies $\|\widetilde{\mathbf{w}}_i - \mathbf{w}_i\|_2 \leq \Delta/10$. It is not hard to verify that

$$\left(1 + \frac{\Delta^2}{50\rho^2}\right)\widetilde{r}_i^2 \geq \left(s_i^2 + \|\widetilde{\mathbf{w}}_i - \mathbf{w}_i\|_2^2\right) \geq \left(1 - \frac{\Delta^2}{50\rho^2}\right)\widetilde{r}_i^2$$

Hence, given

$$n_{L2} \;\; = \;\; \Omega\left(\frac{d\log^2(k/\delta)}{t_{L2}p_{\min}\epsilon^2}\right)$$

i.i.d. tasks each with $t_{L2} = \Omega\left(\log\left(\frac{kd}{p_{\min}\delta\epsilon}\right)/\Delta^4\right)$ examples. We have parameter estimation with accuracy

$$\begin{aligned} \|\widehat{\mathbf{w}}_i - \mathbf{w}_i\|_2 &\leq \epsilon s_i \; , \\ \left|\widehat{s}_i^2 - s_i^2\right| &\leq \frac{\epsilon}{\sqrt{d}} s_i^2 \; , \quad \text{and} \\ |\widehat{p}_i - p_i| &\leq \epsilon\sqrt{t_{L2}/d}\,p_{\min}. \end{aligned}$$

This concludes the proof.

A.1. Proof of Lemma 5.1

Proposition A.1 (Several facts for sub-Gaussian random variables)**.** *Under our data generation model, let* $c_1 > 1$ *denote a sufficiently large constant, let* $\delta \in (0,1)$ *denote the failure probability. We have, with probability* $1-\delta$*, for all* $i \in [n]$*,*

$$\left\|\frac{1}{t}\sum_{j=1}^{t} y_{i,j}\mathbf{x}_{i,j} - \beta_i\right\|_2 \leq c_1 \cdot \sqrt{d} \cdot \rho \cdot \log(nd/\delta) \cdot t^{-1/2}.$$

Remark A.2. *The above about is not tight, and can be optimized to* $\log(\cdot)/t + \log^{1/2}(\cdot)/t^{1/2}$. *Since we don't care about log factors, we only write* $\log(\cdot)/t^{1/2}$ *instead (note that* $t \geq 1$*).*

Proof. For each $i \in [n], j \in [t], k \in [d]$, $y_{i,j}x_{i,j,k}$ is a sub-exponential random variable with sub-exponential norm $\|y_{i,j}x_{i,j,k}\|_{\psi_1} \leq \sqrt{s_i^2 + \|\beta_i\|_2^2} = \rho_i$.

By Bernstein's inequality,

$$\mathbb{P}\left[\left|\frac{1}{t}\sum_{j=1}^{t} y_{i,j}x_{i,j,k} - \beta_{i,k}\right| \geq z\right] \leq 2\exp\left(-c\min\left\{\frac{z^2 t}{\rho_i^2}, \frac{zt}{\rho_i}\right\}\right)$$

for some $c > 0$. Hence we have that with probability $1 - 2\delta,\ \forall\, i \in [n]\,, k \in [d]$,

$$\left|\frac{1}{t}\sum_{j=1}^{t} y_{i,j}x_{i,j,k} - \beta_{i,k}\right| \leq \rho_i \max\left\{\frac{\log(nd/\delta)}{ct}, \sqrt{\frac{\log(nd/\delta)}{ct}}\right\},$$

which implies

$$\left\|\frac{1}{t}\sum_{j=1}^{t} y_{i,j}\mathbf{x}_{i,j} - \beta_i\right\|_2 \leq \sqrt{d}\rho_i \max\left\{\frac{\log(nd/\delta)}{ct}, \sqrt{\frac{\log(nd/\delta)}{ct}}\right\}.$$

□

Proposition A.3. *For any* $\mathbf{v} \in \mathbb{S}^{d-1}$

$$\mathbb{E}\left[\left\langle \mathbf{v}, \frac{1}{t}\sum_{j=1}^{t} y_{i,j}\mathbf{x}_{i,j} - \beta_i\right\rangle^2\right] \leq \mathcal{O}\left(\rho_i^2/t\right).$$

Proof.

$$\begin{aligned}\mathbb{E}\left[\left\langle \mathbf{v}, \frac{1}{t}\sum_{j=1}^{t} y_{i,j}\mathbf{x}_{i,j} - \beta_i\right\rangle^2\right] &= \frac{1}{t^2}\sum_{j=1}^{t}\sum_{j'=1}^{t}\mathbb{E}\left[\mathbf{v}^\top\left(y_{i,j}\mathbf{x}_{i,j} - \beta_i\right)\mathbf{v}^\top\left(y_{i,j'}\mathbf{x}_{i,j'} - \beta_i\right)\right] \\ &= \frac{1}{t^2}\sum_{j=1}^{t}\sum_{j'=1}^{t}\mathbf{v}^\top\,\mathbb{E}\left[\left(y_{i,j}\mathbf{x}_{i,j} - \beta_i\right)\left(y_{i,j'}\mathbf{x}_{i,j'} - \beta_i\right)^\top\right]\mathbf{v}\end{aligned}$$

where

$$\begin{aligned}&\mathbb{E}\left[\left(y_{i,j}\mathbf{x}_{i,j} - \beta_i\right)\left(y_{i,j'}\mathbf{x}_{i,j'} - \beta_i\right)^\top\right] \\ =\ &\mathbb{E}\left[\mathbf{x}_{i,j}\left(\mathbf{x}_{i,j}^\top\beta_i + \epsilon_{i,j}\right)\left(\beta_i^\top\mathbf{x}_{i,j'} + \epsilon_{i,j'}\right)\mathbf{x}_{i,j'}^\top - \left(\mathbf{x}_{i,j}^\top\beta_i + \epsilon_{i,j}\right)\mathbf{x}_{i,j}\beta_i^\top - \left(\mathbf{x}_{i,j'}^\top\beta_i + \epsilon_{i,j'}\right)\mathbf{x}_{i,j'}\beta_i^\top + \beta_i\beta_i^\top\right] \\ =\ &\mathbb{E}\left[\mathbf{x}_{i,j}\mathbf{x}_{i,j}^\top\beta_i\beta_i^\top\mathbf{x}_{i,j'}\mathbf{x}_{i,j'}^\top + \epsilon_{i,j}\epsilon_{i,j'}\mathbf{x}_{i,j}\mathbf{x}_{i,j'}^\top - \left(\mathbf{x}_{i,j}^\top\beta_i\right)^2 - \left(\mathbf{x}_{i,j'}^\top\beta_i\right)^2 + \beta_i\beta_i^\top\right] \\ =\ &\mathbb{E}\left[\mathbf{x}_{i,j}\mathbf{x}_{i,j}^\top\beta_i\beta_i^\top\mathbf{x}_{i,j'}\mathbf{x}_{i,j'}^\top - \beta_i\beta_i^\top\right] + \mathbb{E}\left[\epsilon_{i,j}\epsilon_{i,j'}\mathbf{x}_{i,j}\mathbf{x}_{i,j'}^\top\right].\end{aligned}$$

Therefore, when $j \neq j'$,

$$\mathbb{E}\left[\left(y_{i,j}\mathbf{x}_{i,j} - \beta_i\right)\left(y_{i,j'}\mathbf{x}_{i,j'} - \beta_i\right)^\top\right] = 0.$$

Plugging back we have

$$\begin{aligned}\mathbb{E}\left[\left\langle \mathbf{v}, \frac{1}{t}\sum_{j=1}^{t} y_{i,j}\mathbf{x}_{i,j} - \beta_i\right\rangle^2\right] &= \frac{1}{t^2}\sum_{j=1}^{t}\mathbb{E}\left[\left(\mathbf{v}^\top\mathbf{x}_{i,j}\right)^2\left(\beta_i^\top\mathbf{x}_{i,j}\right)^2 - \left(\mathbf{v}^\top\beta_i\right)^2\right] + \mathbf{v}^\top\,\mathbb{E}\left[\epsilon_{i,j}^2\mathbf{x}_{i,j}\mathbf{x}_{i,j}^\top\right]\mathbf{v} \\ &\leq \frac{1}{t^2}\sum_{j=1}^{t}\mathcal{O}\left(\|\mathbf{v}\|_2^2\,\|\beta_i\|_2^2\right) + \mathcal{O}\left(\mathbf{v}^\top\beta_i\right)^2 + s_i^2\,\|\mathbf{v}\|_2^2 \\ &\leq \mathcal{O}\left(\rho_i^2/t\right).\end{aligned}$$

□

Proposition A.4.

$$\mathbb{E}\left[\left\|\frac{1}{t}\sum_{j=1}^{t} y_{i,j}\mathbf{x}_{i,j} - \beta_i\right\|_2^2\right] \leq \mathcal{O}\left(\rho_i^2 d/t\right)$$

Proof.

$$\begin{aligned}
&\mathbb{E}\left[\left\langle \frac{1}{t}\sum_{j=1}^{t}\left(y_{i,j}\mathbf{x}_{i,j} - \beta_i\right), \frac{1}{t}\sum_{j'=1}^{t}\left(y_{i,j'}\mathbf{x}_{i,j'} - \beta_i\right)\right\rangle\right] \\
&= \frac{1}{t^2}\sum_{j=1}^{t}\sum_{j'=1}^{t}\mathbb{E}\left[y_{i,j}y_{i,j'}\mathbf{x}_{i,j}^\top\mathbf{x}_{i,j'} - \beta_i^\top y_{i,j'}\mathbf{x}_{i,j'} - \beta_i^\top y_{i,j}\mathbf{x}_{i,j} + \beta_i^\top\beta_i\right] \\
&= \frac{1}{t^2}\sum_{j=1}^{t}\sum_{j'=1}^{t}\mathbb{E}\left[y_{i,j}y_{i,j'}\mathbf{x}_{i,j}^\top\mathbf{x}_{i,j'} - \beta_i^\top\beta_i\right] \\
&= \frac{1}{t^2}\sum_{j=1}^{t}\sum_{j'=1}^{t}\mathbb{E}\left[\left(\beta_i^\top\mathbf{x}_{i,j} + \epsilon_{i,j}\right)\left(\beta_i^\top\mathbf{x}_{i,j'} + \epsilon_{i,j'}\right)\mathbf{x}_{i,j}^\top\mathbf{x}_{i,j'} - \|\beta_i\|_2^2\right] \\
&= \frac{1}{t^2}\sum_{j=1}^{t}\sum_{j'=1}^{t}\mathbb{E}\left[\beta_i^\top\mathbf{x}_{i,j}\mathbf{x}_{i,j}^\top\mathbf{x}_{i,j'}\mathbf{x}_{i,j'}^\top\beta_i + \epsilon_{i,j}\epsilon_{i,j'}\mathbf{x}_{i,j}^\top\mathbf{x}_{i,j'} - \|\beta_i\|_2^2\right].
\end{aligned}$$

The above quantity can be split into two terms, one is diagonal term, and the other is off-diagonal term.

If $j \neq j'$, then

$$\mathbb{E}\left[\beta_i^\top\mathbf{x}_{i,j}\mathbf{x}_{i,j}^\top\mathbf{x}_{i,j'}\mathbf{x}_{i,j'}^\top\beta_i + \epsilon_{i,j}\epsilon_{i,j'}\mathbf{x}_{i,j}^\top\mathbf{x}_{i,j'}\right] - \|\beta_i\|_2^2 = 0,$$

and if $j = j'$, then

$$\mathbb{E}\left[\beta_i^\top\mathbf{x}_{i,j}\mathbf{x}_{i,j}^\top\mathbf{x}_{i,j'}\mathbf{x}_{i,j'}^\top\beta_i + \epsilon_{i,j}\epsilon_{i,j'}\mathbf{x}_{i,j}^\top\mathbf{x}_{i,j'} - \|\beta_i\|_2^2\right] = \mathcal{O}\left(d\|\beta_i\|_2^2\right) + \sigma_i^2 d = \mathcal{O}\left(\rho_i^2 d\right).$$

Plugging back we get

$$\begin{aligned}
\mathbb{E}\left[\left\|\frac{1}{t}\sum_{j=1}^{t} y_{i,j}\mathbf{x}_{i,j} - \beta_i\right\|_2^2\right] &\leq \frac{1}{t^2}\cdot t\cdot\mathcal{O}\left(\rho_i^2 d\right) \\
&\leq \mathcal{O}\left(\rho_i^2 d/t\right).
\end{aligned}$$

□

Definition A.5. *Let $c_2 > 1$ denote a sufficiently large constant. We define event $\mathcal{E}$ to be the event that*

$$\forall\, i \in [n], \quad \left\|\frac{1}{t}\sum_{j=1}^{t} y_{i,j}\mathbf{x}_{i,j} - \beta_i\right\|_2 \leq c_2\cdot\sqrt{d}\cdot\rho\cdot\log(nd/\delta)/\sqrt{t},$$

and

$$\forall\, i \in [n], \quad \left\|\frac{1}{t}\sum_{j=t+1}^{t} y_{i,j}\mathbf{x}_{i,j} - \beta_i\right\|_2 \leq c_2\cdot\sqrt{d}\cdot\rho\cdot\log(nd/\delta)/\sqrt{t}.$$

It has been shown in Proposition A.1 that event $\mathcal{E}$ happens with probability δ.

Definition A.6. *For each $i \in [n]$, define matrix $\mathbf{Z}_i \in \mathbb{R}^{d\times d}$ as*

$$\mathbf{Z}_i := \left(\frac{1}{t}\sum_{j=1}^{t} y_{i,j}\mathbf{x}_{i,j}\right)\left(\frac{1}{t}\sum_{j=t+1}^{2t} y_{i,j}\mathbf{x}_{i,j}^\top\right) - \beta_i\beta_i^\top,$$

vector β'_i as

$$\beta'_i := \mathbb{E}\left[\left(\frac{1}{t}\sum_{j=1}^{t} y_{i,j}\mathbf{x}_{i,j}^\top\right)\middle|\mathcal{E}\right],$$

and matrix $\mathbf{Z}'_i \in \mathbb{R}^{d\times d}$ as

$$\mathbf{Z}'_i := \left(\frac{1}{t}\sum_{j=1}^{t} y_{i,j}\mathbf{x}_{i,j}\right)\left(\frac{1}{t}\sum_{j=t+1}^{2t} y_{i,j}\mathbf{x}_{i,j}^\top\right) - \beta'_i\beta_i'^\top.$$

We can show that $\|\beta'_i - \beta_i\|_2$ is small.

Lemma A.7. *With β'_i defined in Definition A.6, it holds that*

$$\|\beta_i - \beta'_i\|_2 \le \mathcal{O}\left(\frac{\sqrt{\delta}\rho_i}{\sqrt{t}}\right)$$

Proof. Notice that by the definition of ℓ_2 norm, it holds that

$$\begin{aligned}\|\beta_i - \beta'_i\|_2 &= \max_{\|\mathbf{v}\|_2=1}\mathbb{E}\left[\left\langle \mathbf{v}, \frac{1}{t}\sum_{j=1}^{t} y_{i,j}\mathbf{x}_{i,j} - \beta_i\right\rangle\middle|\mathcal{E}\right]\\ &= \max_{\|\mathbf{v}\|_2=1}\mathbb{E}\left[\left\langle \mathbf{v}, \frac{1}{t}\sum_{j=1}^{t} y_{i,j}\mathbf{x}_{i,j} - \beta_i\right\rangle\middle|\left\|\frac{1}{t}\sum_{j=1}^{t} y_{i,j}\mathbf{x}_{i,j} - \beta_i\right\|_2 \le c_2\cdot\sqrt{d}\cdot\rho\cdot\log(nd/\delta)/\sqrt{t}\right]\end{aligned}$$

Notice that

$$\mathbb{P}\left[\left\|\frac{1}{t}\sum_{j=1}^{t} y_{i,j}\mathbf{x}_{i,j} - \beta_i\right\|_2 \le c_2\cdot\sqrt{d}\cdot\rho\cdot\log(nd/\delta)/\sqrt{t}\right] \le \delta/n$$

. Define random variable $z_{\mathbf{v}} := \left\langle \mathbf{v}, \frac{1}{t}\sum_{j=1}^{t} y_{i,j}\mathbf{x}_{i,j} - \beta_i\right\rangle$. Applying Fact D.9 with the variance bound Proposition A.3, it holds that

$$\|\beta_i - \beta'_i\|_2 = |\mathbb{E}[z_{\mathbf{v}}|\mathcal{E}]| \le \mathcal{O}\left(\frac{\sqrt{\delta/n}\rho_i}{(1-\delta/n)\sqrt{t}}\right) = \mathcal{O}\left(\frac{\sqrt{\delta}\rho_i}{\sqrt{nt}}\right),$$

which concludes the proof. □

We can upper bound the spectral norm of matrix $\mathbf{Z}'_i$ condition on event $\mathcal{E}$,

Lemma A.8. *Let $\mathbf{Z}_i$ be defined as Definition A.6, let $c_2 > 1$ denote some sufficiently large constant, let $\delta \in (0,1)$ denote the failure probability. Then we have conditon on event $\mathcal{E}$ happens,*

$$\forall\, i \in [n], \quad \|\mathbf{Z}'_i\|_2 \le c_2\cdot d\cdot\rho_i^2\cdot\log^2(nd/\delta)/t$$

Proof. The norm of $\|\mathbf{Z}'_i\|_2$ satisfies

$$\begin{aligned}
\|\mathbf{Z}'_i\|_2 &\leq \left\| \left(\frac{1}{t} \sum_{j=1}^{t} y_{i,j}\mathbf{x}_{i,j} - \beta'_i \right) \left(\frac{1}{t} \sum_{j=t+1}^{2t} y_{i,j}\mathbf{x}_{i,j}^\top \right) \right\|_2 + \left\| \beta'_i \left(\frac{1}{t} \sum_{j=t+1}^{2t} y_{i,j}\mathbf{x}_{i,j}^\top - \beta_i'^\top \right) \right\|_2 \\
&\leq c_1\sqrt{d}\rho_i \log(nd/\delta) t^{-1/2} \cdot \left\| \frac{1}{t} \sum_{j=t+1}^{2t} y_{i,j}\mathbf{x}_{i,j} \right\|_2 + c_1\sqrt{d}\rho_i \log(nd/\delta) t^{-1/2} \cdot \|\beta'_i\|_2 \\
&= c_1\sqrt{d}\rho_i \log(nd/\delta) t^{-1/2} \cdot \left(\left\| \frac{1}{t} \sum_{j=t+1}^{2t} y_{i,j}\mathbf{x}_{i,j} \right\|_2 + \|\beta'_i\|_2 \right) \\
&\leq c_1\sqrt{d}\rho_i \log(nd/\delta) t^{-1/2} \cdot \left(\left\| \frac{1}{t} \sum_{j=t+1}^{2t} y_{i,j}\mathbf{x}_{i,j} - \beta'_i \right\|_2 + 2\,\|\beta'_i\|_2 \right) \\
&\leq c_1\sqrt{d}\rho_i \log(nd/\delta) t^{-1/2} \cdot \left(\mathcal{O}\left(1\right) \cdot \sqrt{d}\rho_i \log(nd/\delta) t^{-1/2} + 2\,\|\beta'_i\|_2 \right) \\
&\leq \mathcal{O}\left(1\right) \cdot d\rho_i^2 \log^2(nd/\delta)/t
\end{aligned}$$

where the second step follows from Proposition A.1, the fourth step follows from triangle inequality, the fifth step follows from Proposition A.1, and the last step follows $\|\beta'_i\|_2 \leq \mathcal{O}\left(\rho_i\right)$.

Rescaling the δ completes the proof. □

We can apply matrix Bernstein inequality under a conditional distribution.

Proposition A.9. *Let $\mathbf{Z}_i$ be defined as Definition A.6. Let $\mathcal{E}$ be defined as Definition A.5. Then we have*

$$\left\| \mathbb{E}\left[\sum_{i=1}^{n} \mathbf{Z}'_i \mathbf{Z}_i'^\top \middle| \mathcal{E} \right] \right\|_2 = \mathcal{O}\left(n\rho^4 d/t\right).$$

Proof.

$$\begin{aligned}
&\left\|\mathbb{E}\left[\mathbf{Z}_i\mathbf{Z}_i^\top\right]\right\|_2 \\
= &\max_{\mathbf{v}\in\mathbb{S}^{d-1}} \left[\mathbb{E}\left[\left(\mathbf{v}^\top \left(\frac{1}{t}\sum_{j=1}^{t} y_{i,j}\mathbf{x}_{i,j} \right) \right)^2 \left\| \frac{1}{t}\sum_{j=t+1}^{2t} y_{i,j}\mathbf{x}_{i,j} \right\|_2^2 - \left(\mathbf{v}^\top\beta_i\right)^2 \|\beta_i\|_2^2 \right] \right] \\
= &\max_{\mathbf{v}\in\mathbb{S}^{d-1}} \left[\mathbb{E}\left[\left(\mathbf{v}^\top \left(\frac{1}{t}\sum_{j=1}^{t} y_{i,j}\mathbf{x}_{i,j} - \beta_i \right) \right)^2 \left\| \frac{1}{t}\sum_{j=t+1}^{2t} y_{i,j}\mathbf{x}_{i,j} \right\|_2^2 \right] + \mathbb{E}\left[\left(\mathbf{v}^\top\beta_i\right)^2 \left\| \left(\frac{1}{t}\sum_{j=t+1}^{2t} y_{i,j}\mathbf{x}_{i,j} \right) - \beta_i \right\|_2^2 \right] \right] \\
\lesssim &(\rho_i^2/t) \cdot (\|\beta_i\|_2^2 + \rho_i^2 d/t) + \|\beta_i\|_2^2\,(\rho_i^2 d/t) \\
\leq &(\rho_i^2/t) \cdot (\rho_i^2 + \rho_i^2 d/t) + \rho_i^2 \cdot (\rho_i^2 d/t) \\
\leq &2\rho_i^4 d/t^2 + \rho_i^4 d/t \\
\leq &3\rho_i^4 d/t.
\end{aligned}$$

where the fourth step follows from $\|\beta_i\|_2 \leq \rho_i$, the fifth step follows $d/t \geq 1$, and the last step follows from $t \geq 1$.

Thus,

$$\left\| \mathbb{E}\left[\sum_{i=1}^{n} \mathbf{Z}_i\mathbf{Z}_i^\top \middle| \mathcal{E} \right] \right\|_2 \leq \frac{1}{\mathbb{P}\left[\mathcal{E}\right]} \left\| \mathbb{E}\left[\sum_{i=1}^{n} \mathbf{Z}_i\mathbf{Z}_i^\top \right] \right\|_2 = \mathcal{O}\left(n\rho^4 d/t\right).$$

where n comes from repeatedly applying triangle inequality. Since

$$\begin{aligned}\left\|\mathbb{E}\left[\sum_{i=1}^n \mathbf{Z}_i\mathbf{Z}_i^\top \middle| \mathcal{E}\right]\right\|_2 &= \left\|\mathbb{E}\left[\sum_{i=1}^n \left(\mathbf{Z}_i' + \beta_i\beta_i^\top - \beta_i'\beta_i'^\top\right)\left(\mathbf{Z}_i'^\top + \beta_i\beta_i^\top - \beta_i'\beta_i'^\top\right)\middle|\mathcal{E}\right]\right\|_2 \\ &= \left\|\mathbb{E}\left[\sum_{i=1}^n \mathbf{Z}_i'\mathbf{Z}_i'^\top\middle|\mathcal{E}\right] + n\left(\beta_i\beta_i^\top - \beta_i'\beta_i'^\top\right)^2\right\|_2,\end{aligned}$$

and

$$\left\|\left(\beta_i\beta_i^\top - \beta_i'\beta_i'^\top\right)^2\right\|_2 = \left\|\beta_i\beta_i^\top - \beta_i'\beta_i'^\top\right\|_2^2 \le \left(\left\|\beta_i\left(\beta_i^\top - \beta_i'^\top\right)\right\|_2 + \left\|\left(\beta_i - \beta_i'\right)\beta_i'^\top\right\|_2\right)^2 \le \mathcal{O}\left(\rho^4\delta/nt\right),$$

by triangle inequality, it holds that

$$\left\|\mathbb{E}\left[\sum_{i=1}^n \mathbf{Z}_i'\mathbf{Z}_i'^\top\middle|\mathcal{E}\right]\right\|_2 \le \mathcal{O}\left(n\rho^4 d/t\right)$$

□

Using the fact thathave $\mathbb{E}\left[\mathbf{Z}_i'|\mathcal{E}\right] = 0$, we can apply matrix Bernstein inequality on $\mathbf{Z}_i|\mathcal{E}$, which will imply the following bounds on $\mathbf{Z}_i$:

Lemma A.10. *Let $\mathbf{Z}_i$ be defined as Definition A.6. For any $\widetilde{\epsilon} \in (0,1)$ and $\delta \in (0,1)$, if*

$$n = \Omega\left(\frac{d}{t}\log^2\left(nd/\delta\epsilon\right)\max\left\{\frac{1}{\widetilde{\epsilon}^2}, \frac{1}{\widetilde{\epsilon}}\log\frac{nd}{\delta\epsilon}\right\}\right),$$

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

$$\left\|\frac{1}{n}\sum_{i=1}^n \mathbf{Z}_i\right\|_2 \le \widetilde{\epsilon}\cdot\rho^2.$$

Proof. Recall that $\mathcal{E}$ is defined as Definition A.5.

Using matrix Bernstein inequality (Proposition D.5), we get for any $z > 0$,

$$\mathbb{P}\left[\left\|\frac{1}{n}\sum_{i=1}^n \mathbf{Z'}_i\right\|_2 \ge z \;\middle|\; \mathcal{E}\right] \le d\cdot\exp\left(-\frac{z^2n/2}{\rho^4 d/t + zcd\rho^2\log^2(nd/\delta)/t}\right).$$

For $z = \widetilde{\epsilon}\rho^2$, we get

$$\mathbb{P}\left[\left\|\frac{1}{n}\sum_{i=1}^n \mathbf{Z'}_i\right\|_2 \ge \widetilde{\epsilon}\rho^2 \;\middle|\; \mathcal{E}\right] \le d\cdot\exp\left(-\frac{\widetilde{\epsilon}^2 n/2}{d/t + \widetilde{\epsilon}cd\log^2(nd/\delta)/t}\right) \tag{14}$$

for some $c > 0$. If we want to bound the right hand side of Equation (14) by δ, it is sufficient to have

$$\begin{aligned}\frac{\widetilde{\epsilon}^2 n/2}{d/t + \widetilde{\epsilon}cd\log^2(nd/\delta)/t} &\ge \log\frac{nd}{\delta} \\ \text{or, } n &\gtrsim \frac{d}{t}\log^2\left(nd/\delta\right)\max\left\{\frac{1}{\widetilde{\epsilon}^2}, \frac{1}{\widetilde{\epsilon}}\log\frac{nd}{\delta}\right\}\end{aligned} \tag{15}$$

Therefore, if $\widetilde{\epsilon}\log(nd/\delta) \gtrsim 1$, we just need $n \gtrsim \frac{d}{\widetilde{\epsilon}t}\log^3\left(nd/\delta\right)$, else we need $n \gtrsim \frac{d}{t\widetilde{\epsilon}^2}\log^2(nd/\delta)$ such that $\mathbb{P}\left[\left\|\frac{1}{n}\sum_{i=1}^n \mathbf{Z'}_i\right\|_2 \ge \widetilde{\epsilon}\rho^2 \;\middle|\; \mathcal{E}\right] \le \delta$. Since

$$\left\|\frac{1}{n}\sum_{i=1}^n \mathbf{Z'}_i - \frac{1}{n}\sum_{i=1}^n \mathbf{Z}_i\right\|_2 = \left\|\frac{1}{n}\sum_{i=1}^n\left(\beta_i\beta_i^\top - \beta_i'\beta_i'^\top\right)\right\|_2 \le \rho^2\sqrt{\delta/nt},$$

we have that

$$\mathbb{P}\left[\left\|\frac{1}{n}\sum_{i=1}^{n}\mathbf{Z}_i\right\|_2 \geq \left(\widetilde{\epsilon}+\sqrt{\delta/nt}\right)\rho^2 \;\middle|\; \mathcal{E}\right] \leq \delta,$$

which implies that

$$\mathbb{P}\left[\left\|\frac{1}{n}\sum_{i=1}^{n}\mathbf{Z}_i\right\|_2 \geq \left(\widetilde{\epsilon}+\sqrt{\delta/nt}\right)\rho^2\right] \leq 2\delta.$$

Since $\sqrt{\delta/nt} \leq \widetilde{\epsilon}$ for n defined in the statement of the lemma, we conclude the proof. □

Lemma A.11. *If* $\mathbf{X} = \frac{1}{n}\sum_{i=1}^{n}\beta_i\beta_i^\top$ *where* $\beta_i = \mathbf{w}_j$ *with probability* p_j*, and* $\mathbf{M} = \sum_{j=1}^{k} p_j\mathbf{w}_j\mathbf{w}_j^\top$ *as its expectation, then for any* $\delta \in (0,1)$ *we have*

$$\mathbb{P}\left[\|\mathbf{X}-\mathbf{M}\|_2 \leq \widetilde{\epsilon}\rho^2\right] \geq 1-\delta. \tag{16}$$

if $n = \Omega\left(\frac{\log^3(k/\delta)}{\widetilde{\epsilon}^2}\right)$.

Proof. Let $\widetilde{p}_j = \frac{1}{n}\sum_{i=1}^{n}\mathbb{1}\left\{\mathbf{w}_j = \beta_i\right\}\ \forall\, j \in [k]$, then $\mathbf{X} = \sum_{j=1}^{k}\widetilde{p}_j\mathbf{w}_j\mathbf{w}_j^\top$. Let $\mathbf{S}_j = (\widetilde{p}_j - p_j)\mathbf{w}_j\mathbf{w}_j^\top\ \forall j \in [k]$, then we have the following for all $j \in [k]$,

$$\begin{aligned}
\mathbb{E}\left[\mathbf{S}_j\right] &= \mathbf{0} \\
\left\|\mathbf{S}_j\right\|_2 &\leq \rho^2\sqrt{\frac{3\log(k/\delta)}{n}} \qquad \text{(from Proposition D.7)} && (17)\\
\left\|\sum_{j=1}^{k}\mathbb{E}\left[\mathbf{S}_j^\top\mathbf{S}_j\right]\right\|_2 &= \left\|\sum_{j=1}^{k}\mathbb{E}\left[(\widetilde{p}_j-p_j)^2\right]\left\|\mathbf{w}_j\right\|_2^2\mathbf{w}_j\mathbf{w}_j^\top\right\|_2 \\
&\leq 3\rho^2\frac{\log(k/\delta)}{n}\left\|\sum_{j=1}^{k}p_j\mathbf{w}_j\mathbf{w}_j^\top\right\|_2 \qquad \text{(from Proposition D.7)} \\
&\leq 3\rho^4\frac{\log(k/\delta)}{n}. && (18)
\end{aligned}$$

Conditioning on the event $\mathcal{E} := \left\{|\widetilde{p}_j - p_j| \leq \sqrt{3\log(k/\delta)/n}\right\}$, from matrix Bernstein D.5 we have

$$\begin{aligned}
&\mathbb{P}\left[\left\|\sum_{j=1}^{k}\mathbf{S}_j\right\|_2 \geq z \;\middle|\; \mathcal{E}\right] \leq 2k\exp\left(\frac{-z^2/2}{3\rho^4\frac{\log(k/\delta)}{n}+\frac{\rho^2 z}{3}\sqrt{\frac{3\log(k/\delta)}{n}}}\right) \\
\implies &\mathbb{P}\left[\left\|\sum_{j=1}^{k}\mathbf{S}_j\right\|_2 \leq 3\rho^2\frac{\log^{3/2}(k/\delta)}{\sqrt{n}} \;\middle|\; \mathcal{E}\right] \geq 1-\delta && (19)
\end{aligned}$$

Since $\mathbb{P}\left[\mathcal{E}\right] \geq 1-\delta$, we have

$$\mathbb{P}\left[\left\|\sum_{j=1}^{k}\mathbf{S}_j\right\|_2 \leq \widetilde{\epsilon}\rho^2\right] \geq 1-\delta \tag{20}$$

for $n = \Omega\left(\frac{\log^3(k/\delta)}{\widetilde{\epsilon}^2}\right)$. □

Lemma A.12. *Given k vectors $\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_k \in \mathbb{R}^d$. For each $i \in [k]$, we define $\mathbf{X}_i = \mathbf{x}_i \mathbf{x}_i^\top$. For every $\gamma \geq 0$, and every PSD matrix $\widehat{\mathbf{M}} \in \mathbb{R}^{d \times d}$ such that*

$$\left\| \widehat{\mathbf{M}} - \sum_{i=1}^{k} \mathbf{X}_i \right\|_2 \quad \leq \quad \gamma, \tag{21}$$

let $\mathbf{U} \in \mathbb{R}^{d \times k}$ be the matrix consists of the top-k singular vectors of $\widehat{\mathbf{M}}$, then for all $i \in [k]$,

$$\left\| \mathbf{x}_i^\top \left(\mathbf{I} - \mathbf{U}\mathbf{U}^\top \right) \right\|_2 \quad \leq \quad \min \left\{ \gamma \|\mathbf{x}_i\|_2 / \sigma_{\min} \, , \, \sqrt{2} \left(\gamma \|\mathbf{x}_i\|_2 \right)^{1/3} \right\} \, ,$$

where $\sigma_{\min}$ is the smallest non-zero singular value of $\sum_{i \in [k]} \mathbf{X}_i$.

Proof. From the gap-free Wedin's theorem in (Allen-Zhu & Li, 2016, Lemma B.3), it follows that

$$\left\| (\mathbf{I} - \mathbf{U}\mathbf{U}^\top) \mathbf{V}_j \right\|_2 \leq \gamma / \sigma_j \, , \tag{22}$$

where $\mathbf{V}_j = [\mathbf{v}_1 \, \ldots \, \mathbf{v}_j]$ is the matrix consisting of the j singular vectors of $\sum_{i' \in [k]} \mathbf{X}_{i'}$ corresponding to the top j singular values, and σ_j is the j-th singular value. To get the first term on the upper bound, notice that $\mathbf{x}_i$ lie on the subspace spanned by $\mathbf{V}_j$ where j is the rank of $\sum_{i' \in [k]} \mathbf{X}_{i'}$. It follows that

$$\left\| \left(\mathbf{I} - \mathbf{U}\mathbf{U}^\top \right) \mathbf{V}_j \mathbf{V}_j^T \mathbf{x}_i \right\|_2 \leq \|\mathbf{x}_i\|_2 \, \gamma / \sigma_j \leq \|\mathbf{x}_i\|_2 \, \gamma / \sigma_{\min}.$$

Next, we optimize over this choice of j to get the tightest bound that does not depend on the singular values.

$$\begin{aligned} \left\| \left(\mathbf{I} - \mathbf{U}\mathbf{U}^\top \right) \mathbf{x}_i \right\|_2^2 &= \left\| \left(\mathbf{I} - \mathbf{U}\mathbf{U}^\top \right) \mathbf{V}_j \mathbf{V}_j^\top \mathbf{x}_i \right\|_2^2 + \left\| \left(\mathbf{I} - \mathbf{U}\mathbf{U}^\top \right) \left(\mathbf{I} - \mathbf{V}_j \mathbf{V}_j^\top \right) \mathbf{x}_i \right\|_2^2 \\ &\leq \left(\gamma^2 / \sigma_j^2 \right) \|\mathbf{x}_i\|_2^2 + \sigma_{j+1} \, , \end{aligned}$$

for any $j \in [k]$ where we used $\left\| \left(\mathbf{I} - \mathbf{V}_j \mathbf{V}_j^\top \right) \mathbf{x}_i \right\|_2^2 \leq \sigma_{j+1}$. This follows from

$$\sigma_{j+1} = \left\| \left(\mathbf{I} - \mathbf{V}_j \mathbf{V}_j^\top \right) \sum_{i' \in [k]} \mathbf{X}_{i'} \left(\mathbf{I} - \mathbf{V}_j \mathbf{V}_j^\top \right) \right\|_2 \geq \left\| \left(\mathbf{I} - \mathbf{V}_j \mathbf{V}_j^\top \right) \mathbf{x}_i \mathbf{x}_i^\top \left(\mathbf{I} - \mathbf{V}_j \mathbf{V}_j^\top \right) \right\|_2 = \left\| \left(\mathbf{I} - \mathbf{V}_j \mathbf{V}_j^\top \right) \mathbf{x}_i \right\|_2^2 .$$

Optimal choice of j minimizes the upper bound, which happens when the two terms are of similar orders. Precisely, we choose j to be the largest index such that $\sigma_j \geq \gamma^{2/3} \|\mathbf{x}_i\|_2^{2/3}$ (we take $j = 0$ if $\sigma_1 \leq \gamma^{2/3} \|\mathbf{x}_i\|_2^{2/3}$). This gives an upper bound of $2\gamma^{2/3} \|\mathbf{x}_i\|_2^{2/3}$. This bound is tighter by a factor of $k^{2/3}$ compared to a similar result from (Li & Liang, 2018, Lemma 5), where this analysis is based on. □

Proof of Lemma 5.1. We combine Lemma A.12 and Lemma A.10 to compute the proof. Let $\epsilon > 0$ be the minimum positive real such that for $\mathbf{x}_i = \sqrt{p_i} \mathbf{w}_i$, $\gamma = \widetilde{\epsilon} \rho^2$, $\sigma_{\min} = \lambda_{\min}$, we have

$$\sqrt{p_i} \left\| \left(\mathbf{I} - \mathbf{U}\mathbf{U}^\top \right) \mathbf{w}_i \right\|_2 \leq \min \left\{ \widetilde{\epsilon} \rho^3 \sqrt{p_i} / \lambda_{\min}, \sqrt{2} \cdot \widetilde{\epsilon}^{1/3} \rho p_i^{1/6} \right\} \leq \epsilon \rho \sqrt{p_i}$$

The above equation implies that

$$\widetilde{\epsilon} = \max \left\{ \frac{\lambda_{\min} \epsilon}{\rho^2}, \frac{p_{\min} \epsilon^3}{2\sqrt{2}} \right\} .$$

Since $\left\| \sum_{i=1}^k \widetilde{p}_i \mathbf{w}_i \mathbf{w}_i^\top - \sum_{i=1}^k p_i \mathbf{w}_i \mathbf{w}_i^\top \right\|_2 + \left\| \widehat{\mathbf{M}} - \sum_{i=1}^k p_i \mathbf{w}_i \mathbf{w}_i^\top \right\|_2 \leq \mathcal{O} \left(\widetilde{\epsilon} \rho^2 \right)$ for

$$n = \Omega \left(\max \left\{ \frac{1}{\widetilde{\epsilon}^2} \log^3 (k/\delta), \frac{d}{t \widetilde{\epsilon}^2} \log^2 \left(nd/\delta \right), \frac{d}{t \widetilde{\epsilon}} \log^3 \left(nd/\delta \right) \right\} \right)$$

from Lemma A.10 and Proposition A.11, we get

$$\left\| \left(\mathbf{I} - \mathbf{U}\mathbf{U}^\top \right) \mathbf{w}_i \right\|_2 \leq \epsilon \rho \qquad \forall \, i \in [k]$$

with probability at least $1 - \delta$. □

A.2. Proof of Lemma 5.2

We start with the following two proposition which shows that the mean of our distance estimator is well separated between the in-cluster tasks and the inter-cluster tasks.

Proposition A.13. *Recall that matrix $\mathbf{U}$ satisfies Equation (8) with error ϵ. If $\Delta \geq 4\rho\epsilon$, then $\forall\, i,j \in [n]$ such that $\beta_i \neq \beta_j$,*

$$\mathbb{E}\left[\left(\widehat{\beta}_i^{(1)} - \widehat{\beta}_j^{(1)}\right)^\top \mathbf{U}\mathbf{U}^\top\mathbf{U}\mathbf{U}^\top \left(\widehat{\beta}_i^{(2)} - \widehat{\beta}_j^{(2)}\right)\right] \geq \Delta^2/4,$$

and $\forall\, i,j \in [n]$ such that $\beta_i = \beta_j$,

$$\mathbb{E}\left[\left(\widehat{\beta}_i^{(1)} - \widehat{\beta}_j^{(1)}\right)^\top \mathbf{U}\mathbf{U}^\top\mathbf{U}\mathbf{U}^\top \left(\widehat{\beta}_i^{(2)} - \widehat{\beta}_j^{(2)}\right)\right] = 0.$$

Proof. If $\beta_i \neq \beta_j$,

$$\begin{aligned}
&\mathbb{E}\left[\left(\widehat{\beta}_i^{(1)} - \widehat{\beta}_j^{(1)}\right)^\top \mathbf{U}\mathbf{U}^\top\mathbf{U}\mathbf{U}^\top \left(\widehat{\beta}_i^{(2)} - \widehat{\beta}_j^{(2)}\right)\right] \\
&= \left\|\mathbf{U}\mathbf{U}^\top(\beta_i - \beta_j)\right\|_2^2 \\
&= \left\|\mathbf{U}\mathbf{U}^\top\beta_i - \beta_i + \beta_i - \beta_j + \beta_j - \mathbf{U}\mathbf{U}^\top\beta_j\right\|_2^2 \\
&\geq \left(\|\beta_i - \beta_j\|_2 - 2\epsilon\rho\right)^2 \\
&\geq \Delta^2/4.
\end{aligned}$$

The proof is trivial for $\beta_i = \beta_j$. □

Proposition A.14.

$$\operatorname{Var}\left[\left(\widehat{\beta}_i^{(1)} - \widehat{\beta}_j^{(1)}\right)^\top \mathbf{U}\mathbf{U}^\top\mathbf{U}\mathbf{U}^\top \left(\widehat{\beta}_i^{(2)} - \widehat{\beta}_j^{(2)}\right)\right] \leq \mathcal{O}\left(\rho^4 \cdot (t+k)/t^2\right).$$

Proof. If $\beta_i \neq \beta_j$, then

$$\begin{aligned}
&\operatorname{Var}\left[\left(\widehat{\beta}_i^{(1)} - \widehat{\beta}_j^{(1)}\right)^\top \mathbf{U}\mathbf{U}^\top\mathbf{U}\mathbf{U}^\top \left(\widehat{\beta}_i^{(2)} - \widehat{\beta}_j^{(2)}\right)\right] \\
&= \mathbb{E}\left[\left(\left(\widehat{\beta}_i^{(1)} - \widehat{\beta}_j^{(1)}\right)^\top \mathbf{U}\mathbf{U}^\top \left(\widehat{\beta}_i^{(2)} - \widehat{\beta}_j^{(2)}\right)\right)^2\right] - \left((\beta_i - \beta_j)^\top \mathbf{U}\mathbf{U}^\top (\beta_i - \beta_j)\right)^2 \\
&= \frac{1}{t^4}\sum_{\substack{a,a'=1\\ b,b'=t+1}}^{t,2t} \mathbb{E}\left[\left((y_{i,a}\mathbf{x}_{i,a} - y_{j,a}\mathbf{x}_{j,a})^\top\mathbf{U}\mathbf{U}^\top(y_{i,b}\mathbf{x}_{i,b} - y_{j,b}\mathbf{x}_{j,b})\right)\left((y_{i,a'}\mathbf{x}_{i,a'} - y_{j,a'}\mathbf{x}_{j,a'})^\top\mathbf{U}\mathbf{U}^\top(y_{i,b'}\mathbf{x}_{i,b'} - y_{j,b'}\mathbf{x}_{j,b'})\right)\right] \\
&\qquad - (\beta_i - \beta_j)^\top\mathbf{U}\mathbf{U}^\top(\beta_i - \beta_j)(\beta_i - \beta_j)^\top\mathbf{U}\mathbf{U}^\top(\beta_i - \beta_j).
\end{aligned}$$

For each term in the summation, we classify it into one of the 3 different cases according to a, b, a', b':

1. If $a \neq a'$ and $b \neq b'$, the term is 0.
2. If $a = a'$ and $b \neq b'$, the term can then be expressed as:

$$\begin{aligned}
&\mathbb{E}\left[\left((y_{i,a}\mathbf{x}_{i,a} - y_{j,a}\mathbf{x}_{j,a})^\top\mathbf{U}\mathbf{U}^\top(y_{i,b}\mathbf{x}_{i,b} - y_{j,b}\mathbf{x}_{j,b})\right)\left((y_{i,a'}\mathbf{x}_{i,a'} - y_{j,a'}\mathbf{x}_{j,a'})^\top\mathbf{U}\mathbf{U}^\top(y_{i,b'}\mathbf{x}_{i,b'} - y_{j,b'}\mathbf{x}_{j,b'})\right)\right] \\
&- (\beta_i - \beta_j)^\top\mathbf{U}\mathbf{U}^\top(\beta_i - \beta_j)(\beta_i - \beta_j)^\top\mathbf{U}\mathbf{U}^\top(\beta_i - \beta_j) \\
&= \mathbb{E}\left[\left((y_{i,a}\mathbf{x}_{i,a} - y_{j,a}\mathbf{x}_{j,a})^\top\mathbf{U}\mathbf{U}^\top(\beta_i - \beta_j)\right)^2\right] - \left((\beta_i - \beta_j)^\top\mathbf{U}\mathbf{U}^\top(\beta_i - \beta_j)\right)^2 \\
&= \mathbb{E}\left[\left(y_{i,a}\mathbf{x}_{i,a}^\top\mathbf{U}\mathbf{U}^\top(\beta_i - \beta_j)\right)^2\right] - \left(\beta_i^\top\mathbf{U}\mathbf{U}^\top(\beta_i - \beta_j)\right)^2 \\
&\qquad + \mathbb{E}\left[\left(y_{j,a}\mathbf{x}_{j,a}^\top\mathbf{U}\mathbf{U}^\top(\beta_i - \beta_j)\right)^2\right] - \left(\beta_j^\top\mathbf{U}\mathbf{U}^\top(\beta_i - \beta_j)\right)^2 \\
&= \mathcal{O}\left(\rho^4\right).
\end{aligned}$$

The last equality follows from the sub-Gaussian assumption of $\mathbf{x}$.

3. If $a \neq a'$ and $b = b'$, this case is symmetric to the last case and $3\sigma_a^2\sigma_{a'}^2$ is an upper bound.

4. If $a = a'$ and $b = b'$, the term can then be expressed as:

$$\begin{aligned}
&\mathbb{E}\left[\left((y_{i,a}\mathbf{x}_{i,a} - y_{j,a}\mathbf{x}_{j,a})^\top \mathbf{U}\mathbf{U}^\top (y_{i,b}\mathbf{x}_{i,b} - y_{j,b}\mathbf{x}_{j,b})\right)^2\right] - \left((\beta_i - \beta_j)^\top \mathbf{U}\mathbf{U}^\top (\beta_i - \beta_j)\right)^2 \\
&= \mathbb{E}\left[y_{i,b}^2((y_{i,a}\mathbf{x}_{i,a} - y_{j,a}\mathbf{x}_{j,a})^\top \mathbf{U}\mathbf{U}^\top \mathbf{x}_{i,b})^2\right] + \mathbb{E}\left[y_{j,b}^2((y_{i,a}\mathbf{x}_{i,a} - y_{j,a}\mathbf{x}_{j,a})^\top \mathbf{U}\mathbf{U}^\top \mathbf{x}_{j,b})^2\right] \\
&\quad - 2\,\mathbb{E}\left[(y_{i,a}\mathbf{x}_{i,a} - y_{j,a}\mathbf{x}_{j,a})^\top \mathbf{U}\mathbf{U}^\top (y_{i,b}\mathbf{x}_{i,b})(y_{i,a}\mathbf{x}_{i,a} - y_{j,a}\mathbf{x}_{j,a})^\top \mathbf{U}\mathbf{U}^\top (y_{j,b}\mathbf{x}_{j,b})\right] \\
&\quad - \left((\beta_i - \beta_j)^\top \mathbf{U}\mathbf{U}^\top (\beta_i - \beta_j)\right)^2 .
\end{aligned}$$

First taking the expectation over $\mathbf{x}_{i,b}, y_{i,b}, \mathbf{x}_{j,b}, y_{j,b}$, we get the following upper bound

$$c_3\rho^2\,\mathbb{E}\left[\left\|(y_{i,a}\mathbf{x}_{i,a} - y_{j,a}\mathbf{x}_{j,a})^\top \mathbf{U}\mathbf{U}^\top\right\|_2^2\right] - 2\,\mathbb{E}\left[(y_{i,a}\mathbf{x}_{i,a} - y_{j,a}\mathbf{x}_{j,a})^\top \mathbf{U}\mathbf{U}^\top \beta_i (y_{i,a}\mathbf{x}_{i,a} - y_{j,a}\mathbf{x}_{j,a})^\top \mathbf{U}\mathbf{U}^\top \beta_j\right]$$

for some $c_3 > 0$. Since

$$\mathbb{E}\left[(y_{i,a}\mathbf{x}_{i,a} - y_{j,a}\mathbf{x}_{j,a})^\top \mathbf{U}\mathbf{U}^\top \beta_i (y_{i,a}\mathbf{x}_{i,a} - y_{j,a}\mathbf{x}_{j,a})^\top \mathbf{U}\mathbf{U}^\top \beta_j\right] \lesssim \rho^2\,\mathbb{E}\left[\left\|(y_{i,a}\mathbf{x}_{i,a} - y_{j,a}\mathbf{x}_{j,a})^\top \mathbf{U}\mathbf{U}^\top\right\|_2^2\right],$$

we have the following upper bound:

$$\begin{aligned}
&\lesssim \mathbb{E}\left[\left\|(y_{i,a}\mathbf{x}_{i,a} - y_{j,a}\mathbf{x}_{j,a})^\top \mathbf{U}\mathbf{U}^\top\right\|_2^2\right] \\
&\lesssim \mathbb{E}\left[\left\|(y_{i,a}\mathbf{x}_{i,a})^\top \mathbf{U}\right\|_2^2\right] + \mathbb{E}\left[\left\|(y_{j,a}\mathbf{x}_{j,a})^\top \mathbf{U}\right\|_2^2\right].
\end{aligned}$$

Since $\mathbb{E}\left[\left((y_{i,a}\mathbf{x}_{i,a})^\top \mathbf{u}_l\right)^2\right] \le \mathcal{O}\left(\rho^2\right)\ \forall\, l \in [k]$, we finally have a $\mathcal{O}\left(k\right)$ upper bound for this case.

The final step is to sum the contributions of these 4 cases. Case 2 and 3 have $\mathcal{O}\left(t^3\right)$ different quadruples (a, b, a', b'). Case 4 has $\mathcal{O}\left(t^2\right)$ different quadruples (a, b, a', b'). Combining the resulting bounds yields an upper bound of:

$$\mathcal{O}\left(\rho^4 \cdot (t + k)/t^2\right). \qquad \square$$

We now have all the required ingredients for the proof of Lemma 5.2

Proof of Lemma 5.2. For each pair i, j, we repeatedly compute

$$\left(\widehat{\beta}_i^{(1)} - \widehat{\beta}_j^{(1)}\right)^\top \mathbf{U}\mathbf{U}^\top\mathbf{U}\mathbf{U}^\top \left(\widehat{\beta}_i^{(2)} - \widehat{\beta}_j^{(2)}\right)$$

$\log(n/\delta)$ times, each with a batch of new sample of size $\rho^2\sqrt{k}/\Delta^2$, and take the median of these estimates. With probability $1 - \delta$, it holds that for all $\beta_i \neq \beta_j$, the median is greater than $c\Delta^2$, and for all $\beta_i = \beta_j$ the median is less than $c\Delta^2$ for some constant $c > 0$. Hence the single-linkage algorithm can correctly identify the k clusters.

Conditioning on the event of perfect clustering, the cluster sizes are distributed according to a multinomial distribution, which from Proposition D.7 can be shown to concentrate as

$$|p_i - \widetilde{p}_i| \le \sqrt{\frac{3\log(k/\delta)}{n} p_i} \le p_i/2$$

with probability at least $1 - \delta$ by our assumption that $n = \Omega\left(\frac{\log(k/\delta)}{p_{\min}}\right)$, which implies that $\widetilde{p}_i \ge p_i/2$.

For each group, we compute the corresponding average of $\mathbf{U}^\top\widehat{\beta}_i$ as

$$\mathbf{U}^\top\widetilde{\mathbf{w}}_l := \frac{1}{n\widetilde{p}_l t} \sum_{i \ni \beta_i = \mathbf{w}_l} \sum_{j=1}^{t} y_{i,j}\mathbf{U}^\top\mathbf{x}_{i,j},$$

which from Proposition A.1 would satisfy

$$\begin{aligned}\left\|\mathbf{U}^\top\left(\widetilde{\mathbf{w}}_l-\mathbf{w}_l\right)\right\|_2 &\lesssim \sqrt{k}\rho_i \max\left\{\frac{\log(k^2/\delta)}{n\widetilde{p}_l t},\sqrt{\frac{\log(k^2/\delta)}{n\widetilde{p}_l t}}\right\}\\ &\le \widetilde{\epsilon}\rho_i.\end{aligned}$$

The last inequality holds due to the condition on n.

The estimate for $r_l^2 \coloneqq s_l^2+\|\mathbf{w}_l-\widetilde{\mathbf{w}}_l\|_2^2 \;\forall\, l\in[k]$ is

$$\widetilde{r}_l^2=\frac{1}{n\widetilde{p}_l t}\sum_{i\ni\beta_i=\mathbf{w}_l}\sum_{j=1}^{t}\left(\mathbf{x}_{i,j}^\top\left(\mathbf{w}_l-\widetilde{\mathbf{w}}_l\right)+\epsilon_{i,j}\right)^2$$

where $\mathbf{x}_{i,j}$ and $y_{i,j}$ are fresh samples from the same tasks. The expectation of $\widetilde{r}_l^2$ can be computed as

$$\begin{aligned}\mathbb{E}\left[\widetilde{r}_l^2\right]&=\frac{1}{n\widetilde{p}_l t}\sum_{i\ni\beta_i=\mathbf{w}_i}\sum_{j=1}^{t}\mathbb{E}\left[\left(\mathbf{x}_{i,j}^\top\left(\mathbf{w}_l-\widetilde{\mathbf{w}}_l\right)+\epsilon_{i,j}\right)^2\right]\\ &=s_l^2+\|\mathbf{w}_l-\widetilde{\mathbf{w}}_l\|_2^2=r_l^2\end{aligned}$$

We can compute the variance of $\widetilde{r}_l^2$ like

$$\begin{aligned}\mathrm{Var}\left[\widetilde{r}_l^2\right]&=\frac{1}{n\widetilde{p}_l t}\sum_{i\ni\beta_i=\mathbf{w}_i}\sum_{j=1}^{t}\mathrm{Var}\left[\left(\mathbf{x}_{i,j}^\top\left(\mathbf{w}_l-\widetilde{\mathbf{w}}_l\right)+\epsilon_{i,j}\right)^2\right]\\ &=\frac{1}{n\widetilde{p}_l t}\sum_{i\ni\beta_i=\mathbf{w}_i}\sum_{j=1}^{t}\left[\mathbb{E}\left[\left(\mathbf{x}_{i,j}^\top\left(\mathbf{w}_l-\widetilde{\mathbf{w}}_l\right)+\epsilon_{i,j}\right)^4\right]-\left(s_l^2+\|\mathbf{w}_l-\widetilde{\mathbf{w}}_l\|_2^2\right)^2\right]\end{aligned}$$

Since $\left(\mathbf{x}_{i,j}^\top\left(\mathbf{w}_l-\widetilde{\mathbf{w}}_l\right)+\epsilon_{i,j}\right)^2$ is a sub-exponential random variable, we can use Bernstein's concentration inequality to get

$$\begin{aligned}&\mathbb{P}\left[\left|\widetilde{r}_l^2-r_l^2\right|>z\right]\le 2\exp\left\{-\min\left\{\frac{z^2 n\widetilde{p}_l t}{r_l^4},\frac{zn\widetilde{p}_l t}{r_l^2}\right\}\right\}\\ &\implies \left|\widetilde{r}_l^2-r_l^2\right|<r_l^2\max\left\{\sqrt{\frac{\log\frac{1}{\delta}}{n\widetilde{p}_l t}},\frac{\log\frac{1}{\delta}}{n\widetilde{p}_l t}\right\}\qquad\text{with probability at least } 1-\delta,\\ &\qquad\le r_l^2\frac{\widetilde{\epsilon}}{\sqrt{k}}\end{aligned}$$

where the last inequality directly follows from the condition on n. □

A.3. Proof of Lemma 5.3

Before proving Lemma 5.3, we first show that with the parameters $\mathbf{w}_i, r_i^2$ estimated with accuracy stated, for all $i\in[k]$ in the condition of Lemma 5.3, we can correctly classify a new task using only $\Omega\left(\log k\right)$ dependency of k on the number of examples t_{out}.

Lemma A.15 (Classification)**.** *Given estimated parameters satisfying $\|\widetilde{\mathbf{w}}_i-\mathbf{w}_i\|_2\le\Delta/10$, $(1-\Delta^2/50)\widetilde{r}_i^2\le s_i^2+\|\widetilde{\mathbf{w}}_i-\mathbf{w}_i\|_2^2\le(1+\Delta^2/50)\widetilde{r}_i^2$ for all $i\in[k]$, and a new task with $t_{\mathrm{out}}\ge\Theta\left(\log(k/\delta)/\Delta^4\right)$ samples whose true regression vector is $\beta=\mathbf{w}_h$, our algorithm predicts h correctly with probability $1-\delta$.*

Proof. Given a new task with t_{out} training examples, $\mathbf{x}_i,\ y_i=\mathbf{w}^\top\mathbf{x}_i+\epsilon_i$ for $i\in[t_{\mathrm{out}}]$ where the true regression vector is $\beta=\mathbf{w}_h$ and the true variance of the noise is $\sigma^2=s_h^2$. Our algorithm compute the the following "log likelihood" like quantity with the estimated parameters, which is defined to be

$$\widehat{l}_i \coloneqq -\sum_{j=1}^{t_{\text{out}}} \left(y_j - \mathbf{x}_j^\top \widetilde{\mathbf{w}}_i\right)^2 / \left(2\widetilde{r}_i^2\right) + t_{\text{out}} \cdot \log\left(1/\widetilde{r}_i\right) \tag{23}$$

$$= -\sum_{j=1}^{t_{\text{out}}} \left(\epsilon_j + \mathbf{x}_j^\top \left(\mathbf{w}_h - \widetilde{\mathbf{w}}_i\right)\right)^2 / \left(2\widetilde{r}_i^2\right) + t_{\text{out}} \cdot \log(1/\widetilde{r}_i),$$

and output the classification as $\arg\max_{i\in[k]} \widehat{l}_i$.

Our proof proceeds by proving a lower bound on the likelihood quantity of the true index $\widehat{l}_h$, and an upper bound on the likelihood quantity of the other indices $\widehat{l}_i$ for $i \in [k]\backslash\{h\}$, and we then argue that the $\widehat{l}_h$ is greater than the other $\widehat{l}_i$'s for $i \in [k]\backslash\{h\}$ with high probability, which implies our algorithm output the correct classification with high probability.

The expectation of $\widehat{l}_h$ is

$$\mathbb{E}\left[\widehat{l}_h\right] = -t_{\text{out}} \cdot \left(s_h^2 + \|\mathbf{w}_h - \widetilde{\mathbf{w}}_h\|_2^2\right) / \left(2\widetilde{r}_h^2\right) + t_{\text{out}} \cdot \log(1/\widetilde{r}_h).$$

Since $\left(\epsilon_j + \mathbf{x}_j^\top \left(\mathbf{w}_h - \widetilde{\mathbf{w}}_h\right)\right)^2 / \left(2\widetilde{r}_h^2\right)$ is a sub-exponential random variable with sub-exponential norm at most $\mathcal{O}\left(\left(s_h^2 + \|\mathbf{w}_h - \widetilde{\mathbf{w}}_h\|_2^2\right)/\widetilde{r}_h^2\right) = \mathcal{O}\left(r_h^2/\widetilde{r}_h^2\right)$, we can apply Bernstein inequality (Vershynin, 2018, Theorem 2.8.1) to $\widehat{l}_h$ and get

$$\mathbb{P}\left[\left|\widehat{l}_h - \mathbb{E}\left[\widehat{l}_h\right]\right| > z\right] \le 2\exp\left\{-c\min\left\{\frac{z^2}{t_{\text{out}} r_h^4/\widetilde{r}_h^4}, \frac{z}{r_h^2/\widetilde{r}_h^2}\right\}\right\},$$

which implies that with probability $1 - \delta/k$,

$$\left|\widehat{l}_h - \mathbb{E}\left[\widehat{l}_h\right]\right| \lesssim r_h^2/\widetilde{r}_h^2 \cdot \max\left\{\sqrt{t_{\text{out}}\log(k/\delta)}, \log(k/\delta)\right\}.$$

Using the fact that $t_{\text{out}} \ge C\log(k/\delta)$ for some $C > 1$, we have that with probability $1 - \delta/k$,

$$\widehat{l}_h \ge -\left(t_{\text{out}} + c\sqrt{t_{\text{out}}\log(k/\delta)}\right) \cdot r_h^2/\left(2\widetilde{r}_h^2\right) + t_{\text{out}} \cdot \log(1/\widetilde{r}_h)$$

for some constant $c > 0$.

For $i \ne h$, the expectation of $\widehat{l}_i$ is at most

$$\mathbb{E}\left[\widehat{l}_i\right] \le -t_{\text{out}} \cdot \left(s_i^2 + \left(\Delta - \|\mathbf{w}_i - \widetilde{\mathbf{w}}_i\|_2\right)^2\right) / \left(2\widetilde{r}_i^2\right) + t_{\text{out}} \cdot \log\left(1/\widetilde{r}_i\right).$$

Since $\left(\epsilon_i + \mathbf{x}_j^\top \left(\mathbf{w}_h - \widetilde{\mathbf{w}}_i\right)\right)^2 / \left(2\widetilde{r}_i^2\right)$ is a sub-exponential random variable with sub-exponential norm at most $\mathcal{O}\left(\left(s_i^2 + \left(\Delta + \|\mathbf{w}_i - \widetilde{\mathbf{w}}_i\|_2\right)^2\right)/\widetilde{r}_i^2\right)$. Again we can apply Bernstein's inequality and get with probability $1 - \delta$

$$\begin{aligned}\widehat{l}_i \le\ & -t_{\text{out}} \cdot \left(s_i^2 + \left(\Delta - \|\mathbf{w}_i - \widetilde{\mathbf{w}}_i\|_2\right)^2\right) / \left(2\widetilde{r}_i^2\right) + t_{\text{out}}\log\left(1/\widetilde{r}_i\right) \\ & + c\sqrt{t_{\text{out}}\log(k/\delta)} \cdot \left(s_i^2 + \left(\Delta + \|\mathbf{w}_i - \widetilde{\mathbf{w}}_i\|_2\right)^2\right) / \left(2\widetilde{r}_i^2\right)\end{aligned}$$

for a constant $c > 0$.

Using our assumption that $\|\mathbf{w}_i - \widetilde{\mathbf{w}}_i\|_2 \le \Delta/10$ for all $i \in [k]$, we get

$$\widehat{l}_i \le \left(-t_{\text{out}} + c'\sqrt{t_{\text{out}}\log(k/\delta)}\right) \cdot \left(s_i^2 + 0.5\Delta^2\right) / \left(2\widetilde{r}_i^2\right) + 0.5t_{\text{out}}\log\left(1/\widetilde{r}_i^2\right)$$

for some constant $c' > 0$. We obtain a worst case bound by taking the maximum over all possible value of $\widetilde{r}_i$ as

$$\widehat{l}_i \le\ -0.5t_{\text{out}} - 0.5t_{\text{out}}\log\left(\left(1 - c'\sqrt{\log(k/\delta)/t_{\text{out}}}\right)\left(s_i^2 + 0.5\Delta^2\right)\right),$$

where we have taken the maximum over all possible values of $\widehat{r}_i$.

Using the assumption that

$$r_h^2/\widetilde{r}_h^2 \leq 1+\Delta^2/50$$

and $t_{\text{out}} \geq C\log(k/\delta)$ for some constant $C>1$, we obtain that

$$\begin{aligned}&-t_{\text{out}}\cdot r_h^2/(2\widetilde{r}_h^2)+0.5t_{\text{out}} \geq t_{\text{out}}\Delta^2/100, \quad \text{and}\\ &-c\sqrt{t_{\text{out}}\log(k/\delta)}\cdot r_h^2/\left(2\widetilde{r}_h^2\right)+0.5t_{\text{out}}\log\left(1-c'\sqrt{\log(k/\delta)/t_{\text{out}}}\right)=\mathcal{O}\left(\sqrt{t_{\text{out}}\log(k/\delta)}\right).\end{aligned}$$

Further notice that

$$\left(1+\Delta^2/5\right)\widetilde{r}_h^2 \leq \frac{\left(1+\Delta^2/5\right)}{1-\Delta^2/50}\left(s_h^2+\Delta^2/100\right) \leq s_h^2+\Delta^2/2.$$

since $s_h^2 \leq 1$, and $\Delta \leq 2$. Plugging in these facts into $\widehat{l}_h-\widehat{l}_i$ and applying the assumption that $\left(s_h^2+\Delta^2/2\right)/\widetilde{r}_h^2 \geq \left(1+\Delta^2/5\right)$ we get

$$\widehat{l}_h-\widehat{l}_i \geq 0.5t_{\text{out}}\log\left(1+\Delta^2/5\right)-t_{\text{out}}\Delta^2/100-\mathcal{O}\left(\sqrt{t_{\text{out}}\log(k/\delta)}\right)$$

By the fact that $\log\left(1+\Delta^2/5\right)-\Delta^2/50 \geq \Delta^2/5000$ for all $\Delta \leq 50$, the above quantity is at least

$$\Theta\left(t_{\text{out}}\Delta^2\right)-\Theta\left(\sqrt{t_{\text{out}}\log(k/\delta)}\right). \tag{24}$$

Since $t_{\text{out}} \geq \Theta\left(\log(k/\delta)/\Delta^4\right)$, we have that with probability δ, for all $i \in [k]\backslash\{h\}$, it holds that $\widehat{l}_h-\widehat{l}_i>0$, which implies the correctness of the classification procedure. □

Proof of Lemma 5.3. Given n i.i.d. samples from our data generation model, by the assumption that $n=\Omega\left(\frac{d\log^2(k/\delta)}{p_{\min}\epsilon^2 t}\right)=\Omega\left(\frac{\log(k/\delta)}{p_{\min}}\right)$ and from Proposition D.7, it holds that the number of tasks such that $\beta=\mathbf{w}_i$ is $n\widehat{p}_i \geq \frac{1}{2}np_i$ with probability at least $1-\delta$. Hence, with this probability, there exists at least $np_i/10$ i.i.d. examples for estimating $\mathbf{w}_i$ and s_i^2. By Proposition D.10, it holds that with probability $1-\delta$, for all $i\in[k]$, our estimation satisfies

$$\begin{aligned}\left\|\widehat{\mathbf{w}}_i-\mathbf{w}_i\right\|_2^2&=\mathcal{O}\left(\frac{\sigma^2\left(d+\log(k/\delta)\right)}{np_i t}\right), \quad \text{and}\\ \left|\widehat{s}_i^2-s_i^2\right|&=\mathcal{O}\left(\frac{\log(k/\delta)}{\sqrt{np_i t-d}}s_i^2\right).\end{aligned}$$

By Proposition D.7, it holds that

$$\left|\widehat{p}_i-p_i\right| \leq \sqrt{\frac{3\log(k/\delta)}{n}p_i}$$

Since $n=\Omega\left(\frac{d\log^2(k/\delta)}{p_{\min}\epsilon^2 t}\right)$, we finally get for all $i\in[k]$

$$\begin{aligned}\left\|\widehat{\mathbf{w}}_i-\mathbf{w}_i\right\|_2 &\leq \epsilon s_i\ ,\\ \left|\widehat{s}_i^2-s_i^2\right| &\leq \frac{\epsilon s_i^2}{\sqrt{d}}\ , \quad \text{and}\\ \left|\widehat{p}_i-p_i\right| &\leq \min\left\{p_{\min}/10, \epsilon p_i\sqrt{t/d}\right\}.\end{aligned}$$

□

B. Proof Theorem 2

We first bound the expected error of the maximum a posterior (*MAP*) estimator.

Lemma B.1. *Given estimated parameters satisfying* $\|\widehat{\mathbf{w}}_i - \mathbf{w}_i\|_2 \le \Delta/10$, $\left(1 - \Delta^2/50\right)\widehat{s}_i^2 \le s_i^2 + \|\widehat{\mathbf{w}}_i - \mathbf{w}_i\|_2^2 \le \left(1 + \Delta^2/50\right)\widehat{s}_i^2$ *for all* $i \in [k]$, *and a new task with* $\tau \ge \Theta\left(\log(k/\delta)/\Delta^4\right)$ *samples* $\mathcal{D} = \{\mathbf{x}_i, y_i\}_{i=1}^{\tau}$. *Define the maximum a posterior (MAP) estimator as*

$$\widehat{\beta}_{\rm MAP}(\mathcal{D}) := \widehat{\mathbf{w}}_{\widehat{i}}$$

where

$$\widehat{i} := \arg\max_{i\in[k]} \left(\sum_{j=1}^{\tau} \frac{-\left(y_j - \widehat{\mathbf{w}}_i^\top \mathbf{x}_j\right)^2}{2\widehat{\sigma}_i^2} + \tau \log\left(1/\widehat{\sigma}_i\right) + \log\left(\widehat{p}_i\right) \right).$$

Then, the expected error of the MAP estimator is bound as

$$\begin{aligned}
&\mathop{\mathbb{E}}_{\mathcal{T}^{\rm new}\sim\mathbb{P}(\mathcal{T})} \mathop{\mathbb{E}}_{\mathcal{D}\sim\mathcal{T}^{\rm new}} \mathop{\mathbb{E}}_{\{\mathbf{x},y\}\sim\mathcal{T}^{\rm new}} \left[\left(\mathbf{x}^\top \widehat{\beta}_{\rm MAP}(\mathcal{D}) - y\right)^2\right] \\
\le&\delta + \sum_{i=1}^{k} p_i \left\|\mathbf{w}_i - \widehat{\mathbf{w}}_i\right\|_2^2 + \sum_{i=1}^{k} p_i s_i^2
\end{aligned}$$

Proof. The proof is very similar to the proof of Lemma A.15. The log of the posterior probability given the training data $\mathcal{D}$ under the estimated parameters is

$$\widehat{l}_i := -\sum_{j=1}^{\tau}\left(y_j - \mathbf{x}_j^\top \widehat{\mathbf{w}}_i\right)^2 / \left(2\widehat{s}_i^2\right) + \tau\cdot\log\left(1/\widehat{s}_i\right) + \log\left(\widehat{p}_i\right), \tag{25}$$

which is different from Equation 23 just by a $\log(1/\widehat{p}_i)$ additive factor. Hence, given that the true regression vector of the new task $\mathcal{T}^{\rm new}$ is $\mathbf{w}_h$, it follows from Equation 24 that $\widehat{l}_h - \widehat{l}_i$ with probability at least $1-\delta$ is greater than

$$\Theta(\tau\Delta^2) - \Theta\left(\sqrt{\tau\log(k/\delta)}\right) + \log\left(\widehat{p}_h/\widehat{p}_i\right),$$

which under the assumption that $|\widehat{p}_i - p_i| \le p_i/10$ is greater than

$$\Theta(\tau\Delta^2) - \Theta\left(\sqrt{\tau\log(k/\delta)}\right) - \log(1/p_h) - \log(10/9). \tag{26}$$

If $p_h \ge \delta/k$, by our assumption that $\tau \ge \Theta\left(\log(k/\delta)/\Delta^4\right)$, it holds that $\widehat{l}_h - \widehat{l}_i > 0$ for all $i \ne h$, and hence the MAP estimator output $\widehat{\mathbf{w}}_h$ with probability at least $1-\delta$. With the remaining less than δ probability, the MAP estimator output $\widehat{\beta}_{\rm MAP} = \widehat{\mathbf{w}}_i$ for some other $i \ne h$ which incurs ℓ_2 error $\|\widehat{\beta}_{\rm MAP} - \mathbf{w}_h\|_2 \le \|\widehat{\beta}_{\rm MAP}\|_2 + \|\mathbf{w}_h\|_2 \le 2$.

If $p_h \le \delta/k$, we pessimistically bound the error of $\widehat{\beta}_{\rm MAP}$ by $\|\widehat{\beta}_{\rm MAP} - \mathbf{w}_h\| \le 2$.

To summarize, notice that

$$\begin{aligned}
&\mathop{\mathbb{E}}_{\mathcal{T}^{\rm new}\sim\mathbb{P}(\mathcal{T})} \mathop{\mathbb{E}}_{\mathcal{D}\sim\mathcal{T}^{\rm new}} \mathop{\mathbb{E}}_{\{\mathbf{x},y\}\sim\mathcal{T}^{\rm new}} \left[\left(\mathbf{x}^\top \widehat{\beta}_{\rm MAP}(\mathcal{D}) - y\right)^2\right] \\
=&\mathop{\mathbb{E}}_{\mathcal{T}^{\rm new}\sim\mathbb{P}(\mathcal{T})} \mathop{\mathbb{E}}_{\mathcal{D}\sim\mathcal{T}^{\rm new}} \left[\left\|\widehat{\beta}_{\rm MAP}(\mathcal{D}) - \mathbf{w}_h\right\|_2^2 + s_h^2\right] \\
\le&\sum_{i=1}^{k} p_i \left(\mathbb{1}\left\{p_i \ge \delta/k\right\}\left(4\delta + (1-\delta)\left\|\mathbf{w}_i - \widehat{\mathbf{w}}_i\right\|_2^2\right)\right) + \sum_{i=1}^{k} 4p_i\mathbb{1}\left\{p_i \le \delta/k\right\} + \sum_{i=1}^{k} p_i s_i^2 \\
\le&4\delta + \sum_{i=1}^{k} p_i\|\mathbf{w}_i - \widehat{\mathbf{w}}_i\|^2 + 4\delta + \sum_{i=1}^{k} p_i s_i^2 \\
=&8\delta + \sum_{i=1}^{k} p_i\|\mathbf{w}_i - \widehat{\mathbf{w}}_i\|^2 + \sum_{i=1}^{k} p_i s_i^2.
\end{aligned}$$

Replacing 8δ by δ concludes the proof. □

Next, we bound the expected error of the posterior mean estimator.

Lemma B.2. *Given estimated parameters satisfying $\left\|\widehat{\mathbf{w}}_i - \mathbf{w}_i\right\|_2 \le \Delta/10$, $s_i^2 + \left\|\widehat{\mathbf{w}}_i - \mathbf{w}_i\right\|_2^2 \le (1+\Delta^2/50)\widehat{s}_i^2$, $s_i^2 + \Delta^2/2 \ge (1+\Delta^2/5)\widehat{s}_i^2$ for all $i \in [k]$, and a new task with $\tau \ge \Theta\left(\log(k/\delta)/\Delta^4\right)$ samples $\mathcal{D} = \{\mathbf{x}_i, y_i\}_{i=1}^{\tau}$. Define the posterior mean estimator as*

$$\widehat{\beta}_{\text{Bayes}}(\mathcal{D}) := \frac{\sum_{i=1}^k \widehat{L}_i \widehat{\mathbf{w}}_i}{\sum_{i=1}^k \widehat{L}_i}$$

where

$$\widehat{L}_i := \exp\left(-\sum_{i=1}^{\tau} \frac{\left(y_j - \mathbf{w}_i^\top \mathbf{x}_j\right)^2}{2\widehat{\sigma}_i^2} + \tau \log(1/\widehat{\sigma}_i) + \log(\widehat{p}_i) \right).$$

Then, the expected error of the posterior mean estimator is bound as

$$\begin{aligned}
&\mathbb{E}_{\mathcal{T}^{\text{new}} \sim \mathbb{P}(\mathcal{T})} \mathbb{E}_{\mathcal{D} \sim \mathcal{T}^{\text{new}}} \mathbb{E}_{\{\mathbf{x},y\} \sim \mathcal{T}^{\text{new}}} \left[\left(\mathbf{x}^\top \widehat{\beta}_{\text{Bayes}}(\mathcal{D}) - y \right)^2 \right] \\
\le &\delta + \sum_{i=1}^k p_i \left\| \mathbf{w}_i - \widehat{\mathbf{w}}_i \right\|_2^2 + \sum_{i=1}^k p_i s_i^2
\end{aligned}$$

Proof. This proof is very similar to the proof of Lemma B.1. Notice that

$$\begin{aligned}
&\mathop{\mathbb{E}}_{\mathcal{T}^{\text{new}} \sim \mathbb{P}(\mathcal{T})} \mathop{\mathbb{E}}_{\mathcal{D} \sim \mathcal{T}^{\text{new}}} \mathop{\mathbb{E}}_{\{\mathbf{x},y\} \sim \mathcal{T}^{\text{new}}} \left[\left(\mathbf{x}^\top \widehat{\beta}_{\text{Bayes}}(\mathcal{D}) - y \right)^2 \right] \\
= &\mathop{\mathbb{E}}_{\mathcal{T}^{\text{new}} \sim \mathbb{P}(\mathcal{T})} \mathop{\mathbb{E}}_{\mathcal{D} \sim \mathcal{T}^{\text{new}}} \left[\left\| \widehat{\beta}_{\text{Bayes}}(\mathcal{D}) - \mathbf{w}_h \right\|_2^2 + s_h^2 \right]
\end{aligned}$$

where $\mathbf{w}_h$ is defined to be the true regression vector of the task $\mathcal{T}^{\text{new}}$.

$$\begin{aligned}
&\left\| \widehat{\beta}_{\text{Bayes}}(\mathcal{D}) - \mathbf{w}_h \right\|_2^2 \\
\le &\left(\left\| \widehat{\mathbf{w}}_h - \mathbf{w}_h \right\|_2 + \left(1 - \frac{\widehat{L}_h}{\sum_{i=1}^k \widehat{L}_i} \right) \left\| \mathbf{w}_h \right\|_2 + \sum_{j \ne h} \frac{\widehat{L}_j}{\sum_{i=1}^k \widehat{L}_i} \left\| \mathbf{w}_j \right\|_2 \right)^2 \\
\le &\left(\left\| \widehat{\mathbf{w}}_h - \mathbf{w}_h \right\|_2 + 2\left(1 - \frac{\widehat{L}_h}{\sum_{i=1}^k \widehat{L}_i} \right) \right)^2 \\
\le &\left(\left\| \widehat{\mathbf{w}}_h - \mathbf{w}_h \right\|_2 + 2\sum_{i \ne h} \widehat{L}_i / \widehat{L}_h \right)^2 \qquad (27)
\end{aligned}$$

Notice that

$$\widehat{L}_i / \widehat{L}_h = \exp(\widehat{l}_i - \widehat{l}_h)$$

where l_i is the logarithm of the posterior distribution as defined in Equation 25. Therefore we can apply Equation 26 and have that with probability δ,

$$\widehat{l}_i - \widehat{l}_h \le -\log(k/\delta)/\Delta^2 \le -\log(k/\delta)$$

for $\tau = \Omega(\log(k/\delta)/\Delta^4)$, which is equivalent to

$$\widehat{L}_i / \widehat{L}_h \le \delta/k.$$

Plugging this into Equation 27 yields for a fixed $\mathcal{T}^{\text{new}}$, with probability $1-\delta$,

$$\begin{aligned}
\left\| \widehat{\beta}_{\text{Bayes}}(\mathcal{D}) - \mathbf{w}_h \right\|_2^2 &\le \left(\left\| \widehat{\mathbf{w}}_h - \mathbf{w}_h \right\|_2 + 2\sum_{i \ne h} \widehat{L}_i / \widehat{L}_h \right)^2 \\
&\le \left\| \widehat{\mathbf{w}}_h - \mathbf{w}_h \right\|_2^2 + 4\delta^2 + 4\delta \left\| \widehat{\mathbf{w}}_h - \mathbf{w}_h \right\|_2 \\
&\le \left\| \widehat{\mathbf{w}}_h - \mathbf{w}_h \right\|_2^2 + 8\delta,
\end{aligned}$$

and the error is at most 4 for the remaining probability δ. Hence we get for a fixed $\mathcal{T}^{\text{new}}$

$$\mathbb{E}_{\mathcal{D}\sim\mathcal{T}^{\text{new}}}\left[\left\|\widehat{\beta}_{\text{Bayes}}(\mathcal{D})-\mathbf{w}_h\right\|_2^2+s_h^2\right]\le\|\widehat{\mathbf{w}}_h-\mathbf{w}_h\|_2^2+s_h^2+12\delta.$$

Finally taking the randomess of $\mathcal{T}^{\text{new}}$ into account, we have

$$\begin{aligned}&\mathbb{E}_{\mathcal{T}^{\text{new}}\sim\mathbb{P}(\mathcal{T})}\mathbb{E}_{\mathcal{D}\sim\mathcal{T}^{\text{new}}}\mathbb{E}_{\{\mathbf{x},y\}\sim\mathcal{T}^{\text{new}}}\left[\left(\mathbf{x}^\top\widehat{\beta}_{\text{Bayes}}(\mathcal{D})-y\right)^2\right]\\ \le&12\delta+\sum_{i=1}^k p_i\left\|\mathbf{w}_i-\widehat{\mathbf{w}}_i\right\|_2^2+\sum_{i=1}^k p_i s_i^2\end{aligned}$$

Replacing 12δ by δ concludes the proof. □

C. Proof of Remark 4.6

We construct a worst case example and analyze the expected error of the Bayes optimal predictor. We choose $s_i=\sigma$, $p_i=1/k$, and $\mathbf{w}_i=\left(\Delta/\sqrt{2}\right)\mathbf{e}_i$ for all $i\in[k]$. Given a new task with τ training examples, we assume Gaussian input $\mathbf{x}_j\sim\mathcal{N}(\mathbf{0},\mathbf{I}_d)\in\mathbb{R}^d$, and Gaussian noise $y_j=\beta^\top\mathbf{x}_j+\epsilon_j\in\mathbb{R}$ with $\epsilon_j\sim\mathcal{N}(0,\sigma^2)$ i.i.d. for all $j\in[\tau]$. Denote the true model parameter by $\beta=\mathbf{w}_h$ for some $h\in[k]$, and the Bayes optimal estimator is

$$\widehat{\beta}=\left[\sum_{i=1}^k L_i\right]^{-1}\sum_{i=1}^k L_i\mathbf{w}_i,$$

where $L_i:=\exp\left(-\frac{1}{2\sigma^2}\sum_{j=1}^\tau(y_j-\mathbf{w}_i^\top\mathbf{x}_j)^2\right)$. The squared ℓ_2 error is lower bounded by

$$\begin{aligned}\left\|\widehat{\beta}-\mathbf{w}_h\right\|_2^2\ge&\left\|\left[\sum_{i=1}^k L_i\right]^{-1}\sum_{i\in[k]\setminus\{h\}}L_i\mathbf{w}_h\right\|_2^2\\ =&\frac{\Delta^2\left(\sum_{i\in[k]\setminus\{h\}}L_i/L_h\right)^2}{2\left(1+\sum_{i\in[k]\setminus\{h\}}L_i/L_h\right)^2}\end{aligned}\tag{28}$$

Let us define $l_i=\log L_i$, which is

$$\begin{aligned}l_i=&-\frac{1}{2\sigma^2}\sum_{j=1}^\tau\left(y_j-\mathbf{x}_j^\top\mathbf{w}_i\right)^2\\ =&-\frac{1}{2\sigma^2}\sum_{j=1}^\tau\left(\epsilon_j+\mathbf{x}_j^\top(\mathbf{w}_h-\mathbf{w}_i)\right)^2\end{aligned}$$

Notice that for all $i\in[k]\setminus\{h\}$, $\mathbb{E}\left[l_i\right]=-\frac{\tau}{2}\left(1+{}^{\Delta^2}\!/{}_{\sigma^2}\right)$. Using Markov's inequality and the fact that $l_i\le 0$, we have that for each fixed $i\in[k]\setminus\{h\}$,

$$\mathbb{P}\left[\,l_i\ge 3\,\mathbb{E}\left[l_i\right]\,\right]\;\ge\;2/3\;.$$

For each $i\in[k]\setminus\{h\}$, define an indicator random variable $I_i=\mathbb{1}\left\{l_i\ge 3\,\mathbb{E}\left[l_i\right]\right\}$. The expectation is lower bounded by

$$\mathbb{E}\left[\sum_{i\in[k]\setminus\{h\}}I_i\right]\ge\frac{2}{3}(k-1)\;.$$

The expectation is upper bounded by

$$\mathbb{E}\left[\sum_{i\in[k]\backslash\{h\}} I_i\right] \leq \mathbb{P}\left[\sum_{i\in[k]\backslash\{h\}} I_i \geq \frac{k-1}{3}\right]\cdot(k-1)$$
$$+\left(1-\mathbb{P}\left[\sum_{i\in[k]\backslash\{h\}} I_i \geq \frac{k-1}{3}\right]\right)\cdot\frac{k-1}{3}.$$

Combining the above two bounds together, we have

$$\mathbb{P}\left[\sum_{i\in[k]\backslash\{h\}} I_i \geq \frac{k-1}{3}\right] \geq 1/2.$$

Hence with probability at least $1/2$,

$$\sum_{i\in[k]\backslash\{h\}} e^{l_i-l_h} \geq \sum_{i\in[k]\backslash\{h\}} e^{l_i} \geq \sum_{i\in[k]\backslash\{h\}} I_i e^{3\,\mathbb{E}[l_i]}$$
$$\geq \frac{k-1}{3} e^{-\frac{3\tau}{2}\left(1+\Delta^2/\sigma^2\right)},$$

which implies that Eq. (28) is greater than $\Delta^2/8$. Hence the expected ℓ_2 error of the Bayes optimal estimator is $\mathbb{E}_{x,\epsilon}\left[(\widehat{y}-y)^2\right] = \mathbb{E}\left[\left(\left(\beta-\widehat{\beta}\right)^\top \mathbf{x}+\epsilon\right)^2\right] = \left\|\beta-\widehat{\beta}\right\|_2^2+\sigma^2 = \Delta^2/8+\sigma^2$.

D. Technical definitions and facts

Definition D.1 (Sub-Gaussian random variable). *A random variable X is said to follow a sub-Gaussian distribution if there exists a constant $K>0$ such that*

$$\mathbb{P}\left[|X|>t\right] \leq 2\exp\left(-t^2/K^2\right) \qquad \forall\, t\geq 0.$$

Definition D.2 (Sub-exponential random variable). *A random variable X is said to follow a sub-exponential distribution if there exists a constant $K>0$ such that*

$$\mathbb{P}\left[|X|>t\right] \leq 2\exp\left(-t/K\right) \qquad \forall\, t\geq 0.$$

Definition D.3 (Sub-exponential norm). *The sub-exponential norm of a random variable X is defined as*

$$\|X\|_{\psi_1} \coloneqq \sup_{p\in\mathbb{N}} p^{-1}\left(\mathbb{E}\left[|X|^p\right]\right)^{1/p}.$$

A random variable is sub-exponential if its sub-exponential norm is finite.

Fact D.4 (Gaussian and sub-Gaussian 4-th moment condition). *Let $\mathbf{v}$ and $\mathbf{u}$ denote two fixed vectors, we have*

$$\mathop{\mathbb{E}}_{\mathbf{x}\sim\mathcal{N}(\mathbf{0},\mathbf{I})}\left[\left(\mathbf{v}^\top\mathbf{x}\right)^2\left(\mathbf{u}^\top\mathbf{x}\right)^2\right] = \|\mathbf{u}\|_2^2\cdot\|\mathbf{v}\|_2^2+2\langle\mathbf{u},\mathbf{v}\rangle^2.$$

If $\mathbf{x}$ is a centered sub-Gaussian random variable with identity second moment, then

$$\mathbb{E}\left[\left(\mathbf{v}^\top\mathbf{x}\right)^2\left(\mathbf{u}^\top\mathbf{x}\right)^2\right] = \mathcal{O}\left(\|\mathbf{u}\|_2^2\cdot\|\mathbf{v}\|_2^2\right).$$

Proposition D.5 (Matrix Bernstein inequality, Theorem 1.6.2 in (Tropp et al., 2015)). *Let $\mathbf{S}_1,\ldots,\mathbf{S}_n$ be independent, centered random matrices with common dimension $d_1\times d_2$, and assume that each one is uniformly bounded $\mathbb{E}\left[\mathbf{S}_k\right]=0$ and $\|\mathbf{S}_k\|_2\leq L\;\forall\, k=1,\ldots,n$.*

Introduce the sum

$$\mathbf{Z} \coloneqq \sum_{k=1}^{n} \mathbf{S}_k$$

and let $v(\mathbf{Z})$ denote the matrix variance statistic of the sum:

$$v(\mathbf{Z}) \coloneqq \max \left\{ \left\| \mathbb{E}\left[\mathbf{Z}\mathbf{Z}^\top\right] \right\|_2, \left\| \mathbb{E}\left[\mathbf{Z}^\top \mathbf{Z}\right] \right\|_2 \right\}$$

Then

$$\mathbb{P}\left[\|\mathbf{Z}\|_2 \geq t\right] \leq (d_1 + d_2) \exp\left\{ \frac{-t^2/2}{v(\mathbf{Z}) + Lt/3} \right\}$$

for all $t \geq 0$.

Fact D.6 (Hoeffding's inequality (Hoeffding, 1963))**.** *Let $X_1, \ldots, X_n$ be independent random variables with bounded interval $0 \leq X_i \leq 1$. Let $\overline{X} = \frac{1}{n}\sum_{i=1}^n X_i$. Then*

$$\mathbb{P}\left[\left|\overline{X} - \mathbb{E}\left[\overline{X}\right]\right| \geq z\right] \leq 2\exp\left\{-2nz^2\right\}.$$

Proposition D.7 (ℓ_∞ deviation bound of multinomial distributions)**.** *Let $\mathbf{p} = \{p_1, \ldots, p_k\}$ be a vector of probabilities (i.e. $p_i \geq 0$ for all $i \in [k]$ and $\sum_{i=1}^k p_i = 1$). Let $\mathbf{x} \sim \mathrm{multinomial}(n, \mathbf{p})$ follow a multinomial distribution with n trials and probability $\mathbf{p}$. Then with probability $1-\delta$, for all $i \in [k]$,*

$$\left|\frac{1}{n}x_i - p_i\right| \leq \sqrt{\frac{3\log(k/\delta)}{n} p_i},$$

which implies

$$\left\|\frac{1}{n}\mathbf{x} - \mathbf{p}\right\|_\infty \leq \sqrt{\frac{3\log(k/\delta)}{n}}.$$

for all $i \in [k]$.

Proof. For each element x_i, applying Chernoff Bound D.8 with $z = \sqrt{\frac{3\log(k/\delta)}{n\,\mathbb{E}\left[\overline{X}\right]}}$ and taking a union bound over all i, we get

$$\left|\frac{1}{n}x_i - p_i\right| \leq \sqrt{\frac{3\log(k/\delta)p_i}{n}}.$$

for all $i \in [k]$. □

Fact D.8 (Chernoff Bound)**.** *Let $X_1, \ldots, X_n$ be independent Bernoulli random variables. Let $\overline{X} = \frac{1}{n}\sum_{i=1}^n X_i$. Then for all $0 < \delta \leq 1$*

$$\mathbb{P}\left[\left|\overline{X} - \mathbb{E}\left[\overline{X}\right]\right| \geq z\,\mathbb{E}\left[\overline{X}\right]\right] \leq \exp\left\{-z^2 n\,\mathbb{E}\left[\overline{X}\right]/3\right\}.$$

Fact D.9 (ϵ-tail bound for distributions with bounded second moment)**.** *Suppose random variable z with probability density function $p(\cdot)$, satisfies $\mathbb{E}\left[z^2\right] \leq \sigma^2$, then for any event $\mathcal{E}$ with $\mathbb{P}\left[\mathcal{E}\right] \geq 1-\epsilon$, it holds that*

$$\left|\mathbb{E}\left[z\right] - \mathbb{P}\left[\mathcal{E}\right]\mathbb{E}\left[z|\mathcal{E}\right]\right| \leq \sqrt{\epsilon}\sigma.$$

Proof. Notice that

$$\begin{aligned}
&\left|\mathbb{E}\left[z\right] - \mathbb{P}\left[\mathcal{E}\right]\mathbb{E}\left[z|\mathcal{E}\right]\right| \\
=&\left|\mathbb{P}\left[\bar{\mathcal{E}}\right]\mathbb{E}\left[z|\bar{\mathcal{E}}\right]\right| \\
=&\left|\int_{-\infty}^{\infty} \mathbb{1}\left\{z \in \bar{\mathcal{E}}\right\} zp(z)dz\right| \\
\leq&\sqrt{\int_{-\infty}^{\infty} \mathbb{1}\left\{z \in \bar{\mathcal{E}}\right\} p(z)dz \cdot \int_{-\infty}^{\infty} z^2 p(z)dz} \quad \text{(Using Cauchy–Schwarz)} \\
\leq&\sqrt{\epsilon}\sigma.
\end{aligned}$$

□

Proposition D.10 (High probability bound on the error of random design linear regression)**.** *Consider the following linear regression problem where we are given n i.i.d. samples*

$$\mathbf{x}_i \sim D\ ,\ y_i = \beta^\top \mathbf{x}_i + \epsilon_i\ ,\ i \in [n]$$

where D is a d-dimensional ($d < n$) sub-Gaussian distribution with constant sub-Gaussian norm, $\mathbb{E}[\mathbf{x}_i] = 0$, $\mathbb{E}\left[\mathbf{x}_i \mathbf{x}_i^\top\right] = \mathbf{I}_d$, and ϵ_i is a sub-Gaussian random variable and satisfies $\mathbb{E}[\epsilon_i] = 0$, $\mathbb{E}\left[\epsilon_i^2\right] = \sigma^2$.

1. *Then, with probability $1-\delta$, the ordinary least square estimator $\widehat{\beta} \coloneqq \arg\min_{\mathbf{w}} \sum_{i=1}^n \left(y_i - \mathbf{w}^\top \mathbf{x}_i\right)^2$ satisfies*

$$\left\|\widehat{\beta} - \beta\right\|_2^2 \le \mathcal{O}\left(\frac{\sigma^2(d + \log(1/\delta))}{n}\right).$$

2. *Define the estimator of the noise $\widehat{\sigma}^2$ as*

$$\widehat{\sigma}^2 \coloneqq \frac{1}{n-d}\sum_{i=1}^n \left(y_i - \widehat{\beta}^\top \mathbf{x}_i\right)^2.$$

Then with probability $1-\delta$, it holds that

$$|\widehat{\sigma}^2 - \sigma^2| \le \frac{\log(1/\delta)}{\sqrt{n-d}}\sigma^2.$$

Proof. (Hsu et al., 2012, Remark 12) shows that in the setting stated in the proposition, with probability $1 - \exp(-t)$, it holds that the least square estimator

$$\left\|\widehat{\beta} - \beta\right\|_2^2 \le \mathcal{O}\left(\frac{\sigma^2\left(d + 2\sqrt{dt} + 2t\right)}{n}\right) + o\left(\frac{1}{n}\right).$$

This implies that with probability $1-\delta$, it holds that

$$\left\|\widehat{\beta} - \beta\right\|_2^2 = \mathcal{O}\left(\frac{\sigma^2(d + \log(1/\delta))}{n}\right).$$

To prove the second part of the proposition, we first show that $\widehat{\sigma}^2$ is an unbiased estimator for σ^2 and then apply Hanson-Wright inequality to show the concentration. Define vector $\mathbf{y} \coloneqq (y_1, \ldots, y_n)$, $\boldsymbol{\epsilon} \coloneqq (\epsilon_1, \ldots, \epsilon_n)$ and matrix $\mathbf{X} \coloneqq \left[\mathbf{x}_1, \ldots, \mathbf{x}_n\right]^\top$. Notice that

$$\begin{aligned}
\mathbb{E}\left[\widehat{\sigma}^2\right] &= \frac{1}{n-d}\mathbb{E}\left[\sum_{i=1}^n \left(y_i - \widehat{\beta}^\top \mathbf{x}_i\right)^2\right] \\
&= \frac{1}{n-d}\mathbb{E}\left[\boldsymbol{\epsilon}^\top \left(\mathbf{I}_n - \mathbf{X}\left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\mathbf{X}^\top\right)\boldsymbol{\epsilon}\right] \\
&= \frac{1}{n-d}\mathbb{E}\left[\operatorname{tr}\left[\mathbf{I}_n - \mathbf{X}\left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\mathbf{X}^\top\right]\right] = \sigma^2,
\end{aligned}$$

where the last equality holds since $\mathbf{X}\left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\mathbf{X}^\top$ has exactly d eigenvalues equal to 1 almost surely. For a fixed $\mathbf{X}$ with rank d, by Hanson-Wright inequality (Vershynin, 2018, Theorem 6.2.1), it holds that

$$\mathbb{P}\left[\left|\widehat{\sigma}^2 - \sigma^2\right| \ge z\right] \le 2\exp\left\{-c\min\left\{(n-d)z^2/\sigma^4, (n-d)z/\sigma^2\right\}\right\},$$

which implies that with probability $1-\delta$

$$\left|\widehat{\sigma}^2 - \sigma^2\right| = \mathcal{O}\left(\frac{\log(1/\delta)}{\sqrt{n-d}}\sigma^2\right). \qquad \square$$

E. Simulations

We set $d = 8k$, $\mathbf{p} = \mathbf{1}_k/k$, $\mathbf{s} = \mathbf{1}_k$, and $\mathcal{P}_\mathbf{x}$ and $\mathcal{P}_\epsilon$ are standard Gaussian distributions.

E.1. Subspace estimation

We compute the subspace estimation error $\rho^{-1} \max_{i\in[k]} \left\| \left(\mathbf{U}\mathbf{U}^\top - \mathbf{I}\right) \mathbf{w}_i \right\|_2$ for various (t_{L1}, n_{L1}) pairs for $k = 16$ and present them in Table 2.

Table 2. Error in subspace estimation for $k = 16$, varying n_{L1} & t_{L1}.

(t_{L1}, n_{L1})	2^{14}	2^{15}	2^{16}	2^{17}	2^{18}	2^{19}	2^{20}
2^1	0.652	0.593	0.403	0.289	0.195	0.132	0.101
2^2	0.383	0.308	0.194	0.129	0.101	0.069	0.05
2^3	0.203	0.153	0.099	0.072	0.052	0.034	0.03

E.2. Clustering

Given a subspace estimation error is ~ 0.1, the clustering step is performed with $n_H = \max\left\{k^{3/2}, 256\right\}$ tasks for various t_H. The minimum t_H such that the clustering accuracy is above 99% for at-least $1-\delta$ fraction of 10 random trials is denoted by $t_{\min}(1-\delta)$. Figure 4, and Table 3 illustrate the dependence of k on $t_{\min}(0.5)$, and $t_{\min}(0.9)$.

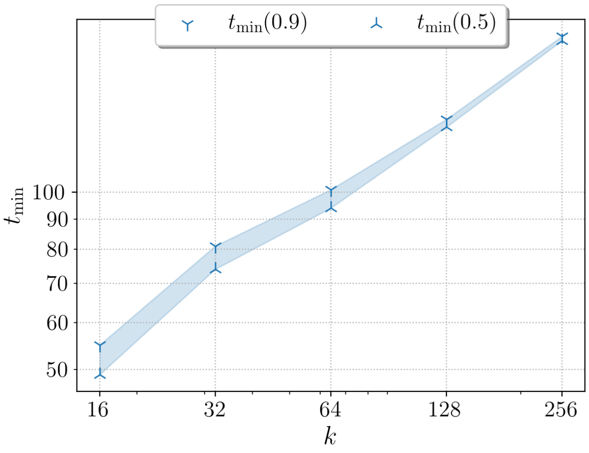


Figure 4. $t_{\min}(0.9)$ and $t_{\min}(0.5)$ for various k

Table 3. $t_{\min}$ for various k, for 99% clustering w.h.p.

k	16	32	64	128	256
$t_{\min}(0.9)$	55	81	101	133	184
$t_{\min}(0.5)$	49	74	94	129	181

E.3. Classification and parameter estimation

Given a subspace estimation error is ~ 0.1, and a clustering accuracy is $> 99\%$, the classification step is performed on $n_{L2} = \max\left\{512, k^{3/2}\right\}$ tasks for variour $t_{L2} \in \mathbb{N}$. The empirical mean of the classification accuracy is computed for every t_{L2}, and illustrated in Figure 6. Similar to the simulations in the clustering step, $t_{\min}(1-\delta)$ is estimated such that the classification accuracy is above 99% for at-least $1-\delta$ fraction times of 10 random trials, and is illustrated in Table 4. With $t_{L2} = t_{\min}(0.9)$, and various $n_{L2} \in \mathbb{N}$, the estimation errors of $\widehat{\mathbf{W}}$, $\widehat{\mathbf{s}}$, and $\widehat{\mathbf{p}}$ are computed as the infimum of ϵ satisfying (12), and is illustrated in Figure 5.

Table 4. $t_{\min}$ for various k, for 99% classification w.h.p.

k	16	32	64	128
$t_{\min}(0.9)$	31	34	36	38
$t_{\min}(0.5)$	28	28	34	36

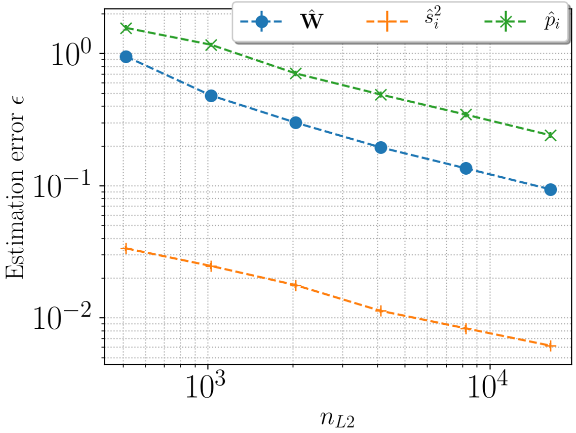


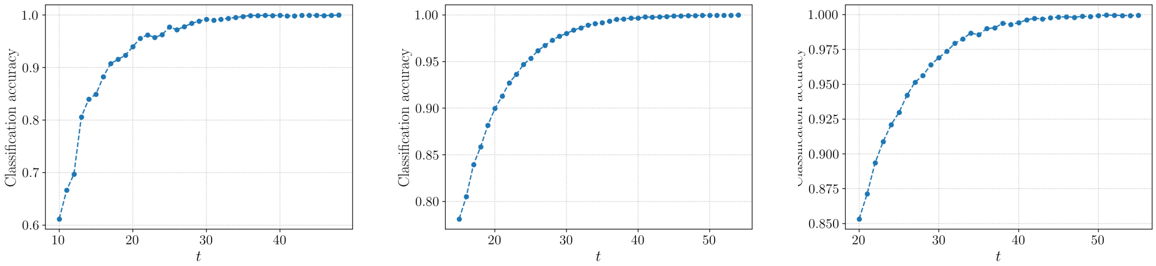
Figure 5. Estimation errors for $k = 32$.

E.4. Prediction

As a continuation of the simulations in this section, we proceed to the prediction step for $k = 32$ and $d = 256$. We use both the estimators: Bayes estimator, and the MAP estimator and illustrate the training and prediction errors in Figure 2. We also compare the prediction error with the vanilla least squares estimator if each task were learnt separately to contrast the gain in meta-learning.

E.5. Comparison for parameter estimation against Expectation Maximization (EM) algorithm

For fair comparisons, we consider our meta dataset for $k = 32$, and $d = 256$ to jointly have n_{L1} tasks with t_{L1} examples, n_H tasks with t_H examples, and n_{L2} tasks with t_{L2} examples as were used in Section E.3. We observe that the convergence of EM algorithm is very sensitive to the initialization, thus we investigate the sensitivity with the following experiment. We initialize $\mathbf{W}^{(0)} = \mathcal{P}_{B_{2,d}(\mathbf{0},1)}\left(\mathbf{W} + \mathbf{Z}\right)$, where $Z_{i,j} \sim \mathcal{N}(0, \gamma^2)\ \forall\ i \in [d]\,, j \in [k]$, $\mathbf{s} = |\mathbf{q}|$, where $\mathbf{q} \sim \mathcal{N}\left(\mathbf{s}, 0.1\mathbf{I}_k\right)$, and $\mathbf{p}^{(0)} = |\mathbf{z}|\,/\,\|\mathbf{z}\|_1$ where $\mathbf{z} \sim \mathcal{N}\left(\mathbf{p}, \mathbf{I}_k/k\right)$. $\mathcal{P}_{\mathcal{X}}(\cdot)$ denotes the projection operator that projects each column of its argument on set $\mathcal{X}$. We observe that EM algorithm fails to converge for $\gamma^2 \geq 0.5$ for this setup unlike our algorithm. When EM converges, we observe similar estimation errors as in Figure 5.



(a) $k = 32$

(b) $k = 64$

(c) $k = 128$

Figure 6. Classification accuracies for various k