A PROOF OF THEOREMS AND TECHNICAL LEMMAS

A.1 PROOF OF LEMMA 2.1

Recall the shorthand $\bar{y}_a = (\sum_{i\in B_a} y_i)/k$, for $a \in [m]$. We have,

$$
\begin{aligned}
\mathcal{L}_{\text{ins}}(\boldsymbol{\theta}) &= \frac{1}{mk}\sum_{a=1}^{m}\sum_{i\in B_a}(\bar{y}_a - f_{\boldsymbol{\theta}}(\boldsymbol{x}_i))^2 \\
&= \frac{1}{mk}\sum_{a=1}^{m}\sum_{i\in B_a}\Big(\bar{y}_a - \frac{1}{k}\sum_{j\in B_a} f_{\boldsymbol{\theta}}(\boldsymbol{x}_j) + \frac{1}{k}\sum_{j\in B_a} f_{\boldsymbol{\theta}}(\boldsymbol{x}_j) - f_{\boldsymbol{\theta}}(\boldsymbol{x}_i)\Big)^2 \\
&= \frac{1}{mk}\sum_{a=1}^{m}\sum_{i\in B_a}\Big(\bar{y}_a - \frac{1}{k}\sum_{j\in B_a} f_{\boldsymbol{\theta}}(\boldsymbol{x}_j)\Big)^2 + \frac{1}{mk}\sum_{a=1}^{m}\sum_{i\in B_a}\Big(\frac{1}{k}\sum_{j\in B_a} f_{\boldsymbol{\theta}}(\boldsymbol{x}_j) - f_{\boldsymbol{\theta}}(\boldsymbol{x}_i)\Big)^2 \\
&\quad + \frac{2}{mk}\sum_{a=1}^{m}\sum_{i\in B_a}\Big(\bar{y}_a - \frac{1}{k}\sum_{j\in B_a} f_{\boldsymbol{\theta}}(\boldsymbol{x}_j)\Big)\Big(\frac{1}{k}\sum_{j\in B_a} f_{\boldsymbol{\theta}}(\boldsymbol{x}_j) - f_{\boldsymbol{\theta}}(\boldsymbol{x}_i)\Big)
\end{aligned}
$$

Note that the first term can be written as

$$
\frac{1}{mk}\sum_{a=1}^{m}\sum_{i\in B_a}\Big(\bar{y}_a - \frac{1}{k}\sum_{j\in B_a} f_{\boldsymbol{\theta}}(\boldsymbol{x}_j)\Big)^2 = \frac{1}{k}\sum_{a=1}^{m}\Big(\bar{y}_a - \frac{1}{k}\sum_{j\in B_a} f_{\boldsymbol{\theta}}(\boldsymbol{x}_j)\Big)^2 = \mathcal{L}_{\text{bag}}(\boldsymbol{\theta})\,.
$$

For the second term, we have

$$
\sum_{i\in B_a}\Big(\frac{1}{k}\sum_{j\in B_a} f_{\boldsymbol{\theta}}(\boldsymbol{x}_j) - f_{\boldsymbol{\theta}}(\boldsymbol{x}_i)\Big)^2 = \frac{1}{k}\sum_{i,j\in B_a}(f_{\boldsymbol{\theta}}(\boldsymbol{x}_i) - f_{\boldsymbol{\theta}}(\boldsymbol{x}_j))^2 = \mathcal{R}(\boldsymbol{\theta})\,,
$$

where we used the following identity for $a_1, \dots, a_k$ and $\bar{a} = (\sum_{i=1}^{k} a_i)/k$:

$$
\sum_{i=1}^{k}(a_i - \bar{a})^2 = \frac{1}{k}\sum_{i,j=1}^{k}(a_i - a_j)^2\,. \tag{11}
$$

Finally, the third term works out at zero because

$$
\begin{aligned}
&\sum_{a=1}^{m}\sum_{i\in B_a}\Big(\bar{y}_a - \frac{1}{k}\sum_{j\in B_a} f_{\boldsymbol{\theta}}(\boldsymbol{x}_j)\Big)\Big(\frac{1}{k}\sum_{j\in B_a} f_{\boldsymbol{\theta}}(\boldsymbol{x}_j) - f_{\boldsymbol{\theta}}(\boldsymbol{x}_i)\Big) \\
&= \sum_{a=1}^{m}\Big(\bar{y}_a - \frac{1}{k}\sum_{j\in B_a} f_{\boldsymbol{\theta}}(\boldsymbol{x}_j)\Big)\Big[\sum_{i\in B_a}\Big(\frac{1}{k}\sum_{j\in B_a} f_{\boldsymbol{\theta}}(\boldsymbol{x}_j) - f_{\boldsymbol{\theta}}(\boldsymbol{x}_i)\Big)\Big] = 0,
\end{aligned}
$$

since

$$
\sum_{i\in B_a}\Big(\frac{1}{k}\sum_{j\in B_a} f_{\boldsymbol{\theta}}(\boldsymbol{x}_j) - f_{\boldsymbol{\theta}}(\boldsymbol{x}_i)\Big) = \sum_{j\in B_a} f_{\boldsymbol{\theta}}(\boldsymbol{x}_j) - \sum_{i\in B_a} f_{\boldsymbol{\theta}}(\boldsymbol{x}_i) = 0\,.
$$

Combining the three terms together we arrive at $\mathcal{L}_{\text{ins}}(\boldsymbol{\theta}) = \mathcal{L}_{\text{bag}}(\boldsymbol{\theta}) + \mathcal{R}(\boldsymbol{\theta})$.

A.2 PROOF OF LEMMA 2.2

We use the shorthand $\bar{f}_a = (\sum_{i\in B_a} f_{\boldsymbol{\theta}}(\boldsymbol{x}_i))/k$, for $a \in [m]$. By Taylor's expansion of the loss ℓ on its second argument we have

$$
\ell(\bar{y}_a, f_{\boldsymbol{\theta}}(\boldsymbol{x}_i)) = \ell(\bar{y}_a, \bar{f}_a) + \frac{\partial}{\partial b}\ell(\bar{y}_a, \bar{f}_a)(f_{\boldsymbol{\theta}}(\boldsymbol{x}_i) - \bar{f}_a) + \frac{1}{2}\frac{\partial^2}{\partial b^2}\ell(\bar{y}_a, f)(f_{\boldsymbol{\theta}}(\boldsymbol{x}_i) - \bar{f}_a)^2\,,
$$

for some f between $\bar{f}_a$ and $f_{\boldsymbol{\theta}}(\boldsymbol{x}_i)$ and ∂/∂_b, ∂^2/∂_b^2 indicate the first and second derivative of $\ell(a, b)$ with respect to the second input b.

Summing both sides of the above equation over $i \in B_a$, the second term works out at zero since $\sum_{i \in B_a}(f_{\boldsymbol{\theta}}(\boldsymbol{x}_i) - \bar{f}_a) = 0$. Using the bound on the second derivative we arrive at

$$\sum_{i \in B_a} \ell(\bar{y}_a, f_{\boldsymbol{\theta}}(\boldsymbol{x}_i)) \le k\ell(\bar{y}_a, \bar{f}_a) + \sum_{i \in B_a} C(f_{\boldsymbol{\theta}}(\boldsymbol{x}_i) - \bar{f}_a)^2 .$$

Next, summing both sides of the above equation over bags $a \in [m]$, and dividing by mk, we get

$$\mathcal{L}_{\text{ins}}(\boldsymbol{\theta}) \le \mathcal{L}_{\text{bag}}(\boldsymbol{\theta}) + \frac{1}{k}\sum_{a=1}^{m}\sum_{i \in B_a} C(f_{\boldsymbol{\theta}}(\boldsymbol{x}_i) - \bar{f}_a)^2 .$$

By invoking identity (11) in the above we arrive at (5).

We are now ready to prove the second part of the statement. If the loss $\ell(\cdot,\cdot)$ is convex in the second input, by the Jensen's inequality we have

$$\frac{1}{k}\sum_{i \in B_a} \ell(\bar{y}_a, f_{\boldsymbol{\theta}}(\boldsymbol{x}_i)) \ge \ell\Big(\bar{y}_a, \frac{1}{k}\sum_{i \in B_a} f_{\boldsymbol{\theta}}(\boldsymbol{x}_i)\Big)$$

Taking the average of both side over the bags $a \in [m]$, we obtain that $\mathcal{L}_{\text{ins}}(\boldsymbol{\theta}) \ge \mathcal{L}_{\text{bag}}(\boldsymbol{\theta})$, which completes the proof of lemma.

B Proof of Theorem 2.5

Recall m as the number of bags, and n as the number of samples. Since the bags are non-overlapping and each of size k, we have $m = n/k$. Define $\boldsymbol{S} \in \mathbb{R}^{m \times n}$, as a matrix the encodes the bagging structure, with $S_{ia} = 1/\sqrt{k}\mathbf{1}_{\{j \in B_a\}}$ where B_a indicates a-th bag, for $a \in [m]$.

We next write the bag-level loss function and the instance-level loss function in terms of $\boldsymbol{S}$ as follows:

$$\mathcal{L}_{\text{bag}}(\boldsymbol{\theta}) = \frac{1}{km}\|\boldsymbol{S}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\theta})\|_2^2 ,$$

$$\mathcal{L}_{\text{ins}}(\boldsymbol{\theta}) = \frac{1}{km}\|\boldsymbol{S}^\top\boldsymbol{S}\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\theta}\|_2^2 .$$

The interpolating loss function (7) then reads as

$$\mathcal{L}_{\text{int}}(\boldsymbol{\theta}) = \frac{1}{mk}\left((1-\rho)\|\boldsymbol{S}\boldsymbol{X}\boldsymbol{\theta} - \boldsymbol{S}\boldsymbol{y}\|_2^2 + \rho\left\|\boldsymbol{X}\boldsymbol{\theta} - \boldsymbol{S}^\top\boldsymbol{S}\boldsymbol{y}\right\|_2^2\right) .$$

This can equivalently be written as

$$\mathcal{L}_{\text{int}}(\boldsymbol{\theta}) = \frac{1}{mk}\left\|\begin{pmatrix}\sqrt{\rho}\boldsymbol{I}\\ \sqrt{1-\rho}\boldsymbol{S}\end{pmatrix}\boldsymbol{X}\boldsymbol{\theta} - \begin{pmatrix}\sqrt{\rho}\boldsymbol{S}^\top\boldsymbol{S}\boldsymbol{y}\\ \sqrt{1-\rho}\boldsymbol{S}\boldsymbol{y}\end{pmatrix}\right\|_2^2 .$$

The minimizer of the above loss admits a closed-from solution given by $\widehat{\boldsymbol{\theta}}_{\text{int}} = \boldsymbol{B}\boldsymbol{y}$, with

$$\boldsymbol{B} = \left(\boldsymbol{X}^\top\left(\rho\boldsymbol{I} + (1-\rho)\boldsymbol{S}^\top\boldsymbol{S}\right)\boldsymbol{X}\right)^{-1}\boldsymbol{X}^\top\boldsymbol{S}^\top\boldsymbol{S} .$$

We define the shorthand $\boldsymbol{E} := \rho\boldsymbol{I} + (1-\rho)\boldsymbol{S}^\top\boldsymbol{S} \in \mathbb{R}^{n \times n}$, which is non-singular for $\rho > 0$, and $\boldsymbol{M} = (\boldsymbol{X}^\top\boldsymbol{E}\boldsymbol{X})^{-1}\boldsymbol{X}^\top \in \mathbb{R}^{d \times n}$. We then have $\boldsymbol{B} = \boldsymbol{M}\boldsymbol{S}^\top\boldsymbol{S}$.

We next recall the bias-variance decomposition (6), where the bias and variance are given by

$$\begin{aligned}\text{Bias}(\widehat{\boldsymbol{\theta}}_{\text{int}}) &= \|(\boldsymbol{B}\boldsymbol{X} - \boldsymbol{I})\boldsymbol{\theta}_0\|_2^2\\ &= \|(\boldsymbol{M}\boldsymbol{S}^\top\boldsymbol{S}\boldsymbol{X} - \boldsymbol{M}\boldsymbol{E}\boldsymbol{X})\boldsymbol{\theta}_0\|_2^2\\ &= \|\boldsymbol{M}(\boldsymbol{S}^\top\boldsymbol{S} - \boldsymbol{E})\boldsymbol{X}\boldsymbol{\theta}_0\|_2^2 , \qquad (12)\\ \text{Var}(\widehat{\boldsymbol{\theta}}_{\text{int}}) &= \sigma^2\|\boldsymbol{M}\boldsymbol{S}^\top\boldsymbol{S}\|_F^2 , \qquad (13)\end{aligned}$$

with $\|\cdot\|_F$ indicating the matrix Frobenius norm.

We continue by treating the bias and the variance separately.

B.1 CALCULATING THE BIAS

Since the distribution of the features matrix $\boldsymbol{X}$ is invariant under rotation, we can assume that $\boldsymbol{\theta}_0 = \|\boldsymbol{\theta}_0\|\boldsymbol{e}_i$, where $\boldsymbol{e}_i \in \mathbb{R}^d$ is the vector with one at i-th entry and zero everywhere else. By taking average on $i \in [d]$ we obtain

$$\begin{aligned}\text{Bias}(\widehat{\boldsymbol{\theta}}_{\text{int}}) &\overset{(d)}{=} \frac{\|\boldsymbol{\theta}_0\|_2^2}{d}\sum_{i\in[d]}\|\boldsymbol{M}(\boldsymbol{S}^\top\boldsymbol{S}-\boldsymbol{E})\boldsymbol{X}\boldsymbol{e}_i\|_2^2\\ &= \frac{\|\boldsymbol{\theta}_0\|_2^2}{d}\operatorname{tr}\Big(\boldsymbol{M}(\boldsymbol{S}^\top\boldsymbol{S}-\boldsymbol{E})\boldsymbol{X}\Big(\sum_{i\in[p]}\boldsymbol{e}_i\boldsymbol{e}_i^\top\Big)\boldsymbol{X}^\top(\boldsymbol{S}^\top\boldsymbol{S}-\boldsymbol{E})\boldsymbol{M}^\top\Big)\\ &= \frac{\|\boldsymbol{\theta}_0\|_2^2}{d}\left\|\boldsymbol{M}(\boldsymbol{S}^\top\boldsymbol{S}-\boldsymbol{E})\boldsymbol{X}\right\|_F^2 .\end{aligned}$$

Let us define $\boldsymbol{\Lambda} \in \mathbb{R}^{n\times n}$ as follows:

$$\begin{aligned}\boldsymbol{\Lambda} &:= -(\boldsymbol{S}^\top\boldsymbol{S}-\boldsymbol{E})\\ &= -(\boldsymbol{S}^\top\boldsymbol{S}-(\rho\boldsymbol{I}+(1-\rho)\boldsymbol{S}^\top\boldsymbol{S}))\\ &= \rho(\boldsymbol{I}-\boldsymbol{S}^\top\boldsymbol{S}) .\end{aligned} \tag{14}$$

The bias can then be written in terms of $\boldsymbol{\Lambda}$ as $\text{Bias}(\boldsymbol{\theta}) = \frac{1}{d}\|\boldsymbol{M}\boldsymbol{\Lambda}\boldsymbol{X}\|_F^2$. In our next lemma, we characterize the asymptotic behavior of the bias.

Lemma B.1. *Under the asymptotic regime of Assumption 2.3, we have*

$$\frac{1}{d}\|\boldsymbol{M}\boldsymbol{\Lambda}\boldsymbol{X}\|_F^2 \overset{(p)}{\to} \alpha_*^2 + \frac{\alpha_*^2}{\frac{(k-1)\psi}{k^2(1-\alpha_*)^2} - (\frac{\alpha_*}{1-\alpha_*})^2\frac{1}{k} - \frac{k-1}{k}} ,$$

where α_* is the nonnegative fixed point of the following equation:

$$\rho + \frac{\psi}{k(1-\alpha_*)} - 1 = \frac{\psi}{k\alpha_*}\rho(k-1) .$$

Since $\|\boldsymbol{\theta}_0\| \to 1$, the result (8) follows from Lemma B.1.

We refer to the supplementary D.1 for the proof of Lemma B.1.

B.2 CALCULATING THE VARIANCE

Since the bags are non-overlapping we have $\boldsymbol{S}\boldsymbol{S}^\top = \boldsymbol{I}_m$. Therefore $\boldsymbol{S}^\top\boldsymbol{S}$ is a projection matrix and can be written as $\boldsymbol{S}^\top\boldsymbol{S} = \boldsymbol{U}\boldsymbol{U}^\top$, with $\boldsymbol{U} \in \mathbb{R}^{n\times m}$ an orthogonal matrix. Recall that the variance is given by $\text{Var}(\widehat{\boldsymbol{\theta}}_{\text{int}}) = \sigma^2\|\boldsymbol{M}\boldsymbol{S}^\top\boldsymbol{S}\|_F^2$. We use the next lemma to characterize the asymptotic behavior of the variance.

Lemma B.2. *Under the asymptotic regime of Assumption 2.3, for any vector $\boldsymbol{a} \in \mathbb{R}^m$ we have*

$$\frac{n}{\|\boldsymbol{a}\|^2}\|\boldsymbol{M}\boldsymbol{U}\boldsymbol{a}\|_2^2 \overset{(p)}{\to} \frac{k}{v_*} ,$$

where v_* is given as the fixed point of the following system of equations in (v, u):

$$\begin{cases}\frac{\psi}{1+u} + \frac{\rho\psi(k-1)}{\rho+u} & = k,\\ \frac{\psi(1+v)}{(1+u)^2} + \frac{\rho^2\psi(k-1)}{(\rho+u)^2} & = k .\end{cases}$$

Proof of Lemma B.2 is given in the supplementary D.2.

We next use the above lemma for each row of $\boldsymbol{U}$ separately (as the vector $\boldsymbol{a}$) and add them together. Using the fact that $\|\boldsymbol{U}\|_F^2 = m$ and $m/n = k$, we get that $\|\boldsymbol{M}\boldsymbol{U}\boldsymbol{U}^\top\|_F^2 \overset{(p)}{\to} 1/v_*$, which completes the variance calculation.

B.3 PROOF OF LEMMA 4.2

We use the idea of (Dwork et al., 2014, Theorem 3.6) to prove this lemma. Given a database $D = (y_1, y_2, \dots, y_n)$, Algorithm 1 (we denote this mapping by $\mathcal{A} : \mathbb{R}^n \to \mathbb{R}^m$) outputs m real numbers $(\tilde{y}_1, \tilde{y}_2, \dots, \tilde{y}_m)$. Given the database D, we define the map $f : \mathbb{R}^n \to \mathbb{R}^m$ by $D \mapsto (\bar{y}_1, \bar{y}_2, \dots, \bar{y}_m)$, which computes the mean of labels in each bag. Fix any pair of neighboring databases D, D' that differ in the label of a single example. We have $\|f(D) - f(D')\|_1 := \sum_{a\in[m]} |f(D)_a - f(D')_a| \le \Delta f := \frac{C\sqrt{\log n}}{k}$. In this argument, we used Assumption 2.4 that assumes non-overlapping bags, and therefore, changing a certain y_i in D leads to a change in only one of $\bar{y}_a$ by at most Δf. Let $p_{\mathcal{A}(D)}(z)$ and $p_{\mathcal{A}(D')}(z)$ denote the probability density function of $\mathcal{A}(D)$ and $\mathcal{A}(D')$. We have

$$\begin{aligned}
\frac{p_{\mathcal{A}(D)}(z)}{p_{\mathcal{A}(D')}(z)} &= \prod_{a\in[m]} \frac{\exp\left(-\frac{\varepsilon|f(D)_a - z_a|}{\Delta f}\right)}{\exp\left(-\frac{\varepsilon|f(D')_a - z_a|}{\Delta f}\right)} \\
&= \prod_{a\in[m]} \exp\left(\frac{\varepsilon\left(|f(D')_a - z_a| - |f(D)_a - z_a|\right)}{\Delta f}\right) \\
&\le \prod_{a\in[m]} \exp\left(\frac{\varepsilon|f(D')_a - f(D)_a|}{\Delta f}\right) \\
&= \exp\left(\frac{\varepsilon\|f(D) - f(D')\|_1}{\Delta f}\right) \\
&\le e^{\varepsilon}\,,
\end{aligned}$$

which completes the proof.

C PROOF OF THEOREM 4.3

Recall that in Algorithm 1, the individual responses are first truncated by $C\sqrt{\log n}$ and then after the aggregate responses are computed, a Laplace noise is added to them to ensure label DP. We define $\mathcal{E}$ is the event that no truncation happens, namely:

$$\mathcal{E} := \mathbf{1}_{\{|y_i| \le C\sqrt{\log n},\ \forall i \in [n]\}}\,. \tag{15}$$

Since $y_i \sim N(0, \|\boldsymbol{\theta}_0\|^2 + \sigma^2)$, $\|\boldsymbol{\theta}_0\| = 1$, by using Gaussian tail bound along with union bounding we arrive at

$$\mathbb{P}(\mathcal{E}) \ge 1 - n\exp\Big(-\frac{C^2}{2(1+\sigma^2)}\log n\Big) = 1 - n^{-c}\,, \tag{16}$$

with $c = \frac{C^2}{2(1+\sigma^2)} - 1 > 0$.

We next bound the risk of estimator $\widehat{\boldsymbol{\theta}}_{\rm int}$ as follows:

$$\text{Risk}(\widehat{\boldsymbol{\theta}}_{\rm int}) = \mathbb{E}[\|\widehat{\boldsymbol{\theta}}_{\rm int} - \boldsymbol{\theta}_0\|^2 \mathbf{1}_{\{\mathcal{E}\}}|\boldsymbol{X}] + \mathbb{E}[\|\widehat{\boldsymbol{\theta}}_{\rm int} - \boldsymbol{\theta}_0\|^2 \mathbf{1}_{\{\mathcal{E}^c\}}|\boldsymbol{X}]\,. \tag{17}$$

For the first term, note that on the instance $\mathcal{E}$ (no truncation), the privatized aggregate responses are just the aggregate responses with an additive zero mean noise with variance $2C^2\log n/(k\varepsilon)^2$. So we can use the analysis in the proof of Theorem 2.5 with the inflated noise variance. Let $\widehat{\boldsymbol{\theta}}^{\rm nt}_{\rm int}$ be the estimator using untruncated responses in Algorithm 1. We then have This gives us

$$\begin{aligned}
\frac{1}{\log n}\mathbb{E}[\|\widehat{\boldsymbol{\theta}}_{\rm int} - \boldsymbol{\theta}_0\|^2 \mathbf{1}_{\{\mathcal{E}\}}|\boldsymbol{X}] &= \frac{1}{\log n}\mathbb{E}[\|\widehat{\boldsymbol{\theta}}^{\rm nt}_{\rm int} - \boldsymbol{\theta}_0\|^2 \mathbf{1}_{\{\mathcal{E}\}}|\boldsymbol{X}] \\
&= \frac{1}{\log n}\mathbb{E}[\|\widehat{\boldsymbol{\theta}}^{\rm nt}_{\rm int} - \boldsymbol{\theta}_0\|^2|\boldsymbol{X}] - \frac{1}{\log n}\mathbb{E}[\|\widehat{\boldsymbol{\theta}}^{\rm nt}_{\rm int} - \boldsymbol{\theta}_0\|^2 \mathbf{1}_{\{\mathcal{E}^c\}}|\boldsymbol{X}] \\
&= \frac{1}{\log n}\,\text{Bias}(\widehat{\boldsymbol{\theta}}^{\rm nt}_{\rm int}) + \frac{1}{\log n}\text{Var}(\widehat{\boldsymbol{\theta}}^{\rm nt}_{\rm int}) - \frac{1}{\log n}\mathbb{E}[\|\widehat{\boldsymbol{\theta}}^{\rm nt}_{\rm int} - \boldsymbol{\theta}_0\|^2 \mathbf{1}_{\{\mathcal{E}^c\}}|\boldsymbol{X}]\,,
\end{aligned} \tag{18}$$

where the bias is given by (8) and variance is given by (9), where σ^2/k is replaced with the inflated variance $\sigma^2/k + 2C^2 \log n/(k\varepsilon)^2$. Since $\mathrm{Bias}(\widehat{\boldsymbol{\theta}}^{\mathrm{nt}}_{\mathrm{int}})$ has a finite limit, the first term above vanishes as $n \to \infty$. For the second term we have

$$\frac{1}{\log n} \mathrm{Var}(\widehat{\boldsymbol{\theta}}^{\mathrm{nt}}_{\mathrm{int}}) \overset{(p)}{\to} \frac{2C^2}{k\varepsilon^2} \frac{1}{v_*} \, .$$

since $\sigma^2/\log n \to 0$. For the third term, by Cauchy–Schwarz inequality we have

$$\mathbb{E}[\|\widehat{\boldsymbol{\theta}}^{\mathrm{nt}}_{\mathrm{int}} - \boldsymbol{\theta}_0\|^2 \mathbf{1}_{\{\mathcal{E}^c\}} | \boldsymbol{X}] \le \mathbb{E}[\|\widehat{\boldsymbol{\theta}}^{\mathrm{nt}}_{\mathrm{int}} - \boldsymbol{\theta}_0\|^4 | \boldsymbol{X}]^{1/2} \, \mathbb{P}(\mathcal{E}^c) \, . \tag{19}$$

Using the high probability bound on the minimum singular value of the Gaussian matrix $\boldsymbol{X}$ (Vershynin, 2018, Theorem 4.6.1), we can show that $\mathbb{E}[\|\widehat{\boldsymbol{\theta}}^{\mathrm{nt}}_{\mathrm{int}} - \boldsymbol{\theta}_0\|^4 | \boldsymbol{X}]$ is bounded in probability and since $\mathbb{P}(\mathcal{E}^c) \le n^{-c}$, we conclude that the third term in (18) also vanishes as $n \to \infty$, in probability. Combining these together we arrive at

$$\frac{1}{\log n} \mathbb{E}[\|\widehat{\boldsymbol{\theta}}_{\mathrm{int}} - \boldsymbol{\theta}_0\|^2 \mathbf{1}_{\{\mathcal{E}\}} | \boldsymbol{X}] \overset{(p)}{\to} \frac{2C^2}{k\varepsilon^2} \frac{1}{v_*} \, . \tag{20}$$

Similar to (19) we can also show that

$$\frac{1}{\log n} \mathbb{E}[\|\widehat{\boldsymbol{\theta}}_{\mathrm{int}} - \boldsymbol{\theta}_0\|^2 \mathbf{1}_{\{\mathcal{E}^c\}} | \boldsymbol{X}] \overset{(p)}{\to} 0 \, ,$$

which along with (20) and (17) implies that

$$\frac{1}{\log n} \mathrm{Risk}(\widehat{\boldsymbol{\theta}}_{\mathrm{int}}) \overset{(p)}{\to} \frac{2C^2}{k\varepsilon^2} \frac{1}{v_*} \, ,$$

completing the proof.

D Proof of Intermediate Lemmas

D.1 Proof of Lemma B.1

Write $\boldsymbol{X} = [\boldsymbol{x}_1, \dots, \boldsymbol{x}_d]$ with $\boldsymbol{x}_i$ representing the i-th column. We then have $\|\boldsymbol{M}\boldsymbol{\Lambda}\boldsymbol{X}\|^2 = \sum_{i=1}^d \|\boldsymbol{M}\boldsymbol{\Lambda}\boldsymbol{x}_i\|^2$. We compute the asymptotic behavior of each of the summand separately. Indeed, by symmetry of the distributions of $\boldsymbol{x}_i$, we will see that all summands converge to the same limit.

Recall that $\boldsymbol{M} = (\boldsymbol{X}^\top \boldsymbol{E} \boldsymbol{X})^{-1} \boldsymbol{X}^\top$. Consider the following optimization problem:

$$\boldsymbol{\alpha}_i = \arg\min_{\boldsymbol{\alpha} \in \mathbb{R}^d} \frac{1}{d} \|\boldsymbol{E}^{1/2} \boldsymbol{X} \boldsymbol{\alpha} - \boldsymbol{E}^{-1/2} \boldsymbol{\Lambda} \boldsymbol{x}_i\|_2^2 \, . \tag{21}$$

It is easy to see that by the KKT condition, $\boldsymbol{\alpha}_i = (\boldsymbol{X}^\top \boldsymbol{E} \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{\Lambda} \boldsymbol{x}_i = \boldsymbol{M}\boldsymbol{\Lambda}\boldsymbol{x}_i$. Therefore, we are interested in characterizing $\|\boldsymbol{\alpha}_i\|$ in the asymptotic regime, described in Assumption 2.3.

We write $\boldsymbol{\alpha}$ as $(\alpha_i, \boldsymbol{\alpha}_{\sim i})$ to separate its i-th entry form the rest. Likewise we write $\boldsymbol{X} = [\boldsymbol{x}_i \boldsymbol{X}_{\sim i}]$ to separate the i-th columns from the rest. We then have

$$\begin{aligned}
&\min_{\boldsymbol{\alpha} \in \mathbb{R}^d} \frac{1}{d} \|\boldsymbol{E}^{1/2} \boldsymbol{X} \boldsymbol{\alpha} - \boldsymbol{E}^{-1/2} \boldsymbol{\Lambda} \boldsymbol{x}_i\|_2^2 \\
&= \min_{\boldsymbol{\alpha} \in \mathbb{R}^p} \frac{1}{d} \|\boldsymbol{E}^{1/2} \boldsymbol{x}_i \alpha_i + \boldsymbol{E}^{1/2} \boldsymbol{X}_{\sim i} \boldsymbol{\alpha}_{\sim i} - \boldsymbol{E}^{-1/2} \boldsymbol{\Lambda} \boldsymbol{x}_i\|_2^2 \\
&= \min_{\boldsymbol{\alpha} \in \mathbb{R}^d} \frac{1}{d} \|\boldsymbol{E}^{1/2} \boldsymbol{X}_{\sim i} \boldsymbol{\alpha}_{\sim i} + (\alpha_i \boldsymbol{E}^{1/2} - \boldsymbol{E}^{-1/2} \boldsymbol{\Lambda}) \boldsymbol{x}_i\|_2^2 \\
&= \min_{\boldsymbol{\alpha} \in \mathbb{R}^d} \max_{\boldsymbol{v} \in \mathbb{R}^n} \frac{2}{d} \left(\boldsymbol{v}^\top (\alpha_i \boldsymbol{E}^{1/2} - \boldsymbol{E}^{-1/2} \boldsymbol{\Lambda}) \boldsymbol{x}_i + \boldsymbol{v}^\top \boldsymbol{E}^{1/2} \boldsymbol{X}_{\sim i} \boldsymbol{\alpha}_{\sim i} - \frac{1}{2} \|\boldsymbol{v}\|_2^2 \right) ,
\end{aligned} \tag{22}$$

where in the last step we used the identity $\max_{\boldsymbol{v}} (\boldsymbol{v}^\top \boldsymbol{x} - \|\boldsymbol{v}\|^2/2) = \|\boldsymbol{x}\|^2/2$ for any vector $\boldsymbol{x}$.

We next note that $\boldsymbol{S}\boldsymbol{S}^\top = \boldsymbol{I}$ since the bags are non-overlapping. Therefore we can write $\boldsymbol{S}^\top\boldsymbol{S} = \boldsymbol{U}\boldsymbol{U}^\top$ for an orthogonal matrix $\boldsymbol{U} \in \mathbb{R}^{n\times m}$. We then have

$$\boldsymbol{E} := \rho\boldsymbol{I} + (1-\rho)\boldsymbol{S}^\top\boldsymbol{S} = \boldsymbol{U}\boldsymbol{U}^\top + \rho\boldsymbol{U}_\perp\boldsymbol{U}_\perp^\top\,, \quad \boldsymbol{\Lambda} = \rho(\boldsymbol{I} - \boldsymbol{S}^\top\boldsymbol{S}) = \rho\boldsymbol{U}_\perp\boldsymbol{U}_\perp^\top.$$

where $\boldsymbol{U}_\perp$ is an orthogonal matrix representing the orthogonal space to the column space of $\boldsymbol{U}$. We next decompose the vector $\boldsymbol{v}$ in the above optimization as $\boldsymbol{v} = \boldsymbol{U}\boldsymbol{v}_1 + \boldsymbol{U}_\perp\boldsymbol{v}_2$ and therefore $\|\boldsymbol{v}\|^2 = \|\boldsymbol{v}_1\|^2 + \|\boldsymbol{v}_2\|^2$.

We introduce the change of variable $\tilde{\boldsymbol{v}} = \boldsymbol{E}^{1/2}\boldsymbol{v}$ in optimization (22). Note that $\tilde{\boldsymbol{v}} = \boldsymbol{U}\boldsymbol{v}_1 + \sqrt{\rho}\boldsymbol{U}_\perp\boldsymbol{v}_2$. Continuing with (22) in terms of $\tilde{\boldsymbol{v}}$ we have

$$\min_{\boldsymbol{\alpha}\in\mathbb{R}^d}\max_{\tilde{\boldsymbol{v}}\in\mathbb{R}^n} \frac{2}{d}\left(\tilde{\boldsymbol{v}}^\top(\alpha_i\boldsymbol{I} - \boldsymbol{E}^{-1}\boldsymbol{\Lambda})\boldsymbol{x}_i + \tilde{\boldsymbol{v}}^\top\boldsymbol{X}_{\sim i}\boldsymbol{\alpha}_{\sim i} - \frac{1}{2}\|\boldsymbol{E}^{-1/2}\tilde{\boldsymbol{v}}\|_2^2\right). \tag{23}$$

To analyze the asymptotic behavior of the solution to the above minimax optimization, we use the Convex-Gaussian-Minimax-Theorem (CGMT) (Thrampoulidis et al., 2015, Theorem 3), which is a power extension of the classical Gordon's Gaussian min-max theorem Gordon (1988), under additional convexity assumptions. According to CGMT, the above optimization is equivalent to the following auxiliary optimization problem:

$$\min_{\boldsymbol{\alpha}\in\mathbb{R}^d}\max_{\tilde{\boldsymbol{v}}\in\mathbb{R}^n} \frac{2}{d}\left(\tilde{\boldsymbol{v}}^\top(\alpha_i\boldsymbol{I} - \boldsymbol{E}^{-1}\boldsymbol{\Lambda})\boldsymbol{x}_i + \|\boldsymbol{\alpha}_{\sim i}\|\tilde{\boldsymbol{v}}^\top\boldsymbol{g} + \|\tilde{\boldsymbol{v}}\|\boldsymbol{h}^\top\boldsymbol{\alpha}_{\sim i} - \frac{1}{2}\|\boldsymbol{E}^{-1/2}\tilde{\boldsymbol{v}}\|_2^2\right), \tag{24}$$

with $\boldsymbol{g} \sim N(0, \boldsymbol{I}_n)$ and $\boldsymbol{h} \sim N(0, \boldsymbol{I}_{d-1})$ independent Gaussian vectors. We next write the above optimization in terms of the components $\boldsymbol{v}_1$ and $\boldsymbol{v}_2$, noting that $\boldsymbol{E}^{-1}\boldsymbol{\Lambda} = \boldsymbol{U}_\perp\boldsymbol{U}_\perp^\top$, as follows:

$$\begin{aligned}\min_{\boldsymbol{\alpha}\in\mathbb{R}^d}\max_{\boldsymbol{v}_1,\boldsymbol{v}_2\in\mathbb{R}^n} \frac{2}{d}\Big(&\alpha_i\boldsymbol{v}_1^\top\boldsymbol{U}^\top\boldsymbol{x}_i + \sqrt{\rho}\boldsymbol{v}_2^\top\boldsymbol{U}_\perp^\top(\alpha_i\boldsymbol{I} - \boldsymbol{U}_\perp\boldsymbol{U}_\perp^\top)\boldsymbol{x}_i + \|\boldsymbol{\alpha}_{\sim i}\|(\boldsymbol{v}_1^\top\boldsymbol{U}^\top\boldsymbol{g} + \sqrt{\rho}\boldsymbol{v}_2^\top\boldsymbol{U}_\perp^\top\boldsymbol{g})\\ &+ \sqrt{\|\boldsymbol{v}_1\|^2 + \rho\|\boldsymbol{v}_2\|^2}\boldsymbol{h}^\top\boldsymbol{\alpha}_{\sim i} - \frac{1}{2}\|\boldsymbol{v}_1\|^2 - \frac{1}{2}\|\boldsymbol{v}_2\|^2\Big).\end{aligned} \tag{25}$$

Define the shorthand

$$\begin{aligned}\boldsymbol{x}_1 &:= \boldsymbol{U}^\top\boldsymbol{x}_i \sim N(0, \boldsymbol{I}_m),\\ \boldsymbol{x}_2 &:= \boldsymbol{U}_\perp^\top\boldsymbol{x}_i \sim N(0, \boldsymbol{I}_{n-m}),\\ \boldsymbol{g}_1 &:= \boldsymbol{U}^\top\boldsymbol{g} \sim N(0, \boldsymbol{I}_m),\\ \boldsymbol{g}_2 &:= \boldsymbol{U}_\perp^\top\boldsymbol{g} \sim N(0, \boldsymbol{I}_{n-m}).\end{aligned}$$

Then optimization (25) can be rewritten as

$$\begin{aligned}\min_{\boldsymbol{\alpha}\in\mathbb{R}^d}\max_{\boldsymbol{v}_1,\boldsymbol{v}_2\in\mathbb{R}^n} \frac{2}{d}\Big(&\alpha_i\boldsymbol{v}_1^\top\boldsymbol{x}_1 + \sqrt{\rho}(\alpha_i - 1)\boldsymbol{v}_2^\top\boldsymbol{x}_2 + \|\boldsymbol{\alpha}_{\sim i}\|(\boldsymbol{v}_1^\top\boldsymbol{g}_1 + \sqrt{\rho}\boldsymbol{v}_2^\top\boldsymbol{g}_2)\\ &+ \sqrt{\|\boldsymbol{v}_1\|^2 + \rho\|\boldsymbol{v}_2\|^2}\boldsymbol{h}^\top\boldsymbol{\alpha}_{\sim i} - \frac{1}{2}\|\boldsymbol{v}_1\|^2 - \frac{1}{2}\|\boldsymbol{v}_2\|^2\Big).\end{aligned} \tag{26}$$

We fix $\|\boldsymbol{v}_1\| = \beta_1$ and $\|\boldsymbol{v}_2\| = \beta_2$ and first optimize over the directions of $\boldsymbol{v}_1$, $\boldsymbol{v}_2$ and then over the norms β_1 and β_2. This brings us to

$$\begin{aligned}\min_{\boldsymbol{\alpha}\in\mathbb{R}^d}\max_{\beta_1,\beta_2\ge 0} \frac{2}{d}\Big(&\beta_1\|\alpha_i\boldsymbol{x}_1 + \|\boldsymbol{\alpha}_{\sim i}\|\boldsymbol{g}_1\| + \beta_2\|\sqrt{\rho}(\alpha_i - 1)\boldsymbol{x}_2 + \|\boldsymbol{\alpha}_{\sim i}\|\sqrt{\rho}\boldsymbol{g}_2\|\\ &+ \sqrt{\beta_1^2 + \rho\beta_2^2}\boldsymbol{h}^\top\boldsymbol{\alpha}_{\sim i} - \frac{1}{2}\beta_1^2 - \frac{1}{2}\beta_2^2\Big).\end{aligned} \tag{27}$$

In order to optimize over $\boldsymbol{\alpha}_{\sim i}$, we first fix its norm to $\eta := \|\boldsymbol{\alpha}_{\sim i}\|$ and optimize over its direction, and then optimize over η, which results in:

$$\begin{aligned}\min_{\eta\ge 0,\alpha_i}\max_{\beta_1,\beta_2\ge 0} \frac{2}{d}\Big(&\beta_1\|\alpha_i\boldsymbol{x}_1 + \eta\boldsymbol{g}_1\| + \beta_2\|\sqrt{\rho}(\alpha_i - 1)\boldsymbol{x}_2 + \eta\sqrt{\rho}\boldsymbol{g}_2\|\\ &+ \eta\sqrt{\beta_1^2 + \rho\beta_2^2}\|\boldsymbol{h}\| - \frac{1}{2}\beta_1^2 - \frac{1}{2}\beta_2^2\Big).\end{aligned} \tag{28}$$

The next step in the CGMT framework is to compute the pointwise limit of the objective functions. Using the concentration of Lipschitz functions of Gaussian vectors we have

$$\frac{1}{\sqrt{d}}\|\alpha_i \boldsymbol{x}_1 + \eta \boldsymbol{g}_1\| \overset{(p)}{\to} \sqrt{(\alpha_i^2+\eta^2)\frac{\psi}{k}}\,,$$
$$\frac{1}{\sqrt{d}}\|\sqrt{\rho}(\alpha_i-1)\boldsymbol{x}_2 + \eta\sqrt{\rho}\boldsymbol{g}_2\| \overset{(p)}{\to} \sqrt{(\rho(\alpha_i-1)^2+\rho\eta^2)\psi\Big(1-\frac{1}{k}\Big)}\,,$$

where we used Assumption 2.3, by which $n/d \to \psi$ and $m = n/k$.

We also have $\frac{1}{\sqrt{d}}\|\boldsymbol{h}\| \overset{(p)}{\to} 1$.

We therefore arrive at the following deterministic optimization problem

$$\min_{\eta\ge 0,\alpha_i}\max_{\beta_1,\beta_2\ge 0}\Bigg(\beta_1\sqrt{(\alpha_i^2+\eta^2)\frac{\psi}{k}} + \beta_2\sqrt{(\rho(\alpha_i-1)^2+\rho\eta^2)\psi\Big(1-\frac{1}{k}\Big)}$$
$$+\sqrt{\beta_1^2+\rho\beta_2^2} - \frac{1}{2}\beta_1^2 - \frac{1}{2}\beta_2^2\Bigg)\,, \tag{29}$$

where we made the change of variables $2\beta_1/\sqrt{d} \to \beta_1$ and $2\beta_2/\sqrt{d} \to \beta_2$.

By writing the stationary conditions for the above optimization, and simplifying the resulting system of equations by solving for β_1, β_2, and substituting for them in the other two equations, we arrive at the following two equations for α_i and η:

$$\begin{cases}\rho + \frac{\psi}{k(1-\alpha_*)} - 1 = \frac{\psi}{k\alpha_*}\rho(k-1)\\ \eta_*^2 + \frac{k\alpha_*^2}{k-1} + \frac{\eta_*^2\alpha_*^2}{(1-\alpha_*)^2(k-1)} = \frac{\psi}{k}\frac{\eta_*^2}{(1-\alpha_*)^2}\,.\end{cases}$$

As the final step, recall that by definition $\eta := \|\boldsymbol{\alpha}_{\sim i}\|$ and therefore, $\|\boldsymbol{\alpha}_i\|^2 \overset{(p)}{\to} \alpha_*^2 + \eta_*^2$. As we see it is independent of the index i and therefore,

$$\frac{1}{d}\|\boldsymbol{M\Lambda X}\|^2 = \frac{1}{d}\sum_{i=1}^d \|\boldsymbol{M\Lambda x}_i\|^2 = \frac{1}{d}\sum_{i=1}^d \alpha_i^2 \overset{(p)}{\to} \alpha_*^2 + \eta_*^2\,.$$

This completes the proof.

D.2 Proof of Lemma B.2

Recall that $\boldsymbol{M} = (\boldsymbol{X}^\top\boldsymbol{E}\boldsymbol{X})^{-1}\boldsymbol{X}^\top$. Consider the following optimization problem:

$$\boldsymbol{\alpha} = \arg\min_{\boldsymbol{\alpha}\in\mathbb{R}^d}\frac{1}{d}\|\boldsymbol{E}^{1/2}\boldsymbol{X}\boldsymbol{\alpha} - \boldsymbol{E}^{-1/2}\boldsymbol{U}\boldsymbol{a}\|_2^2\,. \tag{30}$$

The solution to the above optimization problem has a closed-form solution given by $\boldsymbol{\alpha} = (\boldsymbol{X}^\top\boldsymbol{E}\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{U}\boldsymbol{a} = \boldsymbol{M}\boldsymbol{U}\boldsymbol{a}$. So we are interested in characterizing the norm of the optimal solution to the above optimization problem.

Similar to the proof of Lemma B.1, we use the framework of CGMT to characterize $\|\boldsymbol{\alpha}\|$ in the asymptotic regime described in Assumption 2.3.

Using the identity $\|\boldsymbol{x}\|/2 = \max_{\boldsymbol{v}}(\boldsymbol{v}^\top\boldsymbol{x} - \|\boldsymbol{v}\|^2/2)$, we rewrite the above optimization as:

$$\min_{\boldsymbol{\alpha}\in\mathbb{R}^d}\frac{1}{d}\|\boldsymbol{E}^{1/2}\boldsymbol{X}\boldsymbol{\alpha} - \boldsymbol{E}^{-1/2}\boldsymbol{U}\boldsymbol{a}\|_2^2$$
$$= \min_{\boldsymbol{\alpha}\in\mathbb{R}^d}\max_{\boldsymbol{v}\in\mathbb{R}^n}\frac{2}{d}\left(\boldsymbol{v}^\top\boldsymbol{E}^{1/2}\boldsymbol{X}\boldsymbol{\alpha} - \boldsymbol{v}^\top\boldsymbol{E}^{-1/2}\boldsymbol{U}\boldsymbol{a} - \frac{1}{2}\|\boldsymbol{v}\|_2^2\right)\,, \tag{31}$$

By using Convex-Gaussian-Minimax-Theorem (Thrampoulidis et al., 2015, Theorem 3), the above optimization is equivalent to the following auxiliary optimization problem:

$$\min_{\boldsymbol{\alpha}\in\mathbb{R}^d}\max_{\boldsymbol{v}\in\mathbb{R}^n}\frac{2}{d}\left(\|\boldsymbol{\alpha}\|\boldsymbol{v}^\top\boldsymbol{E}^{1/2}\boldsymbol{g} + \|\boldsymbol{E}^{1/2}\boldsymbol{v}\|\boldsymbol{h}^\top\boldsymbol{\alpha} - \boldsymbol{v}^\top\boldsymbol{E}^{-1/2}\boldsymbol{U}\boldsymbol{a} - \frac{1}{2}\|\boldsymbol{v}\|_2^2\right)\,, \tag{32}$$

with $\boldsymbol{g} \sim N(0, \boldsymbol{I}_n)$ and $\boldsymbol{h} \sim N(0, \boldsymbol{I}_d)$ independent Gaussian vectors.

We also recall that $\boldsymbol{S}^\top \boldsymbol{S} = \boldsymbol{U}\boldsymbol{U}^\top$ and so

$$\boldsymbol{E} := \rho \boldsymbol{I} + (1-\rho)\boldsymbol{S}^\top \boldsymbol{S} = \boldsymbol{U}\boldsymbol{U}^\top + \rho \boldsymbol{U}_\perp \boldsymbol{U}_\perp^\top ,$$

with $\boldsymbol{U}_\perp \in \mathbb{R}^{n\times(n-m)}$ denotes the orthogonal matrix, whose column space is orthogonal to the column space of $\boldsymbol{U}$. We decompose $\boldsymbol{v}$ to its component in the column space of $\boldsymbol{U}$ and $\boldsymbol{U}_\perp$ as

$$\boldsymbol{v} = \boldsymbol{U}\boldsymbol{v}_1 + \boldsymbol{U}_\perp \boldsymbol{v}_2, \quad \|\boldsymbol{v}\|^2 = \|\boldsymbol{v}_1\|^2 + \|\boldsymbol{v}_2\|^2 .$$

Therefore, $\boldsymbol{E}^{1/2}\boldsymbol{v} = \boldsymbol{U}\boldsymbol{v}_1 + \sqrt{\rho}\boldsymbol{U}_\perp \boldsymbol{v}_2$ and so the above optimization (32) can be written as

$$\min_{\boldsymbol{\alpha}\in\mathbb{R}^d} \max_{\boldsymbol{v}_1,\boldsymbol{v}_2\in\mathbb{R}^n} \frac{2}{d}\Big(\|\boldsymbol{\alpha}\|\boldsymbol{v}_1^\top \boldsymbol{U}^\top \boldsymbol{g} + \sqrt{\rho}\|\boldsymbol{\alpha}\|\boldsymbol{v}_2^\top \boldsymbol{U}_\perp^\top \boldsymbol{g} + \sqrt{\|\boldsymbol{v}_1\|^2 + \rho\|\boldsymbol{v}_2\|^2}\boldsymbol{h}^\top \boldsymbol{\alpha} - \boldsymbol{v}_1^\top \boldsymbol{a} - \frac{1}{2}\|\boldsymbol{v}_1\|^2 - \frac{1}{2}\|\boldsymbol{v}_2\|^2 \Big) . \tag{33}$$

We next introduce the following change of variables:

$$\boldsymbol{g}_1 := \boldsymbol{U}^\top \boldsymbol{g} \sim N(0, \boldsymbol{I}_m),$$
$$\boldsymbol{g}_2 := \boldsymbol{U}_\perp^\top \boldsymbol{g} \sim N(0, \boldsymbol{I}_{n-m}).$$

Rewriting the optimization in terms of $\boldsymbol{g}_1$ and $\boldsymbol{g}_2$ we get

$$\min_{\boldsymbol{\alpha}\in\mathbb{R}^d} \max_{\boldsymbol{v}_1,\boldsymbol{v}_2\in\mathbb{R}^n} \frac{2}{d}\Big(\|\boldsymbol{\alpha}\|\boldsymbol{v}_1^\top \boldsymbol{g}_1 + \sqrt{\rho}\|\boldsymbol{\alpha}\|\boldsymbol{v}_2^\top \boldsymbol{g}_2 + \sqrt{\|\boldsymbol{v}_1\|^2 + \rho\|\boldsymbol{v}_2\|^2}\boldsymbol{h}^\top \boldsymbol{\alpha} - \boldsymbol{v}_1^\top \boldsymbol{a} - \frac{1}{2}\|\boldsymbol{v}_1\|^2 - \frac{1}{2}\|\boldsymbol{v}_2\|^2 \Big) . \tag{34}$$

We next do the maximization on $\boldsymbol{v}_1$ and $\boldsymbol{v}_2$ by first fixing the norms to $\beta_1 := \|\boldsymbol{v}_1\|$ and $\beta_2 := \|\boldsymbol{v}_2\|$ and maximize over the directions and then maximize over β_1, β_2. This gives us

$$\min_{\boldsymbol{\alpha}\in\mathbb{R}^d} \max_{\beta_1,\beta_2\ge 0} \frac{2}{d}\Big(\beta_1 \| \|\boldsymbol{\alpha}\|\boldsymbol{g}_1 - \boldsymbol{a}\| + \beta_2\sqrt{\rho}\|\boldsymbol{\alpha}\|\|\boldsymbol{g}_2\| + \sqrt{\beta_1^2 + \rho\beta_2^2}\boldsymbol{h}^\top\boldsymbol{\alpha} - \frac{\beta_1^2+\beta_2^2}{2} \Big) . \tag{35}$$

For minimization over $\boldsymbol{\alpha}$, we first fix its norm to $\eta := \|\boldsymbol{\alpha}\|$ and optimize over its direction, and then over η:

$$\min_{\eta\ge 0} \max_{\beta_1,\beta_2\ge 0} \frac{2}{d}\Big(\beta_1 \|\eta\boldsymbol{g}_1 - \boldsymbol{a}\| + \beta_2\sqrt{\rho}\eta\|\boldsymbol{g}_2\| - \eta\sqrt{\beta_1^2 + \rho\beta_2^2}\ \|\boldsymbol{h}\| - \frac{\beta_1^2+\beta_2^2}{2} \Big) . \tag{36}$$

The next step in the CGMT framework is to compute the pointwise limit of the objective function. By concentration of Lipschitz functions of Gaussian vectors we have

$$\frac{1}{\sqrt{d}}\|\eta\boldsymbol{g}_1 - \boldsymbol{a}\| \xrightarrow{(p)} \sqrt{\frac{\|\boldsymbol{a}\|^2}{d} + \eta^2\frac{\psi}{k}} ,$$
$$\frac{1}{\sqrt{d}}\|\boldsymbol{g}_2\| \xrightarrow{(p)} \sqrt{\psi\Big(1-\frac{1}{k}\Big)},$$
$$\frac{1}{\sqrt{d}}\|\boldsymbol{h}\| \xrightarrow{(p)} 1 ,$$

where we used Assumption 2.3 by which $n/d \to \psi$, and Assumption 2.4 by which $m = n/k$. Using these limits in (36), we arrive at the following deterministic optimization problem:

$$\min_{\eta\ge 0} \max_{\beta_1,\beta_2\ge 0} \beta_1\sqrt{\frac{\|\boldsymbol{a}\|^2}{d} + \eta^2\frac{\psi}{k}} + \beta_2\sqrt{\rho}\eta\sqrt{\psi\Big(1-\frac{1}{k}\Big)} - \eta\sqrt{\beta_1^2+\rho\beta_2^2} \ - \frac{\beta_1^2+\beta_2^2}{2} , \tag{37}$$

where we applied the change of variables $2\beta_1/\sqrt{d} \to \beta_1$ and $2\beta_2/\sqrt{d} \to \beta_2$.

In order to find the optimal solution we solve the stationary conditions. By setting derivative with respect to η to zero we obtain

$$\frac{\beta_1\eta\frac{\psi}{k}}{\sqrt{\frac{\|\boldsymbol{a}\|^2}{d} + \eta^2\frac{\psi}{k}}} + \eta\sqrt{\rho\psi\Big(1-\frac{1}{k}\Big)} - \sqrt{\beta_1^2+\rho\beta_2^2} = 0 . \tag{38}$$

In addition by setting the derivative with respect to β_1 and β_2 to zero, we obtain

$$\begin{aligned} \sqrt{\frac{\|\boldsymbol{a}\|^2}{d} + \eta^2 \frac{\psi}{k}} &= \Big(\frac{\eta}{\sqrt{\beta_1^2 + \rho\beta_2^2}} + 1\Big)\beta_1\,, \\ \eta\sqrt{\rho\psi\Big(1 - \frac{1}{k}\Big)} &= \Big(\frac{\rho\eta}{\sqrt{\beta_1^2 + \rho\beta_2^2}} + 1\Big)\beta_2\,. \end{aligned} \tag{39}$$

By substituting for β_1 and β_2 from (39) into (38) we get

$$\frac{\eta\frac{\psi}{k}}{\eta + c} - 1 + \frac{\rho\eta\psi(1 - \frac{1}{k})}{\rho\eta + c} = 0\,, \tag{40}$$

where $c = \sqrt{\beta_1^2 + \rho\beta_2^2}$.

Also by substituting for β_1 and β_2 from (39) into the definition $c = \sqrt{\beta_1^2 + \rho\beta_2^2}$, we have

$$\frac{\frac{\|\boldsymbol{a}\|^2}{d} + \eta^2\frac{\psi}{k}}{(\eta + c)^2} + \frac{\rho^2\eta^2\psi(1 - \frac{1}{k})}{(\rho\eta + c)^2} = 1\,. \tag{41}$$

We next make the change of variable: $c = \eta u$, and rewriting equations (40 and 41) as follows:

$$\begin{cases} \frac{\psi}{1+u} + \frac{\rho\psi(k-1)}{\rho+u} & = k\,, \\ \frac{\frac{k\|\boldsymbol{a}\|^2}{d\eta^2} + \psi}{(1+u)^2} + \frac{\rho^2\psi(k-1)}{(\rho+u)^2} & = k\,. \end{cases}$$

Defining $v := \frac{k\|\boldsymbol{a}\|^2}{\psi d\eta^2}$ we get the system of equations given in the lemma statement.

As the final step, recall that as we discussed at the beginning of the proof, $\boldsymbol{\alpha}_* = \boldsymbol{MUa}$. Therefore,

$$\frac{n}{\|\boldsymbol{a}\|^2}\|\boldsymbol{MUa}\|_2^2 = \frac{n}{\|\boldsymbol{a}\|^2}\|\boldsymbol{\alpha}_*\|_2^2 \overset{(p)}{\to} \frac{n}{\|\boldsymbol{a}\|^2}\eta_*^2 = \frac{k}{v_*}\,,$$

which completes the proof.