A. Parameters of the Posterior Distribution

Recall that

$$p_n(\boldsymbol{\theta}_n \mid \mathcal{D}) \propto p^{\mathrm{G}}(\boldsymbol{Y}_n \mid \boldsymbol{X}_n\boldsymbol{\theta}_n, \sigma^2\boldsymbol{I})p^{\mathrm{G}}(\boldsymbol{\theta}_n \mid \boldsymbol{\alpha}, \boldsymbol{\Sigma}) \,.$$

We first give a handy proposition for the posterior distribution over the parameters $\boldsymbol{\theta}_n$.

Proposition A.1.

$$\ln p_n(\boldsymbol{\theta}_n) = -\frac{1}{2}(\boldsymbol{\theta}_n - \boldsymbol{\mu})\boldsymbol{\mathcal{T}}^{-1}(\boldsymbol{\theta}_n - \boldsymbol{\mu}) + \mathrm{const}(\boldsymbol{\theta}_n)$$

where covariance is

$$\boldsymbol{\mathcal{T}} = \left(\boldsymbol{\Sigma}^{-1} + \frac{1}{\sigma^2}\boldsymbol{X}_n^\top\boldsymbol{X}_n\right)^{-1}$$

and mean is

$$\boldsymbol{\mu} = \boldsymbol{\mathcal{T}}\left(\boldsymbol{\Sigma}^{-1}\boldsymbol{\alpha} + \frac{1}{\sigma^2}\boldsymbol{X}_n^\top\boldsymbol{Y}_n\right) \,.$$

Proof. The log-likelihood is the following chain of identities:

$$\begin{aligned}
\ln p_n(\boldsymbol{\theta}_n) &= -\frac{\sum_{j=1}^{m_n}(\boldsymbol{x}_{n,j}^\top\boldsymbol{\theta}_n - Y_{n,j})^2}{2\sigma^2} - \frac{1}{2}(\boldsymbol{\theta}_n - \boldsymbol{\alpha})^\top\boldsymbol{\Sigma}^{-1}(\boldsymbol{\theta}_n - \boldsymbol{\alpha}) + \mathrm{const}(\boldsymbol{\theta}_n) \\
&= -\frac{\sum_{j=1}^{m_n}(\boldsymbol{\theta}_n^\top\boldsymbol{x}_{n,j}\boldsymbol{x}_{n,j}^\top\boldsymbol{\theta}_n - 2Y_{n,j}\boldsymbol{x}_{n,j}^\top\boldsymbol{\theta}_n)}{2\sigma^2} \\
&= -\frac{1}{2}(\boldsymbol{\theta}_n^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\theta}_n - 2\boldsymbol{\alpha}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\theta}_n) + \mathrm{const}(\boldsymbol{\theta}_n) \\
&= -\frac{1}{2}\left(\boldsymbol{\theta}_n^\top\boldsymbol{\mathcal{T}}^{-1}\boldsymbol{\theta}_n - 2\left(\boldsymbol{\Sigma}^{-1}\boldsymbol{\alpha} + \frac{1}{\sigma^2}\boldsymbol{X}_n^\top\boldsymbol{Y}_n\right)^\top\boldsymbol{\theta}_n\right) + \mathrm{const}(\boldsymbol{\theta}_n) \\
&= -\frac{1}{2}(\boldsymbol{\theta}_n - \boldsymbol{\mu})\boldsymbol{\mathcal{T}}^{-1}(\boldsymbol{\theta}_n - \boldsymbol{\mu}) + \mathrm{const}(\boldsymbol{\theta}_n) \,.
\end{aligned}$$

□

First note that the first consequence of Proposition A.1 is a MLE for $\boldsymbol{\theta}_n$,

$$\hat{\boldsymbol{\theta}}_n^{\mathrm{MLE}} = \boldsymbol{\mathcal{T}}\left(\boldsymbol{\Sigma}^{-1}\boldsymbol{\alpha} + \frac{1}{\sigma^2}\boldsymbol{X}_n^\top\boldsymbol{Y}_n\right) \,.$$

The second consequence is the following corollary which is obtained by taking $Y = \boldsymbol{\theta}_n^\top\boldsymbol{x} + \varepsilon$ and simply observing that the mean of a $p_n(\boldsymbol{\theta}_n \mid \mathcal{D})$ is $\boldsymbol{\mu}$, and so $\mathbb{E}[Y \mid \mathcal{D}] = \boldsymbol{x}^\top\boldsymbol{\mu}$ while the variance is $\mathbb{V}[Y \mid \mathcal{D}] = \mathbb{E}[(\boldsymbol{x}^\top\boldsymbol{\theta}_n + \varepsilon)^2 \mid \mathcal{D}] - \mathbb{E}[(\boldsymbol{x}^\top\boldsymbol{\theta}_n + \varepsilon) \mid \mathcal{D}]^2 = \boldsymbol{x}^\top\boldsymbol{\mathcal{T}}\boldsymbol{x} + \sigma^2$.

Proposition 4.6 (restated). *Let $Y = \boldsymbol{\theta}_n^\top\boldsymbol{x} + \varepsilon$ for $\varepsilon \sim \mathcal{N}(0, \sigma^2)$ and some $\boldsymbol{x} \in \mathbb{R}^d$. Then,*

$$\mathbb{E}[Y \mid \mathcal{D}] = \boldsymbol{x}^\top\boldsymbol{\mathcal{T}}\left(\boldsymbol{\Sigma}^{-1}\boldsymbol{\alpha} + \frac{1}{\sigma^2}\boldsymbol{X}_n^\top\boldsymbol{Y}_n\right)$$

$$\textit{and} \quad \mathbb{V}[Y \mid \mathcal{D}] = \boldsymbol{x}^\top\boldsymbol{\mathcal{T}}\boldsymbol{x} + \sigma^2 \,.$$

B. Proof of the Lower Bounds

Our task reduces to establishing lower bounds on

$$\mathbb{E}\left[\left(\boldsymbol{x}^\top\boldsymbol{\mathcal{T}}\boldsymbol{\Sigma}^{-1}(\boldsymbol{\alpha} - \hat{\boldsymbol{\alpha}})\right)^2\right] \tag{10}$$

for any choice of estimator $\hat{\boldsymbol{\alpha}}$, which in combination with Lemma 4.5 will prove Theorem 4.1. In the next section we first prove a lower bound for any unbiased estimator relying on the Cramér-Rao inequality. In what follows, in Appendix B.2, we will show a general bound in Lemma B.4 valid for any estimator (possibly biased) using a *hypothesis testing* technique (see, e.g. (Lattimore & Szepesvári, 2018, Chap. 13)). Finally, in Lemma B.5 we prove a high-probability lower bound on Eq. (10).

B.1. Lower Bound for Unbiased Estimator $\hat{\boldsymbol{\alpha}}$

Theorem B.1 (Cramér-Rao inequality)**.** *Suppose that $\boldsymbol{\alpha} \in \mathbb{R}^d$ is an unknown deterministic parameter with a probability density function $f(x \mid \boldsymbol{\alpha})$ and that $\hat{\boldsymbol{\alpha}}$ is an unbiased estimator of $\boldsymbol{\alpha}$. Moreover assume that for all $i, j \in [d]$, $x : f(x \mid \boldsymbol{\alpha}) > 0$, $\frac{\partial^2}{\partial \alpha_i \partial \alpha_j} \ln f(x \mid \boldsymbol{\alpha})$ exists and is finite, and $\frac{\partial^2}{\partial \alpha_i \partial \alpha_j} \int \hat{\boldsymbol{\alpha}} f(x \mid \boldsymbol{\alpha}) \, \mathrm{d}x = \int \hat{\boldsymbol{\alpha}} \left(\frac{\partial^2}{\partial \alpha_i \partial \alpha_j} f(x \mid \boldsymbol{\alpha}) \right) \mathrm{d}x$.*
Then, for the Fisher information matrix defined as

$$\boldsymbol{F} = -\mathbb{E}\left[\nabla_{\boldsymbol{\alpha}} \ln f(X \mid \boldsymbol{\alpha}) \nabla_{\boldsymbol{\alpha}} \ln f(X \mid \boldsymbol{\alpha})^\top \right]$$

we have

$$\mathbb{E}\left[(\hat{\boldsymbol{\alpha}} - \mathbb{E}[\hat{\boldsymbol{\alpha}}])(\hat{\boldsymbol{\alpha}} - \mathbb{E}[\hat{\boldsymbol{\alpha}}])^\top \right] \succeq \boldsymbol{F}^{-1} \, .$$

Lemma B.2. *For any unbiased estimator $\hat{\boldsymbol{\alpha}}$ of $\boldsymbol{\alpha}$ in Eq. (2) we have*

$$\mathbb{E}\left[(\boldsymbol{x}^\top \boldsymbol{\mathcal{T}} \boldsymbol{\Sigma}^{-1} (\boldsymbol{\alpha} - \hat{\boldsymbol{\alpha}})^2) \right] \geq \boldsymbol{x}^\top \boldsymbol{\mathcal{T}} \boldsymbol{\Sigma}^{-1} (\boldsymbol{\Psi}^\top \boldsymbol{K} \boldsymbol{\Psi})^{-1} \boldsymbol{\Sigma}^{-1} \boldsymbol{\mathcal{T}} \boldsymbol{x}. \tag{11}$$

Proof. Recall that according to the equivalence (2) $\boldsymbol{Y} \sim \mathcal{N}(\boldsymbol{\Psi}\boldsymbol{\alpha}, \boldsymbol{K})$ and the unknown parameter is $\boldsymbol{\alpha}$. To compute the Fisher information matrix we first observe that

$$\nabla_{\boldsymbol{\alpha}} \ln p^{\mathrm{G}}\left(\boldsymbol{Y}; \boldsymbol{\Psi}\boldsymbol{\alpha}, \boldsymbol{K}\right) = \boldsymbol{\Psi}^\top \boldsymbol{K}^{-1} (\boldsymbol{Y} - \boldsymbol{\Psi}\boldsymbol{\alpha})$$

and so

$$\begin{aligned}
\boldsymbol{F} &= \mathbb{E}\left[\nabla_{\boldsymbol{\alpha}} \ln p^{\mathrm{G}}\left(\boldsymbol{Y}; \boldsymbol{\Psi}\boldsymbol{\alpha}, \boldsymbol{K}\right) \nabla_{\boldsymbol{\alpha}} \ln p^{\mathrm{G}}\left(\boldsymbol{Y}; \boldsymbol{\Psi}\boldsymbol{\alpha}, \boldsymbol{K}\right)^\top \right] \\
&= \boldsymbol{\Psi}^\top \boldsymbol{K}^{-1} \mathbb{E}\left[(\boldsymbol{Y} - \boldsymbol{\Psi}\boldsymbol{\alpha})(\boldsymbol{Y} - \boldsymbol{\Psi}\boldsymbol{\alpha})^\top \right] \boldsymbol{K}^{-1} \boldsymbol{\Psi} \\
&= \boldsymbol{\Psi}^\top \boldsymbol{K}^{-1} \boldsymbol{\Psi}.
\end{aligned}$$

Thus, by Theorem B.1 we have

$$\mathbb{E}\left[(\boldsymbol{\alpha} - \hat{\boldsymbol{\alpha}})(\boldsymbol{\alpha} - \hat{\boldsymbol{\alpha}})^\top \right] \succeq (\boldsymbol{\Psi}^\top \boldsymbol{K}^{-1} \boldsymbol{\Psi})^{-1} \, .$$

Finally, left-multiplying by $\boldsymbol{x}^\top \boldsymbol{\mathcal{T}} \boldsymbol{\Sigma}^{-1}$ and right-multiplying the above by $\boldsymbol{\Sigma}^{-1} \boldsymbol{\mathcal{T}} \boldsymbol{x}$ gives us the statement. □

B.2. Lower Bound for Any Estimator $\hat{\boldsymbol{\alpha}}$

The proof of is based on the following lemma.

Lemma B.3 (Bretagnolle & Huber 1979)**.** *Let P and Q be probability measures on the same measurable space $(\Omega, \mathcal{F})$, and let $A \in \mathcal{F}$ be an arbitrary event. Then,*

$$P(A) + Q(A^c) \geq \frac{1}{2} \exp(-\mathrm{D}_{\mathrm{KL}}(P, Q)), \tag{12}$$

where $\mathrm{D}_{\mathrm{KL}}(P, Q) = \int_\Omega \ln\left(P(\omega)/Q(\omega)\right) \mathrm{d}P(\omega)$ denotes Kullback-Leibler divergence between P and Q and $A^c = \Omega \setminus A$ is the complement of A.

Lemma B.4. *For any estimator $\hat{\boldsymbol{\alpha}}$ of $\boldsymbol{\alpha}$ in Eq. (2) we have*

$$\mathbb{E}\left[\left(\boldsymbol{x}^\top \boldsymbol{\mathcal{T}} \boldsymbol{\Sigma}^{-1} (\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}) \right)^2 \right] \geq \frac{\boldsymbol{x}^\top \boldsymbol{M} \boldsymbol{x}}{16\sqrt{e}} \, .$$

Proof. Throughout the proof let $\boldsymbol{q} = \boldsymbol{\Sigma}^{-1} \boldsymbol{\mathcal{T}} \boldsymbol{x}$. Consider two meta-learning problems with target distributions $\mathbb{P}$ and $\mathbb{Q}$ characterized by two means: $\boldsymbol{\alpha}_{\mathbb{P}} = \boldsymbol{0}$ and $\boldsymbol{\alpha}_{\mathbb{Q}} = \Delta (\boldsymbol{\Psi}^\top \boldsymbol{K}^{-1} \boldsymbol{\Psi})^{-1} \boldsymbol{q}$ where $\Delta > 0$ is a free parameter to be tuned later on. Thus, according to our established equivalence (2), in these two cases targets are generated by respective models $\mathbb{P} = \mathcal{N}(\boldsymbol{0}, \boldsymbol{K})$ and $\mathbb{Q} = \mathcal{N}(\Delta \boldsymbol{\Psi} (\boldsymbol{\Psi}^\top \boldsymbol{K}^{-1} \boldsymbol{\Psi})^{-1} \boldsymbol{q}, \boldsymbol{K})$.
Recall our abbreviation $\boldsymbol{M} = \boldsymbol{\mathcal{T}} \boldsymbol{\Sigma}^{-1} (\boldsymbol{\Psi}^\top \boldsymbol{K}^{-1} \boldsymbol{\Psi})^{-1} \boldsymbol{\Sigma}^{-1} \boldsymbol{\mathcal{T}}$. Markov's inequality gives

$$\begin{aligned}
\mathbb{E}_{\mathbb{P}}\left[(\hat{\boldsymbol{\alpha}}^\top \boldsymbol{q} - \boldsymbol{\alpha}_{\mathbb{P}}^\top \boldsymbol{q})^2 \right] &= \mathbb{E}_{\mathbb{P}}\left[(\hat{\boldsymbol{\alpha}}^\top \boldsymbol{q})^2 \right] \geq \frac{\Delta^2}{4} \left(\boldsymbol{x}^\top \boldsymbol{M} \boldsymbol{x} \right)^2 \mathbb{P}\left(|\hat{\boldsymbol{\alpha}}^\top \boldsymbol{q}| \geq \frac{\Delta}{2} \boldsymbol{x}^\top \boldsymbol{M} \boldsymbol{x} \right) , \qquad \text{while} \\
\mathbb{E}_{\mathbb{Q}}\left[(\hat{\boldsymbol{\alpha}}^\top \boldsymbol{q} - \boldsymbol{\alpha}_{\mathbb{Q}}^\top \boldsymbol{q})^2 \right] &\geq \frac{\Delta^2}{4} \left(\boldsymbol{x}^\top \boldsymbol{M} \boldsymbol{x} \right)^2 \mathbb{Q}\left(|\boldsymbol{\alpha}_{\mathbb{Q}}^\top \boldsymbol{q} - \hat{\boldsymbol{\alpha}}^\top \boldsymbol{q}| \geq \frac{\Delta}{2} \boldsymbol{x}^\top \boldsymbol{M} \boldsymbol{x} \right) \\
&\geq \frac{\Delta^2}{4} \left(\boldsymbol{x}^\top \boldsymbol{M} \boldsymbol{x} \right)^2 \mathbb{Q}\left(|\hat{\boldsymbol{\alpha}}^\top \boldsymbol{q}| < \frac{\Delta}{2} \boldsymbol{x}^\top \boldsymbol{M} \boldsymbol{x} \right) ,
\end{aligned}$$

where the last inequality comes using the fact that $|a-b| \geq |a| - |b|$ for $a, b \in \mathbb{R}$ and observing that $\boldsymbol{\alpha}_{\mathbb{Q}}^\top \boldsymbol{q} = \boldsymbol{x}^\top \boldsymbol{M}\boldsymbol{x}$. Summing both inequalities above and applying Lemma B.3 we get

$$\begin{aligned}\mathbb{E}_{\mathbb{P}}\left[(\hat{\boldsymbol{\alpha}}^\top \boldsymbol{q} - \boldsymbol{\alpha}_{\mathbb{P}}^\top \boldsymbol{q})^2\right] + \mathbb{E}_{\mathbb{Q}}\left[(\hat{\boldsymbol{\alpha}}^\top \boldsymbol{q} - \boldsymbol{\alpha}_{\mathbb{Q}}^\top \boldsymbol{q})^2\right] &\geq \frac{\Delta^2}{8}\left(\boldsymbol{x}^\top \boldsymbol{M}\boldsymbol{x}\right)^2 \cdot \exp\left(-\mathrm{D}_{\mathrm{KL}}(\mathbb{P}, \mathbb{Q})\right) \\ &\overset{(a)}{=} \frac{\Delta^2}{8}\left(\boldsymbol{x}^\top \boldsymbol{M}\boldsymbol{x}\right)^2 \cdot \exp\left(-\frac{\Delta^2}{2}\boldsymbol{x}^\top \boldsymbol{M}\boldsymbol{x}\right),\end{aligned}$$

where step (a) follows from KL-divergence between multivariate Gaussians with the same covariance matrix. Now, using a basic fact that $2\max\{a, b\} \geq a + b$, we get that for any measure $\mathbb{P}$ given by parameter $\boldsymbol{\alpha}$ we have

$$\mathbb{E}\left[(\hat{\boldsymbol{\alpha}}^\top \boldsymbol{q} - \boldsymbol{\alpha}^\top \boldsymbol{q})^2\right] \geq \frac{\Delta^2}{16}\left(\boldsymbol{x}^\top \boldsymbol{M}\boldsymbol{x}\right)^2 \cdot \exp\left(-\frac{\Delta^2}{2}\boldsymbol{x}^\top \boldsymbol{M}\boldsymbol{x}\right).$$

The statement then follows by choosing $\Delta^2 = (\boldsymbol{x}^\top \boldsymbol{M}\boldsymbol{x})^{-1}$. □

Now we prove a high-probability version of the just given inequality.

Lemma B.5. *For any estimator $\hat{\boldsymbol{\alpha}}$ of $\boldsymbol{\alpha}$ in Eq. (2) and any $\delta \in (0, 1)$ we have*

$$\mathbb{P}\left(\left(\boldsymbol{x}^\top \boldsymbol{\mathcal{T}}\boldsymbol{\Sigma}^{-1}(\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha})\right)^2 \geq \ln\left(\frac{1}{4}\cdot\frac{1}{1-\delta}\right)\boldsymbol{x}^\top \boldsymbol{M}\boldsymbol{x}\right) \geq 1 - \delta\,.$$

Proof. The proof is very similar to the proof of Lemma B.4 except we will not apply Markov's inequality and focus directly on giving a lower bound the deviation probabilities rather than expectations. Thus, similarly as before introduce mean parameters $\boldsymbol{\alpha}_{\mathbb{P}} = \boldsymbol{0}$ and $\boldsymbol{\alpha}_{\mathbb{Q}} = \Delta(\boldsymbol{\Psi}^\top \boldsymbol{K}^{-1}\boldsymbol{\Psi})^{-1}\boldsymbol{q}/\left(\boldsymbol{x}^\top \boldsymbol{M}\boldsymbol{x}\right)$ and their associated probability measures $\mathbb{P} = \mathcal{N}(\boldsymbol{0}, \boldsymbol{K})$ and $\mathbb{Q} = \mathcal{N}\left(\frac{\Delta\boldsymbol{\Psi}(\boldsymbol{\Psi}^\top \boldsymbol{K}^{-1}\boldsymbol{\Psi})^{-1}\boldsymbol{q}}{\boldsymbol{x}^\top \boldsymbol{M}\boldsymbol{x}}, \boldsymbol{K}\right)$.

Note that

$$\begin{aligned}\mathbb{P}\left(|\hat{\boldsymbol{\alpha}}^\top \boldsymbol{q}| \geq \frac{\Delta}{2}\right) &= \mathbb{P}\left(|\boldsymbol{\alpha}_{\mathbb{P}}^\top \boldsymbol{q} - \hat{\boldsymbol{\alpha}}^\top \boldsymbol{q}| \geq \frac{\Delta}{2}\right), \\ \mathbb{Q}\left(|\boldsymbol{\alpha}_{\mathbb{Q}}^\top \boldsymbol{q} - \hat{\boldsymbol{\alpha}}^\top \boldsymbol{q}| \geq \frac{\Delta}{2}\right) &\geq \mathbb{Q}\left(|\hat{\boldsymbol{\alpha}}^\top \boldsymbol{q}| < \frac{\Delta}{2}\right)\end{aligned}$$

and so by using Lemma B.3 we obtain an exponential tail bound

$$\mathbb{P}\left(|\boldsymbol{\alpha}_{\mathbb{P}}^\top \boldsymbol{q} - \hat{\boldsymbol{\alpha}}^\top \boldsymbol{q}| \geq \frac{\Delta}{2}\right) + \mathbb{Q}\left(|\boldsymbol{\alpha}_{\mathbb{Q}}^\top \boldsymbol{q} - \hat{\boldsymbol{\alpha}}^\top \boldsymbol{q}| \geq \frac{\Delta}{2}\right) \geq \exp(-\mathrm{D}_{\mathrm{KL}}(\mathbb{P}\,||\,\mathbb{Q})) = \frac{1}{2}\exp\left(-\frac{\Delta^2}{\boldsymbol{x}^\top \boldsymbol{M}\boldsymbol{x}}\right).$$

Setting the r.h.s. in the above to $2(1-\delta)$ where δ is an error probability, and solving for Δ gives us tuning

$$\Delta^2 = 2\ln\left(\frac{1}{4}\cdot\frac{1}{1-\delta}\right)\boldsymbol{x}^\top \boldsymbol{M}\boldsymbol{x}\,.$$

Thus, we get

$$\mathbb{P}\left(\left(\boldsymbol{\alpha}_{\mathbb{P}}^\top \boldsymbol{q} - \hat{\boldsymbol{\alpha}}^\top \boldsymbol{q}\right)^2 \geq \ln\left(\frac{1}{4}\cdot\frac{1}{1-\delta}\right)\boldsymbol{x}^\top \boldsymbol{M}\boldsymbol{x}\right) + \mathbb{Q}\left(|\boldsymbol{\alpha}_{\mathbb{Q}}^\top \boldsymbol{q} - \hat{\boldsymbol{\alpha}}^\top \boldsymbol{q}| \geq \ln\left(\frac{1}{4}\cdot\frac{1}{1-\delta}\right)\boldsymbol{x}^\top \boldsymbol{M}\boldsymbol{x}\right) \geq 2(1-\delta)$$

and using the fact that $2\max(a, b) \geq a + b$ we get that for any probability measure $\mathbb{P}$ given by parameter $\boldsymbol{\alpha}$ we have

$$\mathbb{P}\left(\left(\boldsymbol{\alpha}_{\mathbb{P}}^\top \boldsymbol{q} - \hat{\boldsymbol{\alpha}}^\top \boldsymbol{q}\right)^2 \geq \ln\left(\frac{1}{4}\cdot\frac{1}{1-\delta}\right)\boldsymbol{x}^\top \boldsymbol{M}\boldsymbol{x}\right) \geq 1 - \delta\,.$$

□

C. Proof of the Upper Bounds

Theorem 4.2 (restated). *For the estimator $\hat{\boldsymbol{\theta}}_n(\hat{\boldsymbol{\alpha}}^{\mathrm{MLE}})$ and for any $\boldsymbol{x} \in \mathbb{R}^d$ we have*

$$\mathbb{E}\left[\mathcal{L}(\boldsymbol{x})\right] = \boldsymbol{x}^\top \boldsymbol{M}\boldsymbol{x} + \boldsymbol{x}^\top \boldsymbol{\mathcal{T}}\boldsymbol{x} + \sigma^2.$$

Moreover for the same estimator, with probability at least $1-\delta, \delta \in (0,1)$ we have

$$\mathcal{L}(\boldsymbol{x}) \le 2\ln\left(\frac{2}{\delta}\right)\boldsymbol{x}^\top \boldsymbol{M}\boldsymbol{x} + \boldsymbol{x}^\top \boldsymbol{\mathcal{T}}\boldsymbol{x} + \sigma^2.$$

Proof. Recall that

$$\hat{\boldsymbol{\alpha}}^{\mathrm{MLE}} = (\boldsymbol{\Psi}^\top \boldsymbol{K}^{-1}\boldsymbol{\Psi})^{-1}\boldsymbol{\Psi}^\top \boldsymbol{K}^{-1}\boldsymbol{Y}\ .$$

The first result follows from Lemma 4.5 where we have to give an identity for

$$\mathbb{E}\left[\left(\boldsymbol{x}^\top \boldsymbol{\mathcal{T}}\boldsymbol{\Sigma}^{-1}(\boldsymbol{\alpha} - \hat{\boldsymbol{\alpha}}^{\mathrm{MLE}})\right)^2\right] \tag{13}$$

and the missing piece is a covariance of the estimator $\hat{\boldsymbol{\alpha}}^{\mathrm{MLE}}$

$$\begin{aligned}
&\mathbb{E}\left[(\boldsymbol{\alpha} - \hat{\boldsymbol{\alpha}}^{\mathrm{MLE}})(\boldsymbol{\alpha} - \hat{\boldsymbol{\alpha}}^{\mathrm{MLE}})^\top\right] \\
&= (\boldsymbol{\Psi}^\top \boldsymbol{K}^{-1}\boldsymbol{\Psi})^{-1}\boldsymbol{\Psi}^\top \boldsymbol{K}^{-1}\mathrm{Cov}(\boldsymbol{Y},\boldsymbol{Y})\boldsymbol{K}^{-1}\boldsymbol{\Psi}(\boldsymbol{\Psi}^\top \boldsymbol{K}^{-1}\boldsymbol{\Psi})^{-1} \\
&= (\boldsymbol{\Psi}^\top \boldsymbol{K}^{-1}\boldsymbol{\Psi})^{-1}\ .
\end{aligned} \tag{14}$$

To prove the second result we have to give a high probability upper bound on Eq. (13).
Let $\boldsymbol{q} = \boldsymbol{\Sigma}^{-1}\boldsymbol{\mathcal{T}}\boldsymbol{x}$ and observe that $\boldsymbol{q}^\top \hat{\boldsymbol{\alpha}}^{\mathrm{MLE}}$ is Gaussian (since $\boldsymbol{Y}$ is composed of Gaussian entries) with mean $\boldsymbol{q}^\top \boldsymbol{\alpha}$ by equivalence (2), and covariance $(\boldsymbol{\Psi}^\top \boldsymbol{K}^{-1}\boldsymbol{\Psi})^{-1}$ by Eq. (14). Then, by Gaussian concentration for any error probability $\delta \in (0,1)$ we have

$$\mathbb{P}\left((\boldsymbol{q}^\top \boldsymbol{\alpha} - \boldsymbol{q}^\top \hat{\boldsymbol{\alpha}}^{\mathrm{MLE}})^2 \ge \sqrt{2\boldsymbol{q}^\top (\boldsymbol{\Psi}^\top \boldsymbol{K}^{-1}\boldsymbol{\Psi})^{-1}\boldsymbol{q}\ln\left(\frac{2}{\delta}\right)}\right) \le \delta$$

which completes the proof. □

D. Derivation of EM Steps

Recall that our goal is to solve

$$\max_{\mathcal{E}'} \int \ln\left(p(\boldsymbol{\vartheta}, \mathcal{D} \,|\, \mathcal{E}')\right) \mathrm{d}p(\boldsymbol{\vartheta} \,|\, \mathcal{D}, \hat{\mathcal{E}}_t)\ .$$

First, we will focus on the integral. The chain rule readily gives

$$\ln p(\boldsymbol{\Theta}, \mathcal{D} \,|\, \mathcal{E}') = \ln p(\boldsymbol{\Theta} \,|\, \mathcal{D}, \mathcal{E}') + \ln p(\boldsymbol{\Theta} \,|\, \mathcal{E}')\ .$$

Using the same reasoning and notation as in the proof of Proposition A.1 we get

$$\begin{aligned}
\int \ln p(\boldsymbol{\vartheta} \,|\, \mathcal{D}, \mathcal{E}')\, \mathrm{d}p(\boldsymbol{\vartheta} \,|\, \mathcal{D}, \hat{\mathcal{E}}_t) &= \sum_{i=1}^{n}\sum_{j=1}^{m_i}\left(\frac{1}{2}\ln\left(\frac{1}{\sigma^2}\right) - \frac{1}{2\sigma^2}\int (Y_{i,j} - \boldsymbol{x}_{i,j}^\top \boldsymbol{\vartheta}_i)^2\, \mathrm{d}p(\boldsymbol{\vartheta}_i \,|\, \mathcal{D}, \hat{\mathcal{E}}_t)\right) + \mathrm{const}(\mathcal{E}') \\
&= \sum_{i=1}^{n}\sum_{j=1}^{m_i}\left(\frac{1}{2}\ln\left(\frac{1}{\sigma^2}\right) - \frac{1}{2\sigma^2}(Y_{i,j} - \boldsymbol{x}_{i,j}^\top \boldsymbol{\mu}_i)^2 - \boldsymbol{x}_{i,j}^\top \boldsymbol{\mathcal{T}}_i \boldsymbol{x}_{i,j}\right) + \mathrm{const}(\mathcal{E}')\,,
\end{aligned}$$

using the fact that $\int (Y_{i,j} - \boldsymbol{x}_{i,j}^\top \boldsymbol{\vartheta}_i)^2\, \mathrm{d}p(\boldsymbol{\vartheta}_i \,|\, \mathcal{D}, \hat{\mathcal{E}}_t) = (Y_{i,j} - \boldsymbol{x}_{i,j}^\top \boldsymbol{\mu}_i)^2 + \boldsymbol{x}_{i,j}^\top \boldsymbol{\mathcal{T}}_i \boldsymbol{x}_{i,j}$ where we took $\boldsymbol{\theta}_i \sim \mathcal{N}(\boldsymbol{\mu}_i, \boldsymbol{\mathcal{T}}_i)$ according to Proposition A.1.
Now we compute the expected log-likelihood of the vector of task parameters:

$$\int \ln p(\boldsymbol{\vartheta} \,|\, \mathcal{E}')\, \mathrm{d}p(\boldsymbol{\vartheta} \,|\, \mathcal{D}, \hat{\mathcal{E}}_t) = \frac{n}{2}\ln\det \boldsymbol{\Sigma}^{-1} - \frac{1}{2}\sum_{i=1}^{n}\int (\boldsymbol{\vartheta}_i - \boldsymbol{\alpha})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{\vartheta}_i - \boldsymbol{\alpha})\, \mathrm{d}p(\boldsymbol{\vartheta}_i \,|\, \mathcal{D}, \hat{\mathcal{E}}_t) + \mathrm{const}(\mathcal{E}')\ .$$

M-step for σ^2. Now, note that since the likelihood of the vector of task variables $\boldsymbol{\Theta}$ does not depend on the parameter σ^2 we can solve for σ^2 based on the first order condition of the problem above. Differentiating the above equation with respect to σ^{-2} (and ignoring the constant) gives

$$\sum_{i=1}^{n}\sum_{j=1}^{m_i}\left(\sigma^2 - \left((Y_{i,j} - \boldsymbol{x}_{i,j}^\top\boldsymbol{\mu}_i)^2 + \boldsymbol{x}_{i,j}^\top\boldsymbol{\mathcal{T}}_i\boldsymbol{x}_{i,j}\right)\right). \tag{15}$$

while setting the derivative to zero gives

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{m_i}\sum_{j=1}^{m_i}\left((Y_{i,j} - \boldsymbol{x}_{i,j}^\top\boldsymbol{\mu}_i)^2 + \boldsymbol{x}_{i,j}^\top\boldsymbol{\mathcal{T}}_i\boldsymbol{x}_{i,j}\right). \tag{16}$$

M-step for $\boldsymbol{\alpha}$. Differentiating the objective w.r.t. $\boldsymbol{\alpha}$ (and ignoring the constant) gives $\sum_{i=1}^{n}\boldsymbol{\Sigma}^{-1}(\mathbb{E}[\boldsymbol{\theta}_i] - \boldsymbol{\alpha})$ from which we get

$$\boldsymbol{\alpha} = \sum_{i=1}^{n}\boldsymbol{\mu}_i\,. \tag{17}$$

M-step for $\boldsymbol{\Sigma}$. Differentiating the expected log-likelihood of the vector of task parameters with respect to $\boldsymbol{A} = \boldsymbol{\Sigma}^{-1}$ gives

$$\sum_{i=1}^{n}\operatorname{tr}(\boldsymbol{\Sigma}d\boldsymbol{A}) - \operatorname{tr}\int\left((\boldsymbol{\vartheta}_i - \boldsymbol{\alpha})(\boldsymbol{\vartheta}_i - \boldsymbol{\alpha})^\top d\boldsymbol{A}\right)\mathrm{d}p(\boldsymbol{\vartheta}_i \mid \mathcal{D}, \hat{\mathcal{E}}_t) \tag{18}$$

from which we get

$$\boldsymbol{\Sigma} = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[(\boldsymbol{\theta}_i - \boldsymbol{\alpha})(\boldsymbol{\theta}_i - \boldsymbol{\alpha})^\top]. \tag{19}$$

Finally, computing the expectation

$$\sum_{i=1}^{n}\left(\mathbb{E}[\boldsymbol{\theta}_i\boldsymbol{\theta}_i^\top] - 2\boldsymbol{\mu}_i\boldsymbol{\alpha}^\top + \boldsymbol{\alpha}\boldsymbol{\alpha}^\top\right) = \sum_{i=1}^{n}\left((\boldsymbol{\mu}_i - \boldsymbol{\alpha})(\boldsymbol{\mu}_i - \boldsymbol{\alpha})^\top + \boldsymbol{\mathcal{T}}_i\right) \tag{20}$$

shows the update for $\boldsymbol{\Sigma}$.

E. Selecting λ in Biased Regression

The parameter λ is selected via random search in the following way. For each of the 50 samples of λ from log-uniform distribution on interval $[0; 100]$ we perform the following procedure to estimate the risk $\hat{L}$. Firstly, we split the training tasks into $K = 10$ groups $\mathcal{S}_1, \ldots, \mathcal{S}_K$ of (approximately) equal size and compute the estimates $\hat{\boldsymbol{\alpha}}_k$ using the data $\mathcal{S}^{\setminus k}$ from all of the groups excluding the group k: $\mathcal{S}^{\setminus k} := \cup_{i\neq k}\mathcal{S}_i$. For each of the estimated values $\hat{\boldsymbol{\alpha}}_k$ we perform adaptation to and testing on the tasks in the group $\mathcal{S}_k$ using the given value of λ. We split the samples of each task data $D_i \in \mathcal{S}_k$ randomly into adaptation and test sets 10 times each time such that the size of adaptation set is close to the size of adaptation sets used with the actual test data. For each of the splits we compute an estimate of the parameter vector $\hat{\boldsymbol{\theta}}_{k,i,l}$ where k is the index of the group which was not used to estimate $\hat{\boldsymbol{\alpha}}_k$, i is the index of a task data $D_i \in \mathcal{S}_k$, l is the index of a random split of the samples in that task into adaptation and test sets. With this parameter vector and using the test set of the task $D_i \in \mathcal{S}_k$ we can also estimate the loss $\hat{L}_{k,i,l}$ after which all the loss values are averaged:

$$\hat{L} = \frac{1}{K}\sum_{k=1}^{K}\frac{1}{|\mathcal{S}^{\setminus k}|}\sum_{i:D_i\in\mathcal{S}^{\setminus k}}\frac{1}{10}\sum_{l=1}^{10}\hat{L}_{k,i,l}.$$

At the end we select the value of λ which lead to the smallest value of $\hat{L}$ using this cross-validation procedure.

F. Supplementary Statements

Proposition F.1. *For $\boldsymbol{M}$ (see Eq. (5)) we have*

$$\boldsymbol{M} = \sigma^4\cdot\left(\boldsymbol{\Sigma}\boldsymbol{X}_n^\top\boldsymbol{X}_n + \sigma^2\boldsymbol{I}\right)^{-1}\boldsymbol{A}^{-1}\left(\boldsymbol{\Sigma}\boldsymbol{X}_n^\top\boldsymbol{X}_n + \sigma^2\boldsymbol{I}\right)^{-1},$$

where we denote

$$\boldsymbol{A} = \sum_{i=1}^{n} \boldsymbol{X}_i^\top (\boldsymbol{X}_i \boldsymbol{\Sigma} \boldsymbol{X}_i^\top + \sigma^2 \boldsymbol{I})^{-1} \boldsymbol{X}_i \,.$$

Proof. Recall that

$$\boldsymbol{M} = \boldsymbol{\mathcal{T}} \boldsymbol{\Sigma}^{-1} \left(\boldsymbol{\Psi}^\top \boldsymbol{K}^{-1} \boldsymbol{\Psi} \right)^{-1} \boldsymbol{\Sigma}^{-1} \boldsymbol{\mathcal{T}}$$

and observe that

$$\boldsymbol{K}^{-1} = \begin{bmatrix} (\boldsymbol{X}_1 \boldsymbol{\Sigma} \boldsymbol{X}_1^\top + \sigma^2 \boldsymbol{I})^{-1} & \boldsymbol{0} & \dots & \boldsymbol{0} \\ \boldsymbol{0} & (\boldsymbol{X}_2 \boldsymbol{\Sigma} \boldsymbol{X}_2^\top + \sigma^2 \boldsymbol{I})^{-1} & \dots & \boldsymbol{0} \\ \vdots & & \ddots & \vdots \\ \boldsymbol{0} & \boldsymbol{0} & \dots & (\boldsymbol{X}_n \boldsymbol{\Sigma} \boldsymbol{X}_n^\top + \sigma^2 \boldsymbol{I})^{-1} \end{bmatrix},$$

which in turn implies

$$\boldsymbol{\Psi}^\top \boldsymbol{K}^{-1} \boldsymbol{\Psi} = \sum_{i=1}^{n} \boldsymbol{X}_i^\top (\boldsymbol{X}_i \boldsymbol{\Sigma} \boldsymbol{X}_i^\top + \sigma^2 \boldsymbol{I})^{-1} \boldsymbol{X}_i \,.$$

On the other hand,

$$\begin{aligned} \boldsymbol{\mathcal{T}} \boldsymbol{\Sigma}^{-1} &= \left(\boldsymbol{\Sigma}^{-1} + \frac{1}{\sigma^2} \boldsymbol{X}_n^\top \boldsymbol{X}_n \right)^{-1} \boldsymbol{\Sigma}^{-1} \\ &= \sigma^2 \left(\sigma^2 \boldsymbol{I} + \boldsymbol{\Sigma} \boldsymbol{X}_n^\top \boldsymbol{X}_n \right)^{-1} \,. \end{aligned}$$

Combining the above gives the statement. □

Lemma F.2. *In the following assume that $\boldsymbol{X}_i^\top \boldsymbol{X}_i = \frac{m_i}{d} \boldsymbol{I}$ for all i. Let $\lambda_j(\boldsymbol{\Sigma})$ be the jth eigenvalue of $\boldsymbol{\Sigma}$. Then,*

$$\lambda_j(\boldsymbol{M}) = \sigma^4 \cdot \frac{d^2}{\left(m_n \lambda_j(\boldsymbol{\Sigma}) + d\sigma^2\right)^2} \cdot \frac{HM\left(\lambda_j(\boldsymbol{\Sigma}) + \frac{d\sigma^2}{m_i}\right)_{i=1}^{n}}{n} \,,$$

where $HM(z_i)_{i=1}^n$ denotes the harmonic mean of sequence $(z_i)_{i=1}^n$. Moreover,

$$\lambda_j(\boldsymbol{\mathcal{T}}) = \frac{d\sigma^2 \lambda_j(\boldsymbol{\Sigma})}{d\sigma^2 + m_n \lambda_j(\boldsymbol{\Sigma})} \,.$$

Finally, the eigenvectors of $\boldsymbol{M}$ and $\boldsymbol{\mathcal{T}}$ coincide with the eigenvectors of $\boldsymbol{\Sigma}$.

Proof. We first characterize eigenvalues of matrix $\boldsymbol{M}$. By Proposition F.1,

$$\boldsymbol{M} = \sigma^4 \cdot \left(\boldsymbol{\Sigma} \boldsymbol{X}_n^\top \boldsymbol{X}_n + \sigma^2 \boldsymbol{I} \right)^{-1} \boldsymbol{A}^{-1} \left(\boldsymbol{\Sigma} \boldsymbol{X}_n^\top \boldsymbol{X}_n + \sigma^2 \boldsymbol{I} \right)^{-1} \,.$$

We start with $\boldsymbol{A}^{-1}$, and by the spectral theorem, $\boldsymbol{\Sigma} = \boldsymbol{U} \boldsymbol{\Lambda} \boldsymbol{U}^\top$ for some unitary $\boldsymbol{U}$ and diagonal $\boldsymbol{\Lambda}$:

$$\begin{aligned} \boldsymbol{A}^{-1} = \left(\sum_{i=1}^{n} \boldsymbol{X}_i^\top (\boldsymbol{X}_i \boldsymbol{\Sigma} \boldsymbol{X}_i^\top + \sigma^2 \boldsymbol{I})^{-1} \boldsymbol{X}_i \right)^{-1} &= \left(\sum_{i=1}^{n} (\boldsymbol{\Sigma} \boldsymbol{X}_i^\top \boldsymbol{X}_i + \sigma^2 \boldsymbol{I})^{-1} \boldsymbol{X}_i^\top \boldsymbol{X}_i \right)^{-1} \\ &= \left(\sum_{i=1}^{n} \left(\boldsymbol{\Sigma} \cdot \frac{m_i}{d} + \sigma^2 \boldsymbol{I} \right)^{-1} \frac{m_i}{d} \right)^{-1} \\ &= \left(\sum_{i=1}^{n} \left(\boldsymbol{U} \boldsymbol{\Lambda} \boldsymbol{U}^\top \cdot \frac{m_i}{d} + \sigma^2 \boldsymbol{I} \right)^{-1} \frac{m_i}{d} \right)^{-1} \\ &= \boldsymbol{U} \left(\sum_{i=1}^{n} \left(\boldsymbol{\Lambda} + \frac{d\sigma^2}{m_i} \cdot \boldsymbol{I} \right)^{-1} \right)^{-1} \boldsymbol{U}^\top \,. \end{aligned}$$

Now,

$$\begin{aligned}\left(\boldsymbol{\Sigma}\boldsymbol{X}_n^\top\boldsymbol{X}_n+\sigma^2\boldsymbol{I}\right)^{-1} &= \left(\boldsymbol{\Sigma}\cdot\frac{m_n}{d}+\sigma^2\boldsymbol{I}\right)^{-1}\\ &= \left(\boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^\top\cdot\frac{m_n}{d}+\sigma^2\boldsymbol{I}\right)^{-1}\\ &= \boldsymbol{U}\left(\boldsymbol{\Lambda}\cdot\frac{m_n}{d}+\sigma^2\boldsymbol{I}\right)^{-1}\boldsymbol{U}^\top.\end{aligned}$$

Thus,

$$\boldsymbol{M}=\boldsymbol{U}\left(\left(\boldsymbol{\Lambda}\cdot\frac{m_n}{d}+\sigma^2\boldsymbol{I}\right)^2\sum_{i=1}^n\left(\boldsymbol{\Lambda}+\frac{d\sigma^2}{m_i}\right)^{-1}\right)^{-1}\boldsymbol{U}^\top.$$

and moreover the jth eigenvalue of $\boldsymbol{M}$ is

$$\begin{aligned}\lambda_j(\boldsymbol{M}) &= \frac{1}{\left(\frac{m_n}{d}\lambda_j(\boldsymbol{\Sigma})+\sigma^2\right)^2}\cdot\frac{1}{\sum_{i=1}^n\frac{1}{\lambda_j(\boldsymbol{\Sigma})+\frac{d\sigma^2}{m_i}}}\\ &= \frac{1}{\left(\frac{m_n}{d}\lambda_j(\boldsymbol{\Sigma})+\sigma^2\right)^2}\cdot\frac{\mathrm{HM}\left(\lambda_j(\boldsymbol{\Sigma})+\frac{d\sigma^2}{m_i}\right)_{i=1}^n}{n}.\end{aligned}$$

where recall that $\mathrm{HM}(z_i)_{i=1}^n$ denotes the harmonic mean of sequence $(z_i)_{i=1}^n$.
Using the same arguments as above

$$\begin{aligned}\boldsymbol{\mathcal{T}} &= \left(\boldsymbol{\Sigma}^{-1}+\frac{1}{\sigma^2}\boldsymbol{X}_n^\top\boldsymbol{X}_n\right)^{-1}\\ &= \left(\boldsymbol{U}\boldsymbol{\Lambda}^{-1}\boldsymbol{U}^\top+\frac{m_n}{d\sigma^2}\right)^{-1}\end{aligned}$$

and so

$$\lambda_j(\boldsymbol{\mathcal{T}})=\frac{1}{\frac{1}{\lambda_j(\boldsymbol{\Sigma})}+\frac{m_n}{d\sigma^2}}=\frac{d\sigma^2\lambda_j(\boldsymbol{\Sigma})}{d\sigma^2+m_n\lambda_j(\boldsymbol{\Sigma})}.$$

Finally, in both cases of $\boldsymbol{M}$ and $\boldsymbol{\mathcal{T}}$ we observe that their eigenvectors are eigenvectors of $\boldsymbol{\Sigma}$. □

Corollaries 4.3 and 4.4 (restated). *In the following assume that $\boldsymbol{X}_i^\top\boldsymbol{X}_i=\frac{m_i}{d}\boldsymbol{I}$ for all i. For $\boldsymbol{\Sigma}=\tau^2\boldsymbol{I}$, any $\boldsymbol{x}\in\mathbb{R}^d$, and any $c>0$,*

$$c\boldsymbol{x}^\top\boldsymbol{M}\boldsymbol{x}+\boldsymbol{x}^\top\boldsymbol{\mathcal{T}}\boldsymbol{x}+\sigma^2=c\cdot\frac{H_{\tau^2}}{n}\cdot\frac{d^2\sigma^4}{\left(\tau^2m_n+d\sigma^2\right)^2}\cdot\|\boldsymbol{x}\|^2+\frac{d\sigma^2\tau^2}{\tau^2m_n+d\sigma^2}\cdot\|\boldsymbol{x}\|^2+\sigma^2,$$

where H_{τ^2} is a harmonic mean of the sequence $\left(\tau^2+\frac{d\sigma^2}{m_i}\right)_{i=1}^n$.

Moreover, let $\boldsymbol{\Sigma}$ be a PSD matrix of rank $s\le d$ with eigenvalues $\lambda_1\ge\ldots\ge\lambda_s>0$. Then for any $\boldsymbol{x}\in\mathbb{R}^d$ and any $c>0$,

$$c\boldsymbol{x}^\top\boldsymbol{M}\boldsymbol{x}+\boldsymbol{x}^\top\boldsymbol{\mathcal{T}}\boldsymbol{x}+\sigma^2\ge c\cdot\frac{H_{\lambda_s}}{n}\cdot\frac{d^2\sigma^4}{\left(\lambda_1m_n+d\sigma^2\right)^2}\cdot\|\boldsymbol{x}\|^2_{\boldsymbol{P}_s^\top\boldsymbol{P}_s}+\frac{d\sigma^2\lambda_s}{\lambda_sm_n+d\sigma^2}\cdot\|\boldsymbol{x}\|^2_{\boldsymbol{P}_s^\top\boldsymbol{P}_s}+\sigma^2$$

where $\boldsymbol{P}_s=[\boldsymbol{u}_1,\ldots,\boldsymbol{u}_s]^\top$ and $(\boldsymbol{u}_j)_{j=1}^s$ are eigenvectors of $\boldsymbol{\Sigma}$.

Proof. Recalling that by Proposition F.1,

$$\boldsymbol{M}=\sigma^4\cdot\left(\boldsymbol{\Sigma}\boldsymbol{X}_n^\top\boldsymbol{X}_n+\sigma^2\boldsymbol{I}\right)^{-1}\boldsymbol{A}^{-1}\left(\boldsymbol{\Sigma}\boldsymbol{X}_n^\top\boldsymbol{X}_n+\sigma^2\boldsymbol{I}\right)^{-1}.$$

and using Lemma F.2 with $\boldsymbol{\Sigma}=\tau^2\boldsymbol{I}$ we get the first result.

Now we turn to the low-rank case. We start by considering a PSD matrix $\boldsymbol{\Sigma}_\varepsilon$ with s eigenvalues $\lambda_1 \geq \ldots \geq \lambda_s > 0$ and remaining $d-s$ are $\varepsilon > 0$. Denote also by $\boldsymbol{M}_\varepsilon$, $\boldsymbol{\mathcal{T}}_\varepsilon$ matrices w.r.t. $\boldsymbol{\Sigma}_\varepsilon$. The idea is to lower bound $\boldsymbol{x}^\top \boldsymbol{M}_\varepsilon \boldsymbol{x}$ and $\boldsymbol{x}^\top \boldsymbol{\mathcal{T}}_\varepsilon \boldsymbol{x}$ and then analyze a limiting behavior as $\varepsilon \to 0$.

By Lemma F.2, $\boldsymbol{M}_\varepsilon$, $\boldsymbol{\mathcal{T}}_\varepsilon$, and $\boldsymbol{\Sigma}_\varepsilon$ share the same eigenvectors $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_s$, and so

$$
\begin{aligned}
c\boldsymbol{x}^\top \boldsymbol{M}_\varepsilon \boldsymbol{x} + \boldsymbol{x}^\top \boldsymbol{\mathcal{T}}_\varepsilon \boldsymbol{x} &= c\sum_{j=1}^{d} \left(\boldsymbol{u}_j^\top \boldsymbol{x}\right)^2 \lambda_j(\boldsymbol{M}_\varepsilon) + \sum_{j=1}^{d} \left(\boldsymbol{u}_j^\top \boldsymbol{x}\right)^2 \lambda_j(\boldsymbol{\mathcal{T}}_\varepsilon) \\
&= c \cdot \sum_{j=1}^{s} \frac{H_{\lambda_j}}{n} \cdot \frac{\sigma^4}{\left(\lambda_j \frac{m_n}{d} + \sigma^2\right)^2} \left(\boldsymbol{u}_j^\top \boldsymbol{x}\right)^2 + c \cdot \underbrace{\frac{H_\varepsilon}{n} \cdot \frac{\sigma^4}{\left(\varepsilon \frac{m_n}{d} + \sigma^2\right)^2}}_{(a)} \left(\sum_{j=s+1}^{d} \left(\boldsymbol{u}_j^\top \boldsymbol{x}\right)^2\right) \\
&\quad + \sum_{j=1}^{s} \frac{\sigma^2 \lambda_j}{\lambda_j \frac{m_n}{d} + \sigma^2} \left(\boldsymbol{u}_j^\top \boldsymbol{x}\right)^2 + \frac{\sigma^2 \varepsilon}{\varepsilon \frac{m_n}{d} + \sigma^2} \left(\sum_{j=s+1}^{d} \left(\boldsymbol{u}_j^\top \boldsymbol{x}\right)^2\right) .
\end{aligned}
$$

Now,

$$
\begin{aligned}
&\lim_{\varepsilon \to 0} \left(c\boldsymbol{x}^\top \boldsymbol{M}_\varepsilon \boldsymbol{x} + \boldsymbol{x}^\top \boldsymbol{\mathcal{T}}_\varepsilon \boldsymbol{x}\right) \\
&= c \cdot \sum_{j=1}^{s} \frac{H_{\lambda_j}}{n} \cdot \frac{\sigma^4}{\left(\lambda_j \frac{m_n}{d} + \sigma^2\right)^2} \left(\boldsymbol{u}_j^\top \boldsymbol{x}\right)^2 + \frac{d\sigma^2}{M} \sum_{j=s+1}^{d} \left(\boldsymbol{u}_j^\top \boldsymbol{x}\right)^2 + \sum_{j=1}^{s} \frac{\sigma^2 \lambda_j}{\lambda_j \frac{m_n}{d} + \sigma^2} \left(\boldsymbol{u}_j^\top \boldsymbol{x}\right)^2 \\
&\geq c \cdot \frac{H_{\lambda_s}}{n} \cdot \frac{\sigma^4}{\left(\lambda_1 \frac{m_n}{d} + \sigma^2\right)^2} \sum_{j=1}^{s} \left(\boldsymbol{u}_j^\top \boldsymbol{x}\right)^2 + \frac{\sigma^2 \lambda_s}{\lambda_s \frac{m_n}{d} + \sigma^2} \sum_{j=1}^{s} \left(\boldsymbol{u}_j^\top \boldsymbol{x}\right)^2 ,
\end{aligned}
$$

where we note that the limit of term (a) is handled as

$$
\lim_{\varepsilon \to 0} \frac{1}{\sum_{i=1}^{n} \frac{1}{\varepsilon + \frac{d\sigma^2}{m_i}}} \cdot \frac{\sigma^4}{\left(\varepsilon \frac{m_n}{d} + \sigma^2\right)^2} = \frac{d\sigma^2}{M} \geq 0 .
$$

□

G. Further Experimental Details and Results

In this section we provide extra figures for our experimental results. Fig. 1 is complemented with Fig. 7, adding a second example in addition to the one shown in the previous figure.

For completeness, the pseudocode of MoM is given in Algorithm 2.

Algorithm 2 MoM Estimator for Learning Linear Features of (Tripuraneni et al., 2020)

Input: $\left((\boldsymbol{x}_{1,j}, y_{1,j})\right)_{j=1}^{m_1}, \ldots, \left((\boldsymbol{x}_{m_{n-1},j}, y_{m_{n-1},j})\right)_{j=1}^{m_{i-1}}$ — training examples from $n-1$ past tasks, s — problem rank.

$\boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^\top \leftarrow \text{SVD}\left(\frac{1}{M-m_n} \sum_{i=1}^{n-1} \sum_{j=1}^{m_n} y_{i,j}^2 \boldsymbol{x}_{i,j} \boldsymbol{x}_{i,j}^\top\right)$

$\hat{\boldsymbol{B}} \leftarrow [D_{1,1}\boldsymbol{u}_1, \ldots, D_{s,s}\boldsymbol{u}_s]$

return $\hat{\boldsymbol{B}}$

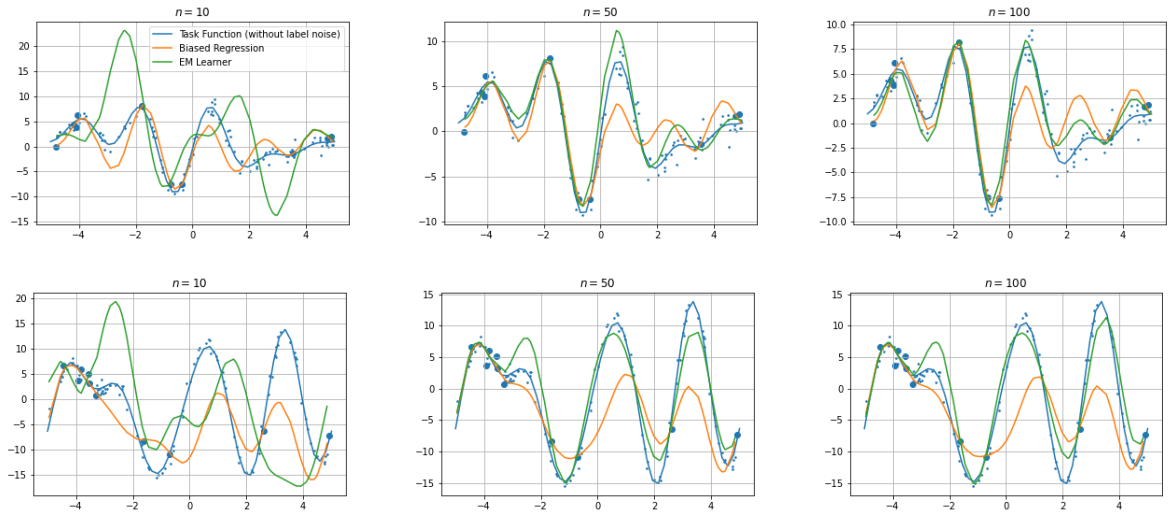

Figure 7. Two sets of examples of predictions on the synthetic, 'Fourier' meta-learning problem. Top and bottom rows correspond to different (random) instances; the top row in fact replicates Fig. 1. Training data is shown in bold, small dots show test data. We also show the predictions for two learners (at every input) and the target function. The column correspond to outputs obtained training on $n \in \{10, 50, 100\}$ tasks.