A. Appendix Overview

The Appendix is structured as follows

- Section B proves the result for linear representation learning with the tr-tr objective, Theorem 5.1, that shows that "bad" full rank solutions exist arbitrarily close to the optimal value of the tr-tr objective $\mathcal{L}^{\text{tr-tr}}_{\lambda,\text{rep}}$. The proof works by arguing that smaller values of λ are desirable for the tr-tr objective, and this can be simulated by effectively making the norm of the representation layer very high. Furthermore a higher rank representation is preferable over lower rank ones to fit the noise in the labels in the training data better.
- Section C proves the main result for linear representation learning with tr-val objective, Theorem 5.4, that proves that the optimal solutions to the tr-val objective $\mathcal{L}^{\text{tr-val}}_{\lambda,\text{rep}}(\boldsymbol{A}; (n_1, n_2))$ for most n_1 and σ will be *low-rank* representations that are also *expressive enough*. The result is also extended to solutions that are τ-optimal in the $\mathcal{L}^{\text{tr-val}}_{\lambda,\text{rep}}$ objective for a small enough τ. The crux of the proof for this result is in Theorem C.1 that provides a closed form expression for the tr-val objective that disentangles the expressivity and the low-rankness of the representation.
- Section D presents additional experimental details and results, including results for the MiniImageNet dataset.

B. More on Train-Train split

B.1. Proof of main result

Theorem 5.1. *For every $\lambda, n > 0$, for every $\tau > 0$, there exists a "bad" representation layer $\boldsymbol{A}_{bad} \in \mathbb{R}^{d\times d}$ that satisfies $\mathcal{L}^{tr\text{-}tr}_{\lambda,rep}(\boldsymbol{A}_{bad}; n) \le \inf_{\boldsymbol{A}\in\mathbb{R}^{d\times d}} \mathcal{L}^{tr\text{-}tr}_{\lambda,rep}(\boldsymbol{A}; n) + \tau$, but has the following lower bound on meta-testing loss*

$$\inf_{\bar{\lambda}>0} \mathcal{L}^{test}_{\bar{\lambda},rep}(\boldsymbol{A}_{bad}; \bar{n}_1) \ge \sigma^2 + \min\left\{1 - \frac{\bar{n}_1}{d(1+\sigma^2)}, \frac{d\sigma^2}{(\bar{n}_1 + d\sigma^2)}\right\}$$

Proof. For most of the proof, we will leave out the n in the expression for $\mathcal{L}^{\text{tr-tr}}_{\lambda,\text{rep}}$, i.e. we will denote the tr-tr loss as $\mathcal{L}^{\text{tr-tr}}_{\lambda,\text{rep}}(\boldsymbol{A})$. We prove this result by using Lemma 5.2 first, and later prove this lemma.

Lemma 5.2. *For every $\lambda > 0$ and $\boldsymbol{A} \in \mathbb{R}^{d\times d}$ with rank r,*

$$\mathcal{L}^{tr\text{-}tr}_{\lambda,rep}(\boldsymbol{A}; n) \ge \lim_{\kappa\to\infty} \mathcal{L}^{tr\text{-}tr}_{\lambda,rep}(\kappa\boldsymbol{A}; n) \ge \sigma^2 \frac{(n-r)_+}{n}$$
$$\& \quad \lim_{\kappa\to\infty} \mathcal{L}^{tr\text{-}tr}_{\lambda,rep}(\kappa\boldsymbol{I}_d; n) = \sigma^2 \frac{(n-d)_+}{n}$$

This tells us that for any matrix $\boldsymbol{A}$, $\mathcal{L}^{\text{tr-tr}}_{\lambda,\text{rep}}(\boldsymbol{A}) \ge \sigma^2 \frac{(n-\text{rank}(\boldsymbol{A}))_+}{n} \ge \sigma^2 \frac{(n-d)_+}{n}$. Also the lower bound of $\sigma^2 \frac{(n-d)_+}{n}$ can be achieved by $\kappa\boldsymbol{I}_d$ in the limit of $\kappa \to \infty$, thus $\sigma^2 \frac{(n-d)_+}{n} = \inf_{\boldsymbol{A}\in\mathbb{R}^{d\times d}} \mathcal{L}^{\text{tr-tr}}_{\lambda,\text{rep}}(\boldsymbol{A})$. Also since $\lim_{\kappa\to\infty} \kappa\boldsymbol{I}_d = \sigma^2 \frac{(n-d)_+}{n} = \inf_{\boldsymbol{A}\in\mathbb{R}^{d\times d}} \mathcal{L}^{\text{tr-tr}}_{\lambda,\text{rep}}(\boldsymbol{A})$, for a large enough $\kappa = \kappa(\tau)$, $\mathcal{L}^{\text{tr-tr}}_{\lambda,\text{rep}}(\kappa(\tau)\boldsymbol{I}_d)$ can be made lesser than $\inf_{\boldsymbol{A}\in\mathbb{R}^{d\times d}} \mathcal{L}^{\text{tr-tr}}_{\lambda,\text{rep}}(\boldsymbol{A}) + \tau$. Thus we pick $\boldsymbol{A}_{\text{bad}} = \kappa(\tau)\boldsymbol{I}_d$. With this choice, the new task is essentially linear regression in d-dimension with isotropic Gaussians.

To show that $\boldsymbol{A}_{\text{bad}}$ is indeed bad, we will use the lower bound for ridge regression on isotropic Gaussian linear regression from Theorem 4.2(a) in Saunshi et al. (2020). They show that the excess risk for $\boldsymbol{I}_d$ (and thus $\kappa(\tau)\boldsymbol{I}_d$), regardless of the choice of regularizer $\bar{\lambda}$, for a new task $\rho_{\boldsymbol{v}}$ will be lower bounded by

$$\inf_{\bar{\lambda}>0} \mathop{\mathbb{E}}_{S\sim\rho_{\boldsymbol{v}}^{\bar{n}_1}} \left[\|\boldsymbol{A}\boldsymbol{w}_{\bar{\lambda}}(\boldsymbol{A}; S) - \boldsymbol{v}\|^2\right] \ge \begin{cases} \frac{d\|\boldsymbol{v}\|^2\sigma^2}{\bar{n}_1\|\boldsymbol{v}\|+\sigma^2 d} & \text{if } \bar{n}_1 \ge d \\ \frac{\bar{n}_1}{d}\frac{\|\boldsymbol{v}\|^2\sigma^2}{\|\boldsymbol{v}\|+\sigma^2} + \frac{d-\bar{n}_1}{d}\|\boldsymbol{v}\|^2 & \text{if } \bar{n}_1 < d \end{cases}$$

Their proof can be easily modified to replace $\|\boldsymbol{v}\|^2$ with $\mathop{\mathbb{E}}_{\boldsymbol{v}\sim\bar{\mu}} \|\boldsymbol{v}\|^2$. The lower bound can be simplified for the the $\bar{n}_1 < d$ case to $\frac{\bar{n}_1}{d}\frac{\|\boldsymbol{v}\|^2\sigma^2}{\|\boldsymbol{v}\|+\sigma^2} + \frac{d-\bar{n}_1}{d}\|\boldsymbol{v}\|^2 = \|\boldsymbol{v}\|^2 - \frac{\bar{n}_1}{d}\frac{\|\boldsymbol{v}\|^2}{\|\boldsymbol{v}\|^2+\sigma^2}$. Plugging in $\|\boldsymbol{v}\| = 1$ completes the proof. □

We now prove the lemma

Lemma 5.2. *For every $\lambda > 0$ and $\boldsymbol{A} \in \mathbb{R}^{d\times d}$ with rank r,*

$$\mathcal{L}^{\textit{tr-tr}}_{\lambda,\textit{rep}}(\boldsymbol{A}; n) \geq \lim_{\kappa\to\infty} \mathcal{L}^{\textit{tr-tr}}_{\lambda,\textit{rep}}(\kappa\boldsymbol{A}; n) \geq \sigma^2 \frac{(n-r)_+}{n}$$
$$\& \quad \lim_{\kappa\to\infty} \mathcal{L}^{\textit{tr-tr}}_{\lambda,\textit{rep}}(\kappa\boldsymbol{I}_d; n) = \sigma^2 \frac{(n-d)_+}{n}$$

Proof. We first prove that having $\lambda = 0$ will lead to the smallest loss $\mathcal{L}^{\text{tr-tr}}_{\lambda,\text{rep}}(\boldsymbol{A})$ for every $\boldsymbol{A}$. We then observe that $\lambda = 0$ can be simulated by increasing the norm of $\boldsymbol{A}$. These claims mathematically mean that, (a) $\mathcal{L}^{\text{tr-tr}}_{\lambda,\text{rep}}(\boldsymbol{A}) \geq \mathcal{L}^{\text{tr-tr}}_{\lambda',\text{rep}}(\boldsymbol{A})$ whenever $\lambda \geq \lambda'$) and (b) $\mathcal{L}^{\text{tr-tr}}_{\lambda,\text{rep}}(\kappa\boldsymbol{A}) = \mathcal{L}^{\text{tr-tr}}_{\frac{\lambda}{\kappa^2},\text{rep}}(\boldsymbol{A})$. This will give us that $\lim_{\kappa\to\infty} \mathcal{L}^{\text{tr-tr}}_{\lambda,\text{rep}}(\kappa\boldsymbol{A}) = \lim_{\lambda\to 0} \mathcal{L}^{\text{tr-tr}}_{0,\text{rep}}(\kappa\boldsymbol{A}) \geq \mathcal{L}^{\text{tr-tr}}_{\lambda,\text{rep}}(\boldsymbol{A})$. Intuitively, $\mathcal{L}^{\text{tr-tr}}_{\lambda,\text{rep}}$ is trying to learn a linear classifier on top of data that is linear transformed by $\boldsymbol{A}$ with the goal of fitting the same data well.

Lemma B.1. *For any representation layer $\boldsymbol{A} \in \mathbb{R}^{d\times d}$ and $\lambda > 0$, we have the following*

$$\lim_{\kappa\to\infty} \mathcal{L}^{\textit{tr-tr}}_{\lambda,\textit{rep}}(\kappa\boldsymbol{A}) = \lim_{\lambda\to 0} \mathcal{L}^{\textit{tr-tr}}_{\lambda,\textit{rep}}(\boldsymbol{A}) \leq \mathcal{L}^{\textit{tr-tr}}_{\lambda,\textit{rep}}(\boldsymbol{A})$$

Fitting the data is better when there is less restriction on the norm of the classifier, which in this case means when λ is smaller. Furthermore, increasing the norm of the representation layer $\boldsymbol{A}$ effectively reduces the impact the regularizer will have. We first prove this lemma later, first we use it to prove Lemma 5.2 that shows that the loss for low rank matrices will be high.

Lemma B.1 already shows that $\mathcal{L}^{\text{tr-tr}}_{\lambda,\text{rep}}(\boldsymbol{A}; n) \geq \lim_{\kappa\to\infty} \mathcal{L}^{\text{tr-tr}}_{\lambda,\text{rep}}(\kappa\boldsymbol{A}; n)$. Also using Lemma B.1, we can replace $\lim_{\kappa\to\infty} \mathcal{L}^{\text{tr-tr}}_{\lambda,\text{rep}}(\kappa\boldsymbol{A})$ with $\lim_{\lambda\to 0} \mathcal{L}^{\text{tr-tr}}_{\lambda,\text{rep}}(\boldsymbol{A})$. Using Equation (9) and from central limit theorem, we have

$$\mathcal{L}^{\text{tr-tr}}_{\lambda,\text{rep}}(\boldsymbol{A}) = \lim_{T\to\infty} \widehat{\mathcal{L}}^{\text{tr-tr}}(\boldsymbol{A}) = \mathop{\mathbb{E}}_{\rho_{\boldsymbol{v}}\sim\mu} \left[\mathop{\mathbb{E}}_{S\sim\rho_{\boldsymbol{v}}^n} \left[\frac{1}{n} \|\boldsymbol{X}\boldsymbol{A}\boldsymbol{w}_\lambda(\boldsymbol{A}; S) - \boldsymbol{Y}\|^2 \right] \right] \tag{14}$$
$$= \mathop{\mathbb{E}}_{\boldsymbol{v}\sim\mathcal{N}(\boldsymbol{0}_d, \boldsymbol{A}^*\boldsymbol{A}^{*\top})} \left[\mathop{\mathbb{E}}_{S\sim\rho_{\boldsymbol{v}}^n} \left[\frac{1}{n} \|\boldsymbol{X}\boldsymbol{A}\boldsymbol{w}_\lambda(\boldsymbol{A}; S) - \boldsymbol{Y}\|^2 \right] \right] \tag{15}$$

This is because $\widehat{\mathcal{L}}^{\text{tr-tr}}_{\lambda,\text{rep}}$ is an average loss for T train tasks, and the limit when $T \to \infty$ it converges to the expectation over the task distribution μ. We first observe that $S \sim \rho_{\boldsymbol{v}}$ is equivalent to sampling $\boldsymbol{X} \sim \mathcal{N}(\boldsymbol{0}_d, \boldsymbol{I}_d)$, $\boldsymbol{\eta} \sim \mathcal{N}(\boldsymbol{0}_n, \sigma^2 I_n)$ which gives us $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{v} + \boldsymbol{\eta}$, where $\boldsymbol{X} \in \mathbb{R}^{n\times d}, \boldsymbol{\eta} \in \mathbb{R}^n, \boldsymbol{Y} \in \mathbb{R}^n$. Using the definition of $\boldsymbol{w}_\lambda(\boldsymbol{A}; S)$ from Equation (8) the standard KKT condition for linear regression, we can write a closed form solution for

$$\boldsymbol{w}_\lambda(\boldsymbol{A}; S) = \mathop{\arg\min}_{\boldsymbol{w}\in\mathbb{R}^d} \frac{1}{n}\|\boldsymbol{X}\boldsymbol{A}\boldsymbol{w} - \boldsymbol{Y}\|^2 + \lambda\|\boldsymbol{w}\|^2 = \left(\frac{\boldsymbol{A}^\top\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{A}}{n} + \lambda\boldsymbol{I}_d \right)^{-1} \frac{\boldsymbol{A}^\top\boldsymbol{X}^\top}{n}\boldsymbol{Y} \tag{16}$$
$$\lim_{\lambda\to 0} \boldsymbol{w}_\lambda(\boldsymbol{A}; S) = \left(\frac{\boldsymbol{A}^\top\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{A}}{n} \right)^\dagger \frac{\boldsymbol{A}^\top\boldsymbol{X}^\top}{n}\boldsymbol{Y} \tag{17}$$

where the last step is folklore that the limit of ridge regression as regularization coefficient goes to 0 is the minimum ℓ_2-norm linear regression solution. Plugging this into Equation (20) and taking the limit

$$\lim_{\lambda\to 0} \mathcal{L}^{\text{tr-tr}}_{\lambda,\text{rep}}(\boldsymbol{A}) = \mathop{\mathbb{E}}_{\boldsymbol{v},S} \left[\frac{1}{n} \left\| \boldsymbol{X}\boldsymbol{A} \left(\lim_{\lambda\to 0} \boldsymbol{w}_\lambda(\boldsymbol{A}; S) \right) - \boldsymbol{Y} \right\|^2 \right]$$
$$= \mathop{\mathbb{E}}_{\boldsymbol{v},S} \left[\frac{1}{n} \left\| \left(\boldsymbol{X}\boldsymbol{A} \left(\frac{\boldsymbol{A}^\top\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{A}}{n} \right)^\dagger \frac{\boldsymbol{A}^\top\boldsymbol{X}^\top}{n} - I_n \right) \boldsymbol{Y} \right\|^2 \right]$$
$$= \frac{1}{n} \mathop{\mathbb{E}}_{\boldsymbol{v},S} \left[\left\| P^\perp_{\boldsymbol{X}\boldsymbol{A}}\boldsymbol{Y} \right\|^2 \right]$$

where for any matrix $\boldsymbol{B} \in \mathbb{R}^{n\times d}$, we denote $P_{\boldsymbol{B}} \in \mathbb{R}^{n\times n} = \boldsymbol{B}\boldsymbol{B}^\dagger$ to denote the projection matrix onto the span of columns of $\boldsymbol{B}$, and $P_{\boldsymbol{B}}^\perp = I_n - P_{\boldsymbol{B}}$ is the projection matrix onto the orthogona subspace. Note that if $\text{rank}(\boldsymbol{B}) = n$, then $P_{\boldsymbol{B}} = I_n$ and $P_{\boldsymbol{B}}^\perp = 0$. Thus the error incurred is the amount of the label $\boldsymbol{Y}$ that the representation $\boldsymbol{X}\boldsymbol{A}$ cannot predict linearly. We further decompose this

$$\begin{aligned}
\lim_{\lambda\to 0}\mathcal{L}^{\text{tr-tr}}_{\lambda,\text{rep}}(\boldsymbol{A}) &= \frac{1}{n}\mathop{\mathbb{E}}_{\boldsymbol{v},S}\left[\left\|P^\perp_{\boldsymbol{X}\boldsymbol{A}}\boldsymbol{Y}\right\|^2\right] = \frac{1}{n}\mathop{\mathbb{E}}_{\boldsymbol{v},S}\left[\left\|P^\perp_{\boldsymbol{X}\boldsymbol{A}}\boldsymbol{X}\boldsymbol{v} + P^\perp_{\boldsymbol{X}\boldsymbol{A}}\boldsymbol{\eta}\right\|^2\right] \\
&= \underbrace{\frac{1}{n}\mathop{\mathbb{E}}_{\boldsymbol{v},S}\left[\left\|P^\perp_{\boldsymbol{X}\boldsymbol{A}}\boldsymbol{X}\boldsymbol{v}\right\|^2\right]}_{\text{fitting signal}} + \underbrace{\frac{1}{n}\mathop{\mathbb{E}}_{\boldsymbol{v},S}\left[\left\|P^\perp_{\boldsymbol{X}\boldsymbol{A}}\boldsymbol{\eta}\right\|^2\right]}_{\text{fitting noise}} + \underbrace{\frac{2}{n}\mathop{\mathbb{E}}_{\boldsymbol{v},S}\left[\boldsymbol{\eta}^\top P^\perp_{\boldsymbol{X}\boldsymbol{A}}\boldsymbol{X}\boldsymbol{v}\right]}_{\text{cross term}} \qquad (18) \\
&= \frac{1}{n}\mathop{\mathbb{E}}_{S}\mathop{\mathbb{E}}_{\boldsymbol{v}\sim\mathcal{N}(\boldsymbol{0}_d,\boldsymbol{A}^*\boldsymbol{A}^{*\top})}\left[\left\|P^\perp_{\boldsymbol{X}\boldsymbol{A}}\boldsymbol{X}\boldsymbol{v}\right\|^2\right] + \mathop{\mathbb{E}}_{S}\frac{1}{n}\text{tr}\left(P^\perp_{\boldsymbol{X}\boldsymbol{A}}\mathop{\mathbb{E}}_{\boldsymbol{\eta}}\left[\boldsymbol{\eta}\boldsymbol{\eta}^\top\right]P^\perp_{\boldsymbol{X}\boldsymbol{A}}\right) \\
&= \underbrace{\frac{1}{n}\mathop{\mathbb{E}}_{S}\left[\left\|P^\perp_{\boldsymbol{X}\boldsymbol{A}}\boldsymbol{X}\boldsymbol{A}^*\right\|^2\right]}_{\alpha(\boldsymbol{A})} + \underbrace{\mathop{\mathbb{E}}_{S}\frac{\sigma^2}{n}\text{tr}\left(P^\perp_{\boldsymbol{X}\boldsymbol{A}}\right)}_{\beta(\boldsymbol{A})} \qquad (19)
\end{aligned}$$

We first note, due to independence of $\boldsymbol{\eta}$ and $\boldsymbol{X}$ that the cross term in Equation (18) is 0 in expectation. Using the distributions for $\boldsymbol{v} \sim \mathcal{N}(\boldsymbol{0}_d, \boldsymbol{A}^*\boldsymbol{A}^{*\top})$ and $\boldsymbol{\eta} \sim \mathcal{N}(\boldsymbol{0}_n, \sigma^2 I_n)$, we get the final expressions Firstly, we note that $\alpha(\boldsymbol{A}) \geq 0$ for every $\boldsymbol{A}$, and a sufficient condition for $\alpha(\boldsymbol{A}) = 0$ is that $\boldsymbol{A}^*$ lies in the span of $\boldsymbol{A}$, i.e. $P^\perp_{\boldsymbol{A}}\boldsymbol{A}^* = 0$. More importantly, it is clear that $\alpha(\boldsymbol{I}_d) = 0$. Next we look at the $\beta(\boldsymbol{A})$ term which is proportional to the trace of $P^\perp_{\boldsymbol{X}\boldsymbol{A}}$ which, for a projection matrix, is also equal to the rank of the matrix. Note that since the rank of $\boldsymbol{X}\boldsymbol{A} \in \mathbb{R}^{d\times n}$ is at most $\min\{n, d\}$. Thus we get

$$\beta(\boldsymbol{A}) = \sigma^2\frac{\text{rank}(P^\perp_{\boldsymbol{X}\boldsymbol{A}})}{n} = \sigma^2\frac{(n - \text{rank}(P^\perp_{\boldsymbol{X}\boldsymbol{A}}))}{n} \geq \sigma^2\frac{(n - \text{rank}(\boldsymbol{A}))_+}{n}$$

where $(x)_+ = \mathbb{1}_{x\geq 0}x$. This along with $\alpha(\boldsymbol{A}) \geq 0$ proves the first part of the result. For the second part $\alpha(\boldsymbol{I}_d) = 0$ along with noticing that $\text{tr}(P^\perp_{\boldsymbol{X}\boldsymbol{I}_d}) = n - \text{rank}(\boldsymbol{X}) = n - \min\{n, d\}$ since a Gaussian matrix is full rank with measure 1, thus giving us $\beta(\boldsymbol{I}_d) = \sigma^2\frac{(n-d)_+}{n}$ and completing the proof. □

Thus Lemma 5.2 shows that $\lim_{\kappa\to\infty}\mathcal{L}^{\text{tr-tr}}_{\lambda,\text{rep}}(\kappa\boldsymbol{I}_d) = \inf_{\boldsymbol{A}}\mathcal{L}^{\text{tr-tr}}_{\lambda,\text{rep}}(\boldsymbol{A})$ and so picking a large enough $\kappa(\tau)$ can always give us $\mathcal{L}^{\text{tr-tr}}_{\lambda,\text{rep}}(\kappa(\tau)\boldsymbol{I}_d) \leq \inf_{\boldsymbol{A}}\mathcal{L}^{\text{tr-tr}}_{\lambda,\text{rep}}(\boldsymbol{A}) + \tau$ which completes the proof of Theorem 5.1. The only thing left to prove is Lemma B.1 which we do below

Proof of Lemma B.1. Using Equation (16) we get the following expression for $\mathcal{L}^{\text{tr-tr}}_{\lambda,\text{rep}}$

$$\begin{aligned}
\mathcal{L}^{\text{tr-tr}}_{\lambda,\text{rep}}(\boldsymbol{A}) &= \mathop{\mathbb{E}}_{\boldsymbol{v},S}\left[\frac{1}{n}\left\|\boldsymbol{X}\boldsymbol{A}\boldsymbol{w}_\lambda(\boldsymbol{A};S) - \boldsymbol{Y}\right\|^2\right] \\
&= \mathop{\mathbb{E}}_{\boldsymbol{v},S}\left[\frac{1}{n}\left\|\left(\boldsymbol{X}\boldsymbol{A}\left(\frac{\boldsymbol{A}^\top\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{A}}{n} + \lambda\boldsymbol{I}_d\right)^{-1}\frac{\boldsymbol{A}^\top\boldsymbol{X}^\top}{n} - I_n\right)\boldsymbol{Y}\right\|^2\right] \qquad (20)
\end{aligned}$$

From this we get

$$\begin{aligned}
\mathcal{L}^{\text{tr-tr}}_{\lambda,\text{rep}}(\kappa\boldsymbol{A}) &= \mathop{\mathbb{E}}_{\boldsymbol{v},S}\left[\frac{1}{n}\left\|\left(\kappa\boldsymbol{X}\boldsymbol{A}\left(\frac{\kappa^2\boldsymbol{A}^\top\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{A}}{n} + \lambda\boldsymbol{I}_d\right)^{-1}\frac{\kappa\boldsymbol{A}^\top\boldsymbol{X}^\top}{n} - I_n\right)\boldsymbol{Y}\right\|^2\right] \\
&= \mathop{\mathbb{E}}_{\boldsymbol{v},S}\left[\frac{1}{n}\left\|\left(\boldsymbol{X}\boldsymbol{A}\left(\frac{\boldsymbol{A}^\top\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{A}}{n} + \frac{\lambda}{\kappa^2}\boldsymbol{I}_d\right)^{-1}\frac{\boldsymbol{A}^\top\boldsymbol{X}^\top}{n} - I_n\right)\boldsymbol{Y}\right\|^2\right] \\
&= \mathcal{L}^{\text{tr-tr}}_{\frac{\lambda}{\kappa^2},\text{rep}}(\boldsymbol{A})
\end{aligned}$$

Thus $\lim_{\kappa\to\infty}\mathcal{L}^{\text{tr-tr}}_{\lambda,\text{rep}}(\kappa\boldsymbol{A}) = \lim_{\kappa\to\infty}\mathcal{L}^{\text{tr-tr}}_{\frac{\lambda}{\kappa^2},\text{rep}}(\boldsymbol{A}) = \lim_{\lambda\to 0}\mathcal{L}^{\text{tr-tr}}_{\lambda,\text{rep}}(\boldsymbol{A})$. Note that we have used the fact that $\mathcal{L}^{\text{tr-tr}}_{\lambda,\text{rep}}$ is continuous in λ for $\lambda > 0$. To prove the remaining result, i.e. $\lim_{\lambda\to 0}\mathcal{L}^{\text{tr-tr}}_{\lambda,\text{rep}}(\boldsymbol{A}) \leq \mathcal{L}^{\text{tr-tr}}_{\lambda,\text{rep}}(\boldsymbol{A})$, we just need to prove that $\mathcal{L}^{\text{tr-tr}}_{\lambda,\text{rep}}(\boldsymbol{A})$ is a

increasing function of λ. Suppose $\frac{\boldsymbol{XA}}{\sqrt{n}} = \boldsymbol{U_X S_X V_X^\top}$ is the singular value decomposition. For any $\lambda' < \lambda$, we can rewrite $\mathcal{L}^{\text{tr-tr}}_{\lambda,\text{rep}}(\boldsymbol{A})$ from Equation (20) as follows

$$\begin{aligned}
\mathcal{L}^{\text{tr-tr}}_{\lambda,\text{rep}}(\boldsymbol{A}) &= \mathop{\mathbb{E}}_{\boldsymbol{v},S}\left[\frac{1}{n}\left\|\left(\boldsymbol{XA}\left(\frac{\boldsymbol{A}^\top\boldsymbol{X}^\top\boldsymbol{XA}}{n}+\lambda\boldsymbol{I}_d\right)^{-1}\frac{\boldsymbol{A}^\top\boldsymbol{X}^\top}{n}-I_n\right)\boldsymbol{Y}\right\|^2\right] \\
&= \mathop{\mathbb{E}}_{\boldsymbol{v},S}\left[\frac{1}{n}\left\|\left(\boldsymbol{V_X S_X U_X^\top}\left(\boldsymbol{U_X S_X V_X^\top V_X S_X U_X^\top}+\lambda\boldsymbol{I}_d\right)^{-1}\boldsymbol{U_X S_X V_X^\top}-I_n\right)\boldsymbol{Y}\right\|^2\right] \\
&= \mathop{\mathbb{E}}_{\boldsymbol{v},S}\left[\frac{1}{n}\left\|\boldsymbol{V_X}\left(\boldsymbol{S_X}\left(\boldsymbol{S_X^2}+\lambda\boldsymbol{I}_d\right)^{-1}\boldsymbol{S_X}-I_n\right)\boldsymbol{V_X^\top Y}\right\|^2\right] \\
&= \mathop{\mathbb{E}}_{\boldsymbol{v},S}\left[\frac{1}{n}\sum_{i=1}^{d}\left[\left(\frac{\boldsymbol{S_X}[i]^2}{\lambda+\boldsymbol{S_X}[i]^2}-1\right)^2\boldsymbol{V_X^\top Y}[i]^2\right]\right] \\
&= \mathop{\mathbb{E}}_{\boldsymbol{v},S}\left[\frac{1}{n}\sum_{i=1}^{d}\left[\left(\frac{\lambda}{\lambda+\boldsymbol{S_X}[i]^2}\right)^2\boldsymbol{V_X^\top Y}[i]^2\right]\right] \\
&\geq \mathop{\mathbb{E}}_{\boldsymbol{v},S}\left[\frac{1}{n}\sum_{i=1}^{d}\left[\left(\frac{\lambda'}{\lambda'+\boldsymbol{S_X}[i]^2}\right)^2\boldsymbol{V_X^\top Y}[i]^2\right]\right] \\
&= \mathcal{L}^{\text{tr-tr}}_{\lambda',\text{rep}}(\boldsymbol{A})
\end{aligned}$$

where the only inequality in the above sequence follows from the observation that $\frac{\lambda}{\lambda+a}$ is an increasing function for $a, \lambda > 0$. This completes the proof. □

C. More on Train-Validation split

C.1. Proof of main results

We first prove the result that for $n_1 = \bar{n}_1$ and $\lambda = \bar{\lambda}$, $\mathcal{L}^{\text{tr-val}}_{\lambda,\text{rep}} \equiv \mathcal{L}^{\text{test}}_{\bar{\lambda},\text{rep}}$.

Proposition 5.3. *$\mathcal{L}^{tr\text{-}val}_{\lambda,rep}(\cdot;(n_1,n_2))$ and $\mathcal{L}^{test}_{\bar{\lambda},rep}(\cdot;\bar{n}_1)$ are equivalent if $\bar{n}_1 = n_1$ and $\bar{\lambda} = \lambda$*

Proof. We again note using the central limit theorem and Equation (9) that

$$\begin{aligned}
\mathcal{L}^{\text{tr-val}}_{\lambda,\text{rep}}(\boldsymbol{A}) = \lim_{T\to\infty}\widehat{\mathcal{L}}^{\text{tr-val}}_{\lambda,\text{rep}}(\boldsymbol{A}) &= \mathop{\mathbb{E}}_{\rho_{\boldsymbol{v}}\sim\bar{\mu}}\left[\mathop{\mathbb{E}}_{(S^{\text{tr}},S^{\text{val}})\sim\rho_{\boldsymbol{v}}^n}\left[\frac{1}{n_2}\left\|\boldsymbol{X}^{\text{val}}\boldsymbol{A}\boldsymbol{w}_\lambda(\boldsymbol{A};S^{\text{tr}})-\boldsymbol{Y}^{\text{val}}\right\|^2\right]\right] \\
&=^{(a)} \mathop{\mathbb{E}}_{\rho_{\boldsymbol{v}}\sim\bar{\mu}}\left[\mathop{\mathbb{E}}_{S^{\text{tr}}\sim\rho_{\boldsymbol{v}}^{n_1}}\left[\mathop{\mathbb{E}}_{S^{\text{val}}\sim\rho_{\boldsymbol{v}}^{n_2}}\frac{1}{n_2}\left\|\boldsymbol{X}^{\text{val}}\boldsymbol{A}\boldsymbol{w}_\lambda(\boldsymbol{A};S^{\text{tr}})-\boldsymbol{Y}^{\text{val}}\right\|^2\right]\right] \\
&= \mathop{\mathbb{E}}_{\rho_{\boldsymbol{v}}\sim\bar{\mu}}\left[\mathop{\mathbb{E}}_{S^{\text{tr}}\sim\rho_{\boldsymbol{v}}^{n_1}}\left[\mathop{\mathbb{E}}_{(\boldsymbol{x},y)\sim\rho_{\boldsymbol{v}}}\left(\boldsymbol{x}^\top\boldsymbol{A}\boldsymbol{w}_\lambda(\boldsymbol{A};S^{\text{tr}})-y\right)^2\right]\right] \\
&= \mathop{\mathbb{E}}_{\rho_{\boldsymbol{v}}\sim\bar{\mu}}\left[\mathop{\mathbb{E}}_{S^{\text{tr}}\sim\rho_{\boldsymbol{v}}^{n_1}}\left[\left\|\boldsymbol{A}\boldsymbol{w}_\lambda(\boldsymbol{A};S^{\text{tr}})-\boldsymbol{v}\right\|^2\right]\right]+\sigma^2 \\
&=^{(b)} \mathcal{L}^{\text{test}}_{\lambda,\text{rep}}(\boldsymbol{A};n_1)
\end{aligned}$$

where (a) follows by noticing that sample $S\sim\rho_{\boldsymbol{v}}^n$ and splitting randomly into S^{tr} and S^{val} is equivalent to independently sampling $S^{\text{tr}}\sim\rho_{\boldsymbol{v}}^{n_1}$ and $S^{\text{val}}\sim\rho_{\boldsymbol{v}}^{n_2}$ and (b) follows from the definition of $\mathcal{L}^{\text{test}}$ from Equation (11).

□

We now prove the main benefit of the tr-val split: learning of low rank linear representations. For that, we use this general result that computes closed form expression for $\mathcal{L}^{\text{tr-val}}_{\lambda,\text{rep}}$.

Theorem C.1. *Let $\lambda = 0$. For a first layer $\boldsymbol{A} \in \mathbb{R}^{d\times d}$ with $r = rank(\boldsymbol{A})$, let $\boldsymbol{A} = \boldsymbol{U}\boldsymbol{S}\boldsymbol{V}^\top$ be the SVD, where $\boldsymbol{U}, \boldsymbol{V} \in \mathbb{R}^{d\times r}$ and $\boldsymbol{S} \in \mathbb{R}^{r\times r}$. Furthermore let $P_{\boldsymbol{U}} = \boldsymbol{U}\boldsymbol{U}^\top$ denote the span of columns of $\boldsymbol{U}$ (and thus $\boldsymbol{A}$) and let $P_{\boldsymbol{U}}^\perp = I_d - P_{\boldsymbol{U}}$. Then the tr-val objective has the following form*

$$\mathcal{L}^{\text{tr-val}}_{\lambda,rep}(\boldsymbol{A};(n_1,n_2)) - \sigma^2 = \begin{cases} (1+\alpha(n_1,r))\|P_{\boldsymbol{U}}^\perp \boldsymbol{A}^*\|^2 + \alpha(n_1,r)\sigma^2 & \text{if } r < n_1 - 1 \\ (\frac{n_1}{r} + \alpha(r,n_1))\|P_{\boldsymbol{U}}^\perp \boldsymbol{A}^*\|^2 + \frac{r-n_1}{r} + \beta(\boldsymbol{S}) + \alpha(r,n_1)\sigma^2 & \text{if } r > n_1 + 1 \\ \infty & \text{otherwise} \end{cases}$$

where $\beta(\boldsymbol{S}) \geq 0$ and $\beta(\boldsymbol{S}) = 0$ when $\boldsymbol{S} = \kappa \boldsymbol{I}_r$ for some $\kappa > 0$. Also α is defined as

$$\alpha(a,b) = \frac{b}{a-b-1} \tag{21}$$

We prove this theorem in Section C.2 First we prove the main result by using this theorem. We note that similar results can be shown for different regimes of k, d, n_1 and σ. The result below is for one reasonable regime where $n_1 = \Omega(k)$ and $\sigma = \mathcal{O}(1)$. Similar results can be obtained if we further assume that $k \ll d$ with weaker conditions on σ, we leave that for future work. Furthermore, this result can be extended to τ-optimal solutions to the tr-val objective rather than just the optimal solution.

Theorem 5.4. *Let $\lambda = 0$. Suppose $n_1 \geq 2k+2$ and $\sigma^2 \in (0, \frac{n_1-k-1}{3k})$, then any optimal solution $\boldsymbol{A}_{best} \in \arg\min_{\boldsymbol{A}\in\mathbb{R}^{d\times d}} \mathcal{L}^{\text{tr-val}}_{\lambda,rep}(\boldsymbol{A};(n_1,n_2))$ will satisfy*

$$rank(\boldsymbol{A}_{best}) = k, \quad P_{\boldsymbol{A}_{best}}\boldsymbol{A}^* = \boldsymbol{A}^*, \quad \mathcal{L}^{\text{tr-val}}_{\lambda,rep}(\boldsymbol{A}_{best};(n_1,n_2)) - \sigma^2 = \sigma^2 \frac{k}{n_1-k-1}$$

where $P_{\boldsymbol{A}} = \boldsymbol{A}\boldsymbol{A}^\dagger$ is the projection matrix onto the columnspace of $\boldsymbol{A}$. For any matrix $\boldsymbol{A}_{good} \in \mathbb{R}^{d\times d}$ that is τ-optimal, i.e. it satisfies $\mathcal{L}^{\text{tr-val}}_{\lambda,rep}(\boldsymbol{A}_{good};(n_1,n_2)) \leq \inf_{\boldsymbol{A}\in\mathbb{R}^{d\times d}} \mathcal{L}^{\text{tr-val}}_{\lambda,rep}(\boldsymbol{A};(n_1,n_2)) + \tau$ for $\tau \in (0, \frac{\sigma^2}{n_1-k-1})$, then we have

$$rank(\boldsymbol{A}_{good}) = k, \quad \|P_{\boldsymbol{A}_{good}}\boldsymbol{A}^* - \boldsymbol{A}^*\|^2 \leq \tau$$

The meta-testing performance of $\boldsymbol{A}_{good}$ on a new task with $\bar{n}_1 > 2k+2$ samples satisfies

$$\inf_{\bar{\lambda}\geq 0} \mathcal{L}^{test}_{\bar{\lambda},rep}(\boldsymbol{A}_{good};\bar{n}_1) - \sigma^2 \leq 2\tau + \sigma^2 \frac{2k}{\bar{n}_1}$$

Proof. Suppose the optimal value of $\mathcal{L}^{\text{tr-val}}_{\lambda,\text{rep}}(\boldsymbol{A};(n_1,n_2))$ is $\mathcal{L}^*$, i.e.

$$\mathcal{L}^* = \inf_{\boldsymbol{A}\in\mathbb{R}^{d\times d}} \mathcal{L}^{\text{tr-val}}_{\lambda,\text{rep}}(\boldsymbol{A};(n_1,n_2)) \tag{22}$$

Let $\boldsymbol{A}_{\text{good}} \in \mathbb{R}^{d\times d}$ be the "good" matrix that is τ-optimal, i.e. $\mathcal{L}^{\text{tr-val}}_{\lambda,\text{rep}}(\boldsymbol{A}_{\text{good}};(n_1,n_2)) - \mathcal{L}^* \leq \tau$. Note that the result for $\boldsymbol{A}_{\text{best}}$ follows from the result for $\tau = 0$.

We use the expression for $\mathcal{L}^{\text{tr-val}}_{\lambda,\text{rep}}(\cdot;(n_1,n_2)$ from Theorem C.1 to find properties for $\boldsymbol{A}_{\text{good}}$ that can ensure that it is τ-optimal. Consider a representation $\boldsymbol{A} \in \mathbb{R}^{d\times d}$ and let $\boldsymbol{A} = \boldsymbol{U}\boldsymbol{S}\boldsymbol{V}^\top$ be its SVD with $\boldsymbol{U}, \boldsymbol{V} \in \mathbb{R}^{d\times r}$ and $\boldsymbol{S} \in \mathbb{R}^{r\times r}$ where $r = \text{rank}(\boldsymbol{A})$. We consider 3 cases for r: $r < k$, $k \leq r \leq n_1$ and $r > n_1$, and find properties that can result in $\boldsymbol{A}$ being τ-optimal in each of the 3 cases. For the ranges of n_1, σ, τ in the theorem statement, it will turn out that when $r < k$ or $r > n_1$, $\boldsymbol{A}$ must satisfy $\mathcal{L}^{\text{tr-val}}_{\lambda,\text{rep}}(\boldsymbol{A};(n_1,n_2) > \mathcal{L}^* + \tau$, thus concluding that $\text{rank}(\boldsymbol{A}_{\text{good}})$ cannot be in these ranges.

To do so we analyze the optimal value of $\mathcal{L}^{\text{tr-val}}_{\lambda,\text{rep}}(\cdot;(n_1,n_2))$ that can be achieved for all three cases. We first start analyzing the most promising case of $k \leq r \leq n_1$.

Case 1: $k \le r \le n_1$ For this case we can use the $r < n_1 - 1$ regime from Theorem C.1. Note that for $r \in \{n_1 - 1, n_1\}$, $\mathcal{L}^{\text{tr-val}}_{\lambda,\text{rep}}(\boldsymbol{A}; (n_1, n_2))$ is unbounded. Thus we get

$$\mathcal{L}^{\text{tr-val}}_{\lambda,\text{rep}}(\boldsymbol{A}; (n_1, n_2)) - \sigma^2 = (1 + \alpha(n_1, r))\|P^{\perp}_{\boldsymbol{U}}\boldsymbol{A}^*\|^2 + \alpha(n_1, r)\sigma^2 \tag{23}$$

where $\alpha(n_1, r)$ is defined in Equation (34). Note that $\|P^{\perp}_{\boldsymbol{U}}\boldsymbol{A}^*\| \ge 0$ and in fact equality can be achieved for every r in this case since $r = \text{rank}(\boldsymbol{U}) \ge k = \text{rank}(\boldsymbol{A}^*)$. Furthermore $\alpha(n_1, r) = \frac{r}{n_1 - r - 1}$ is an increasing function of r, and thus $\alpha(n_1, r) \ge \alpha(n_1, k)$, which can also be achieved by picking $r = k$. Thus for the range of $k \le r \le n_1$, we get

$$\mathcal{L}^* - \sigma^2 \le \min_{\substack{\boldsymbol{A} \text{ s.t.} \\ k \le r \le n_1}} \mathcal{L}^{\text{tr-val}}_{\lambda,\text{rep}}(\boldsymbol{A}; (n_1, n_2)) - \sigma^2 = \sigma^2\alpha(n_1, k) = \sigma^2\frac{k}{n_1 - k - 1} \tag{24}$$

with the minimum being achieved when $\text{rank}(\boldsymbol{A}) = k$ and $\|P^{\perp}_{\boldsymbol{U}}\boldsymbol{A}^*\|$, which is the same as $P_{\boldsymbol{U}}\boldsymbol{A}^* = \boldsymbol{A}^*$.

Suppose $r_{\text{good}} = \text{rank}(\boldsymbol{A}_{\text{good}})$ lies in this range, i.e. $k \le r_{\text{good}} \le n_1$. Using the fact that $\boldsymbol{A}_{\text{good}}$ is τ-optimal, we can show an upper bound the r_{good} and $\|P_{\boldsymbol{A}_{\text{good}}}\boldsymbol{A}^* - \boldsymbol{A}^*\|$. From the τ-optimality condition and Equation (24), we get

$$\begin{aligned}
\tau &\ge \mathcal{L}^{\text{tr-val}}_{\lambda,\text{rep}}(\boldsymbol{A}_{\text{good}}; (n_1, n_2)) - \mathcal{L}^* \ge \mathcal{L}^{\text{tr-val}}_{\lambda,\text{rep}}(\boldsymbol{A}_{\text{good}}; (n_1, n_2)) - \sigma^2\alpha(n_1, k) && (25)\\
&\ge^{(a)} (1 + \alpha(n_1, r))\|P^{\perp}_{\boldsymbol{U}}\boldsymbol{A}^*\|^2 + \sigma^2\alpha(n_1, r_{\text{good}}) - \sigma^2\alpha(n_1, k) && (26)\\
&= (1 + \alpha(n_1, r))\|P^{\perp}_{\boldsymbol{A}_{\text{good}}}\boldsymbol{A}^*\|^2 + \sigma^2(\alpha(n_1, r_{\text{good}}) - \alpha(n_1, k)) && (27)\\
&=^{(b)} \|P^{\perp}_{\boldsymbol{A}_{\text{good}}}\boldsymbol{A}^*\|^2 + \sigma^2\alpha'(n_1, r')(r_{\text{good}} - k) = \|P^{\perp}_{\boldsymbol{A}_{\text{good}}}\boldsymbol{A}^*\|^2 + \sigma^2\frac{r'(n_1 - 1)}{(n_1 - r' - 1)^2}(r_{\text{good}} - k) && (28)\\
&\ge^{(c)} \|P^{\perp}_{\boldsymbol{A}_{\text{good}}}\boldsymbol{A}^*\|^2 + \sigma^2\frac{k(n_1 - 1)}{(n_1 - k - 1)^2}(r_{\text{good}} - k) \ge \|P^{\perp}_{\boldsymbol{A}_{\text{good}}}\boldsymbol{A}^*\|^2 + \sigma^2\frac{k}{(n_1 - k - 1)}(r_{\text{good}} - k) && (29)
\end{aligned}$$

where (a) follows from Equation (23) instantiated for $\boldsymbol{A}_{\text{good}}$, (b) follows from the mean value theorem applied to the continuous function $\alpha(n_1, \cdot)$ with $r' \in [k, r_{\text{good}}]$, (c) follows by $k > r'$ and the monotonocity of $\alpha'(n_1, \cdot)$. Thus from the inequality in Equation (29), we can get an upper bound on r_{good} as follows

$$\begin{aligned}
\tau \ge \sigma^2\frac{k}{(n_1 - k - 1)}(r_{\text{good}} - k) &\implies r_{\text{good}} \le k + \frac{\tau(n_1 - k - 1)}{\sigma^2 k} \le^{(a)} k + \frac{1}{k} && (30)\\
&\implies r_{\text{good}} = k && (31)
\end{aligned}$$

where (a) follows from the upper bound on τ in the theorem statement. We can also get an upper bound on $\|P^{\perp}_{\boldsymbol{A}_{\text{good}}}\boldsymbol{A}^*\|$ using Equation (29) as $\|P^{\perp}_{\boldsymbol{A}_{\text{good}}}\boldsymbol{A}^*\|^2 \le \tau$. Thus any $\boldsymbol{A}_{\text{good}}$ with $k \le r_{\text{good}} \le n_1$ that is τ-optimal will satisfy $r_{\text{good}} = k$ and $\|P^{\perp}_{\boldsymbol{A}_{\text{good}}}\boldsymbol{A}^*\|^2 \le \tau$.

Additionally the minimum achievable value for this range of rank is $\sigma^2\frac{k}{n_1 - k - 1}$ fromr Equation (24). We now analyze the other two cases, where we will show that no $\boldsymbol{A}$ can even be τ-optimal.

Case 2: $r < k$ In this case, since $k < n_1$, we are still in the $r < n_1$ regime. The key point here is that when $r < k$, it is impossible for $\boldsymbol{A}$ to span all of $\boldsymbol{A}^*$. In fact for rank r, it is clear that $\boldsymbol{A}$ can cover only at most r out of k directions from $\boldsymbol{A}^*$. Thus the inexpressiveness term $\|P^{\perp}_{\boldsymbol{U}}\boldsymbol{A}^*\|^2$ will be at least $\frac{k-r}{k}\|\boldsymbol{A}^*\|^2 = \frac{k-r}{k}$. Using the $r < n_1$ expression from Theorem C.1, we get

$$\begin{aligned}
\mathcal{L}^{\text{tr-val}}_{\lambda,\text{rep}}(\boldsymbol{A}; (n_1, n_2)) - \sigma^2 &= (1 + \alpha(n_1, r))\|P^{\perp}_{\boldsymbol{U}}\boldsymbol{A}^*\|^2 + \alpha(n_1, r)\sigma^2\\
&\ge^{(a)} \frac{k - r}{k} + \sigma^2\frac{r}{n_1 - r - 1}\\
&= \frac{k - r}{k} + \sigma^2\frac{r}{n_1 - r - 1} - \sigma^2\frac{k}{n_1 - r - 1} + \sigma^2\frac{k}{n_1 - r - 1} - \sigma^2\frac{k}{n_1 - k - 1} + \sigma^2\frac{k}{n_1 - k - 1}\\
&= \frac{k - r}{k} - \sigma^2\frac{k - r}{n_1 - r - 1} - \sigma^2 k\left(\frac{1}{n_1 - k - 1} - \frac{1}{n_1 - r - 1}\right) + \sigma^2\frac{k}{n_1 - k - 1}\\
&= \frac{k - r}{k} - \sigma^2\frac{k - r}{n_1 - r - 1} - \sigma^2\frac{k(k - r)}{(n_1 - r - 1)(n_1 - k - 1)} + \sigma^2\frac{k}{n_1 - k - 1}
\end{aligned}$$

$$\geq^{(b)} \frac{k-r}{k} - \sigma^2 \frac{k-r}{n_1-k-1} - \sigma^2 \frac{(k-r)}{(n_1-k-1)} + \sigma^2 \frac{k}{n_1-k-1}$$
$$= \frac{k-r}{k} - \sigma^2 \frac{2(k-r)}{n_1-k-1} + \sigma^2 \frac{k}{n_1-k-1} \tag{32}$$

where for (a) we use $\alpha(n_1, r) \geq 0$ and for (b), we use that $k < n_1-k-1 < n_1-r-1$. Since we assume $\sigma^2 < {}^{(n_1-k-1)}/_{2k}$, the difference between the first 2 terms is at least 0. Thus the error when $r < k$ is at least $\mathcal{L}^{\text{tr-val}}_{\lambda,\text{rep}}(\boldsymbol{A}; (n_1, n_2)) - \sigma^2 > \sigma^2 \frac{k}{n_1-k-1}$, which is larger than the previous case. So the optimal solution cannot have $r < k$.

We now check if such an $\boldsymbol{A}$ can be a τ-optimal solution instead. The answer is negative due to the following calculation

$$\mathcal{L}^{\text{tr-val}}_{\lambda,\text{rep}}(\boldsymbol{A}; (n_1, n_2)) - \mathcal{L}^* \geq^{(a)} \mathcal{L}^{\text{tr-val}}_{\lambda,\text{rep}}(\boldsymbol{A}; (n_1, n_2)) - \sigma^2 \frac{k}{n_1-k-1}$$
$$\geq^{(b)} \frac{k-r}{k} - \sigma^2 \frac{2(k-r)}{n_1-k-1}$$
$$\geq^{(c)} \frac{3k\sigma^2}{n_1-k-1} \frac{k-r}{k} - \sigma^2 \frac{2(k-r)}{n_1-k-1}$$
$$= \frac{\sigma^2(k-r)}{n_1-k-1} \geq \frac{\sigma^2}{n_1-k-1} > \tau$$

where (a) follows from Equation (24), (b) follows from Equation (32), (c) follows from $1 \geq \frac{3k\sigma^2}{n_1-k-1}$ from the condition on σ in the theorem statement and the last inequalities follow from $r < k$ condition and the range of τ in the theorem statement respectively. Thus we cannot even have a τ-optimal solution with $r < k$. Next we show the same for the case of $r > n_1$.

Case 3: $r > n_1$ For this case we can use the $r > n_1 + 1$ regime from Theorem C.1. Note that for $r \in \{n_1 + 1, n_1\}$, $\mathcal{L}^{\text{tr-val}}_{\lambda,\text{rep}}(\boldsymbol{A}; (n_1, n_2))$ is unbounded. Thus we have $\mathcal{L}^{\text{tr-val}}_{\lambda,\text{rep}}(\boldsymbol{A}; (n_1, n_2)) - \sigma^2 = (\frac{n_1}{r} + \alpha(r, n_1))\|P^{\perp}_{\boldsymbol{U}}\boldsymbol{A}^*\|^2 + \frac{r-n_1}{r} + \beta(\boldsymbol{S}) + \sigma^2\alpha(r, n_1)$. We again note that since $r > n_1 > k$, we can make $\|P^{\perp}_{\boldsymbol{U}}\boldsymbol{A}^*\|^2 = 0$ by simply picking an "expressive" rank-r subspace $\boldsymbol{U}$. Further more $\beta(\boldsymbol{S}) = 0$ is easy to achieve by making $\boldsymbol{S} = \boldsymbol{I}_r$ which can be done independently of $\boldsymbol{U}$. Thus we can lower bound the error as follows

$$\mathcal{L}^{\text{tr-val}}_{\lambda,\text{rep}}(\boldsymbol{A}; (n_1, n_2)) - \sigma^2 \geq \frac{r-n_1}{r} + \sigma^2\alpha(r, n_1) = 1 - \frac{n_1}{r} + \sigma^2 \frac{n_1}{r-n_1-1} \tag{33}$$

where equality can be achieved by letting columns of $\boldsymbol{A}$ span the subspace $\boldsymbol{A}^*$ and by picking $\boldsymbol{S} = I_r$. We now lower bound this value further to show that we cannot have a τ-optimal solution.

Firstly we note that if $\sigma \geq 1$, then Equation (33) gives us

$$\mathcal{L}^{\text{tr-val}}_{\lambda,\text{rep}}(\boldsymbol{A}; (n_1, n_2)) - \sigma^2 \geq 1 - \frac{n_1}{r} + \frac{n_1}{r-n_1-1} \geq 1 \geq^{(a)} \sigma^2 \frac{3k}{n_1-k-1}$$
$$\geq^{(b)} (\mathcal{L}^* - \sigma^2) + \sigma^2 \frac{2k}{n_1-k-1} \geq^{(c)} (\mathcal{L}^* - \sigma^2) + \tau(n_1-k-1)\frac{2k}{n_1-k-1}$$
$$> (\mathcal{L}^* - \sigma^2) + \tau$$

where (a) follows from the upper bound on σ^2 from the theorem statement, (b) follows from Equation (24) and (c) follows from the upper bound on τ from the theorem statement. Thus we cannot have a τ-optimal solution in this case.

If $\sigma < 1$ on the other hand, we can lower bound the error as

$$\mathcal{L}^{\text{tr-val}}_{\lambda,\text{rep}}(\boldsymbol{A}; (n_1, n_2)) - \sigma^2 = 1 - \frac{n_1}{r} + \sigma^2 \frac{n_1}{r-n_1-1} > 1 - \frac{n_1}{r} + \sigma^2 \frac{n_1}{r-n_1}$$
$$\geq \min_{r>n_1} 1 - \frac{n_1}{r} + \sigma^2 \frac{n_1}{r-n_1}$$

Setting the derivate w.r.t. r to 0 for the above expression, we get that this is minimized at $r = {}^{n_1}/_{(1-\sigma)}$. Plugging in this value and simplifying, we get

$$\mathcal{L}^{\text{tr-val}}_{\lambda,\text{rep}}(\boldsymbol{A}; (n_1, n_2)) - \sigma^2 \geq 1 - (1-\sigma)^2 = 2\sigma - \sigma^2 \geq \sigma^2$$

$$= \sigma^2 \frac{n_1 - 2k - 1}{n_1 - k - 1} + \sigma^2 \frac{k}{n_1 - k - 1} >^{(a)} \frac{\sigma^2}{n_1 - k - 1} + (\mathcal{L}^* - \sigma^2)$$
$$>^{(b)} \tau + (\mathcal{L}^* - \sigma^2)$$

where (a) follows from the condition that $n_1 > 2k + 2$ and (b) follows again from Equation (24). Thus the $r > n_1$ case can never give us a τ-optimal solution, let alone the optimal solution and so $\boldsymbol{A}_{\text{good}}$ cannot have rank $r_{\text{good}} > n_1$ either. Combining all 3 cases, we can conclude that any τ-optimal solution $\boldsymbol{A}_{\text{good}}$ will satisfy $\text{rank}(\boldsymbol{A}_{\text{good}}) = k$ and $\|P^{\perp}_{\boldsymbol{A}_{\text{good}}} \boldsymbol{A}^*\|^2 = \|P_{\boldsymbol{A}_{\text{good}}} \boldsymbol{A}^* - \boldsymbol{A}^*\|^2 \leq \tau$.

For the second part, we use Proposition 5.3 to first get that for $\lambda = \bar{\lambda} = 0$, $\mathcal{L}^{\text{test}}_{\bar{\lambda},\text{rep}}(\boldsymbol{A}_{\text{good}}; \bar{n}_1) = \mathcal{L}^{\text{tr-val}}_{\lambda,\text{rep}}(\boldsymbol{A}_{\text{good}}; (\bar{n}_1, \cdot))$. Since $\bar{n}_1 > 2k + 2$ and $\text{rank}(\boldsymbol{A}_{\text{good}}) = k < \bar{n}_1$, we can use Theorem C.1 to get

$$\mathcal{L}^{\text{test}}_{\bar{\lambda},\text{rep}}(\boldsymbol{A}_{\text{good}}; \bar{n}_1) - \sigma^2 = \mathcal{L}^{\text{tr-val}}_{\lambda,\text{rep}}(\boldsymbol{A}_{\text{good}}; (\bar{n}_1, \cdot)) - \sigma^2 = (1 + \alpha(\bar{n}_1, k))\|P^{\perp}_{\boldsymbol{A}_{\text{good}}} \boldsymbol{A}^*\|^2 + \alpha(\bar{n}_1, k)\sigma^2$$
$$\leq^{(a)} 2\|P^{\perp}_{\boldsymbol{A}_{\text{good}}} \boldsymbol{A}^*\|^2 + \sigma^2 \frac{k}{\bar{n}_1 - k - 1} \leq^{(b)} 2\tau + \sigma^2 \frac{k}{\bar{n}_1 - \bar{n}_1/2} = 2\tau + \sigma^2 \frac{2k}{\bar{n}_1}$$

where (a) follows by noting that $\alpha(\bar{n}_1, k) < 1$ and (b) follows from $k + 1 < \bar{n}_1/2$. This completes the proof of result for τ-optimal solution $\boldsymbol{A}_{\text{good}}$. Setting $\tau = 0$ gives results for optimal solution $\boldsymbol{A}_{\text{best}}$ of $\mathcal{L}^{\text{tr-val}}_{\lambda,\text{rep}}(\cdot, (n_1, n_2))$. □

C.2. Proof of Theorem C.1

We now prove the crucial result that gives a closed form solution for the tr-val objective

Theorem C.1. *Let $\lambda = 0$. For a first layer $\boldsymbol{A} \in \mathbb{R}^{d \times d}$ with $r = \text{rank}(\boldsymbol{A})$, let $\boldsymbol{A} = \boldsymbol{U}\boldsymbol{S}\boldsymbol{V}^\top$ be the SVD, where $\boldsymbol{U}, \boldsymbol{V} \in \mathbb{R}^{d \times r}$ and $\boldsymbol{S} \in \mathbb{R}^{r \times r}$. Furthermore let $P_{\boldsymbol{U}} = \boldsymbol{U}\boldsymbol{U}^\top$ denote the span of columns of $\boldsymbol{U}$ (and thus $\boldsymbol{A}$) and let $P^{\perp}_{\boldsymbol{U}} = I_d - P_{\boldsymbol{U}}$. Then the tr-val objective has the following form*

$$\mathcal{L}^{\textit{tr-val}}_{\lambda,\textit{rep}}(\boldsymbol{A}; (n_1, n_2)) - \sigma^2 = \begin{cases} (1 + \alpha(n_1, r))\|P^{\perp}_{\boldsymbol{U}} \boldsymbol{A}^*\|^2 + \alpha(n_1, r)\sigma^2 & \textit{if } r < n_1 - 1 \\ (\frac{n_1}{r} + \alpha(r, n_1))\|P^{\perp}_{\boldsymbol{U}} \boldsymbol{A}^*\|^2 + \frac{r - n_1}{r} + \beta(\boldsymbol{S}) + \alpha(r, n_1)\sigma^2 & \textit{if } r > n_1 + 1 \\ \infty & \textit{otherwise} \end{cases}$$

where $\beta(\boldsymbol{S}) \geq 0$ and $\beta(\boldsymbol{S}) = 0$ when $\boldsymbol{S} = \kappa \boldsymbol{I}_r$ for some $\kappa > 0$. Also α is defined as

$$\alpha(a, b) = \frac{b}{a - b - 1} \tag{34}$$

Proof. We will prove the result for the cases $r < n_1$ and $r \geq n_1$. Firstly, using Proposition 5.3, we already get that

$$\mathcal{L}^{\text{tr-val}}_{\lambda,\text{rep}}(\boldsymbol{A}; (n_1, n_2)) = \mathcal{L}^{\text{test}}_{\lambda,\text{rep}}(\boldsymbol{A}; n_1) = \sigma^2 + \mathop{\mathbb{E}}_{\rho_{\boldsymbol{v}} \sim \bar{\mu}} \left[\mathop{\mathbb{E}}_{(\boldsymbol{X}, \boldsymbol{Y}) \sim \rho_{\boldsymbol{v}}^{n_1}} \left[\|\boldsymbol{A}\boldsymbol{w}_\lambda(\boldsymbol{A}; (\boldsymbol{X}, \boldsymbol{Y})) - \boldsymbol{v}\|^2 \right] \right]$$
$$= \sigma^2 + \mathop{\mathbb{E}}_{\boldsymbol{v} \sim \mathcal{N}(\boldsymbol{0}_d, \boldsymbol{I}_d)} \left[\mathop{\mathbb{E}}_{\substack{\boldsymbol{X} \sim \mathcal{N}(\boldsymbol{0}_d, \boldsymbol{I}_d)^{n_1}, \\ \boldsymbol{\eta} \sim \mathcal{N}(\boldsymbol{0}_d, \sigma^2 \boldsymbol{I}_n)}} \left[\|\boldsymbol{A}\boldsymbol{w}_\lambda(\boldsymbol{A}; (\boldsymbol{X}, \boldsymbol{X}\boldsymbol{v} + \boldsymbol{\eta})) - \boldsymbol{v}\|^2 \right] \right] \tag{35}$$

We will use this expression in the rest of the proof.

Case 1: $r \leq n_1$ In this case, the rank of the representations $\boldsymbol{X}\boldsymbol{A} \in \mathbb{R}^{n_1 \times d}$ for training data is higher than the number of samples. Thus the unique minimizer for the dataset $(\boldsymbol{X}, \boldsymbol{Y})$ is

$$\lim_{\lambda \to 0} \boldsymbol{w}_\lambda(\boldsymbol{A}; (\boldsymbol{X}, \boldsymbol{Y})) = \mathop{\arg\min}_{\boldsymbol{w} \in \mathbb{R}^d} \ell_{\lambda,\text{rep}}(\boldsymbol{w}; \boldsymbol{A}, (\boldsymbol{X}, \boldsymbol{Y})) = \left(\boldsymbol{A}^\top \boldsymbol{X}^\top \boldsymbol{X} \boldsymbol{A}\right)^\dagger \boldsymbol{A}^\top \boldsymbol{X}^\top \boldsymbol{Y}$$

Plugging this into Equation (35), we get

$$\mathcal{L}^{\text{tr-val}}_{\lambda,\text{rep}}(\boldsymbol{A};(n_1,n_2)) - \sigma^2 = \mathop{\mathbb{E}}_{\boldsymbol{v}\sim\mathcal{N}(\mathbf{0}_d,\boldsymbol{I}_d)}\left[\mathop{\mathbb{E}}_{\substack{\boldsymbol{X}\sim\mathcal{N}(\mathbf{0}_d,\boldsymbol{I}_d)^{n_1},\\ \boldsymbol{\eta}\sim\mathcal{N}(\mathbf{0}_d,\sigma^2\boldsymbol{I}_{n_1})}}\left[\left\|\boldsymbol{A}\left(\boldsymbol{A}^\top\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{A}\right)^\dagger\boldsymbol{A}^\top\boldsymbol{X}^\top(\boldsymbol{X}\boldsymbol{v}+\boldsymbol{\eta})-\boldsymbol{v}\right\|^2\right]\right]$$
$$=^{(a)} \underbrace{\mathop{\mathbb{E}}_{\boldsymbol{v},\boldsymbol{X}}\left[\left\|\boldsymbol{A}\left(\boldsymbol{A}^\top\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{A}\right)^\dagger\boldsymbol{A}^\top\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{v}-\boldsymbol{v}\right\|^2\right]}_{\text{bias}(\boldsymbol{A})} + \underbrace{\mathop{\mathbb{E}}_{\boldsymbol{\eta},\boldsymbol{X}}\left[\left\|\boldsymbol{A}\left(\boldsymbol{A}^\top\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{A}\right)^\dagger\boldsymbol{A}^\top\boldsymbol{X}^\top\boldsymbol{\eta}\right\|^2\right]}_{\text{variance}(\boldsymbol{A})}$$

where (a) follows from the independence of $\boldsymbol{\eta}$ and $(\boldsymbol{X},\boldsymbol{v})$ and that $\mathbb{E}[\boldsymbol{\eta}]=\mathbf{0}_n$. We will analyze the bias and variance terms separately below

Bias: Recall that $\boldsymbol{A}=\boldsymbol{U}\boldsymbol{S}\boldsymbol{V}^\top$ is the singular value decomposition, with $\boldsymbol{U},\boldsymbol{V}\in\mathbb{R}^{d\times r}$ and $\boldsymbol{S}\in\mathbb{R}^{r\times r}$. We set $\boldsymbol{X}_{\boldsymbol{U}}=\boldsymbol{X}\boldsymbol{U}\in\mathbb{R}^{n_1\times r}$ The two key ideas that we will exploit are that $\boldsymbol{X}_{\boldsymbol{U}}\sim\mathcal{N}(\mathbf{0}_d,\boldsymbol{I}_r)^{n_1}$ and that $\boldsymbol{X}_{\boldsymbol{U}}$ is independent of $\boldsymbol{X}P_{\boldsymbol{U}}^\perp$ since $\boldsymbol{X}$ is Gaussian-distributed.

$$\begin{aligned}
\text{bias}(\boldsymbol{A}) &= \mathop{\mathbb{E}}_{\boldsymbol{v},\boldsymbol{X}}\left[\left\|\boldsymbol{A}\left(\boldsymbol{A}^\top\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{A}\right)^\dagger\boldsymbol{A}^\top\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{v}-\boldsymbol{v}\right\|^2\right]\\
&=^{(a)} \mathop{\mathbb{E}}_{\boldsymbol{v},\boldsymbol{X}}\left[\left\|\boldsymbol{U}\boldsymbol{S}\boldsymbol{V}^\top\left(\boldsymbol{V}\boldsymbol{S}\boldsymbol{U}^\top\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{U}\boldsymbol{S}\boldsymbol{V}^\top\right)^\dagger\boldsymbol{V}\boldsymbol{S}\boldsymbol{U}^\top\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{v}-\boldsymbol{v}\right\|^2\right]\\
&=^{(b)} \mathop{\mathbb{E}}_{\boldsymbol{v},\boldsymbol{X}}\left[\left\|\boldsymbol{U}\boldsymbol{S}\left(\boldsymbol{S}(\boldsymbol{X}\boldsymbol{U})^\top\boldsymbol{X}\boldsymbol{U}\boldsymbol{S}\right)^{-1}\boldsymbol{S}(\boldsymbol{X}\boldsymbol{U})^\top\boldsymbol{X}\boldsymbol{v}-\boldsymbol{v}\right\|^2\right]\\
&=^{(c)} \mathop{\mathbb{E}}_{\boldsymbol{v},\boldsymbol{X}}\left[\left\|\boldsymbol{U}\boldsymbol{S}\left(\boldsymbol{S}\boldsymbol{X}_{\boldsymbol{U}}^\top\boldsymbol{X}_{\boldsymbol{U}}\boldsymbol{S}\right)^{-1}\boldsymbol{S}\boldsymbol{X}_{\boldsymbol{U}}^\top\boldsymbol{X}\boldsymbol{v}-\boldsymbol{v}\right\|^2\right] =^{(d)} \mathop{\mathbb{E}}_{\boldsymbol{v},\boldsymbol{X}}\left[\left\|\boldsymbol{U}\boldsymbol{S}\boldsymbol{S}^{-1}(\boldsymbol{X}_{\boldsymbol{U}}^\top\boldsymbol{X}_{\boldsymbol{U}})^{-1}\boldsymbol{S}^{-1}\boldsymbol{S}\boldsymbol{X}_{\boldsymbol{U}}^\top\boldsymbol{X}\boldsymbol{v}-\boldsymbol{v}\right\|^2\right]\\
&=^{(e)} \mathop{\mathbb{E}}_{\boldsymbol{v},\boldsymbol{X}}\left[\left\|\boldsymbol{U}(\boldsymbol{X}_{\boldsymbol{U}}^\top\boldsymbol{X}_{\boldsymbol{U}})^{-1}\boldsymbol{X}_{\boldsymbol{U}}^\top\boldsymbol{X}\boldsymbol{v}-\boldsymbol{v}\right\|^2\right] = \mathop{\mathbb{E}}_{\boldsymbol{v},\boldsymbol{X}}\left[\left\|\boldsymbol{U}(\boldsymbol{X}_{\boldsymbol{U}}^\top\boldsymbol{X}_{\boldsymbol{U}})^{-1}\boldsymbol{X}_{\boldsymbol{U}}^\top\boldsymbol{X}(\boldsymbol{U}\boldsymbol{U}^\top+P_{\boldsymbol{U}}^\perp)\boldsymbol{v}-(\boldsymbol{U}\boldsymbol{U}^\top+P_{\boldsymbol{U}}^\perp)\boldsymbol{v}\right\|^2\right]\\
&=^{(f)} \mathop{\mathbb{E}}_{\boldsymbol{v},\boldsymbol{X}}\left[\left\|\boldsymbol{U}(\boldsymbol{X}_{\boldsymbol{U}}^\top\boldsymbol{X}_{\boldsymbol{U}})^{-1}\boldsymbol{X}_{\boldsymbol{U}}^\top\boldsymbol{X}\boldsymbol{U}\boldsymbol{U}^\top\boldsymbol{v}+\boldsymbol{U}(\boldsymbol{X}_{\boldsymbol{U}}^\top\boldsymbol{X}_{\boldsymbol{U}})^{-1}\boldsymbol{X}_{\boldsymbol{U}}^\top\boldsymbol{X}P_{\boldsymbol{U}}^\perp\boldsymbol{v}-\boldsymbol{U}\boldsymbol{U}^\top\boldsymbol{v}-P_{\boldsymbol{U}}^\perp\boldsymbol{v}\right\|^2\right]\\
&=^{(g)} \mathop{\mathbb{E}}_{\boldsymbol{v},\boldsymbol{X}}\left[\left\|\boldsymbol{U}\boldsymbol{U}^\top\boldsymbol{v}-\boldsymbol{U}\boldsymbol{U}^\top\boldsymbol{v}+\boldsymbol{U}(\boldsymbol{X}_{\boldsymbol{U}}^\top\boldsymbol{X}_{\boldsymbol{U}})^{-1}\boldsymbol{X}_{\boldsymbol{U}}^\top\boldsymbol{X}P_{\boldsymbol{U}}^\perp\boldsymbol{v}\right\|^2+\left\|P_{\boldsymbol{U}}^\perp\boldsymbol{v}\right\|^2\right]\\
&=^{(h)} \mathop{\mathbb{E}}_{\boldsymbol{v},\boldsymbol{X}}\left[\text{tr}\left((P_{\boldsymbol{U}}^\perp\boldsymbol{v})^\top P_{\boldsymbol{U}}^\perp\boldsymbol{X}^\top\boldsymbol{X}_{\boldsymbol{U}}(\boldsymbol{X}_{\boldsymbol{U}}^\top\boldsymbol{X}_{\boldsymbol{U}})^{-2}\boldsymbol{X}_{\boldsymbol{U}}^\top\boldsymbol{X}P_{\boldsymbol{U}}^\perp(P_{\boldsymbol{U}}^\perp\boldsymbol{v})\right)\right]+\left\|P_{\boldsymbol{U}}^\perp\boldsymbol{A}^*\right\|^2\\
&=^{(i)} \mathop{\mathbb{E}}_{\boldsymbol{v},\boldsymbol{X}_{\boldsymbol{U}}}\left[\text{tr}\left(\boldsymbol{X}_{\boldsymbol{U}}(\boldsymbol{X}_{\boldsymbol{U}}^\top\boldsymbol{X}_{\boldsymbol{U}})^{-2}\boldsymbol{X}_{\boldsymbol{U}}^\top\mathop{\mathbb{E}}_{P_{\boldsymbol{U}}^\perp\boldsymbol{X}}\left[\boldsymbol{X}P_{\boldsymbol{U}}^\perp(P_{\boldsymbol{U}}^\perp\boldsymbol{v})(P_{\boldsymbol{U}}^\perp\boldsymbol{v})^\top P_{\boldsymbol{U}}^\perp\boldsymbol{X}^\top\right]\right)\right]+\left\|P_{\boldsymbol{U}}^\perp\boldsymbol{A}^*\right\|^2
\end{aligned}$$

while (a) just uses the SVD of $\boldsymbol{A}$, (b) uses the simple fact that $\boldsymbol{V}^\top(\boldsymbol{V}\boldsymbol{B}\boldsymbol{V}^\top)^\dagger\boldsymbol{V}=\boldsymbol{B}^\dagger$ for an orthogonal matrix $\boldsymbol{V}$. (d) follows from the fact that $\boldsymbol{X}_{\boldsymbol{U}}^\top\boldsymbol{X}_{\boldsymbol{U}}\in\mathbb{R}^{r\times r}$ is full rank with probability 1, and thus invertible. (f) simply decomposes $\boldsymbol{v}=\boldsymbol{U}\boldsymbol{U}^\top\boldsymbol{v}+P_{\boldsymbol{U}}^\perp\boldsymbol{v}$, while (g) follows from the orthogonality of $P_{\boldsymbol{U}}^\perp\boldsymbol{v}$ and any vector in the span of $\boldsymbol{U}$. (h) uses the simple facts that $\|\boldsymbol{a}\|^2=\text{tr}(\boldsymbol{a}\boldsymbol{a}^\top)$ and $P_{\boldsymbol{U}}^\perp P_{\boldsymbol{U}}^\perp=P_{\boldsymbol{U}}^\perp$ and (i) uses the crucial observation that $\boldsymbol{X}_{\boldsymbol{U}}$ is independent of $\boldsymbol{X}P_{\boldsymbol{U}}^\perp$, since for Gaussian distribution, all subspaces are independent of its orthogonal subspace, and that tr is a linear operator.

We now look closer at the term $\boldsymbol{M}=\mathbb{E}\left[\boldsymbol{X}P_{\boldsymbol{U}}^\perp(P_{\boldsymbol{U}}^\perp\boldsymbol{v})(P_{\boldsymbol{U}}^\perp\boldsymbol{v})^\top P_{\boldsymbol{U}}^\perp\boldsymbol{X}^\top\right]$ which is a matrix in $\mathbb{R}^{n_1\times n_1}$.

$$\begin{aligned}
\boldsymbol{M}_{i,j} &= \mathop{\mathbb{E}}_{P_{\boldsymbol{U}}^\perp\boldsymbol{X}}\left[\boldsymbol{x}_i^\top(P_{\boldsymbol{U}}^\perp\boldsymbol{v})(P_{\boldsymbol{U}}^\perp\boldsymbol{v})^\top\boldsymbol{x}_j\right]\\
&= \text{tr}\left((P_{\boldsymbol{U}}^\perp\boldsymbol{v})(P_{\boldsymbol{U}}^\perp\boldsymbol{v})^\top\mathbb{E}\left[\boldsymbol{x}_j\boldsymbol{x}_i^\top\right]\right) = \begin{cases}0 & \text{if } i\neq j\\ \|P_{\boldsymbol{U}}^\perp\boldsymbol{v}\|^2 & \text{if } i=j\end{cases}
\end{aligned}\tag{36}$$

This gives us that $\boldsymbol{M}=\|P_{\boldsymbol{U}}^\perp\boldsymbol{v}\|^2\,\boldsymbol{I}_{n_1}$. We can now complete the computation for bias$(\boldsymbol{A})$.

$$\begin{aligned}
\text{bias}(\boldsymbol{A}) &= \mathop{\mathbb{E}}_{\boldsymbol{v},\boldsymbol{X}_{\boldsymbol{U}}}\left[\text{tr}\left(\boldsymbol{X}_{\boldsymbol{U}}(\boldsymbol{X}_{\boldsymbol{U}}^\top\boldsymbol{X}_{\boldsymbol{U}})^{-2}\boldsymbol{X}_{\boldsymbol{U}}^\top\|P_{\boldsymbol{U}}^\perp\boldsymbol{v}\|^2\right)\right]+\|P_{\boldsymbol{U}}^\perp\boldsymbol{A}^*\|^2\\
&= \mathop{\mathbb{E}}_{\boldsymbol{X}_{\boldsymbol{U}}}\left[\text{tr}\left(\boldsymbol{X}_{\boldsymbol{U}}(\boldsymbol{X}_{\boldsymbol{U}}^\top\boldsymbol{X}_{\boldsymbol{U}})^{-2}\boldsymbol{X}_{\boldsymbol{U}}^\top\right)\right]\mathop{\mathbb{E}}_{\boldsymbol{v}}\left[\|P_{\boldsymbol{U}}^\perp\boldsymbol{v}\|^2\right]+\|P_{\boldsymbol{U}}^\perp\boldsymbol{A}^*\|^2
\end{aligned}$$

$$= \mathop{\mathbb{E}}_{\boldsymbol{X_U}} \left[\text{tr}\left((\boldsymbol{X_U}^\top \boldsymbol{X_U})^{-1}\right)\right] \|P_{\boldsymbol{U}}^{\perp} \boldsymbol{A}^*\|^2 + \|P_{\boldsymbol{U}}^{\perp} \boldsymbol{A}^*\|^2$$

We will deal with the $\mathop{\mathbb{E}}_{\boldsymbol{X_U}} \left[\text{tr}\left((\boldsymbol{X_U}^\top \boldsymbol{X_U})^{-1}\right)\right]$ term later and show that it is equal to $\alpha(n_1, r)$.

Variance: We will use many ideas that were used for the bias term. Again using the SVD, we get

$$\begin{aligned}
\text{variance}(\boldsymbol{A}) &= \mathop{\mathbb{E}}_{\boldsymbol{\eta}, \boldsymbol{X}} \left[\|\boldsymbol{A}\left(\boldsymbol{A}^\top \boldsymbol{X}^\top \boldsymbol{X} \boldsymbol{A}\right)^\dagger \boldsymbol{A}^\top \boldsymbol{X}^\top \boldsymbol{\eta}\|^2\right] \\
&=^{(a)} \mathop{\mathbb{E}}_{\boldsymbol{\eta}, \boldsymbol{X_U}} \left[\|\boldsymbol{U}\boldsymbol{S}\boldsymbol{V}^\top \left(\boldsymbol{V}\boldsymbol{S}\boldsymbol{X_U}^\top \boldsymbol{X_U}\boldsymbol{S}\boldsymbol{V}^\top\right)^\dagger \boldsymbol{V}\boldsymbol{S}\boldsymbol{X_U}^\top \boldsymbol{\eta}\|^2\right] \\
&=^{(b)} \mathop{\mathbb{E}}_{\boldsymbol{\eta}, \boldsymbol{X_U}} \left[\|\boldsymbol{U}\boldsymbol{S}\boldsymbol{S}^{-1}(\boldsymbol{X_U}^\top \boldsymbol{X_U})^{-1}\boldsymbol{S}^{-1}\boldsymbol{S}\boldsymbol{X_U}^\top \boldsymbol{\eta}\|^2\right] =^{(c)} \mathop{\mathbb{E}}_{\boldsymbol{\eta}, \boldsymbol{X_U}} \left[\|(\boldsymbol{X_U}^\top \boldsymbol{X_U})^{-1}\boldsymbol{X_U}^\top \boldsymbol{\eta}\|^2\right] \\
&=^{(d)} \mathop{\mathbb{E}}_{\boldsymbol{\eta}, \boldsymbol{X_U}} \left[\text{tr}\left((\boldsymbol{X_U}^\top \boldsymbol{X_U})^{-1}\boldsymbol{X_U}^\top \boldsymbol{\eta}\boldsymbol{\eta}^\top \boldsymbol{X_U}(\boldsymbol{X_U}^\top \boldsymbol{X_U})^{-1}\right)\right] \\
&=^{(e)} \sigma^2 \mathop{\mathbb{E}}_{\boldsymbol{X_U}} \left[\text{tr}\left((\boldsymbol{X_U}^\top \boldsymbol{X_U})^{-1}\boldsymbol{X_U}^\top \boldsymbol{X_U}(\boldsymbol{X_U}^\top \boldsymbol{X_U})^{-1}\right)\right] \\
&=^{(f)} \sigma^2 \mathop{\mathbb{E}}_{\boldsymbol{X_U}} \left[\text{tr}\left((\boldsymbol{X_U}^\top \boldsymbol{X_U})^{-1}\right)\right]
\end{aligned}$$

Here (a) uses SVD, (b) uses the fact that as before $\boldsymbol{X_U}^\top \boldsymbol{X_U}$ is full rank with probability 1, (d) follows from the norm and trace relationship, (e) follows from the noise distribution $\boldsymbol{\eta} \sim \mathcal{N}(\boldsymbol{0}_{n_1}, \sigma^2 \boldsymbol{I}_{n_1})$ and its independence from $\boldsymbol{X}$.

Thus combining the bias and variance terms, we get

$$\mathcal{L}^{\text{tr-val}}_{\lambda,\text{rep}}(\boldsymbol{A}; (n_1, \boldsymbol{a})) - \sigma^2 = \text{bias}(\boldsymbol{A}) + \text{variance}(\boldsymbol{A}) = \left(1 + \mathop{\mathbb{E}}_{\boldsymbol{X_U}} \left[\text{tr}\left((\boldsymbol{X_U}^\top \boldsymbol{X_U})^{-1}\right)\right]\right) \|P_{\boldsymbol{U}}^{\perp}\|^2 + \mathop{\mathbb{E}}_{\boldsymbol{X_U}} \left[\text{tr}\left((\boldsymbol{X_U}^\top \boldsymbol{X_U})^{-1}\right)\right] \sigma^2$$

Thus the only thing remaining to show is that $\mathop{\mathbb{E}}_{\boldsymbol{X_U}} \left[\text{tr}\left((\boldsymbol{X_U}^\top \boldsymbol{X_U})^{-1}\right)\right] = \alpha(n_1, r)$. To show this, we will use the fact that $(\boldsymbol{X_U}^\top \boldsymbol{X_U})^{-1}$ is from the inverse Wishart distribution, and that $\mathop{\mathbb{E}}_{\boldsymbol{X_U}} (\boldsymbol{X_U}^\top \boldsymbol{X_U})^{-1} = \frac{\boldsymbol{I}_{n_1}}{r - n_1 - 1}$ when $r > n_1 + 1$ and unbounded when $r \in \{n_1, n_1 + 1\}$ (Mardia et al., 1979; Belkin et al., 2020). For $r < n_1 - 1$, this gives us $\mathop{\mathbb{E}}_{\boldsymbol{X_U}} \left[\text{tr}\left((\boldsymbol{X_U}^\top \boldsymbol{X_U})^{-1}\right)\right] = \text{tr}\left(\mathop{\mathbb{E}}_{\boldsymbol{X_U}} \left[(\boldsymbol{X_U}^\top \boldsymbol{X_U})^{-1}\right]\right) = \text{tr}\left(\frac{\boldsymbol{I}_{n_1}}{r - n_1 - 1}\right) = \frac{n_1}{r - n_1 - 1}$, which completes the proof. We now prove the result for $r > n_1$.

Case 2: $r > n_1$ In this case, the rank of the representations $\boldsymbol{X}\boldsymbol{A} \in \mathbb{R}^{n_1 \times d}$ for training data is lower than the number of samples. Thus we can use the dual formulation to get the minimum ℓ_2 norm solution for dataset $(\boldsymbol{X}, \boldsymbol{Y})$

$$\lim_{\lambda \to 0} \boldsymbol{w}_\lambda(\boldsymbol{A}; (\boldsymbol{X}, \boldsymbol{Y})) = \mathop{\arg\min}_{\boldsymbol{X}\boldsymbol{A}\boldsymbol{w} = \boldsymbol{Y}} \|\boldsymbol{w}\|_2 = \boldsymbol{A}^\top \boldsymbol{X}^\top \left(\boldsymbol{X}\boldsymbol{A}\boldsymbol{A}^\top \boldsymbol{X}^\top\right)^{-1} \boldsymbol{Y}$$

Plugging this into Equation (35), we get

$$\begin{aligned}
\mathcal{L}^{\text{tr-val}}_{\lambda,\text{rep}}(\boldsymbol{A}; (n_1, n_2)) - \sigma^2 &= \mathop{\mathbb{E}}_{\boldsymbol{v} \sim \mathcal{N}(\boldsymbol{0}_d, \boldsymbol{I}_d)} \left[\mathop{\mathbb{E}}_{\substack{\boldsymbol{X} \sim \mathcal{N}(\boldsymbol{0}_d, \boldsymbol{I}_d)^{n_1}, \\ \boldsymbol{\eta} \sim \mathcal{N}(\boldsymbol{0}_d, \sigma^2 \boldsymbol{I}_{n_1})}} \left[\|\boldsymbol{A}\boldsymbol{A}^\top \boldsymbol{X}^\top \left(\boldsymbol{X}\boldsymbol{A}\boldsymbol{A}^\top \boldsymbol{X}^\top\right)^{-1} (\boldsymbol{X}\boldsymbol{v} + \boldsymbol{\eta}) - \boldsymbol{v}\|^2\right]\right] \\
&=^{(a)} \underbrace{\mathop{\mathbb{E}}_{\boldsymbol{v}, \boldsymbol{X}} \left[\|\boldsymbol{A}\boldsymbol{A}^\top \boldsymbol{X}^\top \left(\boldsymbol{X}\boldsymbol{A}\boldsymbol{A}^\top \boldsymbol{X}^\top\right)^{-1} \boldsymbol{X}\boldsymbol{v} - \boldsymbol{v}\|^2\right]}_{\text{bias}(\boldsymbol{A})} + \underbrace{\mathop{\mathbb{E}}_{\boldsymbol{\eta}, \boldsymbol{X}} \left[\|\boldsymbol{A}\boldsymbol{A}^\top \boldsymbol{X}^\top \left(\boldsymbol{X}\boldsymbol{A}\boldsymbol{A}^\top \boldsymbol{X}^\top\right)^{-1} \boldsymbol{\eta}\|^2\right]}_{\text{variance}(\boldsymbol{A})}
\end{aligned}$$

We again handle the bias and variance terms separately. Let $\boldsymbol{X}_{\boldsymbol{U}^\perp} = \boldsymbol{X} P_{\boldsymbol{U}}^{\perp}$ and we will again use the fact that $\boldsymbol{X_U} = \boldsymbol{X}\boldsymbol{U}\boldsymbol{U}^\top$ and $\boldsymbol{X}_{\boldsymbol{U}^\perp}$ are independent

Bias:

$$\begin{aligned}
\text{bias}(\boldsymbol{A}) &= \mathop{\mathbb{E}}_{\boldsymbol{v},\boldsymbol{X}}\left[\left\|\boldsymbol{A}\boldsymbol{A}^\top\boldsymbol{X}^\top\left(\boldsymbol{X}\boldsymbol{A}\boldsymbol{A}^\top\boldsymbol{X}^\top\right)^{-1}\boldsymbol{X}\boldsymbol{v}-\boldsymbol{v}\right\|^2\right] \\
&= \mathop{\mathbb{E}}_{\boldsymbol{v},\boldsymbol{X}}\left[\left\|\boldsymbol{U}\boldsymbol{S}^2\boldsymbol{U}^\top\boldsymbol{X}^\top\left(\boldsymbol{X}\boldsymbol{U}\boldsymbol{S}^2\boldsymbol{U}^\top\boldsymbol{X}^\top\right)^{-1}\boldsymbol{X}(\boldsymbol{U}\boldsymbol{U}^\top+P_{\boldsymbol{U}}^\perp)\boldsymbol{v}-(\boldsymbol{U}\boldsymbol{U}^\top+P_{\boldsymbol{U}}^\perp)\boldsymbol{v}\right\|^2\right] \\
&= \mathop{\mathbb{E}}_{\boldsymbol{v},\boldsymbol{X}}\left[\left\|\boldsymbol{U}(\boldsymbol{S}^2\boldsymbol{X}_{\boldsymbol{U}}^\top\left(\boldsymbol{X}_{\boldsymbol{U}}\boldsymbol{S}^2\boldsymbol{X}_{\boldsymbol{U}}^\top\right)^{-1}\boldsymbol{X}_{\boldsymbol{U}}-\boldsymbol{I}_r)\boldsymbol{U}^\top\boldsymbol{v}+\boldsymbol{U}\boldsymbol{S}^2\boldsymbol{X}_{\boldsymbol{U}}^\top\left(\boldsymbol{X}_{\boldsymbol{U}}\boldsymbol{S}^2\boldsymbol{X}_{\boldsymbol{U}}^\top\right)^{-1}\boldsymbol{X}P_{\boldsymbol{U}}^\perp\boldsymbol{v}-P_{\boldsymbol{U}}^\perp\boldsymbol{v}\right\|^2\right] \\
&= \mathop{\mathbb{E}}_{\boldsymbol{v},\boldsymbol{X}}\left[\left\|\boldsymbol{U}(\boldsymbol{S}^2\boldsymbol{X}_{\boldsymbol{U}}^\top\left(\boldsymbol{X}_{\boldsymbol{U}}\boldsymbol{S}^2\boldsymbol{X}_{\boldsymbol{U}}^\top\right)^{-1}\boldsymbol{X}_{\boldsymbol{U}}-\boldsymbol{I}_r)\boldsymbol{U}^\top\boldsymbol{v}+\boldsymbol{U}\boldsymbol{S}^2\boldsymbol{X}_{\boldsymbol{U}}^\top\left(\boldsymbol{X}_{\boldsymbol{U}}\boldsymbol{S}^2\boldsymbol{X}_{\boldsymbol{U}}^\top\right)^{-1}\boldsymbol{X}_{\boldsymbol{U}^\perp}P_{\boldsymbol{U}}^\perp\boldsymbol{v}\right\|^2+\|P_{\boldsymbol{U}}^\perp\boldsymbol{v}\|^2\right] \\
&= \mathop{\mathbb{E}}_{\boldsymbol{v},\boldsymbol{X}}\left[\|\boldsymbol{B}_{\boldsymbol{X}_{\boldsymbol{U}}}\boldsymbol{U}^\top\boldsymbol{v}+\boldsymbol{C}_{\boldsymbol{X}_{\boldsymbol{U}}}\boldsymbol{X}_{\boldsymbol{U}^\perp}P_{\boldsymbol{U}}^\perp\boldsymbol{v}\|^2\right]+\|P_{\boldsymbol{U}}^\perp\boldsymbol{A}^*\|^2
\end{aligned}$$

where $\boldsymbol{B}_{\boldsymbol{X}_{\boldsymbol{U}}}=\boldsymbol{U}(\boldsymbol{S}^2\boldsymbol{X}_{\boldsymbol{U}}^\top\left(\boldsymbol{X}_{\boldsymbol{U}}\boldsymbol{S}^2\boldsymbol{X}_{\boldsymbol{U}}^\top\right)^{-1}\boldsymbol{X}_{\boldsymbol{U}}-\boldsymbol{I}_r)$ and $\boldsymbol{C}_{\boldsymbol{X}_{\boldsymbol{U}}}=\boldsymbol{U}\boldsymbol{S}^2\boldsymbol{X}_{\boldsymbol{U}}^\top\left(\boldsymbol{X}_{\boldsymbol{U}}\boldsymbol{S}^2\boldsymbol{X}_{\boldsymbol{U}}^\top\right)^{-1}$ only depends on the components of $\boldsymbol{X}$ in the direction of $\boldsymbol{U}$, i.e $\boldsymbol{X}_{\boldsymbol{U}}$. The main difference from the $r<n_1$ case is that here $\boldsymbol{B}_{\boldsymbol{X}_{\boldsymbol{U}}}\in\mathbb{R}^{r\times r}$ is not zero, since the rank of $\boldsymbol{X}_{\boldsymbol{U}}$ is $\min\{n_1,r\}=n_1$. We can expand the bias term further

$$\begin{aligned}
\text{bias}(\boldsymbol{A}) &= \mathop{\mathbb{E}}_{\boldsymbol{v},\boldsymbol{X}}\left[\|\boldsymbol{B}_{\boldsymbol{X}_{\boldsymbol{U}}}\boldsymbol{U}^\top\boldsymbol{v}\|^2+\|\boldsymbol{C}_{\boldsymbol{X}_{\boldsymbol{U}}}\boldsymbol{X}_{\boldsymbol{U}^\perp}P_{\boldsymbol{U}}^\perp\boldsymbol{v}\|^2+2\text{tr}\left(\boldsymbol{v}^\top\boldsymbol{U}\boldsymbol{B}_{\boldsymbol{X}_{\boldsymbol{U}}}^\top\boldsymbol{C}_{\boldsymbol{X}_{\boldsymbol{U}}}\boldsymbol{X}_{\boldsymbol{U}^\perp}P_{\boldsymbol{U}}^\perp\boldsymbol{v}\right)\right]+\|P_{\boldsymbol{U}}^\perp\boldsymbol{A}^*\|^2 \\
&=^{(a)} \mathop{\mathbb{E}}_{\boldsymbol{v},\boldsymbol{X}_{\boldsymbol{U}}}\left[\|\boldsymbol{B}_{\boldsymbol{X}_{\boldsymbol{U}}}\boldsymbol{U}^\top\boldsymbol{v}\|^2\right]+\mathop{\mathbb{E}}_{\boldsymbol{v},\boldsymbol{X}}\left[\|\boldsymbol{C}_{\boldsymbol{X}_{\boldsymbol{U}}}\boldsymbol{X}_{\boldsymbol{U}^\perp}P_{\boldsymbol{U}}^\perp\boldsymbol{v}\|^2\right]+2\mathop{\mathbb{E}}_{\boldsymbol{v},\boldsymbol{X}_{\boldsymbol{U}}}\left[\text{tr}\left(\boldsymbol{v}^\top\boldsymbol{U}\boldsymbol{B}_{\boldsymbol{X}_{\boldsymbol{U}}}^\top\boldsymbol{C}_{\boldsymbol{X}_{\boldsymbol{U}}}\mathop{\mathbb{E}}_{\boldsymbol{X}_{\boldsymbol{U}^\perp}}\left[\boldsymbol{X}_{\boldsymbol{U}^\perp}\right]P_{\boldsymbol{U}}^\perp\boldsymbol{v}\right)\right]+\|P_{\boldsymbol{U}}^\perp\boldsymbol{A}^*\|^2 \\
&=^{(b)} \mathop{\mathbb{E}}_{\boldsymbol{v},\boldsymbol{X}_{\boldsymbol{U}}}\left[\|\boldsymbol{B}_{\boldsymbol{X}_{\boldsymbol{U}}}\boldsymbol{U}^\top\boldsymbol{v}\|^2\right]+\mathop{\mathbb{E}}_{\boldsymbol{v},\boldsymbol{X}}\left[\|\boldsymbol{C}_{\boldsymbol{X}_{\boldsymbol{U}}}\boldsymbol{X}_{\boldsymbol{U}^\perp}P_{\boldsymbol{U}}^\perp\boldsymbol{v}\|^2\right]+\|P_{\boldsymbol{U}}^\perp\boldsymbol{A}^*\|^2 \\
&=^{(c)} \mathop{\mathbb{E}}_{\boldsymbol{v},\boldsymbol{X}_{\boldsymbol{U}}}\left[\|\boldsymbol{B}_{\boldsymbol{X}_{\boldsymbol{U}}}\boldsymbol{U}^\top\boldsymbol{v}\|^2\right]+\mathop{\mathbb{E}}_{\boldsymbol{v},\boldsymbol{X}}\left[\text{tr}\left(\boldsymbol{v}^\top P_{\boldsymbol{U}}^\perp\boldsymbol{X}_{\boldsymbol{U}^\perp}^\top\boldsymbol{C}_{\boldsymbol{X}_{\boldsymbol{U}}}^\top\boldsymbol{C}_{\boldsymbol{X}_{\boldsymbol{U}}}\boldsymbol{X}_{\boldsymbol{U}^\perp}P_{\boldsymbol{U}}^\perp\boldsymbol{v}\right)\right]+\|P_{\boldsymbol{U}}^\perp\boldsymbol{A}^*\|^2 \\
&=^{(d)} \mathop{\mathbb{E}}_{\boldsymbol{v},\boldsymbol{X}_{\boldsymbol{U}}}\left[\|\boldsymbol{B}_{\boldsymbol{X}_{\boldsymbol{U}}}\boldsymbol{U}^\top\boldsymbol{v}\|^2\right]+\mathop{\mathbb{E}}_{\boldsymbol{v},\boldsymbol{X}_{\boldsymbol{U}}}\left[\text{tr}\left(\boldsymbol{C}_{\boldsymbol{X}_{\boldsymbol{U}}}^\top\boldsymbol{C}_{\boldsymbol{X}_{\boldsymbol{U}}}\mathop{\mathbb{E}}_{\boldsymbol{X}_{\boldsymbol{U}^\perp}}\left[\boldsymbol{X}_{\boldsymbol{U}^\perp}P_{\boldsymbol{U}}^\perp\boldsymbol{v}\boldsymbol{v}^\top P_{\boldsymbol{U}}^\perp\boldsymbol{X}_{\boldsymbol{U}^\perp}^\top\right]\right)\right]+\|P_{\boldsymbol{U}}^\perp\boldsymbol{A}^*\|^2 \\
&=^{(e)} \mathop{\mathbb{E}}_{\boldsymbol{v},\boldsymbol{X}_{\boldsymbol{U}}}\left[\|\boldsymbol{B}_{\boldsymbol{X}_{\boldsymbol{U}}}\boldsymbol{U}^\top\boldsymbol{v}\|^2\right]+\mathop{\mathbb{E}}_{\boldsymbol{X}_{\boldsymbol{U}}}\left[\text{tr}\left(\boldsymbol{C}_{\boldsymbol{X}_{\boldsymbol{U}}}^\top\boldsymbol{C}_{\boldsymbol{X}_{\boldsymbol{U}}}\right)\right]\|P_{\boldsymbol{U}}^\perp\boldsymbol{A}^*\|^2+\|P_{\boldsymbol{U}}^\perp\boldsymbol{A}^*\|^2
\end{aligned}\tag{37}$$

where (a) uses the fact that $\boldsymbol{X}_{\boldsymbol{U}^\perp}$ is independent of $\boldsymbol{X}_{\boldsymbol{U}}$ and $\boldsymbol{v}$, (b) uses the Gaussianity of $\boldsymbol{X}_{\boldsymbol{U}^\perp}$ and that it has mean $\boldsymbol{0}$, (c) uses $\|\boldsymbol{a}\|^2=\text{tr}(\boldsymbol{a}\boldsymbol{a}^\top)$, (d) uses the independence of $\boldsymbol{X}_{\boldsymbol{U}}$ and $\boldsymbol{X}_{\boldsymbol{U}^\perp}$ and (e) uses the calculation from Equation (36). We first tackle the $\boldsymbol{C}_{\boldsymbol{X}_{\boldsymbol{U}}}$ term

$$\begin{aligned}
\mathop{\mathbb{E}}_{\boldsymbol{X}_{\boldsymbol{U}}}\left[\text{tr}\left(\boldsymbol{C}_{\boldsymbol{X}_{\boldsymbol{U}}}^\top\boldsymbol{C}_{\boldsymbol{X}_{\boldsymbol{U}}}\right)\right] &= \mathop{\mathbb{E}}_{\boldsymbol{X}_{\boldsymbol{U}}}\left[\text{tr}\left(\left(\boldsymbol{X}_{\boldsymbol{U}}\boldsymbol{S}^2\boldsymbol{X}_{\boldsymbol{U}}^\top\right)^{-1}\boldsymbol{X}_{\boldsymbol{U}}\boldsymbol{S}^2\boldsymbol{U}^\top\boldsymbol{U}\boldsymbol{S}^2\boldsymbol{X}_{\boldsymbol{U}}^\top\left(\boldsymbol{X}_{\boldsymbol{U}}\boldsymbol{S}^2\boldsymbol{X}_{\boldsymbol{U}}^\top\right)^{-1}\right)\right] \\
&= \mathop{\mathbb{E}}_{\boldsymbol{X}_{\boldsymbol{U}}}\left[\text{tr}\left(\left(\boldsymbol{X}_{\boldsymbol{U}}\boldsymbol{S}^2\boldsymbol{X}_{\boldsymbol{U}}^\top\right)^{-1}\boldsymbol{X}_{\boldsymbol{U}}\boldsymbol{S}^4\boldsymbol{X}_{\boldsymbol{U}}^\top\left(\boldsymbol{X}_{\boldsymbol{U}}\boldsymbol{S}^2\boldsymbol{X}_{\boldsymbol{U}}^\top\right)^{-1}\right)\right]
\end{aligned}\tag{38}$$

We will encounter this function again in the variance term and we will tackle this later. At a high level, we will show that this term is going to be at least as large as the term for $\boldsymbol{S}=\boldsymbol{I}_r$, which reduces to $\mathop{\mathbb{E}}_{\boldsymbol{X}_{\boldsymbol{U}}}\left[\text{tr}\left(\left(\boldsymbol{X}_{\boldsymbol{U}}\boldsymbol{X}_{\boldsymbol{U}}^\top\right)^{-1}\right)\right]$ which has a closed form expression, again using inverse Wishart distribution. For now we will first deal with the $\boldsymbol{B}_{\boldsymbol{X}_{\boldsymbol{U}}}$ term.

$$\begin{aligned}
\mathop{\mathbb{E}}_{\boldsymbol{v},\boldsymbol{X}_{\boldsymbol{U}}}\left[\|\boldsymbol{B}_{\boldsymbol{X}_{\boldsymbol{U}}}\boldsymbol{U}^\top\boldsymbol{v}\|^2\right] &= \mathop{\mathbb{E}}_{\substack{\boldsymbol{u}\sim\mathcal{N}(\boldsymbol{0}_r,\boldsymbol{U}^\top\boldsymbol{A}^*\boldsymbol{A}^{*\top}\boldsymbol{U}),\\ \boldsymbol{X}_{\boldsymbol{U}}}}\left[\|\boldsymbol{B}_{\boldsymbol{X}_{\boldsymbol{U}}}\boldsymbol{u}\|^2\right]=\mathop{\mathbb{E}}_{\boldsymbol{u},\boldsymbol{X}_{\boldsymbol{U}}}\left[\|\boldsymbol{U}(\boldsymbol{S}^2\boldsymbol{X}_{\boldsymbol{U}}^\top\left(\boldsymbol{X}_{\boldsymbol{U}}\boldsymbol{S}^2\boldsymbol{X}_{\boldsymbol{U}}^\top\right)^{-1}\boldsymbol{X}_{\boldsymbol{U}}-\boldsymbol{I}_r)\boldsymbol{u}\|^2\right] \\
&= \mathop{\mathbb{E}}_{\boldsymbol{u},\boldsymbol{X}_{\boldsymbol{U}}}\left[\|(\boldsymbol{S}^2\boldsymbol{X}_{\boldsymbol{U}}^\top\left(\boldsymbol{X}_{\boldsymbol{U}}\boldsymbol{S}^2\boldsymbol{X}_{\boldsymbol{U}}^\top\right)^{-1}\boldsymbol{X}_{\boldsymbol{U}}-\boldsymbol{I}_r)\boldsymbol{u}\|^2\right]
\end{aligned}$$

We are now going to exploit the symmetry in Gaussian distribution once again. Recall that $\boldsymbol{X}_{\boldsymbol{U}}\sim\mathcal{N}(\boldsymbol{0}_r,\boldsymbol{I}_r)$. We notice that $\boldsymbol{X}_{\boldsymbol{U}}\equiv_D\boldsymbol{X}_{\boldsymbol{U}}\boldsymbol{P}\boldsymbol{D}$, where $\boldsymbol{P}$ is a random permutation matrix and $\boldsymbol{D}$ is a diagonal matrix with random entries in ±1. Essentially this is saying that randomly shuffling the coordinates and multiplying each coordinate by a random

sign results in the same isotropic Gaussian distribution. We observe that $\boldsymbol{P}\boldsymbol{D}\boldsymbol{S}^2\boldsymbol{D}\boldsymbol{P}^\top$ for diagonal matrix $\boldsymbol{S}$ and that $\boldsymbol{P}\boldsymbol{P}^\top = \boldsymbol{P}^\top\boldsymbol{P} = \boldsymbol{D}^2 = \boldsymbol{I}_r$ rewrite the above expectation.

$$
\begin{aligned}
&\mathop{\mathbb{E}}_{\boldsymbol{v},\boldsymbol{X}_U}\left[\|\boldsymbol{B}_{\boldsymbol{X}_U}\boldsymbol{U}^\top\boldsymbol{v}\|^2\right] = \mathop{\mathbb{E}}_{\boldsymbol{u},\boldsymbol{X}_U}\left[\|(\boldsymbol{S}^2\boldsymbol{X}_U^\top\left(\boldsymbol{X}_U\boldsymbol{S}^2\boldsymbol{X}_U^\top\right)^{-1}\boldsymbol{X}_U - \boldsymbol{I}_r)\boldsymbol{u}\|^2\right] \\
&= \mathop{\mathbb{E}}_{\boldsymbol{u},\boldsymbol{X}_U,\boldsymbol{P},\boldsymbol{D}}\left[\|(\boldsymbol{S}^2\boldsymbol{D}\boldsymbol{P}^\top\boldsymbol{X}_U^\top\left(\boldsymbol{X}_U\boldsymbol{P}\boldsymbol{D}\boldsymbol{S}^2\boldsymbol{D}\boldsymbol{P}^\top\boldsymbol{X}_U^\top\right)^{-1}\boldsymbol{X}_U\boldsymbol{P}\boldsymbol{D} - \boldsymbol{I}_r)\boldsymbol{u}\|^2\right] \\
&= \mathop{\mathbb{E}}_{\boldsymbol{u},\boldsymbol{X}_U,\boldsymbol{P},\boldsymbol{D}}\left[\|(\boldsymbol{S}^2\boldsymbol{D}\boldsymbol{P}^\top\boldsymbol{X}_U^\top\left(\boldsymbol{X}_U\boldsymbol{S}^2\boldsymbol{X}_U^\top\right)^{-1}\boldsymbol{X}_U\boldsymbol{P}\boldsymbol{D} - \boldsymbol{D}\boldsymbol{P}^\top\boldsymbol{P}\boldsymbol{D})\boldsymbol{u}\|^2\right] \\
&= \mathop{\mathbb{E}}_{\boldsymbol{u},\boldsymbol{X}_U,\boldsymbol{P},\boldsymbol{D}}\left[\|\boldsymbol{D}\boldsymbol{P}^\top(\boldsymbol{P}\boldsymbol{D}\boldsymbol{S}^2\boldsymbol{D}\boldsymbol{P}^\top\boldsymbol{X}_U^\top\left(\boldsymbol{X}_U\boldsymbol{S}^2\boldsymbol{X}_U^\top\right)^{-1}\boldsymbol{X}_U - \boldsymbol{I}_r)\boldsymbol{P}\boldsymbol{D}\boldsymbol{u}\|^2\right] \\
&= \mathop{\mathbb{E}}_{\boldsymbol{u},\boldsymbol{X}_U,\boldsymbol{P},\boldsymbol{D}}\left[\|\boldsymbol{D}\boldsymbol{P}^\top(\boldsymbol{S}^2\boldsymbol{X}_U^\top\left(\boldsymbol{X}_U\boldsymbol{S}^2\boldsymbol{X}_U^\top\right)^{-1}\boldsymbol{X}_U - \boldsymbol{I}_r)\boldsymbol{P}\boldsymbol{D}\boldsymbol{u}\|^2\right] \\
&= \mathop{\mathbb{E}}_{\boldsymbol{u},\boldsymbol{X}_U,\boldsymbol{P},\boldsymbol{D}}\left[\|(\boldsymbol{S}^2\boldsymbol{X}_U^\top\left(\boldsymbol{X}_U\boldsymbol{S}^2\boldsymbol{X}_U^\top\right)^{-1}\boldsymbol{X}_U - \boldsymbol{I}_r)\boldsymbol{P}\boldsymbol{D}\boldsymbol{u}\|^2\right] \\
&= \mathop{\mathbb{E}}_{\boldsymbol{u},\boldsymbol{X}_U}\left[\mathrm{tr}\left((\boldsymbol{S}^2\boldsymbol{X}_U^\top\left(\boldsymbol{X}_U\boldsymbol{S}^2\boldsymbol{X}_U^\top\right)^{-1}\boldsymbol{X}_U - \boldsymbol{I}_r)\mathop{\mathbb{E}}_{\boldsymbol{P},\boldsymbol{D}}\left[\boldsymbol{P}\boldsymbol{D}\boldsymbol{u}\boldsymbol{u}^\top\boldsymbol{D}\boldsymbol{P}^\top\right](\boldsymbol{S}^2\boldsymbol{X}_U^\top\left(\boldsymbol{X}_U\boldsymbol{S}^2\boldsymbol{X}_U^\top\right)^{-1}\boldsymbol{X}_U - \boldsymbol{I}_r)^\top\right)\right]
\end{aligned}
$$

It is not hard to see that randomly multiplying coordinates of $\boldsymbol{u}$ by ± 1 and then shuffling the coordinates will lead to $\mathop{\mathbb{E}}_{\boldsymbol{P},\boldsymbol{D}}\left[\boldsymbol{P}\boldsymbol{D}\boldsymbol{u}\boldsymbol{u}^\top\boldsymbol{D}\boldsymbol{P}^\top\right] = \frac{\|\boldsymbol{u}\|^2}{r}\boldsymbol{I}_r$.

$$
\begin{aligned}
\mathop{\mathbb{E}}_{\boldsymbol{v},\boldsymbol{X}_U}\left[\|\boldsymbol{B}_{\boldsymbol{X}_U}\boldsymbol{U}^\top\boldsymbol{v}\|^2\right] &= \mathop{\mathbb{E}}_{\boldsymbol{u}\sim\mathcal{N}(\boldsymbol{0}_r,\boldsymbol{U}^\top\boldsymbol{A}^*\boldsymbol{A}^{*\top}\boldsymbol{U})}\left[\frac{\|\boldsymbol{u}\|^2}{r}\right]\mathop{\mathbb{E}}_{\boldsymbol{X}_U}\left[\|\boldsymbol{S}^2\boldsymbol{X}_U^\top\left(\boldsymbol{X}_U\boldsymbol{S}^2\boldsymbol{X}_U^\top\right)^{-1}\boldsymbol{X}_U - \boldsymbol{I}_r\|^2\right] \\
&= \frac{\|P_U\boldsymbol{A}^*\|^2}{r}\mathop{\mathbb{E}}_{\boldsymbol{X}_U}\left[\|\boldsymbol{S}^2\boldsymbol{X}_U^\top\left(\boldsymbol{X}_U\boldsymbol{S}^2\boldsymbol{X}_U^\top\right)^{-1}\boldsymbol{X}_U P_{\boldsymbol{X}_U} - P_{\boldsymbol{X}_U} - P_{\boldsymbol{X}_U}^\perp\|^2\right] \\
&= \frac{\|P_U\boldsymbol{A}^*\|^2}{r}\mathop{\mathbb{E}}_{\boldsymbol{X}_U}\left[\|\boldsymbol{S}^2\boldsymbol{X}_U^\top\left(\boldsymbol{X}_U\boldsymbol{S}^2\boldsymbol{X}_U^\top\right)^{-1}\boldsymbol{X}_U - P_{\boldsymbol{X}_U}\|^2 + \|P_{\boldsymbol{X}_U}^\perp\|^2\right] \\
&= \beta_1(\boldsymbol{S}) + \frac{\|P_U\boldsymbol{A}^*\|^2}{r}\mathop{\mathbb{E}}_{\boldsymbol{X}_U}\left[\mathrm{rank}(P_{\boldsymbol{X}_U}^\perp)\right] \\
&= \beta_1(\boldsymbol{S}) + \frac{r-n_1}{r}\|P_U\boldsymbol{A}^*\|^2
\end{aligned} \tag{39}
$$

where $\beta_1(\boldsymbol{S}) = \frac{\|P_U\boldsymbol{A}^*\|^2}{r}\mathop{\mathbb{E}}_{\boldsymbol{X}_U}\left[\|\boldsymbol{S}^2\boldsymbol{X}_U^\top\left(\boldsymbol{X}_U\boldsymbol{S}^2\boldsymbol{X}_U^\top\right)^{-1}\boldsymbol{X}_U - P_{\boldsymbol{X}_U}\|^2\right]$ is a function that satisfies $\beta_1(\boldsymbol{S}) \geq 0$ and $\beta_1(\kappa\boldsymbol{I}_r) = 0$ for any κ. Also we used that $\mathrm{rank}(P_{\boldsymbol{X}_U}^\perp) = r - \mathrm{rank}(\boldsymbol{X}_U) = r - n_1$ with probability 1.

Plugging Equations (38) and (39) into Equation (37), we get the following final expression for the bias

$$
\begin{aligned}
\mathrm{bias}(\boldsymbol{A}) &= \left(1 + \mathop{\mathbb{E}}_{\boldsymbol{X}_U}\left[\mathrm{tr}\left(\left(\boldsymbol{X}_U\boldsymbol{S}^2\boldsymbol{X}_U^\top\right)^{-1}\boldsymbol{X}_U\boldsymbol{S}^4\boldsymbol{X}_U^\top\left(\boldsymbol{X}_U\boldsymbol{S}^2\boldsymbol{X}_U^\top\right)^{-1}\right)\right]\right)\|P_U^\perp\boldsymbol{A}^*\|^2 + \frac{r-n_1}{r}\|P_U\boldsymbol{A}^*\|^2 + \beta_1(\boldsymbol{S}) \\
&= \left(1 + \mathop{\mathbb{E}}_{\boldsymbol{X}_U}\left[\mathrm{tr}\left(\left(\boldsymbol{X}_U\boldsymbol{S}^2\boldsymbol{X}_U^\top\right)^{-1}\boldsymbol{X}_U\boldsymbol{S}^4\boldsymbol{X}_U^\top\left(\boldsymbol{X}_U\boldsymbol{S}^2\boldsymbol{X}_U^\top\right)^{-1}\right)\right]\right)\|P_U^\perp\boldsymbol{A}^*\|^2 + \frac{r-n_1}{r}\left(1 - \|P_U^\perp\boldsymbol{A}^*\|^2\right) + \beta_1(\boldsymbol{S}) \\
&= \left(1 + \mathop{\mathbb{E}}_{\boldsymbol{X}_U}\left[\mathrm{tr}\left(\left(\boldsymbol{X}_U\boldsymbol{S}^2\boldsymbol{X}_U^\top\right)^{-1}\boldsymbol{X}_U\boldsymbol{S}^4\boldsymbol{X}_U^\top\left(\boldsymbol{X}_U\boldsymbol{S}^2\boldsymbol{X}_U^\top\right)^{-1}\right)\right]\right)\|P_U^\perp\boldsymbol{A}^*\|^2 - \frac{r-n_1}{r}\Bigg)\|P_U^\perp\boldsymbol{A}^*\|^2 + \frac{r-n_1}{r} + \beta_1(\boldsymbol{S}) \\
&= \left(\frac{n_1}{r} + \mathop{\mathbb{E}}_{\boldsymbol{X}_U}\left[\mathrm{tr}\left(\left(\boldsymbol{X}_U\boldsymbol{S}^2\boldsymbol{X}_U^\top\right)^{-1}\boldsymbol{X}_U\boldsymbol{S}^4\boldsymbol{X}_U^\top\left(\boldsymbol{X}_U\boldsymbol{S}^2\boldsymbol{X}_U^\top\right)^{-1}\right)\right]\right)\|P_U^\perp\boldsymbol{A}^*\|^2 + \frac{r-n_1}{r} + \beta_1(\boldsymbol{S})
\end{aligned} \tag{40}
$$

Variance: We now move to the variance term

$$
\begin{aligned}
\mathrm{variance}(\boldsymbol{A}) &= \mathop{\mathbb{E}}_{\boldsymbol{\eta},\boldsymbol{X}}\left[\|\boldsymbol{A}\boldsymbol{A}^\top\boldsymbol{X}^\top\left(\boldsymbol{X}\boldsymbol{A}\boldsymbol{A}^\top\boldsymbol{X}^\top\right)^{-1}\boldsymbol{\eta}\|^2\right] \\
&= \sigma^2\mathop{\mathbb{E}}_{\boldsymbol{X}}\left[\mathrm{tr}\left(\boldsymbol{A}\boldsymbol{A}^\top\boldsymbol{X}^\top\left(\boldsymbol{X}\boldsymbol{A}\boldsymbol{A}^\top\boldsymbol{X}^\top\right)^{-2}\boldsymbol{X}\boldsymbol{A}\boldsymbol{A}^\top\right)\right]
\end{aligned}
$$

$$
\begin{aligned}
&= \sigma^2 \mathop{\mathbb{E}}_{\boldsymbol{X}} \left[\operatorname{tr} \left(\boldsymbol{U}\boldsymbol{S}^2\boldsymbol{U}^\top \boldsymbol{X}^\top \left(\boldsymbol{X}\boldsymbol{U}\boldsymbol{S}^2\boldsymbol{U}^\top \boldsymbol{X}^\top \right)^{-2} \boldsymbol{X}\boldsymbol{U}\boldsymbol{S}^2\boldsymbol{U}^\top \right) \right] \\
&= \sigma^2 \mathop{\mathbb{E}}_{\boldsymbol{X}_{\boldsymbol{U}}} \left[\operatorname{tr} \left(\boldsymbol{S}^2 \boldsymbol{X}_{\boldsymbol{U}}^\top \left(\boldsymbol{X}_{\boldsymbol{U}} \boldsymbol{S}^2 \boldsymbol{X}_{\boldsymbol{U}}^\top \right)^{-2} \boldsymbol{X}_{\boldsymbol{U}} \boldsymbol{S}^2 \right) \right] \\
&= \sigma^2 \mathop{\mathbb{E}}_{\boldsymbol{X}_{\boldsymbol{U}}} \left[\operatorname{tr} \left(\left(\boldsymbol{X}_{\boldsymbol{U}} \boldsymbol{S}^2 \boldsymbol{X}_{\boldsymbol{U}}^\top \right)^{-1} \boldsymbol{X}_{\boldsymbol{U}} \boldsymbol{S}^4 \boldsymbol{X}_{\boldsymbol{U}}^\top \left(\boldsymbol{X}_{\boldsymbol{U}} \boldsymbol{S}^2 \boldsymbol{X}_{\boldsymbol{U}}^\top \right)^{-1} \right) \right] \qquad (41)
\end{aligned}
$$

To further simply both, the bias and variance terms, we need the following result

Lemma C.2. *For $\boldsymbol{X}_{\boldsymbol{U}} \sim \mathcal{N}(\boldsymbol{0}_r, \boldsymbol{I}_r)^{n_1}$ and diagonal matrix $\boldsymbol{S} \in \mathbb{R}^{r \times r}$*

$$
\mathop{\mathbb{E}}_{\boldsymbol{X}_{\boldsymbol{U}}} \left[tr \left(\left(\boldsymbol{X}_{\boldsymbol{U}} \boldsymbol{S}^2 \boldsymbol{X}_{\boldsymbol{U}}^\top \right)^{-1} \boldsymbol{X}_{\boldsymbol{U}} \boldsymbol{S}^4 \boldsymbol{X}_{\boldsymbol{U}}^\top \left(\boldsymbol{X}_{\boldsymbol{U}} \boldsymbol{S}^2 \boldsymbol{X}_{\boldsymbol{U}}^\top \right)^{-1} \right) \right] = \begin{cases} \alpha(r, n_1) + \beta_2(\boldsymbol{S}) & \text{if } r > n_1 + 1 \\ \infty & \text{if } r \in \{n_1, n_1 + 1\} \end{cases}
$$

where $\alpha(r, n_1)$ is defined in Equation (34) and $\beta_2(\boldsymbol{S}) \geq 0$ and $\beta_2(\kappa \boldsymbol{I}_r) = 0$ for any $\kappa > 0$.

Using Lemma C.2 and plugging it into Equations (40) and (41), we get

$$
\mathcal{L}^{\text{tr-val}}_{\lambda,\text{rep}}(\boldsymbol{A}; (n_1, \boldsymbol{a})) - \sigma^2 = \text{bias}(\boldsymbol{A}) + \text{variance}(\boldsymbol{A}) = \left(\frac{n_1}{r} + \alpha(r, n_1) \right) \|P_{\boldsymbol{U}}^{\perp} \boldsymbol{A}^*\|^2 + \frac{r - n_1}{r} + \sigma^2 \alpha(r, n_1) + \beta(\boldsymbol{S})
$$

where $\beta(\boldsymbol{S}) = \beta_1(\boldsymbol{S}) + \beta_2(\boldsymbol{S}) \left(\|P_{\boldsymbol{U}}^{\perp} \boldsymbol{A}^*\|^2 + \sigma^2 \right)$ is a non-negative function that is 0 at $\kappa \boldsymbol{I}_r$ for all $\kappa > 0$.

We now complete the proof by proving Lemma C.2

Proof of Lemma C.2. Let the L.H.S. be $\gamma(\boldsymbol{S}) = \mathop{\mathbb{E}}_{\boldsymbol{X}_{\boldsymbol{U}}} \left[\operatorname{tr} \left(\left(\boldsymbol{X}_{\boldsymbol{U}} \boldsymbol{S}^2 \boldsymbol{X}_{\boldsymbol{U}}^\top \right)^{-1} \boldsymbol{X}_{\boldsymbol{U}} \boldsymbol{S}^4 \boldsymbol{X}_{\boldsymbol{U}}^\top \left(\boldsymbol{X}_{\boldsymbol{U}} \boldsymbol{S}^2 \boldsymbol{X}_{\boldsymbol{U}}^\top \right)^{-1} \right) \right]$. We first show that equality holds for $\boldsymbol{S} = \kappa \boldsymbol{I}_d$. In this case, we have $\gamma(\kappa \boldsymbol{I}_r) = \mathop{\mathbb{E}}_{\boldsymbol{X}_{\boldsymbol{U}}} \left[\operatorname{tr} \left(\left(\boldsymbol{X}_{\boldsymbol{U}} \boldsymbol{X}_{\boldsymbol{U}}^\top \right)^{-1} \boldsymbol{X}_{\boldsymbol{U}} \boldsymbol{X}_{\boldsymbol{U}}^\top \left(\boldsymbol{X}_{\boldsymbol{U}} \boldsymbol{X}_{\boldsymbol{U}}^\top \right)^{-1} \right) \right] = \mathop{\mathbb{E}}_{\boldsymbol{X}_{\boldsymbol{U}}} \left[\operatorname{tr} \left(\left(\boldsymbol{X}_{\boldsymbol{U}} \boldsymbol{X}_{\boldsymbol{U}}^\top \right)^{-1} \right) \right]$. Using the closed-form expression for inverse Wishart distribution once again, we conclude

$$
\gamma(\kappa \boldsymbol{I}_r) = \mathop{\mathbb{E}}_{\boldsymbol{X}_{\boldsymbol{U}}} \left[\operatorname{tr} \left(\left(\boldsymbol{X}_{\boldsymbol{U}} \boldsymbol{X}_{\boldsymbol{U}}^\top \right)^{-1} \right) \right] = \operatorname{tr} \left(\mathop{\mathbb{E}}_{\boldsymbol{X}_{\boldsymbol{U}}} \left[\left(\boldsymbol{X}_{\boldsymbol{U}} \boldsymbol{X}_{\boldsymbol{U}}^\top \right)^{-1} \right] \right) = \begin{cases} \operatorname{tr} \left(\frac{\boldsymbol{I}_{n_1}}{r - n_1 - 1} \right) = \frac{n_1}{r - n_1 - 1} & \text{if } r > n_1 + 1 \\ \infty & \text{otherwise} \end{cases}
$$

Thus we get $\gamma(\kappa \boldsymbol{I}_r) = \alpha(r, n_1)$ when $r > n_1 + 1$ and unbounded otherwise.

We will show for an arbitrary diagonal matrix $\boldsymbol{S}$ that the value is at least as large as $\boldsymbol{I}_r$, i.e. $\gamma(\boldsymbol{S}) \geq \gamma(\kappa \boldsymbol{I}_r)$, which will complete the proof. We first observe that $\boldsymbol{X}_{\boldsymbol{U}} \boldsymbol{S}^4 \boldsymbol{X}_{\boldsymbol{U}}^\top = \boldsymbol{X}_{\boldsymbol{U}} \boldsymbol{S}^2 \boldsymbol{S}^2 \boldsymbol{X}_{\boldsymbol{U}}^\top \succcurlyeq \boldsymbol{X}_{\boldsymbol{U}} \boldsymbol{S}^2 P_{\boldsymbol{X}_{\boldsymbol{U}}} \boldsymbol{S}^2 \boldsymbol{X}_{\boldsymbol{U}}^\top = \boldsymbol{X}_{\boldsymbol{U}} \boldsymbol{S}^2 \boldsymbol{X}_{\boldsymbol{U}}^\top \left(\boldsymbol{X}_{\boldsymbol{U}} \boldsymbol{X}_{\boldsymbol{U}}^\top \right)^{-1} \boldsymbol{X}_{\boldsymbol{U}} \boldsymbol{S}^2 \boldsymbol{X}_{\boldsymbol{U}}^\top$. Using this, we get

$$
\begin{aligned}
\gamma(\boldsymbol{S}) &= \mathop{\mathbb{E}}_{\boldsymbol{X}_{\boldsymbol{U}}} \left[\operatorname{tr} \left(\left(\boldsymbol{X}_{\boldsymbol{U}} \boldsymbol{S}^2 \boldsymbol{X}_{\boldsymbol{U}}^\top \right)^{-1} \boldsymbol{X}_{\boldsymbol{U}} \boldsymbol{S}^4 \boldsymbol{X}_{\boldsymbol{U}}^\top \left(\boldsymbol{X}_{\boldsymbol{U}} \boldsymbol{S}^2 \boldsymbol{X}_{\boldsymbol{U}}^\top \right)^{-1} \right) \right] \\
&\geq \mathop{\mathbb{E}}_{\boldsymbol{X}_{\boldsymbol{U}}} \left[\operatorname{tr} \left(\left(\boldsymbol{X}_{\boldsymbol{U}} \boldsymbol{S}^2 \boldsymbol{X}_{\boldsymbol{U}}^\top \right)^{-1} \boldsymbol{X}_{\boldsymbol{U}} \boldsymbol{S}^2 \boldsymbol{X}_{\boldsymbol{U}}^\top \left(\boldsymbol{X}_{\boldsymbol{U}} \boldsymbol{X}_{\boldsymbol{U}}^\top \right)^{-1} \boldsymbol{X}_{\boldsymbol{U}} \boldsymbol{S}^2 \boldsymbol{X}_{\boldsymbol{U}}^\top \left(\boldsymbol{X}_{\boldsymbol{U}} \boldsymbol{S}^2 \boldsymbol{X}_{\boldsymbol{U}}^\top \right)^{-1} \right) \right] \\
&\geq \mathop{\mathbb{E}}_{\boldsymbol{X}_{\boldsymbol{U}}} \left[\operatorname{tr} \left(\left(\boldsymbol{X}_{\boldsymbol{U}} \boldsymbol{X}_{\boldsymbol{U}}^\top \right)^{-1} \right) \right] \\
&= \gamma(\boldsymbol{I}_r)
\end{aligned}
$$

□

□

D. Experiments

This section contains experimental details and also additional experiments on more datasets and settings. The code for all experiments will be made public.

Algorithm 1 RepLearn(θ^{rep}, $TaskLoader$, $variant$)

Parameters: Regularization (λ), Outer steps (T_{out}), outer lr (η_{out}), batch size (b), Inner steps (T_{in}), inner lr (η_{in})
Input: Representation model θ^{rep}, $TaskLoader$, $variant$ {tr-tr or tr-val}
$\theta^{\text{rep}}(0) \leftarrow RandInit$

for $t = 0$ **to** $T_{\text{out}} - 1$ **do**
  $TaskBatch \leftarrow TaskLoader(batch\_size = b)$

  for $i = 1$ **to** b **do**
    $S \leftarrow TaskBatch[i]$ {Dataset of size n}

    if $variant$ is 'tr-val' **then**
      $(S^{\text{tr}}, S^{\text{val}}) \leftarrow_{\text{split}} S$ {Split into tr-val sets of sizes $n_1 + n_2 = n$}
      $\widetilde{\boldsymbol{W}} \leftarrow_{\text{stop gradient}} \text{InnerLoop}(\theta^{\text{rep}}(t), S^{\text{tr}}, \lambda)$ {Ignore the dependence of $\widetilde{\boldsymbol{W}}$ on $\theta^{\text{rep}}(t)$}
      $\nabla_i \leftarrow \nabla_{\theta^{\text{rep}}} \ell_{\text{rep}}(\widetilde{\boldsymbol{W}}; \theta^{\text{rep}}, S^{\text{val}})|_{\theta^{\text{rep}}(t)}$

    else if $variant$ is 'tr-tr' **then**
      $\widetilde{\boldsymbol{W}} \leftarrow_{\text{stop gradient}} \text{InnerLoop}(\theta^{\text{rep}}(t), S, \lambda)$
      $\nabla_i \leftarrow \nabla_{\theta^{\text{rep}}} \ell_{\text{rep}}(\widetilde{\boldsymbol{W}}; \theta^{\text{rep}}, S)|_{\theta^{\text{rep}}(t)}$ {Defined in Equation (42)}

    end if

    $\nabla = \frac{1}{b} \sum_{i=1}^{b} [\nabla_i]$
    $\theta^{\text{rep}}(t+1) \leftarrow \textbf{Adam}(\theta^{\text{rep}}(t), \nabla, \eta_{\text{in}})$
  end for
end for
return $\theta^{\text{rep}}(T_{\text{out}})$

Algorithm 2 InnerLoop(θ^{rep}, S)

Parameters: Regularization (λ), Inner steps (T_{in}), inner lr (η_{in})
Input: Representation model θ^{rep}, S
$\boldsymbol{W}(0) \leftarrow \mathbf{0}_{d \times k}$
for $t = 0$ **to** $T_{\text{in}} - 1$ **do**
  $\nabla \leftarrow \nabla_{\boldsymbol{W}} \{\ell_{\text{rep}}(\boldsymbol{W}; \theta^{\text{rep}}, S) + \frac{\lambda}{2} \|\boldsymbol{W}\|_F^2\}|_{\boldsymbol{W}(t)}$
  $\boldsymbol{W}(t+1) \leftarrow \textbf{SGD}(\boldsymbol{W}(t), \nabla, \eta_{\text{in}}, \text{momentum} = 0.9)$
end for
return $\boldsymbol{W}(T_{\text{in}})$

D.1. RepLearn Algorithm and Datasets

RepLearn We describe the inner and outer loops for the RepLearn algorithm that we use for experiments in Algorithms 2 and 1 respectively. Recall the definitions for inner and outer losses.

$$\ell_{\text{rep}}(\boldsymbol{W}; \theta^{\text{rep}}, S) = \mathop{\mathbb{E}}_{(\boldsymbol{x},y)\sim S}[\ell(\boldsymbol{W}^\top f_{\theta^{\text{rep}}}(\boldsymbol{x}), y)] \tag{42}$$

$$\ell_{\lambda,\text{rep}}(\boldsymbol{W}; \theta^{\text{rep}}, S) = \mathop{\mathbb{E}}_{(\boldsymbol{x},y)\sim S}[\ell(\boldsymbol{W}^\top f_{\theta^{\text{rep}}}(\boldsymbol{x}), y)] + \frac{\lambda}{2}\|\boldsymbol{W}\|_2^2 \tag{43}$$

$$\mathcal{A}_{\lambda,\text{rep}}(\theta^{\text{rep}}; S) = \arg\min_{\boldsymbol{W}} \ell_{\lambda,\text{rep}}(\boldsymbol{W}; \theta^{\text{rep}}, S) \tag{44}$$

tr-tr Inner loop: For dataset S and current initialization θ_i^{rep}, run T_{in} of gradient descent (with momentum 0.9) with

learning rate η_{in} on $\ell_{\lambda,\text{rep}}(\boldsymbol{W}; \theta^{\text{rep}}, S)$ to get an approximation $\widetilde{\boldsymbol{W}}_{\lambda,\text{rep}}(\theta^{\text{rep}}; S) \approx \mathcal{A}_{\lambda,\text{rep}}(\theta^{\text{rep}}; S)$. Compute the gradient: $\nabla^{\text{tr-val}}(\theta^{\text{rep}}) = \nabla \ell_{\lambda,\text{rep}}(\widetilde{\boldsymbol{W}}_{\lambda,\text{rep}}\left(\theta_i^{\text{rep}}, S\right); \cdot, S)\,|_{\theta_i^{\text{rep}}}$

Outer loop: Run Adam with learning rate η_{out} (other parameters at default value) with batch size b for T_{out} steps by using the gradient

Meta-testing: Tune T_{in} and $\bar{\lambda}$ using validation tasks

Datasets We conduct experiments on the Omniglot (Lake et al., 2015) and MiniImageNet (Vinyals et al., 2016) datasets. The Omniglot dataset consists of 1623 different handwritten characters from 50 different alphabets. Each character was hand drawn by 20 different people. The original Omniglot dataset was split into a background set comprised of 30 alphabets and an evaluation set of 20 alphabets. We use the split recommended by Vinyals et al. (2016), which contains of a training split of 1028 characters, a validation split of 172 characters, and test split of 423 characters. Vinyals et al. (2016) construct the MiniImageNet dataset by sampling 100 random classes from ImageNet. We use 64 classes for training, 16 for validation, and 20 for testing. We use torchmeta (Deleu et al., 2019) to load datasets. All our evaluations in meta-test time are conducted in the transductive setting.

D.2. Omniglot Experiments

Replearn on Omniglot We use a batch size $b = 32$. We use the standard 4-layer convolutional backbone with each layer having 64 output filters followed by batch normalization and ReLU activations. We resize the images to be 28x28 and apply 90, 180, and 270 degree rotations to augment the data, as in prior work, during training and evaluation. We train for $T_{\text{out}} = 30000$ meta-steps and use $T_{\text{in}} = 100$ inner steps regardless of model, and use an inner learning rate $\eta_{\text{in}} = 0.05$, unless there is a failure of optimization, in which case we reduce the learning rate to 0.01. We use an outer learning rate $\eta_{\text{out}} = 0.001$. We evaluate on 600 tasks at meta-test time. At meta-test time, for each model, we pick the best $\bar{\lambda}$ and inner step size based on the validation set, where we explore $\bar{\lambda} \in [0, .1, .3, 1.0, 2.0, 3.0, 10.0, 100.0]$ and inner step size in $[50, 100, 200]$ and evaluate on the test set.

Increasing width of a FC network on Omniglot We examine the performance gap between tr-val versus tr-tr as we increase the expressive power of the representation network. We use a baseline 4 hidden layer fully connected network (FCN) inspired by Finn et al. (2017), but with 64ℓ nodes at all 4 layers for different values of $\ell \in \{1, 4, 8, 16, 32\}$. We set inner learning rates to be .05, except for $\ell \in \{16, 32\}$ where we found a smaller inner learning rate of .01 was needed for convergence.

iMAML We use the original author code[10] as a starting point, which creates a convolutional neural network with four convolutional layers, followed by Batch Normalization and ReLU activations. We modify their code to add a tr-tr variant by using the combined data for the inner loop and the outer loop updates. We apply 90, 180, and 270 degree rotations to Omniglot data, resizing each image to 28x28 pixels. We use 5 conjugate gradient steps. For meta-testing we pick the best $\bar{\lambda} \in \{0, .1, .3, 1.0, 2.0, 3.0, 10.0, 100.0\}$ and $n_{\text{in}} \in .\{8, 16, 32, 64\}$ that maximizes accuracy on the validation set. All other hyper-parameters at meta-test time are equal to the values used during training. We use an outer learning rate of 1e-3. We train for 30000 outer steps for all models tested, and set the number of inner steps n_steps = 16 for 5-way 1-shot and 25 for 20-way 1-shot. We investigate the performance of tr-val versus tr-tr by examining different settings of the regularization parameter λ. We report our results for tr-val versus tr-tr for Omniglot 5-way 1-shot and 20-way 1-shot in Table 6. We find that tr-val significantly outperforms tr-tr in all settings, and the gap is much larger than for RepLearn.

iMAML representations and t-SNE We train a model on Omniglot 5-way 1-shot using iMAML with batch size 32, inner learning rate .05, and meta steps 30000, again using the original author code convolutional neural network. We use $\lambda = 2.0$ for tr-val and $\lambda = 1.0$ for tr-tr. We use the tr-val and tr-tr CNN models to get image representations for input images, and then perform t-SNE on 10 randomly selected classes. We report t-SNE and singular value decay results for iMAML in Figure 4. We also plot the t-SNE representations for varying values of the perplexity in Figure 2. As in the case of RepLearn, the tr-val representations are much better clustered in the tr-tr representations. Furthermore the tr-val representations have a sharper drop in singular values, suggesting that they have lower effective rank than tr-tr representations.

[10] https://github.com/aravindr93/imaml_dev

Table 6. meta-test accuracies in % for a CNN trained with iMAML for tr-val versus tr-tr on Omniglot 5-way 1-shot and Omniglot 20-way 1-shot. We find a huge gap in performance between tr-val and tr-tr, even larger than that for RepLearn.

	5-way 1-shot	20-way 1-shot
$\lambda = 2.0$ tr-val	97.90 ± 0.58	91.0 ± 0.54
$\lambda = 10.0$ tr-tr	49.22 ± 1.83	14.45 ± 0.61
$\lambda = 2.0$ tr-tr	43.71 ± 1.92	16.18 ± 0.65
$\lambda = 1.0$ tr-tr	47.18 ± 1.90	16.96 ± 0.66
$\lambda = 0.3$ tr-tr	48.61 ± 1.90	16.50 ± 0.64
$\lambda = 0.1$ tr-tr	48.60 ± 1.93	17.70 ± 0.69
$\lambda = 0.0$ tr-tr	49.21 ± 1.92	18.30 ± 0.67

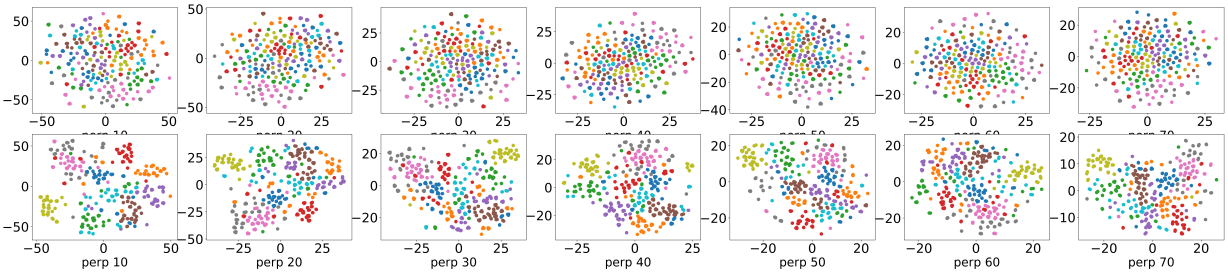


Figure 2. t-SNE plots for tr-val versus tr-tr for varying values of the perplexity for a CNN model trained with iMAML on Omniglot 5-way 1-shot. Plots shown for 10 randomly selected classes from the test split of Omniglot. Top row depicts tr-tr and bottom row depicts tr-val. We find that the tr-val representations appear to be more clustered for all values of perplexity of t-SNE.

Rank and Expressivity For a fully connected model of width factor $\ell = 32$ trained with RepLearn on Omniglot 5-way 1-shot, we conduct linear regression twice. The first regression predicts the tr-tr representations given the tr-val representations, and the second predicts the tr-val representations given the tr-tr representations. We find that the R^2 scores were 0.0973 and 0.0967, respectively. Thus the two sets of representations can express each other well enough, even though tr-val representations have lower effective rank.

We conduct the same experiment for a CNN model trained with iMAML on Omniglot 5-way 1-shot, and find that the R^2 scores were 0.0103 and 0.171, respectively, thus suggesting that tr-val representations are a bit more expressive than tr-tr representations, in addition to having lower effective rank.

Adding explicit regularization to RepLearn For a CNN model trained with RepLearn on Omniglot 5-way 1-shot, we add explicit regularization to tr-tr by adding the Frobenius norm of the representation to the loss. We report the accuracy in percentage evaluated on 1200 tasks in Table 8 in the top row. We find that this significantly improves the performance of the tr-tr method, compared to the tr-tr models without explicit regularization in Table 10, or the bottom row of Table 8. This fits our intuition that the tr-tr method requires some form of regularization to learn low rank representations and to have guaranteed good performance on new tasks.

D.3. MiniImageNet Experiments

RepLearn on MiniImageNet As standard (Rajeswaran et al., 2019), we resize the data to 84x84 pixel images and apply 90, 180, and 270 degree rotations and use a batch size of 16 during training. We use a convolutional neural network with four convolutional layers, with output filter sizes of 32,, 64, 128, and 128 each followed by batch normalization and ReLU activations. We investigate the performance of tr-val versus tr-tr for RepLearn on the MiniImageNet 5-way 1-shot and 5-way 5-shot setting and report our results in Table 7. The findings are very similar to those from the Omniglot dataset, suggesting that our insights hold across multiple benchmark datasets.

Table 7. Tr-val versus tr-tr meta-test accuracies in % for a CNN model trained with RepLearn on MiniImageNet 5-way 1-shot and 5-way 5-shot for varying values of the regularization parameter, λ. The final value of the tr-tr objective is depicted in the last two columns for MiniImageNet 5-way 1-shot and MiniImageNet 5-way 5-shot, respectively. The tr-tr models make the tr-tr loss very small, which is what they were trained to minimize. Thus their failure on few-shot learning is not due to failure of optimization.

	5-way 1-shot	5-way 5-shot	tr-tr loss 5-way 1-shot	tr-tr loss 5-way 5-shot
$\lambda = 0.0$ tr-val	46.16 ± 1.67	65.36 ± 0.91	0.01	0.01
$\lambda = 0.0$ tr-tr	25.53 ± 1.43	33.49 ± 0.82	1.1e-8	2.1e-6
$\lambda = 0.1$ tr-tr	24.69 ± 1.32	34.91 ± 0.85	3.5e-8	5.5e-7
$\lambda = 1.0$ tr-tr	25.88 ± 1.45	40.19 ± 1.12	1.9e-6	9.3e-5

Table 8. Meta-test accuracies in % for tr-tr RepLearn trained on Omniglot 5-way 1-shot with explicit Frobenius norm regularization added to the representations.

	$\lambda = 0.0$ (tr-tr)	$\lambda = 0.1$ (tr-tr)	$\lambda = 3.0$(tr-tr)
with representation regularization	94.73 ± 0.55	95.05 ± 0.55	94.74 ± 0.55
no representation regularization	67.78 ± 1.60	67.53 ± 1.66	89.00 ± 1.08

Table 9. Accuracies in % of representations parameterized by CNN networks of varying number of filters on MiniImageNet. Representations trained using tr-val objective consistently outperforms those learned using tr-tr objective, and the gap increases as width increases.

capacity = $num\_filters^*\ell$	MiniImageNet 5-way 1-shot tr-val	MiniImageNet 5-way 1-shot tr-tr	Supervised 20-way tr-val	Supervised 20-way tr-tr
$\ell = 0.5$	46.66 ± 1.69	26.25 ± 1.45	1.	1.
$\ell = 1$	48.44 ± 1.62	26.81 ± 1.44	1.	1.
$\ell = 4$	$52.22 \pm 1.68.$	24.66 ± 1.26	1.	1.
$\ell = 8$	52.25 ± 1.71	25.28 ± 1.37	1.	1.

Table 10. Tr-val versus tr-tr meta-test accuracies in % for a CNN model trained with RepLearn on Omniglot 5-way 1-shot for varying values of the regularization parameter, λ.

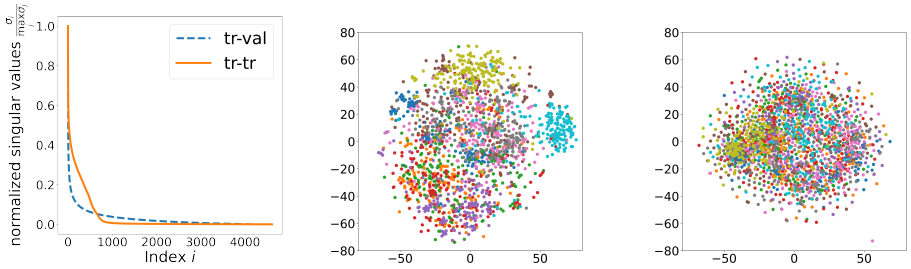
	5-way 1-shot tr-val	5-way 1-shot tr-tr
$\lambda = 0.0$	97.25 ± 0.57	67.78 ± 1.60
$\lambda = 0.1$	97.34 ± 0.59	67.53 ± 1.64
$\lambda = 0.3$	97.59 ± 0.55	66.06 ± 1.67
$\lambda = 1.0$	97.66 ± 0.52	87.25 ± 1.13
$\lambda = 3.0$	97.19 ± 0.59	89.00 ± 1.08
$\lambda = 10.0$	96.50 ± 0.61	85.41 ± 1.22

Increasing capacity of a CNN model on MiniImageNet We start with a CNN model trained on MiniImageNet with four convolutional layers with 32, 64, 128, and 128 filters, respectively. Each convolution is followed by batch normalization and ReLU activations. We train with RepLearn. We increase the capacity by increasing the number of filters by a capacity factor, ℓ, so that the convolutional layers contain 32ℓ, 64ℓ, 128ℓ, and 128ℓ output filters, respectively. We depict our results in Table 9. We find that increasing the network capacity improves the performance of tr-val representations, but slightly hurts tr-tr performance, just like the findings for Omniglot dataset with fully-connected networks. Thus the tr-val method is more robust to architecture choice/capacity and datasets.

t-SNE on MiniImageNet We take the baseline convolutional neural network with capacity factor $\ell = 1$ from the previous section, and conduction t-SNE on the representations produced by the tr-val versus the tr-tr model. We report our results in Figure 3.

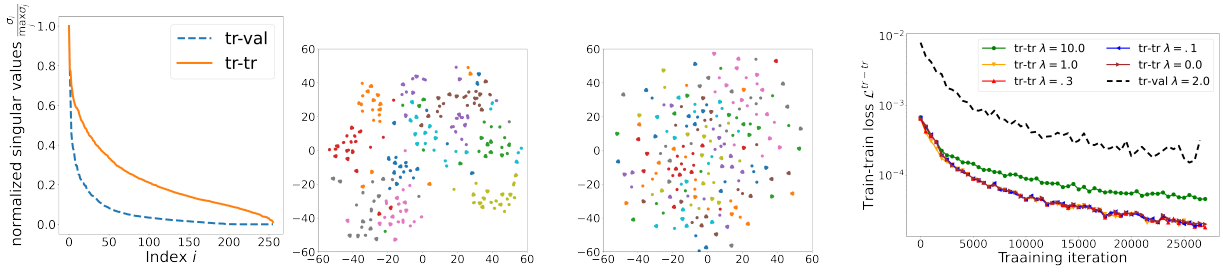
Table 11. Tr-val versus tr-tr meta-test accuracies in % for a CNN model trained with RepLearn on Omniglot 20-way 1-shot for varying values of the regularization parameter, λ.

	20-way 1-shot trval	20-way 1-shot tr-tr
$\lambda = 0.0$	92.26 ± 0.45	49.00 ± 0.88
$\lambda = 0.1$	92.38 ± 0.44	50.84 ± 0.85
$\lambda = 0.3$	92.21 ± 0.47	55.38 ± 0.92
$\lambda = 1.0$	92.44 ± 0.47	84.14 ± 0.63
$\lambda = 3.0$	92.70 ± 0.48	88.20 ± 0.55
$\lambda = 10.0$	91.50 ± 0.48	85.85 ± 0.58



(a) (b) (c)

Figure 3. Plot of singular values and t-SNE plots for MiniImageNet tr-val versus tr-tr. **(a)**: Singular values for representations produced by tr-val versus tr-tr of a CNN model trained with RepLearn on MiniImagement. **(b)**: t-SNE plot of representations produced by tr-val CNN model trained with RepLearn on MiniImageNet for 10 randomly selected classes of test split. **(c)**: t-SNE plot of representations produced by tr-tr CNN model trained with RepLearn on MiniImageNet for the same 10 randomly selected classes of test split. We find that the tr-val representations appear to be more clustered than the tr-tr representations.



(a) (b) (c) (d)

Figure 4. Plot of singular values and t-SNE plots for iMAML tr-val versus tr-tr. **(a)**: Singular values for representations produced by tr-val versus tr-tr of a CNN model trained with iMAML on Omniglot 5-way 1-shot. **(b)**: t-SNE plot of representations produced by tr-val CNN model trained with iMAML on Omniglot 5-way 1-shot for 10 randomly selected classes of test split. **(c)**: t-SNE plot of representations produced by tr-tr CNN trained with iMAML model on Omniglot 5-way 1-shot for the same 10 randomly selected classes of test split. We find that the tr-val representations appear to be more clustered than the tr-tr representations. **(d)**: Tr-tr loss values for tr-tr versus tr-val models.