A. Appendix Overview

Appendix is organized as follows:

In Appendix B we prove the lower bounds for convex meta-learning.

- Appendix B.1 has the proofs for stronger versions of the main lower bound result, Theorem 4.1, that shows lower bounds for the within-task methods of $\texttt{GD}^{\lambda}_{reg}$ and $\texttt{GD}^{\eta,t_0}_{step}$ (MAML).
- Appendix B.2 has proofs for closed form solutions found by $\texttt{GD}^{\lambda}_{reg}$ and $\texttt{GD}^{\eta,t_0}_{step}$ given n samples for a new task. These results are useful to prove the aforementioned theorems.
- Appendix B.3 contains proofs for auxiliary lemmas.

In Appendix C we prove the upper bounds for non-convex meta-learning.

- Appendix C.1 formalizes the steps mentioned in the proof sketch for `Reptile` from Section 6.1. It has the bulk of the proofs about the dynamics of gradient-based algorithms.
- Appendix C.2 proves the main upper bound theorems, Theorem 5.1 and Theorem 5.2.

In Appendix D we prove the tightness of current distance-based convex meta-learning lower bounds (Theorem 7.1).

B. Convex proofs

B.1. Lower bounds

Before proving the lower bounds, we present the following lemma about the closed form solutions found by $\texttt{GD}^{\lambda}_{reg}$ and $\texttt{GD}^{\eta,t_0}_{step}$ starting from an initialization $\mathbf{w}_0$; the proof of this can be found in Appendix B.2.

Note that every $S = \{(\boldsymbol{x}_i, y_i)\} \sim \rho^n_{\boldsymbol{v}}$ is unique determined by a matrix $\boldsymbol{X} \in \mathbb{R}^{n\times d}$ and a noise vector $\xi \in \mathbb{R}^n$, where the i^{th} row of $\boldsymbol{X}$ is $\boldsymbol{x}_i$ and $\xi_i = y_i - \boldsymbol{v}^\top \boldsymbol{x}_i$ are i.i.d. samples from $\mathcal{N}(0, \sigma^2)$.

Lemma B.1. *Let $S = (\boldsymbol{X}, \xi)$ be a sample from $\rho_{\boldsymbol{v}}$, $\boldsymbol{X} \in \mathbb{R}^{n\times d}, \xi \in \mathbb{R}^n$. Let $\Sigma_{\boldsymbol{X}} = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{x}_i\boldsymbol{x}_i^\top \in \mathbb{R}^{d\times d}$*

$$GD^{\lambda}_{reg}(S;\mathbf{w}_0) = (I_d - (\Sigma_{\boldsymbol{X}} + \lambda I_d)^\dagger(\Sigma_{\boldsymbol{X}} + \lambda I_d))\mathbf{w}_0 + (\Sigma_{\boldsymbol{X}} + \lambda I_d)^\dagger \Sigma_{\boldsymbol{X}}\boldsymbol{v} + \frac{1}{n}(\Sigma_{\boldsymbol{X}} + \lambda I_d)^\dagger \boldsymbol{X}^\top \xi$$

$$GD^{\eta,t_0}_{step}(S;\mathbf{w}_0) = (I_d - \eta\Sigma_{\boldsymbol{X}})^{t_0}\mathbf{w}_0 + (I_d - (I_d - \eta\Sigma_{\boldsymbol{X}})^{t_0})\Sigma^\dagger_{\boldsymbol{X}}\Sigma_{\boldsymbol{X}}\boldsymbol{v} + \frac{1}{n}(I_d - (I_d - \eta\Sigma_{\boldsymbol{X}})^{t_0})\Sigma^\dagger_{\boldsymbol{X}}\boldsymbol{X}^\top\xi$$

Here we use $\boldsymbol{A}^\dagger$ to denote the Moore-Penrose pseudo-inverse of matrix $\boldsymbol{A}$. Note that while the inverse exists for all $\lambda > 0$, we use the pseudo-inverse for $\lambda = 0$. Also since $\boldsymbol{X}$ and ξ do not depend on $\boldsymbol{v}$, the only dependence of $\texttt{GD}^{\lambda}_{reg}(S, \mathbf{w}_0)$ and $\texttt{GD}^{\eta,t_0}_{step}(S, \mathbf{w}_0)$ on $\boldsymbol{v}$ is the second term in each of the equations. Since the solutions of both $\texttt{GD}^{\lambda}_{reg}$ and $\texttt{GD}^{\eta,t_0}_{step}$ are linear in $\mathbf{w}_0$, $\boldsymbol{v}$ and ξ, the following lemma will be useful; the proof can be found in Appendix B.3.

Lemma B.2. *For $S = (\boldsymbol{X}, \xi)$ sampled from $\rho_{\boldsymbol{v}}$, $\boldsymbol{X} \in \mathbb{R}^{n\times d}, \xi \in \mathbb{R}^n$, if $\texttt{Alg}(S;\mathbf{w}_0) = \boldsymbol{A}_{\boldsymbol{X}}\mathbf{w}_0 + \boldsymbol{B}_{\boldsymbol{X}}\boldsymbol{v} + \boldsymbol{C}_{\boldsymbol{X}}\xi$ for $\boldsymbol{A}_{\boldsymbol{X}}, \boldsymbol{B}_{\boldsymbol{X}} \succcurlyeq 0$, then*

$$\mathcal{E}_n(\texttt{Alg}(\cdot;\mathbf{w}_0), \mu_{\mathbf{w}_*}) \geq \underbrace{\mathop{\mathbb{E}}_{\boldsymbol{X}}[\|(I_d - \boldsymbol{B}_{\boldsymbol{X}})\mathbf{w}_*\|^2]}_{bias(\texttt{Alg})} + \underbrace{\mathop{\mathbb{E}}_{\boldsymbol{X}}[\sigma^2 \mathrm{tr}(\boldsymbol{C}_{\boldsymbol{X}}^\top \boldsymbol{C}_{\boldsymbol{X}})]}_{var(\texttt{Alg})}$$

where $bias(\texttt{Alg})$ is the error in predicting $\mathbf{w}_*$ and $var(\texttt{Alg})$ is error due to noise in labels in the training data S.

Note that the lower bound on excess risk does not depend on the initialization $\mathbf{w}_0$ or the matrix $\boldsymbol{A}_{\boldsymbol{X}}$. We are now ready to prove the following stronger version of Theorem 4.1.

Proving Theorem 4.1: We now prove strengthened versions of the theorem for $\texttt{GD}^{\lambda}_{reg}$ and $\texttt{GD}^{\eta,t_0}_{step}$ separately. From Definition 7, we have $n_\epsilon(\texttt{Alg}, \mu_{\mathbf{w}_*}) = \min_{n\in\mathbb{N}} : \mathcal{E}_n(\texttt{Alg}, \mu_{\mathbf{w}_*}) \le \epsilon$ is the minimum number of samples needed to achieve excess risk at most ϵ.

(Theorem 4.1(a)). *For every $\mathbf{w}_0 \in \mathbb{R}^d$, number of samples needed to have ϵ excess risk on a new task is*

$$\mathcal{E}_n(GD^{\lambda}_{reg}(\cdot;\mathbf{w}_0), \mu_{\mathbf{w}_*}) \ge \begin{cases} \frac{d\|\mathbf{w}_*\|^2\sigma^2}{\|\mathbf{w}_*\|^2 n+\sigma^2 d} & \text{if } n \ge d \\ \\ \frac{n}{d}\frac{\|\mathbf{w}_*\|^2\sigma^2}{\|\mathbf{w}_*\|^2+\sigma^2} + \frac{(d-n)}{d}\|\mathbf{w}_*\|^2 & \text{if } n < d \end{cases}$$

Furthermore if $\|\mathbf{w}_*\| = \sigma = r \ge 1$ and $\epsilon \in \left(0, \frac{r^2}{2}\right)$, then the number of samples needed to achieve excess error of ϵ is

$$\min_{\lambda\ge 0} \; n_\epsilon(GD^{\lambda}_{reg}(\cdot;\mathbf{w}_0), \mu_{\mathbf{w}_*}) \ge \frac{dr^2}{2\epsilon}$$

Proof of Theorem 4.1(a). Consider m samples S from $\rho_{s\mathbf{w}_*}$, where $s \in \{\pm 1\}$. As observed earlier, sampling $S \sim \rho_{s\mathbf{w}_*}$ corresponds to sampling $\boldsymbol{X} \sim \mathcal{N}^n(0, I_d)$ and $\xi \sim \mathcal{N}(0, \sigma^2 I_n)$. From Lemma B.1, we get

$$\texttt{GD}^{\lambda}_{reg}(S;\mathbf{w}_0) = (I_d - (\Sigma_{\boldsymbol{X}} + \lambda I_d)^\dagger(\Sigma_{\boldsymbol{X}} + \lambda I_d))\mathbf{w}_0 + (\Sigma_{\boldsymbol{X}} + \lambda I_d)^\dagger \Sigma_{\boldsymbol{X}} \boldsymbol{v} + \frac{1}{n}(\Sigma_{\boldsymbol{X}} + \lambda I_d)^\dagger \boldsymbol{X}^\top \xi$$

Instantiating Lemma B.2 with $\boldsymbol{A}_{\boldsymbol{X}} = (I_d - (\Sigma_{\boldsymbol{X}} + \lambda I_d)^\dagger(\Sigma_{\boldsymbol{X}} + \lambda I_d))$, $\boldsymbol{B}_{\boldsymbol{X}} = (\Sigma_{\boldsymbol{X}} + \lambda I_d)^\dagger \Sigma_{\boldsymbol{X}}$ and $\boldsymbol{C}_{\boldsymbol{X}} = \frac{1}{n}(\Sigma_{\boldsymbol{X}} + \lambda I_d)^\dagger \boldsymbol{X}^\top$, we get

$$\begin{aligned}
\mathcal{E}_n(\texttt{GD}^{\lambda}_{reg}(\cdot;\mathbf{w}_0), \mu_{\mathbf{w}_*}) &= \mathop{\mathbb{E}}_{\boldsymbol{X}}\left[\|(I_d - (\Sigma_{\boldsymbol{X}} + \lambda I_d)^\dagger \Sigma_{\boldsymbol{X}})\mathbf{w}_*\|^2\right] + \mathop{\mathbb{E}}_{\boldsymbol{X}}\left[\sigma^2 \mathrm{tr}\left(\frac{\boldsymbol{X}}{n}(\Sigma_{\boldsymbol{X}} + \lambda I_d)^{\dagger^2}\frac{\boldsymbol{X}^\top}{n}\right)\right] \\
&=^{(a)} \mathop{\mathbb{E}}_{\boldsymbol{X}}\left[\|(I_d - (\Sigma_{\boldsymbol{X}} + \lambda I_d)^\dagger \Sigma_{\boldsymbol{X}})\mathbf{w}_*\|^2\right] + \frac{\sigma^2}{n^2}\mathop{\mathbb{E}}_{\boldsymbol{X}}\left[\mathrm{tr}((\Sigma_{\boldsymbol{X}} + \lambda I_d)^{\dagger^2}\boldsymbol{X}^\top \boldsymbol{X})\right] \\
&= \underbrace{\mathop{\mathbb{E}}_{\boldsymbol{X}}\left[\|(I_d - (\Sigma_{\boldsymbol{X}} + \lambda I_d)^\dagger \Sigma_{\boldsymbol{X}})\mathbf{w}_*\|^2\right]}_{bias(\texttt{GD}^{\lambda}_{reg})} + \underbrace{\frac{\sigma^2}{n}\mathop{\mathbb{E}}_{\boldsymbol{X}}\left[\mathrm{tr}((\Sigma_{\boldsymbol{X}} + \lambda I_d)^{\dagger^2}\Sigma_{\boldsymbol{X}})\right]}_{var(\texttt{GD}^{\lambda}_{reg})}
\end{aligned}$$

where (a) follows from property about trace that $\mathrm{tr}(AB) = \mathrm{tr}(BA)$ and the definition of $\Sigma_{\boldsymbol{X}}$. We now lower bound the bias and variance terms separately. Let $\Sigma_{\boldsymbol{X}} = \boldsymbol{V}\boldsymbol{S}\boldsymbol{V}^\top$ be the full SVD, where $\boldsymbol{S} = diag(s_1, \ldots, s_d)$ is a diagonal matrix such that $s_1 \ge s_2 \ge \cdots \ge s_d \ge 0$. Let $\boldsymbol{v}_i$ be the i^{th} column of $\boldsymbol{V}$. Note that $\boldsymbol{V}^\top\boldsymbol{V} = \boldsymbol{V}\boldsymbol{V}^\top = I_d$.

Bias: The bias term can be handled by first noticing the following

$$\begin{aligned}
bias(\texttt{GD}^{\lambda}_{reg}) &= {\mathbf{w}_*}^\top(I_d - (\Sigma_{\boldsymbol{X}} + \lambda I_d)^\dagger \Sigma_{\boldsymbol{X}})^2 \mathbf{w}_* = {\mathbf{w}_*}^\top(\boldsymbol{V}\boldsymbol{V}^\top - (\boldsymbol{V}\boldsymbol{S}\boldsymbol{V}^\top + \lambda \boldsymbol{V}\boldsymbol{V}^\top)^\dagger \boldsymbol{V}\boldsymbol{S}\boldsymbol{V}^\top)^2 \mathbf{w}_* \\
&= {\mathbf{w}_*}^\top(\boldsymbol{V}\boldsymbol{V}^\top - \boldsymbol{V}(\boldsymbol{S} + \lambda I_d)^\dagger \boldsymbol{V}^\top \boldsymbol{V}\boldsymbol{S}\boldsymbol{V}^\top)^2 \mathbf{w}_* = {\mathbf{w}_*}^\top \boldsymbol{V}(I_d - (\boldsymbol{S} + \lambda I_d)^\dagger \boldsymbol{S})^2 \boldsymbol{V}^\top \mathbf{w}_* \\
&= \sum_{i=1}^d h(s_i, \lambda)({\mathbf{w}_*}^\top \boldsymbol{v}_i)^2, \text{ where} \\
h(s,\lambda) &= \begin{cases} \frac{\lambda^2}{(s+\lambda)^2} & \text{if } s > 0 \\ 1 & \text{if } s = 0 \end{cases}
\end{aligned}$$

We can split the expectation w.r.t. $\boldsymbol{X}$ into expectation w.r.t. $\boldsymbol{S}$ and the conditional expectation of $\boldsymbol{V}$ given $\boldsymbol{S}$. A crucial observation is that since the distribution of the rows of $\boldsymbol{X}$ is isotropic gaussian, no direction in space is special. Thus, conditioned on $\boldsymbol{S}$, the distribution of $\boldsymbol{v}_i$ should be identical for all i and we must have that $\mathop{\mathbb{E}}_{\boldsymbol{X}}[\boldsymbol{v}_i|\boldsymbol{S}] = 0$ and $\mathop{\mathbb{E}}_{\boldsymbol{X}}[\boldsymbol{v}_i\boldsymbol{v}_i^\top|\boldsymbol{S}] = CI_d$ for some constant C. The constant C can be calculated by noting that $\|\boldsymbol{v}_i\| = 1$. So we get $1 = \mathop{\mathbb{E}}_{\boldsymbol{X}}[\boldsymbol{v}_i^\top \boldsymbol{v}_i|\boldsymbol{S}] = \mathrm{tr}\left(\mathop{\mathbb{E}}_{\boldsymbol{X}}[\boldsymbol{v}_i\boldsymbol{v}_i^\top|\boldsymbol{S}]\right) = C\mathrm{tr}(I_d) = Cd$, giving $C = \frac{1}{d}$. Then the bias is

$$bias(\texttt{GD}^{\lambda}_{reg}) = \mathop{\mathbb{E}}_{\boldsymbol{S}}\left[\mathop{\mathbb{E}}_{\boldsymbol{V}}\left[\sum_{i=1}^d h(s_i,\lambda)({\mathbf{w}_*}^\top \boldsymbol{v}_i)^2 \middle| \boldsymbol{S}\right]\right] = \mathop{\mathbb{E}}_{\boldsymbol{S}}\left[\sum_{i=1}^d h(s_i,\lambda)\mathop{\mathbb{E}}_{\boldsymbol{v}_i}\left[({\mathbf{w}_*}^\top \boldsymbol{v}_i)^2 \middle| \boldsymbol{S}\right]\right]$$

$$= \mathop{\mathbb{E}}_{\boldsymbol{S}}\left[\sum_{i=1}^{d} h(s_i,\lambda) \mathop{\mathbb{E}}_{\boldsymbol{v}_i}\left[{\mathbf{w}_*}^\top \boldsymbol{v}_i\boldsymbol{v}_i^\top \mathbf{w}_* \middle| \boldsymbol{S}\right]\right] = \mathop{\mathbb{E}}_{\boldsymbol{S}}\left[\sum_{i=1}^{d} h(s_i,\lambda) {\mathbf{w}_*}^\top \mathop{\mathbb{E}}_{\boldsymbol{v}_i}\left[\boldsymbol{v}_i\boldsymbol{v}_i^\top \middle| \boldsymbol{S}\right] \mathbf{w}_*\right]$$
$$= \mathop{\mathbb{E}}_{\boldsymbol{S}}\left[\sum_{i=1}^{d} h(s_i,\lambda)\frac{\|\mathbf{w}_*\|^2}{d}\right] = \frac{\|\mathbf{w}_*\|^2}{d}\mathop{\mathbb{E}}_{\boldsymbol{S}}\left[\sum_{i=1}^{d} h(s_i,\lambda)\right] \tag{12}$$

Variance: We now look at the variance term

$$var(\mathtt{GD}^{\lambda}_{reg}) = \mathop{\mathbb{E}}_{\boldsymbol{X}}\left[\frac{\sigma^2}{n}\mathrm{tr}((\Sigma_{\boldsymbol{X}}+\lambda I_d)^{\dagger^2}\Sigma_{\boldsymbol{X}})\right] = \mathop{\mathbb{E}}_{\boldsymbol{X}}\left[\frac{\sigma^2}{n}\mathrm{tr}((\boldsymbol{V}\boldsymbol{S}\boldsymbol{V}^\top+\lambda\boldsymbol{V}\boldsymbol{V}^\top)^{\dagger^2}\boldsymbol{V}\boldsymbol{S}\boldsymbol{V}^\top)\right]$$
$$= \mathop{\mathbb{E}}_{\boldsymbol{X}}\left[\frac{\sigma^2}{n}\mathrm{tr}(\boldsymbol{V}(\boldsymbol{S}+\lambda I_d)^{\dagger^2}\boldsymbol{V}^\top\boldsymbol{V}\boldsymbol{S}\boldsymbol{V}^\top)\right] = \mathop{\mathbb{E}}_{\boldsymbol{X}}\left[\frac{\sigma^2}{n}\mathrm{tr}(\boldsymbol{V}(\boldsymbol{S}+\lambda I_d)^{\dagger^2}\boldsymbol{S}\boldsymbol{V}^\top)\right]$$
$$= \mathop{\mathbb{E}}_{\boldsymbol{X}}\left[\frac{\sigma^2}{n}\mathrm{tr}((\boldsymbol{S}+\lambda I_d)^{\dagger^2}\boldsymbol{S}\boldsymbol{V}^\top\boldsymbol{V})\right] = \mathop{\mathbb{E}}_{\boldsymbol{S}}\left[\frac{\sigma^2}{n}\mathrm{tr}((\boldsymbol{S}+\lambda I_d)^{\dagger^2}\boldsymbol{S})\right]$$
$$= \frac{\sigma^2}{n}\mathop{\mathbb{E}}_{\boldsymbol{S}}\left[\sum_{i=1}^{d} g(s_i,\lambda)\right], \text{ where}$$
$$g(s,\lambda) = \begin{cases} \frac{s}{(s+\lambda)^2} & \text{if } s>0 \\ 0 & \text{if } s_i = 0 \end{cases}$$

Thus we get the following lower bound for the excess risk

$$\mathcal{E}_n(\mathtt{GD}^{\lambda}_{reg}(\cdot;\mathbf{w}_0),\mu_{\mathbf{w}_*}) = bias(\mathtt{GD}^{\lambda}_{reg}) + var(\mathtt{GD}^{\lambda}_{reg}) \geq \frac{\|\mathbf{w}_*\|^2}{d}\mathop{\mathbb{E}}_{\boldsymbol{S}}\left[\sum_{i=1}^{d} h(s_i,\lambda)\right] + \frac{\sigma^2}{n}\mathop{\mathbb{E}}_{\boldsymbol{S}}\left[\sum_{i=1}^{d} g(s_i,\lambda)\right]$$
$$= \mathop{\mathbb{E}}_{\boldsymbol{S}}\left[\sum_{i=1}^{d} \frac{\|\mathbf{w}_*\|^2}{d}h(s_i,\lambda) + \frac{\sigma^2}{n}g(s_i,\lambda)\right]$$

We will show that $\frac{\|\mathbf{w}_*\|^2}{d}h(s,\lambda) + \frac{\sigma^2}{n}g(s,\lambda) \geq \frac{\|\mathbf{w}_*\|^2\sigma^2}{\|\mathbf{w}_*\|^2 ns+\sigma^2 d}$. While this is evident when $s=0$, since $h(0,\lambda)=1$ and $g(0,\lambda)=0$, for $s>0$ the left hand side reduces to $\frac{\|\mathbf{w}_*\|^2}{d}\frac{\lambda^2}{(\lambda+s)^2} + \frac{\sigma^2}{ns}\frac{s^2}{(\lambda+s)^2}$. This is of the form $\alpha a^2+\beta b^2$ where $\alpha = \frac{\|\mathbf{w}_*\|^2}{d}, \beta = \frac{\sigma^2}{ns}$ and $a = \frac{\lambda}{(\lambda+s)}, b = \frac{s}{(\lambda+s)}$ satisfy $a+b=1$. The following simple lemma (proof in Appendix B.3) will help us prove the desired inequality.

Lemma B.3. *For $\alpha,\beta \geq 0$, we have*

$$\min_{a,b\ s.t.\ a+b=1} \alpha a^2+\beta b^2 = \frac{\alpha\beta}{\alpha+\beta}$$

Using the above lemma, we get $\mathcal{E}_n(\mathtt{GD}^{\lambda}_{reg}(\cdot;\mathbf{w}_0),\mu_{\mathbf{w}_*}) \geq \mathop{\mathbb{E}}_{\boldsymbol{S}}\left[\sum_{i=1}^{d} f(s_i)\right]$, where $f(s) = \frac{\|\mathbf{w}_*\|^2\sigma^2}{\|\mathbf{w}_*\|^2 ns+\sigma^2 d}$. The following lemma (proof in Appendix B.3) is a simple application of Jensen's inequality and aids us in completing the proof

Lemma B.4. *For a function convex function $f(\cdot):\mathbb{R}\to\mathbb{R}$, we have*

$$\mathop{\mathbb{E}}_{\boldsymbol{S}}\left[\sum_{i=1}^{d} f(s_i)\right] \geq \begin{cases} df(1) & \text{if } n \geq d \\ nf(\frac{d}{n}) + (d-n)f(0) & \text{if } n < d \end{cases}$$

where the expectation is over $\boldsymbol{S}$ is for the distribution of eigenvalues of $\Sigma_{\boldsymbol{X}}$ when $\boldsymbol{X}\sim\mathcal{N}(0,I_d)^n$.

By noticing that $f(\cdot,\lambda)$ is convex in the first argument, using Lemma B.4 and the fact that $f(0) = \frac{\|\mathbf{w}_*\|^2}{d}$ and $f(1) = \frac{\|\mathbf{w}_*\|^2\sigma^2}{\|\mathbf{w}_*\|^2 n+\sigma^2 d}$ and $f(\frac{d}{n}) = \frac{\|\mathbf{w}_*\|^2\sigma^2}{\|\mathbf{w}_*\|^2 d+\sigma^2 d}$, we get

$$\mathcal{E}_n(\mathtt{GD}^{\lambda}_{reg}(\cdot;\mathbf{w}_0),\mu_{\mathbf{w}_*}) \geq \begin{cases} df(1) = \frac{d\|\mathbf{w}_*\|^2\sigma^2}{\|\mathbf{w}_*\|^2 n+\sigma^2 d} & \text{if } n \geq d \\ nf(\frac{d}{n}) + (d-n)f(0) = n\frac{\|\mathbf{w}_*\|^2\sigma^2}{\|\mathbf{w}_*\|^2 d+\sigma^2 d} + (d-n)\frac{\|\mathbf{w}_*\|^2}{d} & \text{if } n < d \end{cases}$$

which completes the proof for the first part of the theorem.

For the second part where $\|\mathbf{w}_*\| = \sigma = r \geq 1$ and $\epsilon \in \left(0, \frac{r^2}{2}\right)$, it is not difficult to see that $\mathcal{E}_n(\mathtt{GD}^{\lambda}_{reg}(\cdot, \mathbf{w}_0), \mu_{\mathbf{w}_*}) \geq \frac{dr^2}{n+d}$, $\forall n > 0$. To find the minimum $n \geq d$ such that $\mathcal{E}_n(\mathtt{GD}^{\lambda}_{reg}(\cdot, \mathbf{w}_0), \mu_{\mathbf{w}_*}) \leq \epsilon$, we observe the following

$$\epsilon \geq \mathcal{E}_n(\mathtt{GD}^{\lambda}_{reg}(\cdot, \mathbf{w}_0), \mu_{\mathbf{w}_*}) \geq \frac{dr^2}{n+d} \implies n \geq d\left(\frac{r^2}{\epsilon} - 1\right) \geq^{(a)} \frac{dr^2}{2\epsilon}$$

where (a) uses $\epsilon \leq \frac{r^2}{2}$. This gives us $n_\epsilon(\mathtt{GD}^{\lambda}_{reg}(\cdot, \mathbf{w}_0)) \geq \frac{dr^2}{2\epsilon}$ as desired. □

We now prove the result for $\mathtt{GD}^{\eta,t_0}_{step}$.

(Theorem 4.1(b)). *For every $\mathbf{w}_0 \in \mathbb{R}^d$, number of samples needed to have ϵ excess risk on a new task is*

$$\mathcal{E}_n(GD^{\eta,t_0}_{step}(\cdot; \mathbf{w}_0), \mu_{\mathbf{w}_*}) \geq \begin{cases} \frac{d\|\mathbf{w}_*\|^2\sigma^2}{\|\mathbf{w}_*\|^2 n + \sigma^2 d} & \text{if } n \geq d \\ \\ \frac{n}{d}\frac{\|\mathbf{w}_*\|^2\sigma^2}{\|\mathbf{w}_*\|^2+\sigma^2} + \frac{(d-n)}{d}\|\mathbf{w}_*\|^2 & \text{if } n < d \end{cases}$$

Furthermore if $\|\mathbf{w}_*\| = \sigma = r \geq 1$ and $\epsilon \in \left(0, \frac{r^2}{2}\right)$, then the number of samples needed to achieve excess error of ϵ is

$$\min_{\eta \geq 0, t_0 \in \mathbb{N}} \; n_\epsilon(GD^{\eta,t_0}_{step}(\cdot; \mathbf{w}_0), \mu_{\mathbf{w}_*}) \geq \frac{dr^2}{2\epsilon}$$

Proof of Theorem 4.1(b). From Lemma B.1 we have

$$\mathtt{GD}^{\eta,t_0}_{step}(S; \mathbf{w}_0) = (I_d - \eta\Sigma_{\boldsymbol{X}})^{t_0}\mathbf{w}_0 + (I_d - (I_d - \eta\Sigma_{\boldsymbol{X}})^{t_0})\Sigma^{\dagger}_{\boldsymbol{X}}\Sigma_{\boldsymbol{X}}\boldsymbol{v} + \frac{1}{n}(I_d - (I_d - \eta\Sigma_{\boldsymbol{X}})^{t_0})\Sigma^{\dagger}_{\boldsymbol{X}}\boldsymbol{X}^\top\xi$$

Instantiating Lemma B.2 with $\boldsymbol{A}_{\boldsymbol{X}} = (I_d - \eta\Sigma_{\boldsymbol{X}})^{t_0}$, $\boldsymbol{B}_{\boldsymbol{X}} = (I_d - (I_d - \eta\Sigma_{\boldsymbol{X}})^{t_0})\Sigma^{\dagger}_{\boldsymbol{X}}\Sigma_{\boldsymbol{X}}$ and $\boldsymbol{C}_{\boldsymbol{X}} = \frac{1}{n}(I_d - (I_d - \eta\Sigma_{\boldsymbol{X}})^{t_0})\Sigma^{\dagger}_{\boldsymbol{X}}\boldsymbol{X}^\top$, we get from a similar calculation to $\mathtt{GD}^{\lambda}_{reg}$

$$\begin{aligned}
&\mathcal{E}_n(\mathtt{GD}^{\eta,t_0}_{step}(\cdot; \mathbf{w}_0), \mu_{\mathbf{w}_*}) \\
&= \mathop{\mathbb{E}}_{\boldsymbol{X}}\left[\|(I_d - (I_d - (I_d - \eta\Sigma_{\boldsymbol{X}})^{t_0})\Sigma^{\dagger}_{\boldsymbol{X}}\Sigma_{\boldsymbol{X}})\mathbf{w}_*\|^2\right] + \mathop{\mathbb{E}}_{\boldsymbol{X}}\frac{1}{n^2}\left[\sigma^2\mathrm{tr}\left(\boldsymbol{X}\Sigma^{\dagger}_{\boldsymbol{X}}(I_d - (I_d - \eta\Sigma_{\boldsymbol{X}})^{t_0})^2\Sigma^{\dagger}_{\boldsymbol{X}}\boldsymbol{X}^\top\right)\right] \\
&= \mathop{\mathbb{E}}_{\boldsymbol{X}}\left[\|(I_d - \Sigma^{\dagger}_{\boldsymbol{X}}\Sigma_{\boldsymbol{X}} + (I_d - \eta\Sigma_{\boldsymbol{X}})^{t_0}\Sigma^{\dagger}_{\boldsymbol{X}}\Sigma_{\boldsymbol{X}})\mathbf{w}_*\|^2\right] + \frac{\sigma^2}{n^2}\mathop{\mathbb{E}}_{\boldsymbol{X}}\left[\mathrm{tr}\left((I_d - (I_d - \eta\Sigma_{\boldsymbol{X}})^{t_0})^2\Sigma^{\dagger}_{\boldsymbol{X}}\boldsymbol{X}^\top\boldsymbol{X}\Sigma^{\dagger}_{\boldsymbol{X}}\right)\right] \\
&= \mathop{\mathbb{E}}_{\boldsymbol{X}}\left[\|(I_d - \Sigma^{\dagger}_{\boldsymbol{X}}\Sigma_{\boldsymbol{X}} + (I_d - \eta\Sigma_{\boldsymbol{X}})^{t_0}\Sigma^{\dagger}_{\boldsymbol{X}}\Sigma_{\boldsymbol{X}})\mathbf{w}_*\|^2\right] + \frac{\sigma^2}{n}\mathop{\mathbb{E}}_{\boldsymbol{X}}\left[\mathrm{tr}\left((I_d - (I_d - \eta\Sigma_{\boldsymbol{X}})^{t_0})^2\Sigma^{\dagger}_{\boldsymbol{X}}\Sigma_{\boldsymbol{X}}\Sigma^{\dagger}_{\boldsymbol{X}}\right)\right] \\
&= \underbrace{\mathop{\mathbb{E}}_{\boldsymbol{X}}\left[\|(I_d - \Sigma^{\dagger}_{\boldsymbol{X}}\Sigma_{\boldsymbol{X}} + (I_d - \eta\Sigma_{\boldsymbol{X}})^{t_0}\Sigma^{\dagger}_{\boldsymbol{X}}\Sigma_{\boldsymbol{X}})\mathbf{w}_*\|^2\right]}_{bias(\mathtt{GD}^{\eta,t_0}_{step})} + \underbrace{\frac{\sigma^2}{n}\mathop{\mathbb{E}}_{\boldsymbol{X}}\left[\mathrm{tr}\left((I_d - (I_d - \eta\Sigma_{\boldsymbol{X}})^{t_0})^2\Sigma^{\dagger}_{\boldsymbol{X}}\right)\right]}_{var(\mathtt{GD}^{\eta,t_0}_{step})}
\end{aligned}$$

We separately analyze the bias and variance terms

Bias: The bias term can be handled similarly by noticing that

$$\begin{aligned}
bias(\mathtt{GD}^{\eta,t_0}_{step}) &= \mathop{\mathbb{E}}_{\boldsymbol{X}}\left[\mathbf{w}_*{}^\top(I_d - \Sigma^{\dagger}_{\boldsymbol{X}}\Sigma_{\boldsymbol{X}} + (I_d - \eta\Sigma_{\boldsymbol{X}})^{t_0}\Sigma^{\dagger}_{\boldsymbol{X}}\Sigma_{\boldsymbol{X}})^2\mathbf{w}_*\right] \\
&= \mathop{\mathbb{E}}_{\boldsymbol{V},\boldsymbol{S}}\left[\mathbf{w}_*{}^\top(\boldsymbol{V}\boldsymbol{V}^\top - \boldsymbol{V}\boldsymbol{S}^{\dagger}\boldsymbol{V}^\top\boldsymbol{V}\boldsymbol{S}\boldsymbol{V}^\top + (\boldsymbol{V}\boldsymbol{V}^\top - \eta\boldsymbol{V}\boldsymbol{S}\boldsymbol{V}^\top)^{t_0}\boldsymbol{V}\boldsymbol{S}^{\dagger}\boldsymbol{V}^\top\boldsymbol{V}\boldsymbol{S}\boldsymbol{V}^\top)^2\mathbf{w}_*\right] \\
&= \mathop{\mathbb{E}}_{\boldsymbol{S}}\mathop{\mathbb{E}}_{\boldsymbol{V}}\left[\mathbf{w}_*{}^\top\boldsymbol{V}(I_d - \boldsymbol{S}^{\dagger}\boldsymbol{S} + (I_d - \eta\boldsymbol{S})^{t_0}\boldsymbol{S}^{\dagger}\boldsymbol{S})^2\boldsymbol{V}^\top\mathbf{w}_*\right] \\
&= \mathop{\mathbb{E}}_{\boldsymbol{S}}\mathop{\mathbb{E}}_{\boldsymbol{V}}\left[\sum_{i=1}^{d} h(s_i, \eta, t_0)(\mathbf{w}_*{}^\top\boldsymbol{v}_i)^2\right] = \mathop{\mathbb{E}}_{\boldsymbol{S}}\left[\sum_{i=1}^{d} h(s_i, \eta, t_0)\mathop{\mathbb{E}}_{\boldsymbol{v}_i}(\mathbf{w}_*{}^\top\boldsymbol{v}_i)^2\right]
\end{aligned}$$

$$= \frac{\|\mathbf{w}_*\|^2}{d} \mathop{\mathbb{E}}_{\boldsymbol{S}} \left[\sum_{i=1}^{d} h(s_i, \eta, t_0) \right], \text{ where}$$
$$h(s, \eta, t_0) = (1 - \eta s)^{2t_0}$$

Variance: We now look at the variance term

$$var(\mathtt{GD}_{step}^{\eta,t_0}) = \frac{\sigma^2}{n} \mathop{\mathbb{E}}_{\boldsymbol{X}} \left[\mathrm{tr}((I_d - (I_d - \eta\Sigma_{\boldsymbol{X}})^{t_0})^2 \Sigma_{\boldsymbol{X}}^{\dagger}) \right] = \frac{\sigma^2}{n} \mathop{\mathbb{E}}_{\boldsymbol{X}} \left[\mathrm{tr}((\boldsymbol{V}\boldsymbol{V}^\top - (\boldsymbol{V}\boldsymbol{V}^\top - \eta \boldsymbol{V}\boldsymbol{S}\boldsymbol{V}^\top)^{t_0})^2 \boldsymbol{V}\boldsymbol{S}^{\dagger}\boldsymbol{V}^\top) \right]$$
$$= \frac{\sigma^2}{n} \mathop{\mathbb{E}}_{\boldsymbol{X}} \left[\mathrm{tr}(\boldsymbol{V}(I_d - (I_d - \eta\boldsymbol{S})^{t_0})^2 \boldsymbol{V}^\top \boldsymbol{V}\boldsymbol{S}^{\dagger}\boldsymbol{V}^\top) \right] = \frac{\sigma^2}{n} \mathop{\mathbb{E}}_{\boldsymbol{X}} \left[\mathrm{tr}(\boldsymbol{V}(I_d - (I_d - \eta\boldsymbol{S})^{t_0})^2 \boldsymbol{S}^{\dagger}\boldsymbol{V}^\top) \right]$$
$$= \frac{\sigma^2}{n} \mathop{\mathbb{E}}_{\boldsymbol{S}} \left[\mathrm{tr}((I_d - (I_d - \eta\boldsymbol{S})^{t_0})^2 \boldsymbol{S}^{\dagger}) \right] = \frac{\sigma^2}{n} \mathop{\mathbb{E}}_{\boldsymbol{S}} \left[\sum_{i=1}^{d} g(s_i, \eta, t_0) \right], \text{ where}$$
$$g(s, \eta, t_0) = \begin{cases} \frac{(1-(1-\eta s)^{t_0})^2}{s} & \text{if } s > 0 \\ 0 & \text{if } s = 0 \end{cases}$$

Thus we get the following lower bound for the excess risk

$$\mathcal{E}_n(\mathtt{GD}_{step}^{\eta,t_0}(\cdot; \mathbf{w}_0), \mu_{\mathbf{w}_*}) = bias(\mathtt{GD}_{step}^{\eta,t_0}) + var(\mathtt{GD}_{step}^{\eta,t_0}) \geq \frac{\|\mathbf{w}_*\|^2}{d} \mathop{\mathbb{E}}_{\boldsymbol{S}} \left[\sum_{i=1}^{d} h(s_i, \eta, t_0) \right] + \frac{\sigma^2}{n} \mathop{\mathbb{E}}_{\boldsymbol{S}} \left[\sum_{i=1}^{d} g(s_i, \eta, t_0) \right]$$
$$= \mathop{\mathbb{E}}_{\boldsymbol{S}} \left[\sum_{i=1}^{d} \frac{\|\mathbf{w}_*\|^2}{d} h(s_i, \eta, t_0) + \frac{\sigma^2}{n} g(s_i, \eta, t_0) \right]$$

We will again show that $\frac{\|\mathbf{w}_*\|^2}{d} h(s, \eta, t_0) + \frac{\sigma^2}{n} g(s, \eta, t_0) \geq \frac{\|\mathbf{w}_*\|^2 \sigma^2}{\|\mathbf{w}_*\|^2 ns + \sigma^2 d}$. Again, this is obvious for $s = 0$ from the definitions of h and g. For $s > 0$, we can write $\frac{\|\mathbf{w}_*\|^2}{d} h(s, \eta, t_0) + \frac{\sigma^2}{n} g(s, \eta, t_0) = \frac{\|\mathbf{w}_*\|^2}{d} (1 - \eta s)^{2t_0} + \frac{\sigma^2}{sn} (1 - (1 - \eta s)^{t_0})^2$, which is again of the form $\alpha a^2 + \beta b^2$ with $a = (1 - \eta s)^{t_0}, b = (1 - (1 - \eta s)^{t_0})$ satisfying $a + b = 1$. Thus Lemma B.3 gives us the desired inequality, which directly implies $\mathcal{E}_n(\mathtt{GD}_{reg}^{\lambda}(\cdot; \mathbf{w}_0), \mu_{\mathbf{w}_*}) \geq \mathop{\mathbb{E}}_{\boldsymbol{S}} \left[\sum_{i=1}^{d} f(s_i) \right]$, where $f(s) = \frac{\|\mathbf{w}_*\|^2 \sigma^2}{\|\mathbf{w}_*\|^2 ns + \sigma^2 d}$. This is exactly the same lower bound as in the proof of $\mathtt{GD}_{reg}^{\lambda}$, and thus the theorem follows from identical arguments. □

B.2. Closed form solutions

We now prove Lemma B.1. Before that, we will state and prove the following simple lemmas about linear dynamics that will be useful later.

Lemma B.5. *For a symmetric psd matrix $\boldsymbol{M} \in \mathbb{R}^{d \times d}$, let $\boldsymbol{M} = \boldsymbol{B}\boldsymbol{S}\boldsymbol{B}^{-1}$ be its diagonalization. For $\boldsymbol{b} \in \mathbb{R}^d$ that is in the range of $\boldsymbol{M}$, the solution to the system $\frac{d\mathbf{w}_t}{dt} = -\boldsymbol{M}\mathbf{w}_t + \boldsymbol{b}$ starting from $\mathbf{w}_0$ is*

$$\mathbf{w}_t = \boldsymbol{B}e^{-t\boldsymbol{S}}\boldsymbol{B}^{-1}\mathbf{w}_0 + \boldsymbol{B}(I_d - e^{-t\boldsymbol{S}})\boldsymbol{S}^{\dagger}\boldsymbol{B}^{-1}\boldsymbol{b}$$
$$\mathbf{w}_\infty = (I_d - \boldsymbol{M}^{\dagger}\boldsymbol{M})\mathbf{w}_0 + \boldsymbol{M}^{\dagger}\boldsymbol{b}$$

where for a diagonal matrix $\boldsymbol{S} = diag(s_1, \ldots, s_d)$, $e^{-t\boldsymbol{S}}$ is defined as $diag(e^{-ts_1}, \ldots, e^{-ts_d})$ and $\boldsymbol{S}^{\dagger}$ is a diagonal matrix with $\boldsymbol{S}^{\dagger}(i,i) = s_i^{-1}$ if $s_i > 0$, otherwise $\boldsymbol{S}^{\dagger}(i,i) = 0$.

Proof. Since $\boldsymbol{b}$ is the range of $\boldsymbol{M}$, let $\boldsymbol{b} = \boldsymbol{M}\bar{\boldsymbol{b}}$. The dynamics $\frac{d\mathbf{w}_t}{dt} = -\boldsymbol{B}\boldsymbol{S}\boldsymbol{B}^{-1}\mathbf{w}_t + \boldsymbol{B}\boldsymbol{S}\boldsymbol{B}^{-1}\bar{\boldsymbol{b}}$ can be rewritten as $\frac{d(\boldsymbol{B}^{-1}\mathbf{w}_t)}{dt} = -\boldsymbol{S}(\boldsymbol{B}^{-1}\mathbf{w}_t + \boldsymbol{B}^{-1}\bar{\boldsymbol{b}})$. Setting $\tilde{\mathbf{w}}_t = \boldsymbol{B}^{-1}\mathbf{w}_t$ and $\tilde{\boldsymbol{b}} = \boldsymbol{B}^{-1}\bar{\boldsymbol{b}}$, we get $\frac{d\tilde{\mathbf{w}}_t}{dt} = -\boldsymbol{S}(\tilde{\mathbf{w}}_t + \tilde{\boldsymbol{b}})$. Since $\boldsymbol{S} = diag(s_1, \ldots, s_d)$ is a diagonal matrix, we can decouple the dynamics

$$\frac{d\tilde{\mathbf{w}}_t(i)}{dt} = -s_i(\tilde{\mathbf{w}}_t(i) - \tilde{\boldsymbol{b}}(i)), \ \forall i \in [d]$$

These scalar dynamics can be solved and it can be verified easily that $\tilde{\mathbf{w}}_t(i) = e^{-ts_i}\tilde{\mathbf{w}}_0(i) + (1 - e^{-ts_i})\tilde{\boldsymbol{b}}(i)$. By observing that $\bar{\boldsymbol{b}} = \boldsymbol{M}^{\dagger}\boldsymbol{b}$ and $\tilde{\boldsymbol{b}} = \boldsymbol{B}^{-1}\bar{\boldsymbol{b}} = \boldsymbol{B}^{-1}\boldsymbol{M}^{\dagger}\boldsymbol{b} = \boldsymbol{S}^{\dagger}\boldsymbol{B}^{-1}\boldsymbol{b}$, we can summarize the dynamics as $\tilde{\mathbf{w}}_t = e^{-t\boldsymbol{S}}\tilde{\mathbf{w}}_0 + (I_d -$

$e^{-t\boldsymbol{S}})\boldsymbol{S}^\dagger\boldsymbol{B}^{-1}\boldsymbol{b}$. Using $\tilde{\mathbf{w}}_0 = \boldsymbol{B}^{-1}\mathbf{w}_0$ and $\mathbf{w}_t = \boldsymbol{B}\tilde{\mathbf{w}}_t$, multiplying by $\boldsymbol{B}$ on both sides completes the first part of the proof, i.e. $\mathbf{w}_t = \boldsymbol{B}e^{-t\boldsymbol{S}}\boldsymbol{B}^{-1}\mathbf{w}_0 + \boldsymbol{B}(I_d - e^{-t\boldsymbol{S}})\boldsymbol{S}^\dagger\boldsymbol{B}^{-1}\boldsymbol{b}$. Furthermore, as $t \to \infty$, we see that $e^{-t\boldsymbol{S}} \to diag(\mathbf{1}\{s_1 = 0\}, \dots, \mathbf{1}\{s_d = 0\})$ since for $s_i \neq 0$, $e^{-ts_i} \to 0$ while if $s_i = 0$ then $e^{-ts_i} = 1$ for every $t \in \mathbb{R}$. This completes the second part of the proof. □

Lemma B.6. *For a symmetric psd matrix $\boldsymbol{M} \in \mathbb{R}^{d\times d}$, let $\boldsymbol{M} = \boldsymbol{B}\boldsymbol{S}\boldsymbol{B}^{-1}$ be its diagonalization. For $\boldsymbol{b} \in \mathbb{R}^d$ that is in the range of $\boldsymbol{M}$, the solution to the system $\mathbf{w}_{t+1} - \mathbf{w}_t = -\eta(\boldsymbol{M}\mathbf{w}_t - \boldsymbol{b})$ starting from $\mathbf{w}_0$ is*

$$\begin{aligned}\mathbf{w}_t &= \boldsymbol{B}(I_d - \eta\boldsymbol{S})^t\boldsymbol{B}^{-1}\mathbf{w}_0 + \boldsymbol{B}(I_d - (I_d - \eta\boldsymbol{S})^t)\boldsymbol{S}^\dagger\boldsymbol{B}^{-1}\boldsymbol{b}\\ &= (I_d - \eta\boldsymbol{M})^t\mathbf{w}_0 + (I_d - (I_d - \eta\boldsymbol{M})^t)\boldsymbol{M}^\dagger\boldsymbol{b}\end{aligned}$$

Proof. Since $\boldsymbol{b}$ is the range of $\boldsymbol{M}$, let $\boldsymbol{b} = \boldsymbol{M}\bar{\boldsymbol{b}}$. The dynamics $\mathbf{w}_{t+1} - \mathbf{w}_t = -\eta\boldsymbol{B}\boldsymbol{S}\boldsymbol{B}^{-1}\mathbf{w}_t + \eta\boldsymbol{B}\boldsymbol{S}\boldsymbol{B}^{-1}\bar{\boldsymbol{b}}$ can be rewritten as $\boldsymbol{B}^{-1}\mathbf{w}_{t+1} = -(I_d - \eta\boldsymbol{S})\boldsymbol{B}^{-1}\mathbf{w}_t + \eta\boldsymbol{S}\boldsymbol{B}^{-1}\bar{\boldsymbol{b}}$. Setting $\tilde{\mathbf{w}}_t = \boldsymbol{B}^{-1}\mathbf{w}_t$ and $\tilde{\boldsymbol{b}} = \boldsymbol{B}^{-1}\bar{\boldsymbol{b}}$, we get $\tilde{\mathbf{w}}_{t+1} = -(I_d - \eta\boldsymbol{S})\tilde{\mathbf{w}}_t + \eta\boldsymbol{S}\tilde{\boldsymbol{b}}$. Since $\boldsymbol{S} = diag(s_1, \dots, s_d)$ is a diagonal matrix, we can decouple the dynamics, for every $i \in [d]$,

$$\begin{aligned}\tilde{\mathbf{w}}_{t+1}(i) &= -(1 - \eta s_i)\tilde{\mathbf{w}}_t(i) + \eta s_i\tilde{\boldsymbol{b}}(i)\\ &= -(1 - \eta s_i)^{t+1}\tilde{\mathbf{w}}_0(i) + \eta s_i\left(\sum_{j=0}^{t}(1 - \eta s_i)^j\right)\tilde{\boldsymbol{b}}(i)\end{aligned}$$

This can be simplified to eventually get $\tilde{\mathbf{w}}_t(i) = (1 - \eta s_i)^t\tilde{\mathbf{w}}_0(i) + (1 - (1 - \eta s_i)^t)\tilde{\boldsymbol{b}}(i)$. By observing that $\bar{\boldsymbol{b}} = \boldsymbol{M}^\dagger\boldsymbol{b}$ and $\tilde{\boldsymbol{b}} = \boldsymbol{B}^{-1}\bar{\boldsymbol{b}} = \boldsymbol{B}^{-1}\boldsymbol{M}^\dagger\boldsymbol{b} = \boldsymbol{S}^\dagger\boldsymbol{B}^{-1}\boldsymbol{b}$, we can summarize the dynamics as $\tilde{\mathbf{w}}_t = (I_d - \eta\boldsymbol{S})^t\tilde{\mathbf{w}}_0 + (I_d - (I_d - \eta\boldsymbol{S})^t)\boldsymbol{S}^\dagger\boldsymbol{B}^{-1}\boldsymbol{b}$. Using $\tilde{\mathbf{w}}_0 = \boldsymbol{B}^{-1}\mathbf{w}_0$ and $\mathbf{w}_t = \boldsymbol{B}\tilde{\mathbf{w}}_t$, multiplying by $\boldsymbol{B}$ on both sides completes the proof, i.e. $\mathbf{w}_t = \boldsymbol{B}(I_d - \eta\boldsymbol{S})^t\boldsymbol{B}^{-1}\mathbf{w}_0 + \boldsymbol{B}(I_d - (I_d - \eta\boldsymbol{S})^t)\boldsymbol{S}^\dagger\boldsymbol{B}^{-1}\boldsymbol{b}$. By observing that $\boldsymbol{B}(I_d - \eta\boldsymbol{S})^t\boldsymbol{B}^{-1} = (I_d - \eta\boldsymbol{B}\boldsymbol{S}\boldsymbol{B}^{-1})^t = (I_d - \eta\boldsymbol{M})^t$, we get

$$\begin{aligned}\mathbf{w}_t &= \boldsymbol{B}(I_d - \eta\boldsymbol{S})^t\boldsymbol{B}^{-1}\mathbf{w}_0 + \boldsymbol{B}(I_d - (I_d - \eta\boldsymbol{S})^t)\boldsymbol{B}^{-1}\boldsymbol{B}\boldsymbol{S}^\dagger\boldsymbol{B}^{-1}\boldsymbol{b}\\ &= (I_d - \eta\boldsymbol{M})^t\mathbf{w}_0 + (\boldsymbol{B}\boldsymbol{B}^{-1} - \boldsymbol{B}(I_d - \eta\boldsymbol{S})^t\boldsymbol{B}^{-1})\boldsymbol{B}\boldsymbol{S}^\dagger\boldsymbol{B}^{-1}\boldsymbol{b}\\ &= (I_d - \eta\boldsymbol{M})^t\mathbf{w}_0 + (I_d - (I_d - \eta\boldsymbol{M})^t)\boldsymbol{M}^\dagger\boldsymbol{b}\end{aligned}$$

which completes the proof □

Proving Lemma B.1: We restate the statement of the Lemma B.1 here for convenience.

(Lemma B.1). *Let $S = (\boldsymbol{X}, \xi)$ be a sample from $\rho_{\boldsymbol{v}}$, $\boldsymbol{X} \in \mathbb{R}^{n\times d}, \xi \in \mathbb{R}^n$. Let $\Sigma_{\boldsymbol{X}} = \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{x}_i\boldsymbol{x}_i^\top \in \mathbb{R}^{d\times d}$*

$$\mathtt{GD}^{\lambda}_{reg}(S; \mathbf{w}_0) = (I_d - (\Sigma_{\boldsymbol{X}} + \lambda I_d)^\dagger(\Sigma_{\boldsymbol{X}} + \lambda I_d))\mathbf{w}_0 + (\Sigma_{\boldsymbol{X}} + \lambda I_d)^\dagger\Sigma_{\boldsymbol{X}}\boldsymbol{v} + \frac{1}{n}(\Sigma_{\boldsymbol{X}} + \lambda I_d)^\dagger\boldsymbol{X}^\top\xi$$

$$\mathtt{GD}^{\eta,t_0}_{step}(S; \mathbf{w}_0) = (I_d - \eta\Sigma_{\boldsymbol{X}})^{t_0}\mathbf{w}_0 + (I_d - (I_d - \eta\Sigma_{\boldsymbol{X}})^{t_0})\Sigma_{\boldsymbol{X}}^\dagger\Sigma_{\boldsymbol{X}}\boldsymbol{v} + \frac{1}{n}(I_d - (I_d - \eta\Sigma_{\boldsymbol{X}})^{t_0})\Sigma_{\boldsymbol{X}}^\dagger\boldsymbol{X}^\top\xi$$

Proof of Lemma B.1. We first prove the result for $\mathtt{GD}^{\lambda}_{reg}$. Recall the definition of the regularized loss from Equation 6 and the dynamics for $\mathtt{GD}^{\lambda}_{reg}$

$$\ell_{S,\lambda}(\mathbf{w}) = \frac{1}{n}\sum_{i=1}^{n}(\mathbf{w}^\top\boldsymbol{x}_i - y_i)^2 + \frac{\lambda}{2}\|\mathbf{w}\|^2; \quad \frac{d\mathbf{w}_t}{dt} = -\nabla_{\mathbf{w}}\ell_{S,\lambda}(\mathbf{w}_t)$$

where $y_i = \boldsymbol{v}^\top\boldsymbol{x}_i + \xi_i$. The gradient of $\ell_{S,\lambda}$ is

$$\begin{aligned}\nabla_{\mathbf{w}}\ell_{S,\lambda}(\mathbf{w}) &= \frac{1}{n}\sum_{i=1}^{n}(\mathbf{w}^\top\boldsymbol{x}_i - y_i)\boldsymbol{x}_i + \lambda\mathbf{w} = \frac{1}{n}\sum_{i=1}^{n}(\mathbf{w}^\top\boldsymbol{x}_i - \boldsymbol{v}^\top\boldsymbol{x}_i - \xi_i)\boldsymbol{x}_i + \lambda\mathbf{w}\\ &= \left(\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{x}_i\boldsymbol{x}_i^\top + \lambda I_d\right)\mathbf{w} - \left(\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{x}_i\boldsymbol{x}_i^\top\right)\boldsymbol{v} - \frac{1}{n}\boldsymbol{X}^\top\xi\end{aligned}$$

$$= (\Sigma_{\boldsymbol{X}} + \lambda I_d)\mathbf{w} - \Sigma_{\boldsymbol{X}}\boldsymbol{v} - \frac{1}{n}\boldsymbol{X}^\top \xi$$

If $\boldsymbol{M} = (\Sigma_{\boldsymbol{X}} + \lambda I_d)$ and $\boldsymbol{b} = \Sigma_{\boldsymbol{X}}\boldsymbol{v} + \frac{1}{n}\boldsymbol{X}^\top\xi$, then $\nabla_{\mathbf{w}}\ell_{S,\lambda}(\mathbf{w}) = \boldsymbol{M}\mathbf{w} - \boldsymbol{b}$ and the dynamics are $\frac{d\mathbf{w}_t}{dt} = -\boldsymbol{M}\mathbf{w}_t + \boldsymbol{b}$. Note that $\boldsymbol{b}$ is in the range of $\boldsymbol{M}$ for every $\lambda \geq 0$; this is obvious for $\lambda > 0$ when $\boldsymbol{M}$ is full rank, but even $\lambda = 0$, since $\boldsymbol{b}$ lies in the span of rows of $\boldsymbol{X}$, it lies in the span of $\Sigma_{\boldsymbol{X}} = \frac{1}{n}\boldsymbol{X}^\top\boldsymbol{X}$. Thus by Lemma B.5, we get that $\mathbf{w}_\infty = (I_d - \boldsymbol{M}^\dagger\boldsymbol{M})\mathbf{w}_0 + \boldsymbol{M}^\dagger\boldsymbol{b}$. Plugging in values of $\boldsymbol{M}$ and $\boldsymbol{b}$ gives the desired closed form for $\texttt{GD}^{\lambda}_{reg}$.

We now derive the closed form solution for $\texttt{GD}^{\eta,t_0}_{step}$. Recall the dynamics of $\texttt{GD}^{\eta,t_0}_{step}$

$$\ell_S(\mathbf{w}) = \frac{1}{n}\sum_{i=1}^{n}(\mathbf{w}^\top\boldsymbol{x}_i - y_i)^2; \;\; \mathbf{w}_{t+1} = \mathbf{w}_t - \eta\nabla_{\mathbf{w}}\ell_S(\mathbf{w}_t)$$

where again $y_i = \boldsymbol{v}^\top\boldsymbol{x}_i + \xi_i$. The gradient of ℓ_S is

$$\begin{aligned}
\nabla_{\mathbf{w}}\ell_S(\mathbf{w}) &= \frac{1}{n}\sum_{i=1}^{n}(\mathbf{w}^\top\boldsymbol{x}_i - y_i)\boldsymbol{x}_i = \frac{1}{n}\sum_{i=1}^{n}(\mathbf{w}^\top\boldsymbol{x}_i - \boldsymbol{v}^\top\boldsymbol{x}_i - \xi_i)\boldsymbol{x}_i \\
&= \left(\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{x}_i\boldsymbol{x}_i^\top\right)\mathbf{w} - \left(\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{x}_i\boldsymbol{x}_i^\top\right)\boldsymbol{v} - \frac{1}{n}\boldsymbol{X}^\top\xi \\
&= \Sigma_{\boldsymbol{X}}\mathbf{w} - \Sigma_{\boldsymbol{X}}\boldsymbol{v} - \frac{1}{n}\boldsymbol{X}^\top\xi
\end{aligned}$$

Setting $\boldsymbol{M} = \Sigma_{\boldsymbol{X}}$ and $\boldsymbol{b} = \Sigma_{\boldsymbol{X}}\boldsymbol{v} + \frac{1}{n}\boldsymbol{X}^\top\xi$, we get $\nabla_{\mathbf{w}}\ell_S(\mathbf{w}) = \boldsymbol{M}\mathbf{w} - \boldsymbol{b}$ and the dynamics are $\mathbf{w}_{t+1} - \mathbf{w}_t = -\eta(\boldsymbol{M}\mathbf{w}_t - \boldsymbol{b})$. Again since $\boldsymbol{b}$ is in the span of $\boldsymbol{M}$, we can use Lemma B.1 to get $\mathbf{w}_{t_0} = (I_d - \eta\boldsymbol{M})^{t_0}\mathbf{w}_0 + (I_d - (I_d - \eta\boldsymbol{M})^{t_0})\boldsymbol{M}^\dagger\boldsymbol{b}$. Plugging in the values of $\boldsymbol{M}$ and $\boldsymbol{b}$ completes the proof for $\texttt{GD}^{\eta,t_0}_{step}$. □

B.3. Other proofs

Proving Lemma B.2

Proof of Lemma B.2. We start by looking at the loss for $s\mathbf{w}_*$ for $s \in \{\pm 1\}$

$$\begin{aligned}
\mathop{\mathbb{E}}_{S\sim\rho_{s\mathbf{w}_*}}\left[\ell_{s\mathbf{w}_*}(\texttt{Alg}(\cdot;\mathbf{w}_0)) - \sigma^2\right] &= \mathop{\mathbb{E}}_{S\sim\rho_{s\mathbf{w}_*}}\left[\|\texttt{Alg}(S;\mathbf{w}_0)) - s\mathbf{w}_*\|^2\right] \\
&= \mathop{\mathbb{E}}_{\substack{\boldsymbol{X}\sim\mathcal{N}^n(0,I_d)\\ \xi\sim\mathcal{N}(0,\sigma^2 I_m)}}\left[\|\boldsymbol{A}_{\boldsymbol{X}}\mathbf{w}_0 + \boldsymbol{B}_{\boldsymbol{X}}(s\mathbf{w}_*) + \boldsymbol{C}_{\boldsymbol{X}}\xi - s\mathbf{w}_*\|^2\right] \\
&= \mathop{\mathbb{E}}_{\boldsymbol{X},\xi}\|\boldsymbol{A}_{\boldsymbol{X}}\mathbf{w}_0 - (I_d - \boldsymbol{B}_{\boldsymbol{X}})s\mathbf{w}_* + \boldsymbol{C}_{\boldsymbol{X}}\xi\|^2 \\
&=^{(a)} \mathop{\mathbb{E}}_{\boldsymbol{X}}\|(\boldsymbol{A}_{\boldsymbol{X}}\mathbf{w}_0 - (I_d - \boldsymbol{B}_{\boldsymbol{X}})s\mathbf{w}_*\|^2 + \mathop{\mathbb{E}}_{\boldsymbol{X},\xi}\|\boldsymbol{C}_{\boldsymbol{X}}\xi\|^2 \\
&=^{(b)} \mathop{\mathbb{E}}_{\boldsymbol{X}}\|(s\boldsymbol{A}_{\boldsymbol{X}}\mathbf{w}_0 - (I_d - \boldsymbol{B}_{\boldsymbol{X}})\mathbf{w}_*\|^2 + \mathop{\mathbb{E}}_{\boldsymbol{X},\xi}\|\boldsymbol{C}_{\boldsymbol{X}}\xi\|^2
\end{aligned}$$

where (a) uses the fact that $\boldsymbol{X}$ and ξ are independent and $\mathop{\mathbb{E}}_{\xi}\xi = 0$ and (b) uses $s^2 = 1$. Thus we get,

$$\begin{aligned}
\mathcal{E}_n(\texttt{GD}^{\lambda}_{reg}(\cdot;\mathbf{w}_0),\mu_{\mathbf{w}_*}) &= \mathop{\mathbb{E}}_{s\sim\{\pm1\}}\left[\mathop{\mathbb{E}}_{S\sim\rho_{s\mathbf{w}_*}}\left[\ell_{s\mathbf{w}_*}(\texttt{GD}^{\lambda}_{reg}(\cdot;\mathbf{w}_0))\right] - \sigma^2\right. \\
&= \mathop{\mathbb{E}}_{s\sim\{\pm1\}}\left[\mathop{\mathbb{E}}_{S\sim\rho_{s\mathbf{w}_*}}\left[\ell_{s\mathbf{w}_*}(\texttt{GD}^{\lambda}_{reg}(\cdot;\mathbf{w}_0)) - \sigma^2\right]\right] \\
&= \mathop{\mathbb{E}}_{s\sim\{\pm1\}}\mathop{\mathbb{E}}_{\boldsymbol{X},\xi}\left[\|(s\boldsymbol{A}_{\boldsymbol{X}}\mathbf{w}_0 - (I_d - \boldsymbol{B}_{\boldsymbol{X}})\mathbf{w}_*\|^2 + \|\boldsymbol{C}_{\boldsymbol{X}}\xi\|^2\right] \\
&= \mathop{\mathbb{E}}_{\boldsymbol{X}}\mathop{\mathbb{E}}_{s\sim\{\pm1\}}\left[\|(s\boldsymbol{A}_{\boldsymbol{X}}\mathbf{w}_0 - (I_d - \boldsymbol{B}_{\boldsymbol{X}})\mathbf{w}_*\|^2\right] + \mathop{\mathbb{E}}_{\boldsymbol{X},\xi}\left[\xi^\top\boldsymbol{C}_{\boldsymbol{X}}^\top\boldsymbol{C}_{\boldsymbol{X}}\xi\right]
\end{aligned}$$

$$\geq^{(a)} \mathop{\mathbb{E}}_{\boldsymbol{X}}\left[\|(\mathop{\mathbb{E}}_{s\sim\{\pm1\}}[s]\boldsymbol{A}_{\boldsymbol{X}}\mathbf{w}_0-(I_d-\boldsymbol{B}_{\boldsymbol{X}})\mathbf{w}_*\|^2\right]+\mathop{\mathbb{E}}_{\boldsymbol{X},\xi}\left[\xi^\top\boldsymbol{C}_{\boldsymbol{X}}^\top\boldsymbol{C}_{\boldsymbol{X}}\xi\right]$$
$$=^{(b)} \mathop{\mathbb{E}}_{\boldsymbol{X}}\left[\|(I_d-\boldsymbol{B}_{\boldsymbol{X}})\mathbf{w}_*\|^2\right]+\mathop{\mathbb{E}}_{\boldsymbol{X}}\left[\mathop{\mathbb{E}}_{\xi}\mathrm{tr}(\boldsymbol{C}_{\boldsymbol{X}}^\top\boldsymbol{C}_{\boldsymbol{X}}\xi\xi^\top)\right]$$
$$=^{(c)} \mathop{\mathbb{E}}_{\boldsymbol{X}}\left[\|(I_d-\boldsymbol{B}_{\boldsymbol{X}})\mathbf{w}_*\|^2\right]+\mathop{\mathbb{E}}_{\boldsymbol{X}}[\sigma^2\mathrm{tr}(\boldsymbol{C}_{\boldsymbol{X}}^\top\boldsymbol{C}_{\boldsymbol{X}})]$$

where (a) is true by convexity of the quadratic function in s, (b) uses $\xi^\top\boldsymbol{P}\xi=\mathrm{tr}(\boldsymbol{P}\xi\xi^\top)$ for any $d\times d$ matrix $\boldsymbol{P}$, (c) uses the linearity of tr operator and the fact that $\mathop{\mathbb{E}}_{\xi}\xi\xi^\top=\sigma^2 I_m$ when $\xi\sim\mathcal{N}(0,\sigma^2 I_m)$. This completes the proof. Note that we only needed first and second moment conditions on ξ to prove this lemma. □

Proving Lemma B.4

Proof of Lemma B.4. If $n\geq d$, we just follow the steps below that heavily use Jensen's inequality due to the convexity of f.

$$\mathop{\mathbb{E}}_{\boldsymbol{S}}\left[\sum_{i=1}^d f(s_i)\right]=\mathop{\mathbb{E}}_{\boldsymbol{S}}\left[d\mathop{\mathbb{E}}_{i\sim[d]}[f(s_i)]\right]\geq^{(a)} d\mathop{\mathbb{E}}_{\boldsymbol{S}}\left[f\left(\mathop{\mathbb{E}}_{i\sim[d]}[s_i]\right)\right]=d\mathop{\mathbb{E}}_{\boldsymbol{S}}\left[f\left(\frac{1}{d}\mathrm{tr}(\boldsymbol{S})\right)\right]=^{(b)} d\mathop{\mathbb{E}}_{\boldsymbol{X}}\left[f\left(\frac{1}{d}\mathrm{tr}(\Sigma_{\boldsymbol{X}})\right)\right]$$
$$\geq^{(c)} df\left(\frac{1}{d}\mathop{\mathbb{E}}_{\boldsymbol{X}}[\mathrm{tr}(\Sigma_{\boldsymbol{X}})]\right)=df\left(\frac{1}{d}\mathrm{tr}(\mathop{\mathbb{E}}_{\boldsymbol{X}}\Sigma_{\boldsymbol{X}})\right)=^{(d)} df\left(\frac{1}{d}\mathrm{tr}(I_d)\right)=df\,(1)$$

where (a) follows from Jensen's inequality, (b) follows from the fact that $\mathrm{tr}(\boldsymbol{S})=\mathrm{tr}(\boldsymbol{S}\boldsymbol{V}^\top\boldsymbol{V})=\mathrm{tr}(\boldsymbol{V}\boldsymbol{S}\boldsymbol{V}^\top)=\mathrm{tr}(\Sigma_{\boldsymbol{X}})$, (c) follows from Jensen's inequality and (d) follows from $\mathop{\mathbb{E}}_{\boldsymbol{X}}\Sigma_{\boldsymbol{X}}=\mathop{\mathbb{E}}_{\boldsymbol{X}}\frac{1}{n}\sum_{i=1}^n\boldsymbol{x}_i\boldsymbol{x}_i^\top=\frac{1}{n}\sum_{i=1}^n\mathop{\mathbb{E}}_{\boldsymbol{X}}\boldsymbol{x}_i\boldsymbol{x}_i^\top=\frac{1}{n}\sum_{i=1}^n I_d=I_d$.

When $n<d$, we know that $\boldsymbol{X}$ (and hence $\Sigma_{\boldsymbol{X}}$) has rank at most $n<d$, thus the $d-n$ smallest eigenvalues are 0, i.e. $s_i=0$ for $n+1\leq i\leq d$. Note that $\sum_{i=1}^d s_i=\sum_{i=1}^n s_i=\mathrm{tr}(\boldsymbol{S})$. Following the steps below,

$$\mathop{\mathbb{E}}_{\boldsymbol{S}}\left[\sum_{i=1}^d f(s_i)\right]=\mathop{\mathbb{E}}_{\boldsymbol{S}}\left[\sum_{i=1}^n f(s_i)\right]+\mathop{\mathbb{E}}_{\boldsymbol{S}}\left[\sum_{i=n+1}^d f(s_i)\right]=\mathop{\mathbb{E}}_{\boldsymbol{S}}\left[n\mathop{\mathbb{E}}_{i\sim[n]}[f(s_i)]\right]+\mathop{\mathbb{E}}_{\boldsymbol{S}}\left[\sum_{i=n+1}^d f(0)\right]$$
$$\geq\mathop{\mathbb{E}}_{\boldsymbol{S}}\left[nf(\mathop{\mathbb{E}}_{i\sim[n]}[s_i])\right]+(d-n)f(0)=\mathop{\mathbb{E}}_{\boldsymbol{S}}\left[nf\left(\frac{1}{n}\mathrm{tr}(\boldsymbol{S})\right)\right]+(d-n)f(0)$$
$$=\mathop{\mathbb{E}}_{\boldsymbol{X}}\left[nf\left(\frac{1}{n}\mathrm{tr}(\Sigma_{\boldsymbol{X}})\right)\right]+(d-n)f(0)\geq nf\left(\frac{1}{n}\mathrm{tr}(\mathop{\mathbb{E}}_{\boldsymbol{X}}\Sigma_{\boldsymbol{X}})\right)+(d-n)f(0)$$
$$=nf\left(\frac{d}{n}\right)+(d-n)f(0)$$

□

C. Non-convex proofs

C.1. Theorems and Lemmas for Reptile

Let $\bar{\boldsymbol{A}}_{i+1},\bar{\mathbf{w}}_{i+1}=\mathtt{GD}_{pop}(\ell_{\rho_{i+1}},(\boldsymbol{A}_i,\mathbf{w}_i))$ be the solution for task ρ_{i+1} that is found by gradient descent starting from current initialization. Thus the reptile update is $\boldsymbol{A}_{i+1}=(1-\tau)\boldsymbol{A}_i+\tau\bar{\boldsymbol{A}}_{i+1}$ and $\mathbf{w}_{i+1}=(1-\tau)\mathbf{w}_i+\tau\bar{\mathbf{w}}_{i+1}$. Let $\bar{\mathbf{w}}_*=\mathbf{w}_*/\|\mathbf{w}_*\|$ be the unit vector and let $r=\|\mathbf{w}_*\|$.

Lemma C.1. *Given a sequence of tasks $\rho_{1:T}$ where $\rho_i=\rho_{s_i\mathbf{w}_*}$ for $s_i\in\{\pm1\}$. Starting with $\boldsymbol{A}_0=\kappa I_d,\mathbf{w}=\mathbf{0}_d$, then the initialization learned by* `Reptile` *satisfies the following at every step $i\in[T]$*

$$\boldsymbol{A}_i=(a_i-\kappa)\bar{\mathbf{w}}_*\bar{\mathbf{w}}_*+\kappa I_d,\ \mathbf{w}_i=b_i\bar{\mathbf{w}}_*$$

where

$$a_0=\kappa,\ b_0=0$$

$$c_i = a_i^2 - b_i^2,\ a_{i+1} = (1-\tau)a_i + \tau\bar{a}_{i+1},\ b_{i+1} = (1-\tau)b_i + \tau\ s_{i+1}\bar{b}_{i+1}$$
$$\bar{a}_{i+1} = \sqrt{\frac{c_i + \sqrt{4r^2 + c_i^2}}{2}}, \bar{b}_{i+1} = \sqrt{\frac{-c_i + \sqrt{4r^2 + c_i^2}}{2}}$$

The following key lemma about the solution of gradient flow for a single task starting from an initialization is crucial to prove the above lemma.

Lemma C.2. *Starting from $\boldsymbol{A}(0) = (a(0) - \kappa)\bar{\mathbf{w}}_*\bar{\mathbf{w}}_*^\top + \kappa I_d$, $\mathbf{w}(0) = b(0)\bar{\mathbf{w}}_*$, with $a(0) > b(0)$, the solution of gradient flow on loss $\ell_{s\mathbf{w}_*}$ for $s \in \{\pm 1\}$, is $\bar{\boldsymbol{A}}, \bar{\mathbf{w}} = GD_{pop}(\ell_{s\mathbf{w}_*}, (\boldsymbol{A}, \mathbf{w}))$, where*

$$\bar{\boldsymbol{A}} = (\bar{a} - \kappa)\bar{\mathbf{w}}_*\bar{\mathbf{w}}_*^\top + \kappa I_d,\ \bar{\mathbf{w}} = \bar{b}\bar{\mathbf{w}}_*\text{ , where}$$
$$\bar{a} = \sqrt{\frac{c + \sqrt{4r^2 + c^2}}{2}},\ \bar{b} = s\sqrt{\frac{-c + \sqrt{4r^2 + c^2}}{2}},\ c = a(0)^2 - b(0)^2$$

Proof of Lemma C.1. We prove this using a simple induction by assuming Lemma C.2. It is clear for $i = 0$ that $a_0 = \kappa$ and $b_0 = 0$. Suppose $\boldsymbol{A}_i = (a_i - \kappa)\bar{\mathbf{w}}_*\bar{\mathbf{w}}_* + \kappa I_d,\ \mathbf{w}_i = b_i\bar{\mathbf{w}}_*$. From Lemma C.2, we get that $\bar{\boldsymbol{A}}_{i+1} = (\bar{a}_i - \kappa)\bar{\mathbf{w}}_*\bar{\mathbf{w}}_* + \kappa I_d,\ \mathbf{w}_{i+1} = s_{i+1}\bar{b}_i\bar{\mathbf{w}}_*$. Doing the interpolation step completes the proof. □

Proof of Lemma C.2. The proof uses ideas from Saxe et al. (2014; 2019); Gidel et al. (2019), where the dynamics of linear networks is analyzed in the case where the subspace of the initialization is aligned with the target $s\bar{\mathbf{w}}_*$. We provide a proof of this lemma by borrowing the key ideas those works. Let $\boldsymbol{U} \in \mathbb{R}^{d\times d}$ be an orthonormal matrix, i.e. $\boldsymbol{U}^\top\boldsymbol{U} = I_d$, whose first column is $\bar{\mathbf{w}}_*$. Thus we can rewrite $\boldsymbol{A}(0) = \boldsymbol{U}\Lambda_1(0)\boldsymbol{U}^\top$, where $\Lambda_1(0) \in \mathbb{R}^{d\times d}$ is a diagonal matrix that looks like $\Lambda_1(0) = diag(a(0), \kappa, \ldots, \kappa)$, $\mathbf{w}(0) = \boldsymbol{U}\Lambda_2(0)$, where $\Lambda_2(0) = (b(0), 0, \ldots, 0) \in \mathbb{R}^d$ and $s\mathbf{w}_* = \boldsymbol{U}\Lambda_*$, where $\Lambda_* = (sr, 0, \ldots, 0) \in \mathbb{R}^d$. The loss to run gradient flow on is $\ell_{s\mathbf{w}_*}(\boldsymbol{A}, \mathbf{w}) = \|\boldsymbol{A}^\top\mathbf{w} - s\mathbf{w}_*\|^2$. Dynamics of gradient flow is

$$\frac{d\boldsymbol{A}(t)}{dt} = -\nabla_{\boldsymbol{A}}\ell_{s\mathbf{w}_*}(\boldsymbol{A}(t), \mathbf{w}(t)) = s\mathbf{w}(t){\mathbf{w}_*}^\top - \mathbf{w}(t)\mathbf{w}(t)^\top\boldsymbol{A}(t)$$
$$\frac{d\mathbf{w}(t)}{dt} = -\nabla_{\mathbf{w}}\ell_{s\mathbf{w}_*}(\boldsymbol{A}(t), \mathbf{w}(t)) = s\boldsymbol{A}(t)\mathbf{w}_* - \boldsymbol{A}(t)\boldsymbol{A}(t)^\top\mathbf{w}(t)$$

Just like Saxe et al. (2014); Gidel et al. (2019), we define $\Lambda_1(t) = \boldsymbol{U}^\top\boldsymbol{A}(t)\boldsymbol{U}$, $\Lambda_2(t) = \boldsymbol{U}^\top\mathbf{w}(t)$, $\Lambda_* = \boldsymbol{U}^\top\mathbf{w}_*$. Thus

$$\frac{d\Lambda_1(t)}{dt} = \Lambda_2(t)\Lambda_*^\top - \Lambda_2(t)\Lambda_2(t)^\top\Lambda_1(t)$$
$$\frac{d\Lambda_2(t)}{dt} = \Lambda_1(t)\Lambda_* - \Lambda_1(t)\Lambda_1(t)^\top\Lambda_2(t)$$

By a similar argument, we see that the time derivative of Λ_1 is non-zero only for the first diagonal entry while the derivative of Λ_2 is non-zero only for the first entry. Thus the entire dynamics can be summarized by the dynamics of two scalar values

$$\frac{da(t)}{dt} = b(t)sr - b(t)^2a(t)$$
$$\frac{db(t)}{dt} = a(t)sr - a(t)^2b(t)$$

Using the hyperbolic change of coordinates of $(a(t), b(t)) = (\sqrt{c}\cosh(\theta/2), \sqrt{c}\sinh(\theta/2))$ and the analysis in Appendix A from Saxe et al. (2014), we have that the fixed point of the dynamics is at $\bar{\theta} = \sinh^{-1}(2rs/c)$, thus giving the solutions

$$\bar{a} = \sqrt{c}\cosh(\bar{\theta}/2) = \sqrt{\frac{c + \sqrt{4r^2 + c^2}}{2}}$$
$$\bar{b} = \sqrt{c}\sinh(\bar{\theta}/2) = \sqrt{\frac{-c + \sqrt{4r^2 + c^2}}{2}}$$

□

We now prove the key theorem that shows how the reptile update amplifies the component of the first layer in the direction of $\mathbf{w}_*$. Precisely, it shows that with high probability over sampling of the training tasks, a_T from Lemma C.1 is large for appropriate choice of T and τ.

Theorem C.3. *Suppose $\{a_i, b_i, c_i\}_{i=1}^T$ follow the dynamics from Lemma C.1 with $\{s_1, \ldots, s_T\} \sim \{\pm 1\}^T$. Then with probability at least $1-\delta$, $a_T \geq \min\left\{\frac{\sqrt{r}}{2\sqrt{\tau \log(T/\delta)}}, \sqrt{r}\frac{(\tau T)^{1/4}}{2}\right\}$. Picking $\tau = T^{-1/3}\log(2T/\delta)^{-2/3}$, we get that $a_T \geq \frac{\sqrt{r}T^{1/6}\log(2T/\delta)^{-1/6}}{2} = \tilde{\Omega}(\sqrt{r}T^{1/6})$*

Proof. The proof has 3 mains steps

- **Step 1:** a_i is non-decreasing and the increment in a_i is a decreasing function of $|a_i b_i|$. Also $|a_i b_i| \leq r$.
- **Step 2:** With high probability, $|b_i|$ is small
- **Step 3:** Either $|a_i b_i|$ is small, which gives an increment in a_i, otherwise, or $a_i = \Omega(1/|b_i|)$ is large since $|b_i|$ is small

Step 1: We first prove that a_i is non-decreasing, which happens if $\bar{a}_{i+1} \geq a_i$ for every i.

$$\begin{aligned}\bar{a}_{i+1}^2 - a_i^2 &= \frac{c_i + \sqrt{4r^2 + c_i^2}}{2} - a_i^2 = \frac{a_i^2 - b_i^2 + \sqrt{4r^2 + (a_i^2 - b_i^2)^2} - 2a_i^2}{2} \\ &= \frac{\sqrt{4r^2 + (a_i^2 - b_i^2)^2} - (a_i^2 + b_i^2)}{2} = \frac{\sqrt{4(r^2 - a_i^2 b_i^2) + (a_i^2 + b_i^2)^2} - (a_i^2 + b_i^2)}{2}\end{aligned}$$

Thus $|a_i b_i| < r$ will ensure that a_i is non-decreasing. We show that using induction, $|a_0 b_0| = 0$ and assume $|a_i b_i| \leq r$. Notice that since $\bar{a}_{i+1}^2 - \bar{b}_{i+1} = a_i^2 - b_i^2$, we have that $(|a_i| - |\bar{a}_{i+1}|)(|b_i| - |\bar{b}_{i+1}|) \geq 0$.

$$\begin{aligned}|a_{i+1}b_{i+1}| &= |(1-\tau)^2 a_i b_i + \tau^2 \bar{a}_{i+1}\bar{b}_{i+1} + \tau(1-\tau)a_i\bar{b}_{i+1} + \tau(1-\tau)b_i\bar{a}_{i+1} \\ &\leq (1-\tau)^2|a_i b_i| + \tau^2|\bar{a}_{i+1}\bar{b}_{i+1}| + \tau(1-\tau)[|a_i||\bar{b}_{i+1}| + |b_i||\bar{a}_{i+1}|] \\ &\leq (1-\tau)^2 r + \tau^2 r + \tau(1-\tau)[|a_i||b_i| + |\bar{a}_{i+1}||\bar{b}_{i+1}|] + \tau(1-\tau)(|a_i| - |\bar{a}_{i+1}|)(|\bar{b}_{i+1}| - |b_i|) \\ &\leq (1-\tau)^2 r + \tau^2 r + \tau(1-\tau)[r + r] + 0 = r\end{aligned}$$

Thus finishing the first step in the proof.

Step 2: We now move to the second step about $|b_i|$ being small.

Proposition C.4. *With probability at least $1-\delta$ over $\{s_1, \ldots, s_T\}$, $|b_i| \leq \sqrt{2r\tau\log(2T/\delta)}$, for every $i \in [T]$*

From the dynamics, we have $X_{i+1} = b_{i+1} - (1-\tau)b_i = \tau s_{i+1}\bar{b}_{i+1}$. Note that $\bar{b}_{i+1}$ depends only on $s_{1:i}$ and $|\bar{b}_{i+1}| \leq \sqrt{r}$, thus conditioned on $s_{1:i}$, X_{i+1} is $\tau\sqrt{r}$ sub-gaussian and $\mathbb{E}[X_{i+1}|s_{1:i}] = \tau\,\mathbb{E}[s_{i+1}\bar{b}_{i+1}|s_{1:i}] = 0 = \tau\bar{b}_{i+1}\,\mathbb{E}[s_{i+1}|s_{1:i}] = 0$.

It is easy to verify that we can rewrite $b_{i+1} = X_{i+1} + (1-\tau)X_i + (1-\tau)^2 X_{i-1} + \cdots + (1-\tau)^i X_1$, where we also use the fact that $b_0 = 0$. Using Markov's inequality we get

$$\begin{aligned}\Pr(b_{i+1} > \nu) = \Pr(e^{tb_{i+1}} > e^{t\nu}) &\leq e^{-t\nu}\,\mathbb{E}\,e^{tb_{i+1}} = e^{-t\nu}\,\mathbb{E}\,e^{t\sum_{j=0}^{i+1}(1-\tau)^j X_{i+1-j}} \\ &= e^{-t\nu}\,\mathbb{E}\prod_{j=0}^{i+1} e^{t(1-\tau)^j X_{i+1-j}} = e^{-t\nu}\prod_{j=0}^{i}\mathbb{E}[e^{t(1-\tau)^j X_{i+1-j}}|s_{1:i-j}] \\ &\leq^{(a)} e^{-t\nu}\prod_{j=0}^{i} e^{\frac{t^2\tau^2 r(1-\tau)^{2j}}{2}} = e^{-t\nu}e^{\frac{t^2\tau^2 r\sum_{j=0}^{i}(1-\tau)^{2j}}{2}} \\ &\leq e^{-t\nu}e^{\frac{t^2\tau^2 r\sum_{j=0}^{\infty}(1-\tau)^{2j}}{2}} = e^{-t\nu}e^{\frac{t^2\tau^2 r}{2(1-(1-\tau)^2)}} = e^{-t\nu}e^{\frac{t^2\tau^2 r}{2(2\tau-\tau^2)}}\end{aligned}$$

$$\leq^{(b)} e^{-t\nu} e^{\frac{t^2\tau r}{2}}$$

Where for (a) we use the fact that $(1-\tau)^j X_{i+1-j}$ is zero mean and $(1-\tau)^{2j}\tau^2 r$-subgaussian when conditioned on $s_{1:i-j}$, and for (b) we use $\tau < 1$. Picking the optimal value of $t = \frac{\nu}{\tau r}$, we get $\Pr(b_{i+1} > \nu) \leq e^{-\frac{\nu^2}{2\tau r}}$. By using the symmetry of b_{i+1} (since the sequence $\{-s_1, \ldots, -s_T\}$ will give $-b_{i+1}$ instead), we get that $\Pr(b_{i+1} < -\nu) \leq e^{-\frac{\nu^2}{2\tau r}}$ and by union bound we get that $\Pr(\forall i \in [T],\ |b_i| > \nu) \leq 2Te^{-\frac{\nu^2}{2\tau r}}$. Setting $\nu = \sqrt{2r\tau \log(\frac{2T}{\delta})}$, we get $\Pr(\forall i \in [T],\ |b_i| > \nu) \leq \delta$

Step 3: Let $\gamma = \sqrt{2\tau \log(\frac{2T}{\delta})}$; from step 2 we have $|b_i| \leq \sqrt{r}\gamma, \forall i \in [T]$. An easy induction can also show that $|b_i| \leq \sqrt{r}$. To show a_T is large, we assume that $a_T < \alpha$ for some α and see how large T can be without leading to a contradiction. We also assume that $\alpha \geq 1$, this assumptions will be justified in the end. Since a_i is non-decreasing, we also get that $a_i \leq \alpha\sqrt{r}, \forall i \in [T]$. If $a_i b_i \geq \frac{r}{\sqrt{2}}$ for any i, then we have $a_i \geq \frac{r}{\sqrt{2}b_i} \geq \frac{\sqrt{r}}{\sqrt{2}\gamma}$ which would finish the proof. If $a_i b_i < \frac{r}{\sqrt{2}}$ for every i, then we will prove that there is at least a constant increment in a_i. Let $\Delta_i = \bar{a}_{i+1} - a_i$; as shown in step 1, $\Delta_i \geq 0$.

$$\begin{aligned}
(\Delta_i + a_i)^2 - a_i^2 &= \frac{c_i + \sqrt{4r^2 + c_i^2}}{2} - a_i^2 = \frac{a_i^2 - b_i^2 + \sqrt{4r^2 + (a_i^2 - b_i^2)^2}}{2} - a_i^2 \\
&= \frac{\sqrt{4(r^2 - a_i^2 b_i^2) + (a_i^2 + b_i^2)^2} - (a_i^2 + b_i^2)}{2} \\
&\geq^{(a)} \frac{\sqrt{2r^2 + (a_i^2 + b_i^2)^2} - (a_i^2 + b_i^2)}{2} \\
&\geq^{(b)} \frac{\sqrt{2r^2 + r^2(\alpha^2+1)^2} - r(\alpha^2+1)}{2} = r\frac{\sqrt{2 + (\alpha^2+1)^2} - (\alpha^2+1)}{2} \\
&= r\frac{1}{\sqrt{2 + (\alpha^2+1)^2} + (\alpha^2+1)} \geq \frac{r}{\sqrt{2(\alpha^2+1)^2} + (\alpha^2+1)} \\
&= \frac{r}{(\sqrt{2}+1)(\alpha^2+1)}
\end{aligned}$$

where (a) follows because $|a_i b_i| < \frac{r}{\sqrt{2}}$ and (b) follows from the fact that $\sqrt{x + y^2} - y \geq \sqrt{x + z^2} - z$ whenever $y < z$, where x here is $2r^2$, y is $a_i^2 + b_i^2$ and z is $r\alpha^2 + r$. Thus we get

$$\begin{aligned}
\Delta_i &\geq \sqrt{\frac{r}{(\sqrt{2}+1)(\alpha^2+1)} + a_i^2} - a_i \geq^{(a)} \sqrt{\frac{r}{(\sqrt{2}+1)(\alpha^2+1)} + r\alpha^2} - \sqrt{r}\alpha \\
&= \sqrt{r}\left[\sqrt{\frac{1}{(\sqrt{2}+1)(\alpha^2+1)} + \alpha^2} - \alpha\right] = \sqrt{r}\frac{\frac{1}{(\sqrt{2}+1)(\alpha^2+1)}}{\sqrt{\frac{1}{(\sqrt{2}+1)(\alpha^2+1)} + \alpha^2} + \alpha} \\
&\geq \frac{\sqrt{r}}{(\sqrt{2}+1)(\alpha^2+1)^{3/2}} := \Delta
\end{aligned}$$

From the dynamics, $a_{i+1} = a_i + \tau(\bar{a}_{i+1} - a_i) \geq a_i + \tau\Delta_i \geq a_i + \tau\Delta = a_0 + (i+1)\tau\Delta$. Thus $a_T \geq T\tau\Delta$. But we assumed that $a_T \leq \alpha\sqrt{r}$, so we have

$$\sqrt{r}\alpha \geq T\tau\Delta \geq \frac{T\tau\sqrt{r}}{(\sqrt{2}+1)(\alpha^2+1)^{3/2}} \geq \frac{T\tau\sqrt{r}}{(\sqrt{2}+1)(\alpha^2+\alpha^2)^{3/2}} = \frac{T\tau\sqrt{r}}{(\sqrt{2}+4)\alpha^3}$$

Thus we get that $\alpha > \frac{(T\tau)^{1/4}}{2}$. This completes the proof □

We now prove why the initialization learned at the end of Reptile will help with sample complexity of new task. We denote $\boldsymbol{X} \sim \mathcal{N}(0, I_d)^n$ as sampling n i.i.d. vectors from $\mathcal{N}(0, I_d)$ and stacking them into a matrix $\boldsymbol{X} \in \mathbb{R}^{n\times d}$, and $\Sigma_{\boldsymbol{X}} := \frac{1}{n}\boldsymbol{x}_i\boldsymbol{x}_i^\top$

Lemma C.5. *Given a symmetric and invertible $\boldsymbol{A} \in \mathbb{R}^{d\times d}$ as the first layer, the excess risk for learning the second layer is*

$$\mathcal{E}_n(GD2^{\lambda}_{reg}(\cdot; (\boldsymbol{A}, \boldsymbol{0}_d)), \mu_{\mathbf{w}_*}) = \mathop{\mathbb{E}}_{\boldsymbol{X}\sim\mathcal{N}(0,I_d)^n}[\lambda^2 \|\boldsymbol{A}(\boldsymbol{A}\Sigma_{\boldsymbol{X}}\boldsymbol{A} + \lambda I_d)^{-1}\boldsymbol{A}^{-1}\mathbf{w}_*\|^2]$$

$$+\frac{\sigma^2}{n}\mathop{\mathbb{E}}_{\boldsymbol{X}\sim\mathcal{N}(0,I_d)^n}\operatorname{tr}\left(\boldsymbol{A}(\boldsymbol{A}\Sigma_{\boldsymbol{X}}\boldsymbol{A}+\lambda I_d)^{-1}\boldsymbol{A}\Sigma_{\boldsymbol{X}}\boldsymbol{A}(\boldsymbol{A}\Sigma_{\boldsymbol{X}}\boldsymbol{A}+\lambda I_d)^{-1}\boldsymbol{A}\right)$$

Proof of Lemma C.5. By definition, we have

$$\begin{aligned}\mathcal{E}_{n,\lambda}(\boldsymbol{A},\mu_{\mathbf{w}_*}) &= \mathop{\mathbb{E}}_{s\sim\{\pm1\}}\mathop{\mathbb{E}}_{S\sim\rho^n_{s\mathbf{w}_*}}\ell_{s\mathbf{w}_*}(\texttt{GD2}^{\lambda}_{reg}(S;(\boldsymbol{A},\mathbf{0}_d)))-\sigma^2\\ &= \mathop{\mathbb{E}}_{s\sim\{\pm1\}}\mathop{\mathbb{E}}_{S\sim\rho^n_{s\mathbf{w}_*}}\|s\mathbf{w}_*-\texttt{GD2}^{\lambda}_{reg}(S;(\boldsymbol{A},\mathbf{0}_d)))\|^2\end{aligned}$$

We first compute the inner expectation for $s=1$, a similar calculation will work for $s=-1$. First, we state the solution for GD for the regularized loss $\ell_{S,\lambda}$ starting from $\boldsymbol{A}$ and we prove this later. Let $S=(\boldsymbol{X},\boldsymbol{y})$ be all the samples and predictions, where $\boldsymbol{X}\in\mathbb{R}^{n\times d}$ and $\boldsymbol{y}\in\mathbb{R}^n$. Define $\xi=\boldsymbol{y}-\boldsymbol{X}^\top\mathbf{w}_*$ to be the noise in the predictions; by the definition of $\rho_{\mathbf{w}_*}$, we have that $\xi\sim\mathcal{N}(0,\sigma^2 I_n)$. We can now write the solution to $\texttt{GD2}^{\lambda}_{reg}$ by using Lemma B.5 as

$$\texttt{GD2}^{\lambda}_{reg}(S;(\boldsymbol{A},\mathbf{0}_d))=(\boldsymbol{A}\Sigma_{\boldsymbol{X}}\boldsymbol{A}+\lambda I_d)^{-1}(\boldsymbol{A}\Sigma_{\boldsymbol{X}}\mathbf{w}_*+\frac{1}{n}\boldsymbol{A}\boldsymbol{X}^\top\xi)$$

The intuition is that $\ell_{S,\lambda}(\cdot,\boldsymbol{A})$ has a unique solution because of the regularization, and gradient descent converges to that unique solution. Using this, we can compute the excess risk for $\rho_{\mathbf{w}_*}$

$$\begin{aligned}\mathop{\mathbb{E}}_{S\sim\rho^n_{\mathbf{w}_*}}\|\mathbf{w}_*-\boldsymbol{A}\,\texttt{GD2}^{\lambda}_{reg}(S;(\boldsymbol{A},\mathbf{0}_d))\|^2 &= \mathop{\mathbb{E}}_{\substack{\boldsymbol{X}\sim\mathcal{N}(0,I_d)^n\\ \xi\sim\mathcal{N}(0,\sigma I_n)}}\|\mathbf{w}_*-\boldsymbol{A}(\boldsymbol{A}\Sigma_{\boldsymbol{X}}\boldsymbol{A}+\lambda I_d)^{-1}(\boldsymbol{A}\Sigma_{\boldsymbol{X}}\mathbf{w}_*+\frac{1}{n}\boldsymbol{A}\boldsymbol{X}^\top\xi)\|^2\\ &= \mathop{\mathbb{E}}_{\boldsymbol{X}\sim\mathcal{N}(0,I_d)^n}\|\boldsymbol{A}\left(I_d-(\boldsymbol{A}\Sigma_{\boldsymbol{X}}\boldsymbol{A}+\lambda I_d)^{-1}(\boldsymbol{A}\Sigma_{\boldsymbol{X}}\boldsymbol{A})\right)\boldsymbol{A}^{-1}\mathbf{w}_*\|^2\\ &\qquad+\mathop{\mathbb{E}}_{\substack{\boldsymbol{X}\sim\mathcal{N}(0,I_d)^n\\ \xi\sim\mathcal{N}(0,\sigma I_n)}}\|\boldsymbol{A}(\boldsymbol{A}\Sigma_{\boldsymbol{X}}\boldsymbol{A}+\lambda I_d)^{-1}\frac{1}{n}\boldsymbol{A}\boldsymbol{X}^\top\xi\|^2\\ &= \mathop{\mathbb{E}}_{\boldsymbol{X}\sim\mathcal{N}(0,I_d)^n}\lambda^2\|\boldsymbol{A}(\boldsymbol{A}\Sigma_{\boldsymbol{X}}\boldsymbol{A}+\lambda I_d)^{-1}\boldsymbol{A}^{-1}\mathbf{w}_*\|^2\\ &+\frac{\sigma^2}{n}\mathop{\mathbb{E}}_{\boldsymbol{X}\sim\mathcal{N}(0,I_d)^n}\operatorname{tr}\left(\boldsymbol{A}(\boldsymbol{A}\Sigma_{\boldsymbol{X}}\boldsymbol{A}+\lambda I_d)^{-1}\boldsymbol{A}\Sigma_{\boldsymbol{X}}\boldsymbol{A}(\boldsymbol{A}\Sigma_{\boldsymbol{X}}\boldsymbol{A}+\lambda I_d)^{-1}\boldsymbol{A}\right)\end{aligned}$$

□

Lemma C.6. *Suppose $\boldsymbol{A}=(\alpha-\kappa)\bar{\mathbf{w}}_*\bar{\mathbf{w}}_*^\top+\kappa I_d$, where $\alpha\geq\kappa$, then for $\alpha=\operatorname{poly}(\epsilon^{-1},d,\kappa,\|\mathbf{w}_*\|^2)$, $\lambda=\Theta\left(\alpha^{3/2}\right)$ and $n=\Omega(\log(\epsilon^{-1}\|\mathbf{w}_*\|_2))$, we have the following,*

$$\mathop{\mathbb{E}}_{\boldsymbol{X}\sim\mathcal{N}(0,I_d)^n}[\lambda^2\|\boldsymbol{A}(\boldsymbol{A}\Sigma_{\boldsymbol{X}}\boldsymbol{A}+\lambda I_d)^{-1}\boldsymbol{A}^{-1}\mathbf{w}_*\|^2]\leq\epsilon$$

$$\mathop{\mathbb{E}}_{\boldsymbol{X}\sim\mathcal{N}(0,I_d)^n}\operatorname{tr}\left(\boldsymbol{A}(\boldsymbol{A}\Sigma_{\boldsymbol{X}}\boldsymbol{A}+\lambda I_d)^{-1}\boldsymbol{A}\Sigma_{\boldsymbol{X}}\boldsymbol{A}(\boldsymbol{A}\Sigma_{\boldsymbol{X}}\boldsymbol{A}+\lambda I_d)^{-1}\boldsymbol{A}\right)\leq 2+\epsilon$$

Proof. We write the SVD of $\boldsymbol{A}$ as the following,

$$\boldsymbol{A}=U\begin{bmatrix}\alpha & & & \\ & \kappa & & \\ & & \ddots & \\ & & & \kappa\end{bmatrix}U^\top:=UD_{\alpha,\kappa}U^\top$$

where $D_{\alpha,\kappa}:=\alpha\boldsymbol{e}_1\boldsymbol{e}_1^\top+\kappa(I_d-\boldsymbol{e}_1\boldsymbol{e}_1^\top)$, and we know $U^\top\mathbf{w}_*=\|\mathbf{w}_*\|\boldsymbol{e}_1$.

For simplicity, from now on we write $\Sigma:=U^\top\Sigma_{\boldsymbol{X}}U$ which is identically distributed as $\Sigma_{\boldsymbol{X}}$, and we let $v\in\mathbb{R}^d$ denote the top eigenvector of $D_{\alpha,\kappa}\Sigma D_{\alpha,\kappa}$. Now we use an eigenvector perturbation argument to show v is close to $\boldsymbol{e}_1$ if α is much larger than κ. For this purpose, we write $D_{\alpha,\kappa}\Sigma D_{\alpha,\kappa}=\alpha^2\Sigma_{11}\boldsymbol{e}_1\boldsymbol{e}_1^\top+E$ where

$$E:=\kappa D_{\alpha,\kappa}\Sigma(I_d-\boldsymbol{e}_1\boldsymbol{e}_1^\top)+\kappa(I_d-\boldsymbol{e}_1\boldsymbol{e}_1^\top)\Sigma D_{\alpha,\kappa}+\kappa^2(I_d-\boldsymbol{e}_1\boldsymbol{e}_1^\top)\Sigma(I_d-\boldsymbol{e}_1\boldsymbol{e}_1^\top)$$

It is clear that

$$\begin{aligned}\|E\|_F &\le \left(2\kappa\|D_{\alpha,\kappa}\|_2\|I_d - \boldsymbol{e}_1\boldsymbol{e}_1^\top\|_F + \kappa^2\|I_d - \boldsymbol{e}_1\boldsymbol{e}_1^\top\|_F^2\right)\|\Sigma\|_F \\ &\le \left(2\alpha d\kappa + d\kappa^2\right)\operatorname{tr}(\Sigma)\end{aligned}$$

By the Davis-Kahan theorem (Davis & Kahan, 1970), we have

$$\|vv^\top - \boldsymbol{e}_1\boldsymbol{e}_1^\top\|_F \le 2\sqrt{2}\frac{\|E\|_F}{\alpha^2\Sigma_{11}} \le C\frac{\left(2\alpha d\kappa + d\kappa^2\right)\operatorname{tr}(\Sigma)}{\alpha^2\Sigma_{11}}$$

where C is an absolute constant. Furthermore, we can bound the eigenvalues of $D_{\alpha,\kappa}\Sigma D_{\alpha,\kappa}$ using Weyl's inequality:

$$\begin{aligned}\left|\sigma_1(D_{\alpha,\kappa}\Sigma D_{\alpha,\kappa}) - \alpha^2\Sigma_{11}\right| &\le \|E\|_F \le \left(2\alpha d\kappa + d\kappa^2\right)\operatorname{tr}(\Sigma) \\ \forall i : 2 \le i \le d, \quad |\sigma_i(D_{\alpha,\kappa}\Sigma D_{\alpha,\kappa})| &\le \|E\|_F \le \left(2\alpha d\kappa + d\kappa^2\right)\operatorname{tr}(\Sigma)\end{aligned}$$

where σ_1 denotes the largest eigenvalue and σ_i's are the rest. It follows that

$$\begin{aligned}&\lambda^2\|\boldsymbol{A}(\boldsymbol{A}\Sigma_{\boldsymbol{X}}\boldsymbol{A} + \lambda I_d)^{-1}\boldsymbol{A}^{-1}\mathbf{w}_*\|^2 = \lambda^2\|\mathbf{w}_*\|^2\|D_{\alpha,\kappa}(D_{\alpha,\kappa}\Sigma D_{\alpha,\kappa} + \lambda I_d)^{-1}D_{\alpha,\kappa}^{-1}\boldsymbol{e}_1\|^2 \\ &\quad\le \frac{\lambda^2\|\mathbf{w}_*\|^2}{\left(\alpha^2\Sigma_{11} - \left(2\alpha d\kappa + d\kappa^2\right)\operatorname{tr}(\Sigma) + \lambda\right)^2}\|D_{\alpha,\kappa}vv^\top\frac{1}{\alpha}\boldsymbol{e}_1\|^2 + \|D_{\alpha,\kappa}(I_d - vv^\top)\frac{1}{\alpha}\boldsymbol{e}_1\|^2\|\mathbf{w}_*\|^2 \\ &\quad\le \frac{\lambda^2}{\left(\alpha^2\Sigma_{11} - \left(2\alpha d\kappa + d\kappa^2\right)\operatorname{tr}(\Sigma) + \lambda\right)^2}\|\mathbf{w}_*\|^2 + \left(1 + \frac{d\kappa}{\alpha}\right)^2\|vv^\top - \boldsymbol{e}_1\boldsymbol{e}_1^\top\|_F^2\|\mathbf{w}_*\|^2 \\ &\quad\le \frac{\lambda^2}{\left(\alpha^2\Sigma_{11} - \left(2\alpha d\kappa + d\kappa^2\right)\operatorname{tr}(\Sigma) + \lambda\right)^2}\|\mathbf{w}_*\|^2 + \left(1 + \frac{d\kappa}{\alpha}\right)^2\frac{\left(2\alpha d\kappa + d\kappa^2\right)^2\operatorname{tr}(\Sigma)^2}{(\alpha^2\Sigma_{11})^2}\|\mathbf{w}_*\|^2\end{aligned}$$

where C' is another absolute constant.

Finally, we note that $n\Sigma_{ii} \sim \chi^2(n)$, i.e., χ^2 distribution with n degree of freedom for all $i \in [d]$. Thus by standard concentration bound, we have $\Pr\left[\Sigma_{11} \ge 0.9 \wedge \operatorname{tr}(\Sigma) \le 2d\right] \ge 1 - \exp(-\Omega(n))$. To evaluate the expectations, we condition on two events, namely $\Sigma_{11} \ge 0.9 \wedge \operatorname{tr}(\Sigma) \le 2d$ and its complement. Thus in the case where $\alpha = \Omega\left(\max\{\epsilon^{-1}d^4\kappa^4\|\mathbf{w}_*\|^2\}\right)$, $\lambda = \Theta(\alpha^{3/2})$ and $n = \Omega\left(\log(\epsilon^{-1}\|\mathbf{w}_*\|)\right)$, we have

$$\begin{aligned}&\mathop{\mathbb{E}}_{\boldsymbol{X}\sim\mathcal{N}(0,I_d)^n}[\lambda^2\|\boldsymbol{A}(\boldsymbol{A}\Sigma_{\boldsymbol{X}}\boldsymbol{A} + \lambda I_d)^{-1}\boldsymbol{A}^{-1}\mathbf{w}_*\|^2] \\ &\le \|\mathbf{w}_*\|^2\exp(-\Omega(n)) \\ &\quad + \left(1 - \exp(-\Omega(n))\right)\|\mathbf{w}_*\|^2\left(\frac{\lambda^2}{(0.9\cdot\alpha^2 - 4\alpha d^2\kappa - 2d^2\kappa^2 + \lambda)^2} + 4\left(1 + \frac{d\kappa}{\alpha}\right)^2\frac{\left(2\alpha d\kappa + d\kappa^2\right)^2 d^2}{(0.9\cdot\alpha^2)^2}\right) \\ &\le \epsilon\end{aligned}$$

For the second part, we have

$$\begin{aligned}&\mathop{\mathbb{E}}_{\boldsymbol{X}\sim\mathcal{N}(0,I_d)^n}\|\boldsymbol{A}(\boldsymbol{A}\Sigma_{\boldsymbol{X}}\boldsymbol{A} + \lambda I_d)^{-1}\boldsymbol{A}\boldsymbol{X}^\top\|_F^2 \le \mathop{\mathbb{E}}_{\boldsymbol{X}\sim\mathcal{N}(0,I_d)^n}[\|\boldsymbol{A}\|_2^2\operatorname{tr}((\boldsymbol{A}\Sigma_{\boldsymbol{X}}\boldsymbol{A} + \lambda I_d)^{-2}\boldsymbol{A}\Sigma_{\boldsymbol{X}}\boldsymbol{A})] \\ &= \mathop{\mathbb{E}}_{\boldsymbol{X}\sim\mathcal{N}(0,I_d)^n}\left[\alpha^2\sum_{i=1}^d\frac{\sigma_i(D_{\alpha,\kappa}\Sigma D_{\alpha,\kappa})}{\left(\sigma_i(D_{\alpha,\kappa}\Sigma D_{\alpha,\kappa}) + \lambda\right)^2}\right] \\ &\le \mathop{\mathbb{E}}_{\boldsymbol{X}\sim\mathcal{N}(0,I_d)^n}\left[\frac{\sigma_1(D_{\alpha,\kappa}\Sigma D_{\alpha,\kappa})}{\sigma_1(D_{\alpha,\kappa}\Sigma D_{\alpha,\kappa}) + \lambda}\cdot\frac{\alpha^2}{\sigma_1(D_{\alpha,\kappa}\Sigma D_{\alpha,\kappa}) + \lambda} + (d-1)\frac{\alpha^2}{\lambda^2}\|E\|_F\right] \\ &\le \mathop{\mathbb{E}}_{\boldsymbol{X}\sim\mathcal{N}(0,I_d)^n}\left[\frac{\alpha^2}{\sigma_1(D_{\alpha,\kappa}\Sigma D_{\alpha,\kappa}) + \lambda} + (d-1)\frac{\alpha^2}{\lambda^2}\left(2\alpha d\kappa + d\kappa^2\right)\operatorname{tr}(\Sigma)\right] \\ &\le \underbrace{\mathop{\mathbb{E}}_{\boldsymbol{X}\sim\mathcal{N}(0,I_d)^n}\left[\frac{\alpha^2}{\alpha^2\Sigma_{11} - \left(2\alpha d\kappa + d\kappa^2\right)\operatorname{tr}(\Sigma) + \lambda}\right]}_{:=\diamondsuit} + \frac{\alpha^2}{\lambda^2}(2d^3\kappa^2 + d^2\kappa^2)\end{aligned}$$

In order to bound $\diamondsuit$, we first condition on the event of $\mathcal{E} := \Sigma_{11} \geq \frac{1}{\sqrt{\alpha}} \wedge \operatorname{tr}(\Sigma) \leq \frac{d\sqrt{\alpha}}{8}$ which occurs with overwhelming probability. In fact, we have by the standard concentration bound and the CDF of $\chi^2(n)$ distribution, i.e., $\Pr[\Sigma_{11} \leq \frac{1}{\sqrt{\alpha}}] \leq (1/\alpha)^{n/4}$ that

$$\Pr[\mathcal{E}] \geq 1 - \alpha^{-n/4} - \exp(-\Omega(\sqrt{\alpha})$$

It follows that

$$\begin{aligned}
\diamondsuit \leq & \mathop{\mathbb{E}}_{\boldsymbol{X}\sim\mathcal{N}(0,I_d)^n}\left[\frac{\alpha^2}{\alpha^2\Sigma_{11} - (2\alpha d\kappa + d\kappa^2)\operatorname{tr}(\Sigma) + \lambda}\middle|\mathcal{E}\right]\Pr[\mathcal{E}] \\
& + \mathop{\mathbb{E}}_{\boldsymbol{X}\sim\mathcal{N}(0,I_d)^n}\left[\frac{\alpha^2}{\alpha^2\Sigma_{11} - (2\alpha d\kappa + d\kappa^2)\operatorname{tr}(\Sigma) + \lambda}\middle|\neg\mathcal{E}\right](1 - \Pr[\mathcal{E}]) \\
\leq & \mathop{\mathbb{E}}_{\boldsymbol{X}\sim\mathcal{N}(0,I_d)^n}\left[\frac{\alpha^2}{\alpha^2\Sigma_{11} - (2\alpha d\kappa + d\kappa^2)\operatorname{tr}(\Sigma) + \lambda}\middle|\mathcal{E}\right]\Pr[\mathcal{E}] + \alpha^2(1 - \Pr[\mathcal{E}]) \\
\leq & \mathop{\mathbb{E}}_{\boldsymbol{X}\sim\mathcal{N}(0,I_d)^n}\left[\frac{\alpha^2}{\alpha^2\Sigma_{11} - (2\alpha d\kappa + d\kappa^2)\operatorname{tr}(\Sigma) + \lambda}\middle|\mathcal{E}\right] + \alpha^2\left(\alpha^{-n/4} + \exp(-\Omega(\sqrt{\alpha}))\right) \\
\leq & \mathop{\mathbb{E}}_{\boldsymbol{X}\sim\mathcal{N}(0,I_d)^n}\left[\frac{\alpha^2}{0.5\cdot\alpha^2\Sigma_{11} + \lambda}\middle|\mathcal{E}\right]\left(1 - \alpha^{-n/4} - \exp(-\Omega(\sqrt{\alpha}))\right) + \alpha^2\left(\alpha^{-n/4} + \exp(-\Omega(\sqrt{\alpha}))\right) \\
\leq & \mathop{\mathbb{E}}_{\boldsymbol{X}\sim\mathcal{N}(0,I_d)^n}\left[\frac{\alpha^2}{0.5\cdot\alpha^2\Sigma_{11} + \lambda}\middle|\mathcal{E}\right] + \alpha^2\left(\alpha^{-n/4} + \exp(-\Omega(\sqrt{\alpha}))\right) \\
\leq & \mathop{\mathbb{E}}_{\boldsymbol{X}\sim\mathcal{N}(0,I_d)^n}\left[\frac{2}{\Sigma_{11}}\right] + \alpha^2\left(\alpha^{-n/4} + \exp(-\Omega(\sqrt{\alpha}))\right) \\
\leq & \frac{2n}{n-2} + \alpha^2\left(\alpha^{-n/4} + \exp(-\Omega(\sqrt{\alpha}))\right)
\end{aligned}$$

where we use the fact that $\mathbb{E}[2/\Sigma_{11}|\mathcal{E}] \leq \mathbb{E}[2/\Sigma_{11}]$ and the expectation of inverse χ^2 distribution. Putting it together and assuming $\alpha = \Omega\left(\operatorname{poly}(\epsilon^{-1}d^3\kappa^2)\right), \lambda = \alpha^{3/2}$ and $n \geq 10$, we conclude

$$\begin{aligned}
& \mathop{\mathbb{E}}_{\boldsymbol{X}\sim\mathcal{N}(0,I_d)^n} \|\boldsymbol{A}(\boldsymbol{A}\Sigma_{\boldsymbol{X}}\boldsymbol{A} + \lambda I_d)^{-1}\boldsymbol{A}\boldsymbol{X}^\top\|_F^2 \\
& \leq \frac{2n}{n-2} + \alpha^2\left(\alpha^{-n/4} + \exp(-\Omega(\sqrt{\alpha}))\right) + \frac{\alpha^2}{\lambda^2}(2d^3\kappa^2 + d^2\kappa^2) \\
& \leq 2 + \epsilon
\end{aligned}$$

□

C.2. Proof of Main Results

Reptile: We finally prove the main theorem about the success of `Reptile`.

(Theorem 5.1). *Starting with $(\boldsymbol{A}_0, \mathbf{w}_0) = (\kappa I_d, \mathbf{0}_d)$, let $\boldsymbol{A}_T = \texttt{Reptile}(\rho_{1:T}, (\boldsymbol{A}_0, \mathbf{w}_0))$ be the initialization learned using T tasks $\{\rho_1, \ldots, \rho_T\} \sim_{i.i.d.} \mu_{\mathbf{w}_*}^T$. If $T \geq poly(d, r, 1/\epsilon, \log(1/\delta), \kappa)$ and $\tau = \mathcal{O}(T^{-1/3})$, then with probability at least $1 - \delta$ over sampling of T tasks,*

$$\min_{\lambda\geq 0} \mathcal{E}_n(GD2^{\lambda}_{reg}(\cdot; (\boldsymbol{A}_T, \mathbf{0}_d)), \mu_{\mathbf{w}_*}) \leq \epsilon + \frac{cr^2}{n}$$

for a small constant c. Thus with the same probability, we have

$$\min_{\lambda\geq 0} n_\epsilon(GD2^{\lambda}_{reg}(\cdot; \boldsymbol{A}_T, \mathbf{0}_d), \mu_{\mathbf{w}_*}) = \mathcal{O}\left(\frac{r^2}{\epsilon}\right)$$

Proof of Theorem 5.1. The theorem essentially follows from Lemma C.1, Theorem C.3, Lemma C.5 and Lemma C.6. From Lemma C.1 and Theorem C.3, we get that with probability at least $1-\delta$ choosing $\tau = T^{-1/3}\log(2T/\delta)^{-2/3}$ will ensure $\boldsymbol{A}_T = (\alpha-\kappa)\bar{\mathbf{w}}_*\bar{\mathbf{w}}_* + \kappa I_d$ with $\alpha = \Omega(\sqrt{r}T^{1/6})$. Combining Lemma C.5 and Lemma C.6 we know that if $\alpha = \Omega(\text{poly}(\epsilon^{-1}, d, \kappa, r))$, then $\mathcal{E}_n(\texttt{GD2}^{\lambda}_{reg}(\cdot;(\boldsymbol{A},\mathbf{0}_d)),\mu_{\mathbf{w}_*}) \le \frac{\epsilon}{2} + \frac{c\sigma^2}{n} = \frac{\epsilon}{2} + \frac{cr^2}{n}$. To ensure α is this large, we just need that the number of tasks to satisfy $T = poly(\epsilon^{-1}, d, \kappa, r, \log(\delta^{-1}))$ for the appropriate polynomial from Lemma C.6. Thus for $\mathcal{E}_n(\texttt{GD2}^{\lambda}_{reg}(\cdot;(\boldsymbol{A},\mathbf{0}_d)),\mu_{\mathbf{w}_*}) \le \epsilon$, we just need $n = \Omega\left(\frac{r^2}{\epsilon}\right)$ samples for a new task, completing the proof. □

Representation learning: We now prove the main theorem about the success of `RepLearn`.

(Theorem 5.2). *Starting with $(\boldsymbol{A}_0, \mathbf{w}_{0,1:T}) = (\kappa I_d, \mathbf{0}_d, \ldots, \mathbf{0}_d)$, let $\boldsymbol{A}_T = \texttt{RepLearn}(\rho_{1:T}, (\boldsymbol{A}_0, \mathbf{w}_{0,1:T}))$, be the initialization learned using T tasks $\{\rho_1, \ldots, \rho_T\} \sim_{i.i.d.} \mu^T_{\mathbf{w}_*}$. If $T \ge poly(d, r, 1/\epsilon, \log(1/\delta), \kappa)$, then with probability at least $1-\delta$ over sampling of the T tasks,*

$$\min_{\lambda \ge 0} \mathcal{E}_n(\texttt{GD2}^{\lambda}_{reg}(\cdot;(\boldsymbol{A}_T,\mathbf{0}_d)),\mu_{\mathbf{w}_*}) \le \epsilon + \frac{cr^2}{n}$$

for a small constant c. Thus with the same probability, we have

$$\min_{\lambda \ge 0} n_\epsilon(\texttt{GD2}^{\lambda}_{reg}(\cdot;\boldsymbol{A}_T,\mathbf{0}_d),\mu_{\mathbf{w}_*}) = \mathcal{O}\left(\frac{r^2}{\epsilon}\right)$$

Proof of Theorem 5.2. The proof of this is very similar to the proof of Theorem 5.1 above. Just as in that proof, we need to show that for a large enough T, $\boldsymbol{A}_T := \boldsymbol{A}_T^{\texttt{RepLearn}} = (\alpha-\kappa)\mathbf{w}_*\mathbf{w}_*{}^\top + \kappa I_d$ for a large enough α. The theorem will then follow from Lemma C.5 and Lemma C.6 just as in the previous proof. To prove the closed form solution for $\boldsymbol{A}_T$, we use the following lemma that is very similar to Lemma C.2

Lemma C.7. *Starting from $\boldsymbol{A}(0) = (a(0)-\kappa)\bar{\mathbf{w}}_*\bar{\mathbf{w}}_*^\top + \kappa I_d$, $\mathbf{w}_i(0) = \mathbf{0}_d,\ i \in [T]$, with $a(0) > 0$, the solution of gradient flow on loss $\mathcal{L}_{rep}(\boldsymbol{A}, \mathbf{w}_{1:T})$ for $s \in \{\pm 1\}$, is $\bar{\boldsymbol{A}}, \bar{\mathbf{w}}_{1:T}$, where*

$$\bar{\boldsymbol{A}} = (\bar{a}-\kappa)\bar{\mathbf{w}}_*\bar{\mathbf{w}}_*^\top + \kappa I_d,\ \bar{\mathbf{w}}_i = \bar{b}_i\bar{\mathbf{w}}_*\text{ , where}$$
$$\bar{a} = \sqrt{\frac{a(0)^2 + \sqrt{4r^2T + a(0)^4}}{2}},\ \bar{b}_i = s_i\sqrt{\frac{-a(0)^2 + \sqrt{4r^2T + a(0)^4}}{2}}$$

Proof. We first rewrite the representation learning objective using the derivation in Section 6.2 as follows

$$\mathcal{L}_{rep}(\boldsymbol{A}, \mathbf{w}_{1:T}) = \frac{1}{T}\|\boldsymbol{A}^\top\boldsymbol{W} - \boldsymbol{W}_*\|^2 \tag{13}$$

where $\boldsymbol{W} \in \mathbb{R}^{d\times T}$, $\boldsymbol{W}_* \in \mathbb{R}^{d\times T}$ and the i^{th} column of $\boldsymbol{W}$ is $\mathbf{w}_i$ and the i^{th} column of $\boldsymbol{W}_*$ is $s_i\mathbf{w}_*$. Just as in Lemma C.2, we define $\boldsymbol{U}$ to be an orthogonal matrix whose first column is $\bar{\mathbf{w}}_*$. We also define $\boldsymbol{V} \in \mathbb{R}^T$ to be the vector of the signs of the tasks, i.e. $\boldsymbol{V} = \frac{1}{\sqrt{T}}(s_1, \ldots, s_T)$. We can then rewrite $\boldsymbol{A}(0) = \boldsymbol{U}\Lambda_1(0)\boldsymbol{U}^\top$, where $\Lambda_1(0) \in \mathbb{R}^{d\times d}$ is a diagonal matrix that looks like $\Lambda_1(0) = diag(a(0), \kappa, \ldots, \kappa)$, $\boldsymbol{W}(0) = \boldsymbol{U}\Lambda_2(0)\boldsymbol{V}^\top$, where $\Lambda_2(0) = (b(0), 0, \ldots, 0) \in \mathbb{R}^d$ with $b(0) = 0$ and $\boldsymbol{W}_* = \boldsymbol{U}\Lambda_*\boldsymbol{V}^\top$, where $\Lambda_* = (\sqrt{T}r, 0, \ldots, 0) \in \mathbb{R}^d$. Note that $\boldsymbol{U}^\top\boldsymbol{U} = I_d$ and $\boldsymbol{V}^\top\boldsymbol{V} = 1$

The dynamics of gradient flow on $\mathcal{L}_{rep}$ using $Equation\ 13$ is

$$\frac{d\boldsymbol{A}(t)}{dt} = \boldsymbol{W}(t)\boldsymbol{W}_*{}^\top - \boldsymbol{W}(t)\boldsymbol{W}(t)^\top\boldsymbol{A}(t)$$
$$\frac{d\boldsymbol{W}(t)}{dt} = \boldsymbol{A}(t)\boldsymbol{W}_* - \boldsymbol{A}(t)\boldsymbol{A}(t)^\top\boldsymbol{W}(t)$$

By defining $\Lambda_1(t) = \boldsymbol{U}^\top\boldsymbol{A}(t)\boldsymbol{U}$, $\Lambda_2(t) = \boldsymbol{U}^\top\boldsymbol{W}(t)\boldsymbol{V}$, $\Lambda_* = \boldsymbol{U}^\top\boldsymbol{W}_*\boldsymbol{V}$, we can multiply the above dynamics by $\boldsymbol{U}^\top$ on the left and $\boldsymbol{V}$ on the right, and use the properties above to get

$$\frac{d\Lambda_1(t)}{dt} = \Lambda_2(t)\Lambda_*^\top - \Lambda_2(t)\Lambda_2(t)^\top\Lambda_1(t)$$

$$\frac{d\Lambda_2(t)}{dt} = \Lambda_1(t)\Lambda_*^\top - \Lambda_1(t)\Lambda_1(t)^\top\Lambda_2(t)$$

Just like Lemma C.2, this reduces to a scalar dynamics and the solution we get is $\bar{\boldsymbol{A}} = \boldsymbol{U}\bar{\Lambda}_1\boldsymbol{U}^\top$, $\bar{\boldsymbol{W}} = \boldsymbol{U}\bar{\Lambda}_2\boldsymbol{V}^\top$, where $\bar{\Lambda}_1 = diag(\bar{a}, \kappa, \ldots, \kappa), \bar{\Lambda}_2 = (\bar{b}, 0, \ldots, 0)$ and

$$\bar{a} = \sqrt{\frac{a(0)^2 + \sqrt{4r^2T + a(0)^4}}{2}}, \quad \bar{b} = \sqrt{\frac{-a(0)^2 + \sqrt{4r^2T + a(0)^4}}{2}}$$

This completes the proof of the lemma. □

Back to the main theorem, we see from the above lemma that $\alpha = \Omega(\sqrt{r}T^{1/4})$, where $\boldsymbol{A}_T = (\alpha - \kappa)\mathbf{w}_*\mathbf{w}_*{}^\top + \kappa I_d$. So making $T = poly(\epsilon^{-1}, d, \kappa, r, \log(\delta^{-1}))$ large enough will make α large enough to invoke Lemma C.5 and Lemma C.6 to complete the proof, just like in the proof of Theorem 5.1. □

D. Information-Theoretic Lower-Bounds for the Convex Case

Theorem D.1. *For any $G, V > 0$, there exists a domain $\mathcal{Z}$, parameter class $\Theta \subseteq \mathbb{R}^d$ and a distribution μ over tasks such every $\rho \sim \mu$ is a distribution over $\mathcal{Z}$ and $\ell_\rho(\theta) = \mathbb{E}_{z\sim\rho}\ell_z(\theta)$ where $\ell_z : \Theta \to \mathbb{R}$ is convex and G-Lipschitz w.r.t. the Euclidean norm for every $z \in \mathcal{Z}$. Additionally, Θ satisfies*

$$\min_{\phi\in\Theta} \mathbb{E}_{\rho\sim\mu}\|\phi - \mathrm{Proj}_{\Theta^*_\rho}(\phi)\| \le V$$

and

$$\mathcal{E}_n(\texttt{Alg}, \mu) = \Omega\left(GV \min\left\{\frac{1}{\sqrt{n}}, \frac{1}{\sqrt{d}}\right\}\right)$$

for any algorithm $\texttt{Alg} : \mathcal{Z}^n \to \Theta$ that returns a parameter given a training set.

Proof. This result extends the result of Agarwal et al. (2012, Theorem 1) to the case of distributions over functions; all equations and statements referenced in this proof are from that paper. We first define the domain $\mathcal{Z}$, parameter class Θ, meta-distribution μ and the within-task distributions and losses.

Parameter class: We use a ℓ_2 ball of radius V as the class, i.e. $\Theta = \{\theta \in \mathbb{R}^d : \|\theta\| \le V/2\}$.

Domain and loss: We defined $\mathcal{Z}$ to be a tuple of an index and a bit, i.e. $\mathcal{Z} = [d] \times \{0, 1\}$. For a given $z \in \mathcal{Z}$, we define ℓ_z as follows

$$\ell_z(\theta) = \begin{cases} G\left|\theta(i) + \frac{V}{2\sqrt{d}}\right| & \text{if } z = (i, 1), i \in [d] \\ G\left|\theta(i) - \frac{V}{2\sqrt{d}}\right| & \text{if } z = (i, 0), i \in [d] \end{cases}$$

Note that ℓ_z is convex and G-Lipschitz for every $z \in \mathcal{Z}$.

Meta-learning distribution: We define the distribution μ on the vertices of the hypercube $\{\pm 1\}^d$. First we let $\mathcal{V}$ be the $\frac{d}{4}$-packing of the hypercube in the Hamming distance defined in Agarwal et al. (2012). Each task ρ_α is parametrized by a vertex $\alpha \in \mathcal{V}$. To sample a new task $\rho_\alpha \sim \mu$, we sample $\alpha \sim \mathcal{V}$ uniformly and return ρ_α that we define below.

Data distribution: For a given task $\rho_\alpha \sim \mu$, we define a distribution over $\mathcal{Z}$. Sampling $z \sim \rho_\alpha$ is equivalent to first sample an index uniformly at random, $i \sim [d]$, and then independently sampling a bit from a biased Bernoulli distribution $b \sim \mathrm{Ber}\left(\frac{1}{2} + \alpha(i)\delta\right)$, for some $\delta \in (0, 1/4)$, and returning (i, b). Thus the population loss for ρ_α becomes

$$\ell_{\rho_\alpha}(\theta) = \sum_{i=1}^{d}\left(\frac{1}{2} + \alpha(i)\delta\right)\left|\theta(i) + \frac{V}{2\sqrt{d}}\right| + \left(\frac{1}{2} - \alpha(i)\delta\right)\left|\theta(i) + \frac{V}{2\sqrt{d}}\right|$$

It is not difficult to see that the minimizer of the population loss $\theta^*_{\rho_\alpha} \in \mathbb{R}^d$ in fact lies in Θ and is

$$\theta^*_{\rho_\alpha}(i) = \begin{cases} -\frac{V}{2\sqrt{d}} & \text{if } \alpha(i) = 1 \\ \frac{V}{2\sqrt{d}} & \text{if } \alpha(i) = -1 \end{cases}$$

Crucially, we note that since $\theta^*_{\rho_\alpha} \in \Theta$ for every $\alpha \in \mathcal{V}$, we get that

$$\min_{\phi \in \Theta} \mathbb{E}_{\rho \sim \mu} \|\phi - \mathrm{Proj}_{\Theta^*_\rho}(\phi)\| \leq \mathbb{E}_{\rho \sim \mu} \|\mathbf{0}_d - \theta^*_{\rho_\alpha}\| = V$$

Given this setup, we are ready to prove a lower bound for $\mathcal{E}_n(\texttt{Alg}, \mu)$ using the result from Agarwal et al. (2012). We define the class of functions $\mathcal{G}(\delta) = \{\ell_{\rho_\alpha} : \alpha \in \mathcal{V}\}$ and define $g_\alpha = \ell_{\rho_\alpha}$. Note that this is the same definition of $\mathcal{G}(\delta)$ as in Agarwal et al. (2012).

We now follow their proof of Theorem 1, where in addition to the randomness of sampling from the task-distribution ρ_α we must consider the randomness of sampling $\alpha \sim \mathcal{V}$. This manifests only in the application of Lemmas 2 and 3 from their paper. We can modify their proof of Lemma 2 to only assume

$$\mathcal{E}_n(\texttt{Alg}, \mu) = \mathop{\mathbb{E}}_{\alpha \sim \mathcal{V}}[\Delta(\texttt{Alg}, \alpha)] \leq \frac{\psi(\delta)}{9}, \text{ where } \Delta(\texttt{Alg}, \alpha) = \mathop{\mathbb{E}}_{S \sim \rho_\alpha^n}[\ell_{\rho_\alpha}(\texttt{Alg}(S)) - \ell^*_{\rho_\alpha}]$$

instead of Equation 21 which effectively assumes $\max_{\alpha \in \mathcal{V}} \Delta(\texttt{Alg}, \alpha) \leq \frac{\psi(\delta)}{9}$, where $\psi(\delta)$ is defined in Equation 19. We can modify the application of Markov's inequality, to get

$$\mathop{\mathbb{E}}_{\alpha \sim \mathcal{V}} \mathbb{P}_{S \sim \rho_\alpha^n}(\texttt{Alg}(S) \neq \alpha) \leq \mathop{\mathbb{E}}_{\alpha \sim \mathcal{V}} \mathbb{P}_{S \sim \rho_\alpha^n}(\Delta(\texttt{Alg}, \alpha) \geq \psi(\delta)/3) \leq \mathop{\mathbb{E}}_{S \sim \rho_\alpha^n} \frac{\Delta(\texttt{Alg}, \alpha)}{\psi(\delta)/3} \leq \frac{\psi(\delta)/9}{\psi(\delta)/3} \leq 1/3$$

where the first step is the same as in their proof, second step from Markov's inequality and third is from the assumption. The main difference from their proof, just like the assumption, is that we take expectation over $\alpha \in \mathcal{V}$ rather than a maximum.

For Lemma 3, note that the result already includes the randomness of sampling $\alpha \sim \mathrm{Unif}(\mathcal{V})$. Applying these results in the proof of Theorem 1, we use $\delta = \frac{36\varepsilon\sqrt{d}}{GV}$ for target error ε to obtain $n = \Omega(G^2V^2/\varepsilon^2)$ for all $d \geq 11$ and $\varepsilon \leq \frac{GV}{144\sqrt{d}}$, completing the proof. □