A Notation Glossary

Notation	Definition
$[n]$	$\{1, \ldots, n\}$
$\|\cdot\|_2$	Euclidean vector norm
$\|\cdot\|_F$	Frobenius matrix norm
$\langle \cdot, \cdot \rangle$	Euclidean inner product OR Frobenius inner product
$\sigma_i(A)$	i^{th} largest singular value of A
$\lambda_i(A)$	i^{th} largest eigenvalue of A
$\lambda_{\max}(A)$	Largest eigenvalue of A
$\lambda_{\min}(A)$	Smallest eigenvalue of A
$A^{1/2}$	Principal square root of PSD A
$\operatorname{Col} A$	Column span of A
P_A	Projection onto $\operatorname{Col} A$
$P_A^{\perp}$	Projection onto $(\operatorname{Col} A)^{\perp}$
ℓ	Loss function
$\mathcal{L}_{\infty}(g, g^*)$	Population risk of g when the true predictor is g^*
$\mathcal{L}(g, g^*)$	Empirical risk of g when the true predictor is g^* (samples suppressed)
$\mathcal{L}_{\infty}^{\text{ex}}(g, g^*)$	Excess population risk of g when the true predictor is g^*.
$\mathcal{A}_{\mathcal{C}}(\theta)$	Predictors whose parameters lie in $\theta + \mathcal{C}$
$\mathcal{F}^{\otimes T}$	$\{(x_1, \ldots, x_T) \mapsto (f_1(x_1), \ldots, f_T(x_T)) \mid f_1, \ldots, f_T \in \mathcal{F}\}$
$\mathcal{R}_n(\mathcal{H})$	Rademacher complexity of function class $\mathcal{H}$ on n samples

B Proof of Theorem 3.1

In this section, we will prove the performance guarantee in the linear representation setting presented in Theorem 3.1. We first compute a bound on the difference in the spans of the true underlying representation B^* and the representation B_0 obtained from training on the source tasks. Having done so, we then analyze the performance of the best predictor found by projected gradient descent.

For clarity of presentation, we will write $\theta_t^* = B_t^* w_t^* + \delta_t^*$ and $\hat{\theta}_t = (B + \Delta_t) w_t$ throughout this section. Furthermore, let $\hat{\delta}_t = \Delta_t w_t$. Finally, we will be making use of the following covariance concentration results throughout this section, allowing us to connect empirical averages to population averages and vice versa:

Lemma B.1 (Source covariance concentration, Du et al. (2020), Claim A.1)**.** *If $n_{\mathrm{S}} \gg \rho^4[d + \log(T/\delta)]$, then with probability at least $1 - \delta/9$ over the random draw of $n_{\mathrm{S}}T$ source inputs,*

$$0.9\Sigma \preceq \frac{1}{n_{\mathrm{S}}} X_t^\top X_t \preceq 1.1\Sigma$$

for any $t \in [T]$.

Lemma B.2 (Target covariance concentration)**.** *If $n_{\mathrm{T}} \gg \rho^4[d + \log(T/\delta)]$, then with probability at least $1 - \delta/9$ over the random draw of n_{T} target inputs,*

$$0.9\Sigma \preceq \frac{1}{n_{\mathrm{T}}} X^\top X \preceq 1.1\Sigma$$

for any $t \in [T]$.

Proof. The proof is similar to that of Lemma B.1, and is omitted for brevity. □

B.1 Source Guarantees for the Linear Setting

We proceed to analyze the representation B_0 obtained from the source training procedure outlined in (4). Key to the analysis is a bound on the average population loss over the source tasks that the global minimizer of (4) can achieve as a function of n_{S}:

Lemma B.3 (Source training bound)**.** *Let B_0 be a minimizer of (4), and $\{(\Delta_t, w_t)\}_{t\in[T]}$ be minimizers for the inner optimization problem given B_0. If the regularizer coefficients are chosen such that*

$$\lambda \asymp \frac{1}{T\delta_0}\left[\frac{\sigma \left\|\Sigma\right\|_2^{1/2}}{\sqrt{n_{\mathrm{S}}T}}\sqrt{kT + kd\log\kappa n_{\mathrm{S}} + \log\frac{1}{\delta}} + \frac{\sigma}{\sqrt{n_{\mathrm{S}}}}\sqrt{\operatorname{tr}\Sigma\left(1+\log\frac{T}{\delta}\right)}\right],$$

and $\gamma = \delta_0^2\lambda$, then with probability at least $1-\delta/3$,

$$\begin{aligned}
&\frac{1}{T}\sum_{t\in[T]}\left\|\Sigma^{1/2}(\theta_t^* - \hat{\theta}_t)\right\|_2^2 \\
&\lesssim \frac{\sigma^2}{n_{\mathrm{S}}T}\left(kT + kd\log\kappa n_{\mathrm{S}} + \log\frac{1}{\delta}\right) + \frac{\sigma\delta_0\left\|\Sigma\right\|_2^{1/2}}{\sqrt{n_{\mathrm{S}}T}}\sqrt{kT + kd\log\kappa n_{\mathrm{S}} + \log\frac{1}{\delta}} \\
&+ \frac{\sigma\delta_0}{\sqrt{n_{\mathrm{S}}}}\sqrt{\operatorname{tr}\Sigma\left(1+\log\frac{T}{\delta}\right)}.
\end{aligned}$$

Proof. Throughout this proof, we instantiate the high-probability event in Lemma B.1, which occurs with probability at least $1-\delta/9$.

Note that we can express δ_t^* as $[\delta_t^*(w_t^*)^\top / \left\|w_t^*\right\|_2^2]w_t^* = \Delta_t^* w_t^*$. Thus, via the optimality of B_0 and $\{(\Delta_t, w_t)\}_{t\in[T]}$, we can form the basic inequality

$$\begin{aligned}
&\sum_{t\in[T]}\frac{1}{2n_{\mathrm{S}}T}\left\|y_t - X_t\hat{\theta}_t\right\|_2^2 + \frac{\lambda}{2}\left\|\Delta_t\right\|_F^2 + \frac{\gamma}{2}\left\|w_t\right\|_2^2 \\
&\quad\le \sum_{t\in[T]}\frac{1}{2n_{\mathrm{S}}T}\left\|z_t\right\|_2^2 + \frac{\lambda}{2}\left\|\Delta_t^*\right\|_F^2 + \frac{\gamma}{2}\left\|w_t^*\right\|_2^2 \\
&\quad\le \sum_{t\in[T]}\frac{1}{2n_{\mathrm{S}}T}\left\|z_t\right\|_2^2 + \frac{\lambda}{2}\sum_{t\in[T]}\frac{\left\|\delta_t^*\right\|_2^2}{\left\|w_t^*\right\|_2^2} + \frac{\gamma}{2}\sum_{t\in[T]}\left\|w_t^*\right\|_2^2 \\
&\quad\le \sum_{t\in[T]}\frac{1}{2n_{\mathrm{S}}T}\left\|z_t\right\|_2^2 + \left(\frac{\lambda\delta_0^2+\gamma}{2}\right)T \\
&\quad\le \underbrace{\frac{\sigma\delta_0\left\|\Sigma\right\|_2^{1/2}}{\sqrt{nT}}\sqrt{kT + kd\log\kappa n_{\mathrm{S}} + \log\frac{1}{\delta}} + \frac{\sigma\delta_0}{\sqrt{n}}\sqrt{\operatorname{tr}\Sigma\left(1+\log\frac{T}{\delta}\right)}}_{=:C_{n_{\mathrm{S}},T}} \\
&\qquad + \sum_{t\in[T]}\frac{1}{2n_{\mathrm{S}}T}\left\|z_t\right\|_2^2.
\end{aligned}$$

Note that the simplification of the regularizer on the optimum holds since $\left\|w_t^*\right\|_2 = \Theta(1)$ by Assumption 3.2. Equivalently, by rearranging,

$$\begin{aligned}
&\sum_{t\in[T]}\frac{1}{2n_{\mathrm{S}}T}\left\|X_t(\theta_t^* - \hat{\theta}_t)\right\|_2^2 + \frac{\lambda}{2}\left\|\Delta_t\right\|_F^2 + \frac{\gamma}{2}\left\|w_t\right\|_2^2 \\
&\quad\le \sum_{t\in[T]}\frac{1}{2n_{\mathrm{S}}T}\left\langle z_t, X_t(\theta_t^* - \hat{\theta}_t)\right\rangle + C_{n_{\mathrm{S}},T}.
\end{aligned}$$

Finally, by Proposition I.4, the regularizer on Δ_t and w_t can be rewritten as a regularizer on $\hat{\delta}_t$, i.e.

$$\sum_{t\in[T]}\frac{1}{2n_{\mathrm{S}}T}\left\|X_t(\theta_t^* - \hat{\theta}_t)\right\|_2^2 + \sqrt{\lambda\gamma}\left\|\hat{\delta}_t\right\|_2 \le \sum_{t\in[T]}\frac{1}{2n_{\mathrm{S}}T}\left\langle z_t, X_t(\theta_t^* - \hat{\theta}_t)\right\rangle + C_{n_{\mathrm{S}},T}. \quad (8)$$

To proceed, define the set S as

$$\left\{[\alpha_t+\beta_t]_{t\in[T]} \;\middle|\; \operatorname{rank}[\alpha_t]\le 2k, \|\beta_t\|_2 \le \delta_0 + \left\|\hat{\delta}_t\right\|_2, \|X_t(\alpha_t+\beta_t)\|_2 \le \left\|X_t(\theta_t^*-\hat{\theta}_t)\right\|_2\right\},$$

and observe that $[\theta_t^*-\hat{\theta}_t]_{t\in[T]}\in S$, by letting $\alpha_t = B^*w_t^* - B_0w_t$ and $\beta_t = \delta_t^*-\hat{\delta}_t$. We bound the right-hand side of (8) via bounding the supremum of the inner product over S, i.e.

$$\sum_{t\in[T]}\left\langle z_t, X_t(\theta_t^*-\hat{\theta}_t)\right\rangle \le \sup_{[\alpha_t+\beta_t]\in S}\sum_{t\in[T]}\langle z_t, X_t(\alpha_t+\beta_t)\rangle.$$

Now, we decompose the Gaussian width as

$$\sup_{[\alpha_t+\beta_t]\in S}\sum_{t\in[T]}\langle z_t, X_t(\alpha_t+\beta_t)\rangle \le \underbrace{\sup_{[\alpha_t]\in S}\sum_{t\in[T]}\langle z_t, X_t\alpha_t\rangle}_{\text{(I)}} + \underbrace{\sup_{[\beta_t]\in S_2}\sum_{t\in[T]}\langle z_t, X_t\beta_t\rangle}_{\text{(II)}},$$

where $S_2 = \left\{[\beta_t]_{t\in[T]} \;\middle|\; \|\beta_t\|_2 \le \delta_0 + \left\|\hat{\delta}_t\right\|_2\right\}$. Note the abuse of notation in (I), where we say $[\alpha_t]_{t\in[T]}\in S$ if there exists $[\beta_t]_{t\in[T]}$ so that $[\alpha_t+\beta_t]_{t\in[T]}\in S$. This decomposes the Gaussian width into the sum of the Gaussian widths of a low-rank set (I) and a small norm set (II). We proceed to bound both terms accordingly to these two properties.

Bounding the Gaussian width of the low-rank set (I).

To bound the Gaussian width, we first enlarge S to remove β_t from the definition of the feasible set. Fix any (α_t,β_t) pair satisfying the conditions in S, and note that

$$\|X_t\beta_t\|_2 \lesssim \sqrt{n_{\mathrm{S}}}\,\|\Sigma\|_2^{1/2}\left(\delta_0 + \left\|\hat{\delta}_t\right\|_2\right).$$

Therefore, by the reverse triangle inequality,

$$\begin{aligned}&\left|\|X_t\alpha_t\|_2 - \|X_t\beta_t\|_2\right| \le \|X_t(\alpha_t+\beta_t)\|_2\\ &\quad\implies \|X_t\alpha_t\|_2 \le \underbrace{\left\|X_t(\theta_t^*-\hat{\theta}_t)\right\|_2 + \sqrt{n_{\mathrm{S}}}\,\|\Sigma\|_2^{1/2}\left(\delta_0+\left\|\hat{\delta}_t\right\|_2\right)}_{=:\rho_t}.\end{aligned}$$

Consequently, we can enlarge the feasible set to

$$S_1 \coloneqq \{[\alpha_t] \mid \operatorname{rank}[\alpha_t]\le 2k, \|X_t\alpha_t\|_2\le\rho_t\}.$$

Having relaxed the constraints, we now proceed to the main argument.

Since $\operatorname{rank}[\alpha_t]\le 2k$, there exists an orthogonal matrix $V\in\mathbb{R}^{d\times 2k}$ (dependent on α_t) and vectors $r_t\in\mathbb{R}^{2k}$ such that $\alpha_t = Vr_t$. Therefore, the inner product in the Gaussian width would be unchanged if we project z_t onto X_tV, i.e.

$$\sup_{[\alpha_t]\in S_1}\sum_{t\in[T]}\langle z_t, X_t\alpha_t\rangle = \sup_{[\alpha_t]\in S_1}\sum_{t\in[T]}\langle P_{X_tV}z_t, X_tVr_t\rangle.$$

The key idea we will leverage is that if V were chosen independently of z_t, then $P_{X_tV}z_t$ is Gaussian in a $2k$-dimensional space, and thus norm bounded by $\tilde{O}(\sqrt{2k})$ with high probability. However, due to the supremum over α_t, this independence assumption is not satisfied. Nevertheless, we can obtain a *fixed* finite covering of the set of all rank-$2k$ matrices, and ensure that the aforementioned norm bound on $P_{X_tV}z_t$ holds for every V in the covering via a union-bound. By choosing the discretization level of the covering appropriately, we can control the error resulting from approximating the supremum by some element of the covering.

Formally, let $\mathcal{O}^{d\times 2k}$ be the set of orthogonal matrices in $\mathbb{R}^{d\times 2k}$. By Proposition I.1, there exists an ε-covering of $\mathcal{O}^{d\times 2k}$ in the Frobenius norm with at most $(6\sqrt{2k}/\varepsilon)^{2kd}$ elements. Let $\bar{V}$ be an

element of the covering such that $\left\|V-\bar{V}\right\|_F \le \varepsilon$. Then,

$$\begin{aligned}
\langle z_t, X_t\alpha_t\rangle &= \left\langle z_t, X_t\bar{V}r_t\right\rangle + \left\langle z_t, X_t(V-\bar{V})r_t\right\rangle \\
&\le \left\|P_{X_t\bar{V}}z_t\right\|_2 \left\|X_t\bar{V}r_t\right\|_2 + \|z_t\|_2 \left\|X_t(V-\bar{V})r_t\right\|_2 \\
&\le \left\|P_{X_t\bar{V}}z_t\right\|_2 \|X_tVr_t\|_2 + (\left\|P_{X_t\bar{V}}z_t\right\|_2 + \|z_t\|_2) \left\|X_t(V-\bar{V})r_t\right\|_2 \\
&\le \left\|P_{X_t\bar{V}}z_t\right\|_2 \|X_tVr_t\|_2 + (\left\|P_{X_t\bar{V}}z_t\right\|_2 + \|z_t\|_2) \left\|X_t(V-\bar{V})r_t\right\|_2 \\
&\lesssim \left\|P_{X_t\bar{V}}z_t\right\|_2 \|X_t\alpha_t\|_2 + \|z_t\|_2 \left\|X_t(V-\bar{V})r_t\right\|_2,
\end{aligned}$$

and thus

$$\begin{aligned}
\sum_{t\in[T]} \langle z_t, X_tVr_t\rangle \le &\sqrt{\underbrace{\left(\sum_{t\in[T]} \left\|P_{X_t\bar{V}}z_t\right\|_2^2\right)}_{\text{(A)}} \left(\sum_{t\in[T]} \|X_t\alpha_t\|_2^2\right)} \\
&+ \sqrt{\underbrace{\left(\sum_{t\in[T]} \|z_t\|_2^2\right)}_{\text{(B)}} \underbrace{\left(\sum_{t\in[T]} \left\|X_t(V-\bar{V})r_t\right\|_2\right)}_{\text{(C)}}}.
\end{aligned}$$

We will bound (A), (B), and (C) individually.

(A) For a fixed $\bar{V}$, (A) is a chi-squared random variable with kT degrees of freedom scaled by σ^2. However, since $\bar{V}$ depends on V, we need to have a high probability bound for any element of the covering. By using known concentration bounds for chi-squared random variables together with the union-bound, we find that uniformly over the covering,

$$\text{(A)} \le \sigma^2\left(kT + \log\frac{1}{\delta'}\right) \qquad \text{w.p.} \ge 1-\delta'\left(\frac{6\sqrt{2k}}{\varepsilon}\right)^{2kd}.$$

(B) Note that (B) is a chi-squared random variable with $n_{\mathrm{S}}T$ degrees of freedom scaled by σ^2, and thus

$$\text{(B)} \lesssim \sigma^2\left(n_{\mathrm{S}}T + \log\frac{1}{\delta}\right) \qquad \text{w.p.} \ge 1-\frac{\delta}{18}.$$

(C) Since $(1/n_{\mathrm{S}})X^\top X$ is concentrated about Σ via Lemma B.1,

$$\left\|X_t(V-\bar{V})r_t\right\|_2^2 \lesssim n_{\mathrm{S}}\|\Sigma\|_2\varepsilon^2\|r_t\|_2^2 = n_{\mathrm{S}}\|\Sigma\|_2\varepsilon^2\|Vr_t\|_2^2 \lesssim \kappa\varepsilon^2\|X_tVr_t\|_2^2.$$

Putting these bounds together, and setting $\varepsilon = \sqrt{k/\kappa n_{\mathrm{S}}}$ and $\delta' = \delta/[18(6\sqrt{2k}/\varepsilon)^{2kd}]$, we obtain

$$\begin{aligned}
\sum_{t\in[T]} \langle z_t, X_t\alpha_t\rangle &\lesssim \left(\sigma\sqrt{kT+\log\frac{1}{\delta'}} + \varepsilon\sigma\sqrt{\kappa n_{\mathrm{S}}T+\log\frac{1}{\delta}}\right)\sqrt{\left(\sum_{t\in[T]}\|X_tVr_t\|_2^2\right)} \\
&\lesssim \left(\sigma\sqrt{kT+\log\frac{1}{\delta'}}\right)\sqrt{\left(\sum_{t\in[T]}\|X_tVr_t\|_2^2\right)} \qquad (k<n_{\mathrm{S}}, \delta'<\delta) \\
&\lesssim \sigma\sqrt{kT+kd\log\kappa n_{\mathrm{S}}+\log\frac{1}{\delta}}\sqrt{\left(\sum_{t\in[T]}\|X_t\alpha_t\|_2^2\right)}.
\end{aligned}$$

Taking the supremum over S_1, we thus obtain the following bound on the Gaussian width:

$$\sup_{[\alpha_t]\in S_1} \sum_{t\in[T]} \langle z_t, X_t\alpha_t\rangle$$
$$\lesssim \sigma\sqrt{kT + kd\log\kappa n_{\mathrm{S}} + \log\frac{1}{\delta}}\left[\sqrt{\sum_{t\in[T]}\left\|X_t(\theta_t^* - \hat{\theta}_t)\right\|_2^2} + \sqrt{\|\Sigma\|_2\, n_{\mathrm{S}} T}\left(\delta_0 + \left\|\hat{\delta}_t\right\|_2\right)\right].$$

Note that the events for this sub-argument occur with probability at least $1 - \delta/9$.

Bounding the Gaussian width of the low-norm set (II).

Recall that we want to bound

$$\sup_{[\beta_t]\in S_2} \sum_{t\in[T]} \langle z_t, X_t\beta_t\rangle, \quad S_2 \coloneqq \left\{[\beta_t] \;\middle|\; \|\beta_t\|_2 \le \delta_0 + \left\|\hat{\delta}_t\right\|_2\right\}.$$

For any $t \in [T]$,

$$\langle z_t, X_t\beta_t\rangle = \left\langle X_t^\top z_t, \beta_t\right\rangle \le \left(\delta_0 + \left\|\hat{\delta}_t\right\|_2\right)\left\|X_t^\top z_t\right\|_2.$$

Furthermore, by the Hanson-Wright inequality, we have that with probability at least $1 - \delta/9T$,

$$\left\|X_t^\top z_t\right\|_2 \lesssim \sigma\sqrt{n_{\mathrm{S}}\operatorname{tr}\Sigma\left(1 + \log\frac{T}{\delta}\right)}.$$

Putting everything together, we thus find that with probability at least $1 - \delta/9$,

$$\sup_{[\beta_t]\in S_2} \sum_{t\in[T]} \langle z_t, X_t\beta_t\rangle \le \sigma\left(T\delta_0 + \sum_{t\in[T]}\left\|\hat{\delta}_t\right\|_2\right)\sqrt{n_{\mathrm{S}}\operatorname{tr}\Sigma\left(1 + \log\frac{T}{\delta}\right)}.$$

Combining the bounds and concluding.

Having bounded both Gaussian widths, we can thus bound the right-hand side of (8) as

$$\sum_{t\in[T]} \frac{1}{2n_{\mathrm{S}}T}\left\|X_t(\theta_t^* - \hat{\theta}_t)\right\|_2^2 + \left(\sqrt{\lambda\gamma} - \frac{1}{\delta_0 T}C_{n_{\mathrm{S}},T}\right)\left\|\hat{\delta}_t\right\|_2$$
$$\lesssim \frac{\sigma}{\sqrt{nT}}\sqrt{kT + kd\log\kappa n_{\mathrm{S}} + \log\frac{1}{\delta}}\left(\sum_{t\in[T]} \frac{1}{2n_{\mathrm{S}}T}\left\|X_t(\theta_t^* - \hat{\theta}_t)\right\|_2^2\right)^{1/2} + C_{n_{\mathrm{S}},T}.$$

Therefore, as long as $\sqrt{\lambda\gamma} \ge (1/\delta_0 T)C_{n_{\mathrm{S}},T}$,

$$\sum_{t\in[T]} \frac{1}{2n_{\mathrm{S}}T}\left\|X_t(\theta_t^* - \hat{\theta}_t)\right\|_2^2$$
$$\lesssim \frac{\sigma}{\sqrt{n_{\mathrm{S}}T}}\sqrt{kT + kd\log\kappa n_{\mathrm{S}} + \log\frac{1}{\delta}}\left(\sum_{t\in[T]} \frac{1}{2n_{\mathrm{S}}T}\left\|X_t(\theta_t^* - \hat{\theta}_t)\right\|_2^2\right)^{1/2} + C_{n_{\mathrm{S}},T}.$$

Finally, by solving the quadratic inequality using Proposition I.2, we find that

$$\sum_{t\in[T]} \frac{1}{2n_{\mathrm{S}}T}\left\|X_t(\theta_t^* - \hat{\theta}_t)\right\|_2^2$$
$$\lesssim \frac{\sigma^2}{n_{\mathrm{S}}T}\left(kT + kd\log\kappa n_{\mathrm{S}} + \log\frac{1}{\delta}\right) + \frac{\sigma\delta_0\,\|\Sigma\|_2^{1/2}}{\sqrt{n_{\mathrm{S}}T}}\sqrt{kT + kd\log\kappa n_{\mathrm{S}} + \log\frac{1}{\delta}}$$
$$+ \frac{\sigma\delta_0}{\sqrt{n_{\mathrm{S}}}}\sqrt{\operatorname{tr}\Sigma\left(1 + \log\frac{T}{\delta}\right)},$$

from which the desired performance bound follows due to the concentration of the empirical covariance from Lemma B.1. Since the concentration of the source covariance matrices and each of the sub-arguments all hold with probability at least $1 - \delta/9$, it follows that the main claim holds with probability at least $1 - \delta/3$. $\square$

The prior bound is central to the analysis, as the performance of the learner can be tied to how well B_0 spans the correct space. To see why this is the case, note that for large n_{S}, the effect of the noise on the optimization in (4) is negligible. In this regime, no matter which representation B_0 the learner has chosen, the optimal predictor would satisfy $P_{X_t B_0} X_t \hat{\theta}_t \approx P_{X_t B_0} X_t \theta_t^*$ and $P^{\perp}_{X_t B_0} X_t \hat{\theta}_t \approx P^{\perp}_{X_t B_0} X_t \theta_t^*$. Consequently, the performance of the predictors chosen by the learner can be tied to the chosen representation B_0. We formalize this intuition in the following result:

Lemma B.4 (Transfer Lemma)**.** *Under Assumption 3.2, we have that*

$$\left\| P_{\Sigma^{1/2} B_0} \Sigma^{1/2} B^* \right\|_F^2 \lesssim k \left[\frac{1}{T} \sum_{t \in [T]} \left\| \Sigma^{1/2} (\theta_t^* - \hat{\theta}_t) \right\|_2^2 \right].$$

Proof. Throughout this proof, we will write $P \coloneqq P_{\Sigma^{1/2} B_0}$ and $P^{\perp} \coloneqq P^{\perp}_{\Sigma^{1/2} B_0}$ for readability. To proceed, note that we intuitively expect that for a learner that has learned the correct spaces, $P \Sigma^{1/2} \hat{\theta}_t \approx \Sigma^{1/2} B^* w_t^*$ as it is the low-rank component of the estimator, and consequently, $P^{\perp} \Sigma^{1/2} \hat{\theta}_t \approx \Sigma^{1/2} \delta_t^*$. Then, decomposing into the corresponding errors, we have that for any $t \in [T]$,

$$\begin{aligned}
\left\| \Sigma^{1/2} (\theta_t^* - \hat{\theta}_t) \right\|_2^2 &\gtrsim \left\| \Sigma^{1/2} B^* w_t^* - P \Sigma^{1/2} \hat{\theta}_t \right\|_2^2 + \left\| \Sigma^{1/2} \delta_t^* - P^{\perp} \Sigma^{1/2} \hat{\theta}_t \right\|_2^2 \\
&\quad + 2 \left\langle \Sigma^{1/2} B^* w_t^* - P \Sigma^{1/2} \hat{\theta}_t, \Sigma^{1/2} \delta_t^* - P^{\perp} \Sigma^{1/2} \hat{\theta}_t \right\rangle \\
&\geq \left\| P^{\perp} \Sigma^{1/2} B^* w_t^* \right\|_2^2 + \left\| \Sigma^{1/2} \delta_t^* - P^{\perp} \Sigma^{1/2} \hat{\theta}_t \right\|_2^2 \\
&\quad - 2 \left| \left\langle \Sigma^{1/2} B^* w_t^* - P \Sigma^{1/2} \hat{\theta}_t, \Sigma^{1/2} \delta_t^* - P^{\perp} \Sigma^{1/2} \hat{\theta}_t \right\rangle \right|.
\end{aligned}$$

We proceed to bound the inner product above. We do so by observing that if we were to replace $\hat{\theta}_t$ by θ_t^*, then

$$\begin{aligned}
&\left\langle \Sigma^{1/2} B^* w_t^* - P \Sigma^{1/2} \theta_t^*, \Sigma^{1/2} \delta_t^* - P^{\perp} \Sigma^{1/2} \delta_t^* \right\rangle \\
&\quad = \left\langle P^{\perp} \Sigma^{1/2} B^* w_t^* - P \Sigma^{1/2} \delta_t^*, P^{\perp} \Sigma^{1/2} \delta_t^* - P \Sigma^{1/2} B^* w_t^* \right\rangle \\
&\quad = (w_t^*)^{\top} (B^*)^{\top} \Sigma \delta_t^* \\
&\quad = 0.
\end{aligned}$$

To translate this result into a bound on the original inner product, we note that as $n_{\mathrm{S}} \to \infty$, and the impact of the noise and the regularizer on the optimization is diminished, we expect $\hat{\theta}_t$ to learn the projections of $X_t \theta_t^*$ onto $X_t B_0$ and its complement. With these two insights in mind, we decompose the inner product as

$$\begin{aligned}
&\left| \left\langle \Sigma^{1/2} B^* w_t^* - P \Sigma^{1/2} \hat{\theta}_t, \Sigma^{1/2} \delta_t^* - P^{\perp} \Sigma^{1/2} \hat{\theta}_t \right\rangle \right| \\
&\quad \leq \left| \left\langle \Sigma^{1/2} B^* w_t^* - P \Sigma^{1/2} \theta_t^*, P^{\perp} \Sigma^{1/2} (\theta_t^* - \hat{\theta}_t) \right\rangle \right| \\
&\qquad + \left| \left\langle P \Sigma^{1/2} (\theta_t^* - \hat{\theta}_t), \Sigma^{1/2} \delta_t^* - P^{\perp} \Sigma^{1/2} \hat{\theta}_t \right\rangle \right| \\
&\qquad + \underbrace{\left| \left\langle \Sigma^{1/2} B^* w_t^* - P \Sigma^{1/2} \theta_t^*, \Sigma^{1/2} \delta_t^* - P^{\perp} \Sigma^{1/2} \theta_t^* \right\rangle \right|}_{=0} \\
&\quad = \left| \left\langle P^{\perp} \Sigma^{1/2} B^* w_t^*, P^{\perp} \Sigma^{1/2} (\theta_t^* - \hat{\theta}_t) \right\rangle \right| + \left| \left\langle P \Sigma^{1/2} (\theta_t^* - \hat{\theta}_t), \Sigma^{1/2} \delta_t^* - P^{\perp} \Sigma^{1/2} \hat{\theta}_t \right\rangle \right|.
\end{aligned}$$

Putting everything together, we find that

$$\begin{aligned}
&\frac{1}{T}\sum_{t\in[T]}\left\|P^{\perp}\Sigma^{1/2}B^{*}w_t^{*}\right\|_2^2+\frac{1}{T}\sum_{t\in[T]}\left\|\Sigma^{1/2}\delta_t^{*}-P^{\perp}\Sigma^{1/2}\hat{\theta}_t\right\|_2^2\\
&\quad\lesssim\frac{1}{T}\sum_{t\in[T]}\left\|\Sigma^{1/2}(\theta_t^{*}-\hat{\theta}_t)\right\|_2^2+\frac{1}{T}\sum_{t\in[T]}\left|\left\langle\Sigma^{1/2}B^{*}w_t^{*}-P\Sigma^{1/2}\hat{\theta}_t,\Sigma^{1/2}\delta_t^{*}-P^{\perp}\Sigma^{1/2}\hat{\theta}_t\right\rangle\right|\\
&\quad\le\frac{1}{T}\sum_{t\in[T]}\left\|\Sigma^{1/2}(\theta_t^{*}-\hat{\theta}_t)\right\|_2^2+\frac{1}{T}\sum_{t\in[T]}\left|\left\langle P^{\perp}\Sigma^{1/2}B^{*}w_t^{*},P^{\perp}\Sigma^{1/2}(\theta_t^{*}-\hat{\theta}_t)\right\rangle\right|\\
&\qquad+\frac{1}{T}\sum_{t\in[T]}\left|\left\langle P\Sigma^{1/2}(\theta_t^{*}-\hat{\theta}_t),\Sigma^{1/2}\delta_t^{*}-P^{\perp}\Sigma^{1/2}\hat{\theta}_t\right\rangle\right|\\
&\quad\le\frac{1}{T}\sum_{t\in[T]}\left\|\Sigma^{1/2}(\theta_t^{*}-\hat{\theta}_t)\right\|_2^2+\sqrt{\frac{1}{T}\sum_{t\in[T]}\left\|P^{\perp}\Sigma^{1/2}(\theta_t^{*}-\hat{\theta}_t)\right\|_2^2}\sqrt{\frac{1}{T}\sum_{t\in[T]}\left\|P^{\perp}\Sigma^{1/2}B^{*}w_t^{*}\right\|_2^2}\\
&\qquad+\sqrt{\frac{1}{T}\sum_{t\in[T]}\left\|P\Sigma^{1/2}(\theta_t^{*}-\hat{\theta}_t)\right\|_2^2}\sqrt{\frac{1}{T}\sum_{t\in[T]}\left\|\Sigma^{1/2}\delta_t^{*}-P^{\perp}\Sigma^{1/2}\hat{\theta}_t\right\|_2^2}.
\end{aligned}$$

This is a quadratic inequality in the two terms on the left-hand side, and so by applying Proposition I.2,

$$\begin{aligned}
\frac{1}{T}\sum_{t\in[T]}\left\|P^{\perp}\Sigma^{1/2}B^{*}w_t^{*}\right\|_2^2&\lesssim\frac{1}{T}\sum_{t\in[T]}\left\|P^{\perp}\Sigma^{1/2}(\theta_t^{*}-\hat{\theta}_t)\right\|_2^2+\frac{1}{T}\sum_{t\in[T]}\left\|P\Sigma^{1/2}(\theta_t^{*}-\hat{\theta}_t)\right\|_2^2\\
&\quad+\frac{1}{T}\sum_{t\in[T]}\left\|\Sigma^{1/2}(\theta_t^{*}-\hat{\theta}_t)\right\|_2^2\\
&\lesssim\frac{1}{T}\sum_{t\in[T]}\left\|\Sigma^{1/2}(\theta_t^{*}-\hat{\theta}_t)\right\|_2^2,
\end{aligned}$$

where the last line follows by orthogonality. Finally, by Proposition I.3,

$$\frac{1}{T}\sum_{t\in[T]}\left\|P^{\perp}\Sigma^{1/2}B^{*}w_t^{*}\right\|_2^2=\frac{1}{T}\left\|P^{\perp}\Sigma^{1/2}B^{*}W^{*}\right\|_F^2\ge\frac{\sigma_k(W^{*})^2}{T}\left\|P^{\perp}_{\Sigma^{1/2}B_0}\Sigma^{1/2}B^{*}\right\|_F^2,$$

which together with the diversity assumption in Assumption 3.2 yields the final bound

$$\left\|P^{\perp}_{\Sigma^{1/2}B_0}\Sigma^{1/2}B^{*}\right\|_F^2\lesssim k\left[\frac{1}{T}\sum_{t\in[T]}\left\|\Sigma^{1/2}(\theta_t^{*}-\hat{\theta}_t)\right\|_2^2\right].\qquad\square$$

B.2 Target Guarantees for the Linear Setting

Having established a connection between the performance on the source tasks and the difference in the spans of B^{*} and B_0, we can now analyze the performance of the target training procedure. First, we bound the performance of nearly optimal points in $\mathcal{C}_\beta$ for several possible choices of c_1, c_2.

Lemma B.5 (Statistical Rates for $\mathcal{L}_\beta$)**.** *Assume that (Δ, w) is ζ-suboptimal for $\mathcal{L}_\beta$ with the constraint set $\mathcal{C}_\beta=\{(\Delta,w)\mid\|\Delta\|_F\le c_1/\beta,\|w\|_2\le c_2/\beta\}$, i.e.*

$$\mathcal{L}_\beta(\Delta,w)\le\min_{(\Delta',w')\in\mathcal{C}_\beta}\mathcal{L}_\beta(\Delta',w')+\zeta.$$

We write $\hat{\theta}$ for the predictor corresponding to (Δ,w), i.e. $\hat{\theta}=\beta(A_{B_0}+\Delta)(w_0+w)$. Now, let

$$\begin{aligned}
\bar{w}&=(B_0X^{\top}XB_0)^{\dagger}B_0^{\top}X^{\top}X\theta^{*}\\
\bar{\delta}&=\theta^{*}-B_0\bar{w}\\
\tilde{w}&=(B_0^{\top}X^{\top}XB_0)^{\dagger}B_0^{\top}X^{\top}XB^{*}w^{*}.
\end{aligned}$$

Note that $XB_0\bar{w} = P_{XB_0}X\theta^*$, $X\bar{\delta} = P^{\perp}_{XB_0}X\theta^*$, *and* $XB\tilde{w} = P_{XB_0}XB^*w^*$. *Then, assuming* $\beta > \max(\|\bar{w}\|_2, \|\tilde{w}\|_2)$,

$$\frac{1}{2n_{\mathrm{T}}}\left\|X(\theta^* - \hat{\theta})\right\|_2^2 \lesssim \zeta + \begin{cases} \left\|P^{\perp}_{\Sigma^{1/2}B_0}\Sigma^{1/2}B^*w^*\right\|_2^2 & \\ \quad + \dfrac{\sigma^2}{n_{\mathrm{T}}}\left(k + \log\dfrac{1}{\delta}\right) & c_1 = \delta_0, c_2 = \|\tilde{w}\|_2 \\ \quad + \dfrac{\sigma\delta_0}{\sqrt{n_{\mathrm{T}}}}\sqrt{\operatorname{tr}\Sigma\left(1 + \log\dfrac{1}{\delta}\right)} & \\ \dfrac{\sigma^2}{n_{\mathrm{T}}}\left(k + \log\dfrac{1}{\delta}\right) & \\ \quad + \dfrac{\sigma\left\|\bar{\delta}\right\|_2}{\sqrt{n_{\mathrm{T}}}}\sqrt{\operatorname{tr}\Sigma\left(1 + \log\dfrac{1}{\delta}\right)} & c_1 = \left\|\bar{\delta}\right\|_2, c_2 = \|\bar{w}\|_2 \\ \dfrac{\sigma^2}{n_{\mathrm{T}}}\left(d + \log\dfrac{1}{\delta}\right) & c_1 = \|\theta^*\|_2, c_2 = 0 \end{cases}$$

with probability at least $1 - \delta/3$.

Proof. We proceed by proving the three cases separately. Throughout the proof, we instantiate the high-probability event in Lemma B.2, which guarantees that

$$0.9\Sigma \preceq \frac{1}{n_{\mathrm{T}}}X^\top X \preceq 1.1\Sigma.$$

Recall that this event occurs with probability at least $1 - \delta/9$.

$\underline{c_1 = \delta_0, c_2 = \|\tilde{w}\|_2}$

Due to the choice of c_1 and c_2, there exists a parameter in $\mathcal{C}_\beta$ corresponding to the prediction vector $P_{XB_0}X\theta^* + X\delta^*$. Writing the corresponding basic inequality, we thus have that

$$\begin{aligned} \frac{1}{2n_{\mathrm{T}}}\left\|y - X\hat{\theta}\right\|_2^2 - \zeta &\leq \frac{1}{2n_{\mathrm{T}}}\left\|P^{\perp}_{XB_0}XB^*w^*\right\|_2^2 + \frac{1}{n_{\mathrm{T}}}\left\langle z, P^{\perp}_{XB_0}XB^*w^*\right\rangle + \frac{1}{2n_{\mathrm{T}}}\|z\|_2^2 \\ \implies \frac{1}{2n_{\mathrm{T}}}\left\|X(\theta^* - \hat{\theta})\right\|_2^2 &\leq \zeta + \frac{1}{2n_{\mathrm{T}}}\left\|P^{\perp}_{XB_0}XB^*w^*\right\|_2^2 + \frac{1}{n_{\mathrm{T}}}\left\langle z, P^{\perp}_{XB_0}XB^*w^*\right\rangle \\ &\quad - \frac{1}{n_{\mathrm{T}}}\left\langle z, X(\theta^* - \hat{\theta})\right\rangle. \end{aligned}$$

Simplifying further,

$$\begin{aligned} &\frac{1}{n_{\mathrm{T}}}\left\langle z, P^{\perp}_{XB_0}XB^*w^*\right\rangle - \frac{1}{n_{\mathrm{T}}}\left\langle z, X(\theta^* - \hat{\theta})\right\rangle \\ &\quad = \frac{1}{n_{\mathrm{T}}}\left\langle z, P_{XB_0}X(\hat{\theta} - \theta^*)\right\rangle - \frac{1}{n_{\mathrm{T}}}\left\langle z, P^{\perp}_{XB_0}X(\delta^* - \beta\Delta(w_0 + w))\right\rangle. \end{aligned}$$

Now, by the Hanson-Wright inequality, we can bound the last term as

$$\begin{aligned} -\frac{1}{n_{\mathrm{T}}}\left\langle X^\top P^{\perp}_{XB_0}z, \delta^* - \beta\Delta(w_0 + w)\right\rangle &\leq \frac{1}{n_{\mathrm{T}}}\left\|X^\top P^{\perp}_{XB_0}z\right\|_2\|\delta^* + \beta\Delta(w_0 + w)\|_2 \\ &\lesssim \frac{\sigma\delta_0}{\sqrt{n_{\mathrm{T}}}}\sqrt{\operatorname{tr}\Sigma\left(1 + \log\frac{1}{\delta}\right)} \end{aligned}$$

with probability at least $1 - \delta/9$. Therefore, we can rewrite the prior basic inequality as

$$\begin{aligned} \frac{1}{2n_{\mathrm{T}}}\left\|X(\theta^* - \hat{\theta})\right\|_2^2 \lesssim \underbrace{\frac{1}{2n_{\mathrm{T}}}\left\|P^{\perp}_{XB_0}XB^*w^*\right\|_2^2 + \frac{\sigma\delta_0}{\sqrt{n_{\mathrm{T}}}}\sqrt{\operatorname{tr}\Sigma\left(1 + \log\frac{1}{\delta}\right)} + \zeta}_{=:C} \\ + \frac{1}{n_{\mathrm{T}}}\left\langle z, P_{XB_0}X(\theta^* - \hat{\theta})\right\rangle. \end{aligned}$$

Now, note that we can form the quadratic inequality

$$\frac{1}{n_{\mathrm{T}}}\left\|P_{XB_0}X(\theta^*-\hat{\theta})\right\|_2^2 \le \frac{1}{n_{\mathrm{T}}}\left\|P_{XB_0}z\right\|_2\left\|P_{XB_0}X(\theta^*-\hat{\theta})\right\|_2 + C,$$

and thus by applying Proposition I.2 to solve the inequality and noting that $P_{XB_0}z/\sigma$ is distributed as a chi-squared random variable with k degrees of freedom,

$$\frac{1}{n_{\mathrm{T}}}\left\|X(\theta^*-\hat{\theta})\right\|_2^2 \lesssim \zeta + \left\|P^{\perp}_{\Sigma^{1/2}B_0}\Sigma^{1/2}B^*w^*\right\|_2^2 + \frac{\sigma^2}{n_{\mathrm{T}}}\left(k+\log\frac{1}{\delta}\right) + \frac{\sigma\delta_0}{\sqrt{n_{\mathrm{T}}}}\sqrt{\operatorname{tr}\Sigma\left(1+\log\frac{1}{\delta}\right)}$$

with probability at least $1-\delta/9$, which is the bound that we wanted to show.

$\underline{c_1 = \left\|\bar{\delta}\right\|_2, c_2 = \|\bar{w}\|_2}$

Due to the choice of c_1 and c_2, there exists a parameter in $\mathcal{C}_\beta$ corresponding to a predictor that agrees with θ^* on the target samples. Therefore,

$$\begin{aligned}&\frac{1}{2n_{\mathrm{T}}}\left\|X(\theta^*-\hat{\theta})\right\|_2^2 \le \zeta + \frac{1}{2n_{\mathrm{T}}}\|z\|_2^2\\ &\quad\implies \frac{1}{2n_{\mathrm{T}}}\left\|X(\theta^*-\hat{\theta})\right\|_2^2 \le \zeta + \frac{1}{n_{\mathrm{T}}}\left\langle z, P_{XB_0}X(\theta^*-\hat{\theta})\right\rangle + \frac{1}{n_{\mathrm{T}}}\left\langle z, P^{\perp}_{XB_0}X(\theta^*-\hat{\theta})\right\rangle.\end{aligned} \tag{9}$$

We proceed with an argument similar to that used in the source guarantee, albeit simpler since the representation B_0 is fixed (and thus no covering argument is required). Along these lines, we bound the first term using the low-rank of B_0. Via projections,

$$\begin{aligned}&\frac{1}{n_{\mathrm{T}}}\left\|P_{XB_0}X(\theta^*-\hat{\theta})\right\|_2^2\\ &\quad\lesssim \zeta + \frac{1}{n_{\mathrm{T}}}\left\|P_{XB_0}z\right\|_2\left\|P_{XB_0}X(\theta^*-\hat{\theta})\right\|_2 + \frac{1}{n_{\mathrm{T}}}\left\langle z, P^{\perp}_{XB_0}X(\theta^*-\hat{\theta})\right\rangle.\end{aligned}$$

Therefore, by applying Proposition I.2 to solve the quadratic inequality, we have that with probability at least $1-\delta/9$,

$$\frac{1}{2n_{\mathrm{T}}}\left\|X(\theta^*-\hat{\theta})\right\|_2^2 \le \zeta + \frac{\sigma^2}{n_{\mathrm{T}}}\left(k+\log\frac{1}{\delta}\right) + \frac{1}{n_{\mathrm{T}}}\left\langle z, P^{\perp}_{XB_0}X(\theta^*-\hat{\theta})\right\rangle.$$

To bound the second term, we simply make use of the norm constraints defining the feasible set $\mathcal{C}_\beta$, which we note is analogous to the low-norm sub-argument of the source guarantee. Formally, the Hanson-Wright inequality implies that with probability at least $1-\delta/9$,

$$\begin{aligned}\frac{1}{n_{\mathrm{T}}}\left\langle z, P^{\perp}_{XB_0}X(\theta^*-\hat{\theta})\right\rangle &= \frac{1}{\sqrt{n_{\mathrm{T}}}}\left\|\frac{1}{\sqrt{n_{\mathrm{T}}}}X^\top P^{\perp}_{XB_0}z\right\|_2 \left\|\bar{\delta}+\beta\Delta w_0+\beta\Delta w\right\|_2\\ &\lesssim \frac{\sigma\left\|\bar{\delta}\right\|_2}{\sqrt{n_{\mathrm{T}}}}\sqrt{\operatorname{tr}\Sigma\left(1+\log\frac{1}{\delta}\right)}.\end{aligned}$$

Putting everything together, we thus have that

$$\frac{1}{2n_{\mathrm{T}}}\left\|X(\theta^*-\hat{\theta})\right\|_2^2 \lesssim \zeta + \frac{\sigma^2}{n_{\mathrm{T}}}\left(k+\log\frac{1}{\delta}\right) + \frac{\sigma\left\|\bar{\delta}\right\|_2}{\sqrt{n_{\mathrm{T}}}}\sqrt{\operatorname{tr}\Sigma\left(1+\log\frac{1}{\delta}\right)}.$$

$\underline{c_1 = \left\|\theta^*\right\|_2, c_2 = 0}$

Due to the choice of c_1 and c_2, there exists a parameter in $\mathcal{C}_\beta$ corresponding to a predictor that agrees with θ^* on the target samples. Therefore, we can write the basic inequality

$$\frac{1}{n_{\mathrm{T}}}\left\|y-X\hat{\theta}\right\|_2^2 \le \zeta + \frac{1}{n_{\mathrm{T}}}\|z\|_2^2 \implies \frac{1}{n_{\mathrm{T}}}\left\|X(\theta^*-\hat{\theta})\right\|_2^2 \le \zeta + \frac{1}{n_{\mathrm{T}}}\left\langle z, X(\theta^*-\theta)\right\rangle.$$

Now, noting that $P_X z/\sigma$ is a chi-squared random variable with d degrees of freedom, we have that with probability at least $1 - 2\delta/9$,

$$\frac{1}{n_{\mathrm{T}}}\left\langle z, X(\theta^* - \hat{\theta})\right\rangle \leq \frac{1}{n_{\mathrm{T}}}\left\|P_X z\right\|_2 \left\|X(\theta^* - \hat{\theta})\right\|_2 \leq \frac{1}{n_{\mathrm{T}}}\sqrt{d + \log\frac{1}{\delta}}\left\|X(\theta^* - \hat{\theta})\right\|_2.$$

Therefore, by solving the resulting quadratic inequality via Proposition I.2, we obtain the bound

$$\frac{1}{n_{\mathrm{T}}}\left\|X(\theta^* - \hat{\theta})\right\|_2^2 \leq \zeta + \frac{\sigma^2}{n_{\mathrm{T}}}\left(d + \log\frac{1}{\delta}\right).$$

Observe that all relevant high-probability events for each case occur simultaneously with probability at least $1 - \delta/3$, as desired. □

B.3 Optimization Landscape during Target Time Training

Having derived statistical rates on nearly-optimal points for several choices of $\mathcal{C}_\beta$ in the prior section, all that remains to be shown is that projected gradient descent can indeed find such points. In particular, we will demonstrate that for large enough β, the optimization landscape induced by $\mathcal{L}_\beta$ is approximately convex. We do so by demonstrating that the objective satisfies the assumptions outlined in Section H, and thus the accompanying guarantees for projected gradient descent hold.

Lemma B.6 (Approximate linearity of function class)**.** *Let $g_\theta(x) = \beta x^\top(A_{B_0} + \Delta)(w_0 + w)$ for $\theta = (\Delta, w) \in \mathcal{C}_\beta$, where $\mathcal{C}_\beta$ is considered as a subset of $\mathbb{R}^{dk+k}$. Then, assuming the high-probability event in Lemma B.2 holds, then*

$$\sup_{\theta\in\mathcal{C}_\beta}\frac{1}{n_{\mathrm{T}}}\sum_{i\in[n_{\mathrm{T}}]}\left\|\nabla_\theta^2 g_\theta(x_i)\right\|_2^2 \lesssim \beta^2\operatorname{tr}\Sigma \quad \textit{and} \quad \frac{1}{n_{\mathrm{T}}}\sum_{i\in[n_{\mathrm{T}}]}\left\|\nabla_\theta g_0(x_i)\right\|_2^2 \lesssim \beta^2\operatorname{tr}\Sigma.$$

Proof. To bound the average squared Hessian operator norm, note that $\nabla_\theta^2 g_\theta(x_i)[\Delta, w] = \beta x_i^\top \Delta w$, which is independent of θ. Then, by the variational characterization of the operator norm,

$$\begin{aligned}\left\|\nabla_\theta^2 g_\theta(x_i)\right\|_2 &= \sup_{\|\Delta\|_F^2 + \|w\|_2^2 \leq 1}\left|\nabla_\theta^2 g_\theta(x_i)[\Delta, w]\right| = \sup_{\|\Delta\|_F^2 + \|w\|_2^2 \leq 1}\beta\left|x_i^\top \Delta w\right| \\ &\leq \sup_{\|\Delta\|_F^2 + \|w\|_2^2 \leq 1}\beta\left\|x_i\right\|_2\left\|\Delta\right\|_F\left\|w\right\|_2 \\ &\leq \beta\left\|x_i\right\|_2.\end{aligned}$$

and thus $\left\|\nabla_\theta^2 g_\theta(x_i)\right\|_2^2 \lesssim \beta^2\left\|x_i\right\|_2^2$. Consequently,

$$\sup_{\theta\in\mathcal{C}_\beta}\frac{1}{n_{\mathrm{T}}}\sum_{i\in[n_{\mathrm{T}}]}\left\|\nabla_\theta^2 g_\theta(x_i)\right\|_2^2 \leq \frac{\beta^2}{n_{\mathrm{T}}}\operatorname{tr} XX^\top \lesssim \beta^2\operatorname{tr}\Sigma.$$

We now proceed to bound the squared norm of the gradient. Observe that the gradient is given by $\nabla_\theta g_0(x_i) = \beta[A_{B_0}^\top x_i, \operatorname{vec}(x_i w_0^\top)]$, and therefore,

$$\begin{aligned}\frac{1}{n_{\mathrm{T}}}\sum_{i\in[n_{\mathrm{T}}]}\left\|\nabla_\theta g_0(x_i)\right\|_2^2 &= \frac{\beta^2}{n_{\mathrm{T}}}\sum_{i\in[n_{\mathrm{T}}]}\left\|A_{B_0}^\top x_i\right\|_2^2 + \left\|x_i w_0^\top\right\|_F^2 \lesssim \frac{\beta^2}{n_{\mathrm{T}}}\sum_{i\in[n_{\mathrm{T}}]}\left\|x_i\right\|_2^2 = \frac{\beta^2}{n_{\mathrm{T}}}\operatorname{tr} XX^\top \\ &\lesssim \beta^2\operatorname{tr}\Sigma,\end{aligned}$$

where the first inequality uses the fact that $\left\|A_{B_0}^\top x_i\right\|_2^2 = 2\left\|B_0^\top x_i\right\|_2^2 = 2\left\|P_{B_0} x_i\right\|_2^2 \lesssim \left\|x_i\right\|_2^2$, by the definition of A_{B_0} and assumed orthogonality of B_0. □

Lemma B.7 ($\mathcal{L}_\beta$ is Lipschitz)**.** *Let $g_\theta(x) = \beta x^\top(A_{B_0} + \Delta)(w_0 + w)$ for $\theta = (\Delta, w) \in \mathcal{C}_\beta$, where $\mathcal{C}_\beta$ is considered as a subset of $\mathbb{R}^{dk+k}$. Furthermore, assume that the high-probability event in Lemma B.2 holds. Then, with probability at least $1 - \delta/3$ over the draw of n_{T} labels, we have that for any $c_1, c_2 > 0$ and $\beta^2 > c_1^2 + c_2^2$,*

$$\sup_{\theta\in\mathcal{C}_\beta}\left\|\nabla_g\mathcal{L}_\beta(g_\theta(X))\right\|_2^2 \lesssim \frac{1}{n_{\mathrm{T}}}\left[\left\|\Sigma\right\|_2\left(\left\|\theta^*\right\|_2^2 + c_1^2 + c_2^2\right) + \sigma^2\left(1 + \log\frac{1}{\delta}\right)\right].$$

Proof. We assume that $\|z\|_2^2/n_{\mathrm{T}} \lesssim \sigma^2[1+\log(1/\delta)]$, which occurs with probability at least $1-\delta/3$ via standard tail bounds on chi-squared random variables. Then, the gradient of the loss with respect to the predictions is given by

$$\nabla_g \mathcal{L}_\beta(g_\theta(X)) = \frac{1}{n_{\mathrm{T}}}(y - g_\theta(X)) = \frac{1}{n_{\mathrm{T}}}[X(\theta^* - \beta A_{B_0} w - \beta \Delta w_0 - \beta \Delta w) + z].$$

Therefore,

$$\begin{aligned}\sup_{\theta \in \mathcal{C}_\beta} \|\nabla_g \mathcal{L}_\beta(g_\theta(X))\|_2^2 &\lesssim \sup_{\theta \in \mathcal{C}_\beta} \frac{1}{n_{\mathrm{T}}^2} \|X(\theta^* - \beta A_{B_0} w_0 - \beta \Delta w_0 - \beta \Delta w)\|_2^2 + \frac{1}{n_{\mathrm{T}}^2}\|z\|_2^2 \\ &\lesssim \frac{1}{n_{\mathrm{T}}}\left[\|\Sigma\|_2\left(\|\theta^*\|_2^2 + c_1^2 + c_2^2\right) + \sigma^2\left(1 + \log\frac{1}{\delta}\right)\right]. \qquad \square\end{aligned}$$

Lemma B.8. *Let $\tilde{w}$, $\bar{w}$, and $\bar{\delta}$ be defined as in Lemma B.5. Then, assuming that the high-probability event in Lemma B.2 holds, the three quantities above are norm-bounded by $\kappa^{1/2}(\|B^* w^*\|_2 + \delta_0)$, up to constant factors. Furthermore, we can also bound $\|\bar{\delta}\|_2$ as*

$$\|\bar{\delta}\|_2 \leq \delta_0 + \left(\frac{\|w^*\|_2 + \delta_0 \kappa^{1/2}}{\lambda_{\min}^{1/2}}\right)\left\|P^{\perp}_{\Sigma^{1/2}B_0}\Sigma^{1/2}B^*\right\|_F.$$

Proof. Throughout the proof, we will write $\lambda_{\max}$ and $\lambda_{\min}$ as shorthand for $\lambda_{\max}(\Sigma)$ and $\lambda_{\min}(\Sigma)$, respectively. By definition, since B_0 is assumed to be orthonormal, the concentration of the sample covariance in Lemma B.2 implies that

$$\begin{aligned}\|\bar{w}\|_2^2 &\lesssim \frac{1}{n_{\mathrm{T}}\lambda_{\min}}\|XB_0\bar{w}\|_2^2 = \frac{1}{n_{\mathrm{T}}\lambda_{\min}}\|P_{XB_0}X\theta^*\|_2^2 \lesssim \frac{1}{\lambda_{\min}}\left\|\Sigma^{1/2}\theta^*\right\|_2^2 \leq \kappa\|\theta^*\|_2^2 \\ &\leq \kappa(\|B^*w^*\|_2 + \delta_0)^2.\end{aligned}$$

Following similar arguments for $\tilde{w}$ and $\bar{\delta}$, we obtain the same bounds. Thus, we have demonstated that all three quantities are indeed norm-bounded by $\kappa^{1/2}(\|B^*w^*\|_2 + \delta_0)$, up to constant factors.

Finally, we proceed to derive the final bound on $\bar{\delta}$. Note that $X\bar{\delta} = X\delta^* - P_{XB_0}X\delta^* + P^{\perp}_{XB_0}XB^*w^*$. Therefore,

$$\begin{aligned}\|\bar{\delta}\|_2 &\lesssim \delta_0 + \frac{1}{\sqrt{n_{\mathrm{T}}\lambda_{\min}}}\left[\|P_{XB_0}X\delta^*\|_2 + \left\|P^{\perp}_{XB_0}XB^*w^*\right\|_2\right] \\ &\lesssim \delta_0 + \frac{1}{\lambda_{\min}^{1/2}}\left[\left\|P_{\Sigma^{1/2}B_0}\Sigma^{1/2}\delta^*\right\|_2 + \left\|P^{\perp}_{\Sigma^{1/2}B_0}\Sigma^{1/2}B^*w^*\right\|_2\right] \\ &\lesssim \delta_0 + \frac{1}{\lambda_{\min}^{1/2}}\left[\left\|\Sigma^{1/2}\delta^*\right\|_2\left\|P_{\Sigma^{1/2}B_0}P^{\perp}_{\Sigma^{1/2}B^*}\right\|_F + \|w^*\|_2\left\|P^{\perp}_{\Sigma^{1/2}B_0}\Sigma^{1/2}B^*\right\|_F\right]\end{aligned}$$

Now, by applying the properties of the trace operator and Proposition I.3,

$$\begin{aligned}\left\|P_{\Sigma^{1/2}B_0}P^{\perp}_{\Sigma^{1/2}B^*}\right\|_F^2 &= \operatorname{tr} P_{\Sigma^{1/2}B_0}P^{\perp}_{\Sigma^{1/2}B^*}P_{\Sigma^{1/2}B_0} = \operatorname{tr}(I - P_{\Sigma^{1/2}B^*})P_{\Sigma^{1/2}B_0} \\ &= \operatorname{tr} P_{\Sigma^{1/2}B_0} - P_{\Sigma^{1/2}B^*}P_{\Sigma^{1/2}B_0} \\ &\leq \operatorname{tr} P_{\Sigma^{1/2}B^*} - P_{\Sigma^{1/2}B^*}P_{\Sigma^{1/2}B_0} = \operatorname{tr} P_{\Sigma^{1/2}B^*}(I - P_{\Sigma^{1/2}B_0}) \\ &= \operatorname{tr} P_{\Sigma^{1/2}B^*}P^{\perp}_{\Sigma^{1/2}B_0}P_{\Sigma^{1/2}B^*} = \left\|P^{\perp}_{\Sigma^{1/2}B_0}P_{\Sigma^{1/2}B^*}\right\|_F^2 \\ &\leq \frac{1}{\lambda_{\min}}\left\|P^{\perp}_{\Sigma^{1/2}B_0}\Sigma^{1/2}B^*\right\|_F^2,\end{aligned}$$

and thus

$$\begin{aligned}\|\bar{\delta}\|_2 &\lesssim \delta_0 + \frac{1}{\lambda_{\min}^{1/2}}\left[\frac{1}{\lambda_{\min}^{1/2}}\left\|\Sigma^{1/2}\delta^*\right\|_2 + \|w^*\|_2\right]\left\|P^{\perp}_{\Sigma^{1/2}B_0}\Sigma^{1/2}B^*\right\|_F \\ &\leq \delta_0 + \left(\frac{\|w^*\|_2 + \delta_0\kappa^{1/2}}{\lambda_{\min}^{1/2}}\right)\left\|P^{\perp}_{\Sigma^{1/2}B_0}\Sigma^{1/2}B^*\right\|_F. \qquad \square\end{aligned}$$

B.4 Deducing Theorem 3.1

Having proven all the previous results, we can now assemble the main claim in Theorem 3.1. Recall that we have defined the rates

$$
\begin{aligned}
r_{\mathrm{S}}(n_{\mathrm{S}}, T) &\coloneqq \frac{\sigma^2}{n_{\mathrm{S}} T}\left(kT + kd\log\kappa n_{\mathrm{S}} + \log\frac{1}{\delta}\right) + \frac{\sigma\delta_0\left\|\Sigma\right\|_2^{1/2}}{\sqrt{n_{\mathrm{S}}T}}\sqrt{kT + kd\log\kappa n_{\mathrm{S}} + \log\frac{1}{\delta}} \\
&\quad + \frac{\sigma\delta_0}{\sqrt{n_{\mathrm{S}}}}\sqrt{\operatorname{tr}\Sigma\left(1+\log\frac{T}{\delta}\right)} \\
r_{\mathrm{T}}^{(1)}(n_{\mathrm{T}}) &\coloneqq kr_{\mathrm{S}}(n_{\mathrm{S}}, T) + \frac{\sigma^2}{n_{\mathrm{T}}}\left(k + \log\frac{1}{\delta}\right) + \frac{\sigma\delta_0}{\sqrt{n_{\mathrm{T}}}}\sqrt{\operatorname{tr}\Sigma\left(1+\log\frac{1}{\delta}\right)} \\
r_{\mathrm{T}}^{(2)}(n_{\mathrm{T}}) &\coloneqq \frac{\sigma^2}{n_{\mathrm{T}}}\left(k + \log\frac{1}{\delta}\right) \\
&\quad + \frac{\sigma}{\sqrt{n_{\mathrm{T}}}}\left[\delta_0 + \left(\frac{\left\|w^*\right\|_2 + \delta_0\kappa^{1/2}}{\lambda_{\min}^{1/2}(\Sigma)}\right)\sqrt{kr_{\mathrm{S}}(n_{\mathrm{S}}, T)}\right]\sqrt{\operatorname{tr}\Sigma\left(1+\log\frac{1}{\delta}\right)} \\
r_{\mathrm{T}}^{(3)}(n_{\mathrm{T}}) &\coloneqq \frac{\sigma^2}{n_{\mathrm{T}}}\left(d + \log\frac{1}{\delta}\right).
\end{aligned}
$$

Theorem 3.1 (Performance guarantee, linear representations)**.** *Assume that Assumptions 3.1 and 3.2 hold,* $n_{\mathrm{S}} \gg \rho^4(d+\log(T/\delta))$*, and* $n_{\mathrm{T}} \gg \rho^4(d+\log(1/\delta))$*. Then there are* $(\lambda, \gamma, \beta, T_{\mathrm{PGD}}, \eta, c_1, c_2)$ *such that the training procedure in Section 3.2, with high probability, finds* θ *achieving excess risk bounded as*

$$
\mathbb{E}\left[(x^\top\theta^* - x^\top\theta)^2\right] \lesssim \min(r_{\mathrm{T}}^{(1)}(n_{\mathrm{T}}), r_{\mathrm{T}}^{(2)}(n_{\mathrm{T}}), r_{\mathrm{T}}^{(3)}(n_{\mathrm{T}})).
$$

Proof. Assume that during source training, the regularization parameters (λ, γ) are chosen according to Lemma B.3. We instantiate the high-probability events in Lemmas B.3, B.5 and B.7. These events occur altogether with probability at least $1-\delta$. Throughout the proof, we define $r(n) = \min(r_{\mathrm{T}}^{(1)}(n), r_{\mathrm{T}}^{(2)}(n), r_{\mathrm{T}}^{(3)}(n))$.

Given all the events above, we know that the optimization landscape induced by $\mathcal{L}_\beta$ is well-behaved. That is, the function class is approximately linear in the parameters by Lemma B.6 and the loss is Lipschitz by Lemma B.7. Furthermore, for any of the proposed feasible sets in Lemma B.5, the bounds on relevant quantities provided by Lemma B.8 imply that

$$
c_1^2 + c_2^2 \lesssim \underbrace{\kappa(\left\|B^*w^*\right\|_2^2 + \delta_0^2)}_{=:R^2} \implies \sup_{\theta\in\mathcal{C}_\beta}\left\|\nabla_g\mathcal{L}_\beta(g_\theta(X))\right\|_2^2 \lesssim \frac{1}{n_{\mathrm{T}}}\underbrace{\left[\left\|\Sigma\right\|_2 R^2 + \sigma^2\left(1+\log\frac{1}{\delta}\right)\right]}_{=:\alpha^2}.
$$

Therefore, we have demonstrated that $\mathcal{L}_\beta$ satisfies the assumptions in Section H for any $\beta > R^2$. Consequently, by running projected gradient descent with parameters

$$
\beta = \max\left(R^2, \frac{\alpha R^2\sqrt{\operatorname{tr}\Sigma}}{r(n_{\mathrm{T}})}\right) \quad \text{and} \quad T_{\mathrm{PGD}} = \frac{\alpha^2 R^2\operatorname{tr}\Sigma}{r(n_{\mathrm{T}})^2},
$$

and setting η as in Theorem H.1, we can guarantee that projected gradient descent finds an $r(n_{\mathrm{T}})$–suboptimal point. Therefore, by choosing (c_1, c_2) in order to achieve the minimal rate $r(n_{\mathrm{T}})$ in Lemma B.5, we can guarantee that

$$
\min_t \frac{1}{n_{\mathrm{T}}}\left\|X(\theta_t - \theta^*)\right\|_2^2 \lesssim r(n_{\mathrm{T}}).
$$

Finally, due to target covariance concentration as guaranteed by the event in Lemma B.2, we thus have that the excess risk of the best predictor found by the algorithm on the target task is bounded as

$$
\min_t \mathbb{E}\left[(x^\top\theta_t - x^\top\theta_t^*)^2\right] = \min_t\left\|\Sigma^{1/2}(\theta_t - \theta^*)\right\|_2^2 \lesssim \min(r_{\mathrm{T}}^{(1)}(n), r_{\mathrm{T}}^{(2)}(n), r_{\mathrm{T}}^{(3)}(n)). \qquad \square
$$

C Linear Hard Case: Construction, Proofs, and Experiments

C.1 Hard Case Construction

In the following section, we will provide the construction of a task distribution family that is used in proving Theorem 3.2 in the main text.

C.1.1 Formal Construction

We now proceed with a formal construction that satisfies the conditions of Section 3.1. We provide the intermediate results used to prove Theorem 3.2, but leave their proofs for Section C.2 to simplify the presentation in this section.

Fix an $\varepsilon \in (0, 1)$, and let p be a Gaussian distribution on $\mathbb{R}^d$ with block-diagonal covariance Σ

$$\Sigma \coloneqq \begin{bmatrix} \varepsilon I_{d-k} & 0 \\ 0 & I_k \end{bmatrix}.$$

We define $E^*, E_k \subset \mathbb{R}^d$ to be the two eigenspaces of Σ determined by the two blocks, *i.e.*

$$E^* = \operatorname{Col} \begin{bmatrix} \varepsilon I_{d-k} \\ 0 \end{bmatrix} \quad \text{and} \quad E_k = \operatorname{Col} \begin{bmatrix} 0 \\ I_k \end{bmatrix}.$$

Then, for an orthogonal matrix $B \in \mathbb{R}^{d \times k}$ with $\operatorname{Col} B^* \subseteq E^*$, define a corresponding task distribution given by

$$\theta = \frac{1}{\sqrt{2\varepsilon}} Bw + \delta, \tag{10}$$

where w and δ are uniformly sampled from the unit spheres in $\mathbb{R}^k$ and E_k, respectively.

Recall that in the linear representation setting, FROZENREP optimizes the following objective to obtain a representation $\hat{B}$:

$$\hat{B} = \operatorname*{argmin}_B \min_{w_t} \frac{1}{2n_{\mathrm{S}}T} \sum_{t \in [T]} \|y_t - X_t B w_t\|_2^2 .$$

First, we characterize the span of $\hat{B}$ *in the limit of infinite source tasks and data*. Intuitively, $\operatorname{span}\{B^* w_t^*\}$ and $\operatorname{span}\{\Delta_t^* w_t^*\}$ correspond to the green and red "vectors" in Figure 2, respectively, and thus we expect FROZENREP to learn E_k.

Lemma C.1 (FROZENREP learns incorrect space)**.** *Fix an orthogonal matrix $B^* \in \mathbb{R}^{d \times k}$ with $\operatorname{Col} B^* \subseteq E^*$, and assume that we sample task weights from the distribution in (10). Then, with infinitely many tasks and per-task samples ($n_{\mathrm{S}}, T \to \infty$), $\operatorname{Col} \hat{B} = E_k$.*

Although "incorrect", it is unclear *a priori* that learning $\operatorname{Col} E_k$ is undesirable performance-wise. The next result formalizes the resulting degradation in performance due to learning $\operatorname{Col} E_k$.

Theorem C.1 (FROZENREP minimax bound)**.** *For an orthogonal matrix $B^* \in \mathbb{R}^{d \times k}$ whose column space lies in E^*, let S_{B^*} be the set*

$$S_{B^*} \coloneqq \left\{ \frac{1}{\sqrt{2\varepsilon}} B^* w + \delta \;\middle|\; \|w\|_2 \le 1, \|\delta\|_2 \le 1, \delta \in E_k \right\}.$$

We consider the following procedure:

1. *We draw n_{T} samples for target time training, which are collected into a matrix X.*
2. *Player chooses a target-time estimator $\hat{\theta} = \bar{B}\hat{w} + \hat{\delta}$, where $\bar{B}$, $\hat{w}$, and $\hat{\delta}$ are measurable function of (X, y), and $\operatorname{Col} \bar{B} = \operatorname{Col} \hat{B}$.*
3. *Player chooses target-time estimator $\hat{\theta} = \hat{B}\hat{w} + \hat{\delta}$, where $\hat{w}$ and $\hat{\delta}$ are measurable functions of $(\hat{B}, X, y)$.*

4. *Adversary chooses an orthogonal matrix $B^* \in \mathbb{R}^{d\times k}$ satisfying $\operatorname{Col} B^* \subset E_k$, and a target time predictor $\theta^* \in S_{B^*}$.*

5. FROZENREP *returns a representation $\hat{B}$ under the setting of Lemma C.1 with the task distribution determined by B^*.*

6. *Target time samples are generated using $y \sim \mathcal{N}\left(X\theta^*, \sigma^2 I_n\right)$, and the player estimator is evaluated.*

Then, with probability at least $1-\delta$ over the draw of X, we have that

$$\min_{\hat{w},\hat{\delta}} \max_{\substack{B^* \\ \theta^* \in S_{B^*}}} \mathbb{E}\left[\frac{1}{n_T}\left\|X(\theta^* - \hat{B}\hat{w} - \hat{\delta})\right\|_2^2\right] \gtrsim \frac{\sigma^2 d}{n_T},$$

where the expectation is over the randomness in the labels $y \sim \mathcal{N}\left(X\theta^*, \sigma^2 I_{n_\mathrm{T}}\right)$. Note that the minimization over $\hat{w}$ and $\hat{\delta}$ is performed over the set of measurable functions from (X, y) to $\mathbb{R}^k$ and $\mathbb{R}^d$, respectively.

The result above comprises the minimax bound in Theorem 3.2. In contrast, by specializing the ADAPTREP performance bound in Theorem 3.1, we obtain the following corollary:

Corollary C.1. *Set $k = \Theta(1)$, $d \gg k$, and $\varepsilon = k/d$. Furthermore, assume that $n_\mathrm{S} T \gtrsim d^2$, and $n_\mathrm{S} \geq n_\mathrm{T} \asymp d$. Then, for a fixed target task in S_{B^*} as defined in Theorem C.1, with probability at least $1-\delta$ over the draw of samples, the procedure in Section 3.2 achieves excess risk bounded as*

$$\mathbb{E}\left[(x^\top\theta^* - x^\top\hat{\theta})^2\right] \lesssim \min\left(\sigma/\sqrt{n_\mathrm{T}},\ \sigma^2 d/n_\mathrm{T}\right).$$

Thus, we have constructed the desired task distribution family, proving Theorem 3.2.

C.2 Proofs for Section 3.4

Lemma C.1 (FROZENREP learns incorrect space)**.** *Fix an orthogonal matrix $B^* \in \mathbb{R}^{d\times k}$ with $\operatorname{Col} B^* \subseteq E^*$, and assume that we sample task weights from the distribution in (10). Then, with infinitely many tasks and per-task samples $(n_\mathrm{S}, T \to \infty)$, $\operatorname{Col}\hat{B} = E_k$.*

Proof. Fix a B. For any $t \in [T]$, as $n_\mathrm{S} \to \infty$, we have that

$$\begin{aligned}
\min_{w_t} \frac{1}{n_\mathrm{S}}\left\|y_t - X_t B w_t\right\|_2^2 &= \min_{w_t} \frac{1}{n_\mathrm{S}}\left\|X_t(\theta_t^* - Bw_t) + z_t\right\|_2 \\
&= \min_{w_t}\left\|\Sigma^{1/2}(\theta_t^* - Bw_t)\right\|_2^2 + \sigma^2 \\
&= \left\|P^\perp_{\Sigma^{1/2}B}\Sigma^{1/2}\theta_t^*\right\|_2^2 + \sigma^2.
\end{aligned}$$

Therefore, we can rewrite the objective defining $\hat{B}$ as

$$\begin{aligned}
\hat{B} &= \operatorname*{argmin}_B \frac{1}{T}\sum_{t\in[T]}\left\|P^\perp_{\Sigma^{1/2}B}\Sigma^{1/2}\theta_t^*\right\|_2^2 \\
&= \operatorname*{argmin}_B \operatorname{tr} P^\perp_{\Sigma^{1/2}B}\Sigma^{1/2}\left[\frac{1}{T}\sum_{t\in[T]}\theta_t^*(\theta_t^*)^\top\right]\Sigma^{1/2} \\
&= \operatorname*{argmin}_B \operatorname{tr} P^\perp_{\Sigma^{1/2}B}\left[\frac{1}{2}P_{B^*} + P_{E_k}\right] \\
&= \operatorname*{argmin}_B \frac{1}{2}\left\|P^\perp_{\Sigma^{1/2}B}P_{B^*}\right\|_F^2 + \left\|P^\perp_{\Sigma^{1/2}B}P_{E_k}\right\|_F^2,
\end{aligned}$$

where the third equality uses the definition of the task distribution in (10). Now, let S be the space defined as $E_k + \operatorname{Col} B^*$. Observe that

$$\begin{aligned}\left\|P^{\perp}_{\Sigma^{1/2}B}P_S\right\|_F^2 &= \operatorname{tr} P^{\perp}_{\Sigma^{1/2}B}P_S \\ &= \operatorname{tr} P^{\perp}_{\Sigma^{1/2}B}P^*_B + \operatorname{tr} P^{\perp}_{\Sigma^{1/2}B}P_{E_k} \\ &= \left\|P^{\perp}_{\Sigma^{1/2}B}P_{B^*}\right\|_F^2 + \left\|P^{\perp}_{\Sigma^{1/2}B}P_{E_k}\right\|_F^2.\end{aligned}$$

By putting everything together, we thus see that

$$\hat{B} = \operatorname*{argmin}_B \left\|P^{\perp}_{\Sigma^{1/2}B}P_S\right\|_F^2 + \left\|P^{\perp}_{\Sigma^{1/2}B}P_{E_k}\right\|_F^2.$$

This objective is minimized if and only if the span of $\Sigma^{1/2}B$ is exactly E_k. By inverting, this is only possible if $\operatorname{Col} \hat{B} = E_k$. □

Theorem C.1 (FROZENREP minimax bound)**.** *For an orthogonal matrix $B^* \in \mathbb{R}^{d\times k}$ whose column space lies in E^*, let S_{B^*} be the set*

$$S_{B^*} \coloneqq \left\{ \frac{1}{\sqrt{2\varepsilon}} B^* w + \delta \;\middle|\; \|w\|_2 \le 1, \|\delta\|_2 \le 1, \delta \in E_k \right\}.$$

We consider the following procedure:

1. *We draw n_{T} samples for target time training, which are collected into a matrix X.*
2. *Player chooses a target-time estimator $\hat{\theta} = \bar{B}\hat{w} + \hat{\delta}$, where $\bar{B}$, $\hat{w}$, and $\hat{\delta}$ are measurable function of (X, y), and $\operatorname{Col}\bar{B} = \operatorname{Col}\hat{B}$.*
3. *Player chooses target-time estimator $\hat{\theta} = \hat{B}\hat{w}+\hat{\delta}$, where $\hat{w}$ and $\hat{\delta}$ are measurable functions of $(\hat{B}, X, y)$.*
4. *Adversary chooses an orthogonal matrix $B^* \in \mathbb{R}^{d\times k}$ satisfying $\operatorname{Col} B^* \subset E_k$, and a target time predictor $\theta^* \in S_{B^*}$.*
5. FROZENREP *returns a representation $\hat{B}$ under the setting of Lemma C.1 with the task distribution determined by B^*.*
6. *Target time samples are generated using $y \sim \mathcal{N}\left(X\theta^*, \sigma^2 I_n\right)$, and the player estimator is evaluated.*

Then, with probability at least $1-\delta$ over the draw of X, we have that

$$\min_{\hat{w},\hat{\delta}} \max_{\substack{B^* \\ \theta^* \in S_{B^*}}} \mathbb{E}\left[\frac{1}{n_T}\left\|X(\theta^* - \hat{B}\hat{w} - \hat{\delta})\right\|_2^2\right] \gtrsim \frac{\sigma^2 d}{n_T},$$

where the expectation is over the randomness in the labels $y \sim \mathcal{N}\left(X\theta^*, \sigma^2 I_{n_{\mathrm{T}}}\right)$. Note that the minimization over $\hat{w}$ and $\hat{\delta}$ is performed over the set of measurable functions from (X, y) to $\mathbb{R}^k$ and $\mathbb{R}^d$, respectively.

Proof. We instantiate the event guaranteed by Lemma B.2 over the draw of target samples, which guarantees that

$$0.9\Sigma \lesssim \frac{1}{n_T} X^\top X \lesssim 1.1\Sigma,$$

with probability at least $1-\delta$.

Recall that by Lemma C.1, FROZENREP will always find an orthogonal matrix whose column space is E_k, no matter how B^* is chosen. Therefore, we can rewrite the minimax expression as

$$\begin{aligned}\min_{\hat{w},\hat{\delta}} \max_{\substack{B^*,\theta^* \\ \theta^* \in S_{B^*}}} \mathbb{E}\left[\frac{1}{n_T}\left\|X(\theta^* - \hat{B}\hat{w} - \hat{\delta})\right\|_2^2\right] &= \min_{\hat{\theta}} \max_{\substack{B^*,\theta^* \\ \theta^* \in S_{B^*}}} \mathbb{E}\left[\frac{1}{n_T}\left\|X(\theta^* - \hat{\theta})\right\|_2^2\right] \\ &\gtrsim \min_{\hat{\theta}} \max_{\substack{B^*,\theta^* \\ \theta^* \in S_{B^*}}} \mathbb{E}\left[\left\|\theta^* - \hat{\theta}\right\|_\Sigma^2\right],\end{aligned}$$

where $\hat{\theta}$ is simply a measurable function of (X, y) to $\mathbb{R}^d$. To further simplify the problem, observe that if we define the set

$$T = \left\{ \theta \in \mathbb{R}^d \;\middle|\; \|P_{E^*}\theta\|_2^2 \le \frac{1}{2\varepsilon}, \|P_{E_k}\theta\|_2^2 \le 1 \right\} \subseteq S_{B^*}$$

then since the estimator only depends on B^* through θ^*,

$$\min_{\hat{w},\hat{\delta}} \max_{\substack{B^*,\theta^* \\ \theta^* \in S_{B^*}}} \mathbb{E}\left[\frac{1}{n_T} \left\| X(\theta^* - \hat{B}\hat{w} - \hat{\delta}) \right\|_2^2 \right] \gtrsim \min_{\hat{\theta}} \max_{\theta^* \in T} \mathbb{E}\left[\left\| (\theta^* - \hat{\theta}) \right\|_\Sigma^2 \right]. \tag{11}$$

To lower bound the right-hand side of (11), we use local coverings in the Σ-norm and apply the Fano bound for minimax risk. Let $B := \left\{ \theta \in \mathbb{R}^d \;\middle|\; \varepsilon \|P_{E^*}\theta\|_2^2 + \|P_{E_k}\theta\|_2^2 \le 1 \right\}$ be the unit ball in the Σ-norm, so that $\frac{1}{\sqrt{2}}B \subseteq T$. Using a known volumetric argument, there exists a $(1/2)$-packing of B in the Σ-norm with at least 2^d elements. Equivalently, there exists a 2δ-packing of $4\delta B$ with at least 2^d elements – let this packing be S. Then, for any $\theta, \theta' \in S$,

$$\begin{aligned} \mathrm{KL}\left(\mathcal{N}\left(X\theta, \sigma^2 I_{n_T}\right) \middle\| \mathcal{N}\left(X\theta', \sigma^2 I_{n_T}\right) \right) &= \frac{1}{2\sigma^2} \|X(\theta - \theta')\|_2^2 \lesssim \frac{n_T}{2\sigma^2} \|\theta - \theta'\|_\Sigma^2 \\ &\le \frac{32 n_T}{\sigma^2} \delta^2. \end{aligned}$$

Therefore, by Fano's inequality, for any $\delta^2 \le 1/32$ (which ensures that $S \subseteq 4\delta B \subseteq \frac{1}{\sqrt{2}}B \subseteq T$),

$$\begin{aligned} \min_{\hat{\theta}} \max_{\theta^* \in T} \mathbb{E}\left[\frac{1}{n} \left\| (\theta^* - \hat{\theta}) \right\|_\Sigma^2 \right] &\ge \delta^2 \left(1 - \frac{32n}{\sigma^2 d \log 2} \delta^2 - \frac{1}{d} \right) \\ &\ge \delta^2 \left(\frac{3}{4} - \frac{32n}{\sigma^2 d \log 2} \delta^2 \right). \qquad (d \ge 4) \end{aligned}$$

As long as $n_T \ge \frac{\log 2}{2} \sigma^2 d$, we can set $\delta^2 = \left(\frac{\log 2}{64} \right) \frac{\sigma^2 d}{n_T}$, and thus putting everything together, we have that

$$\min_{\hat{w},\hat{\delta}} \max_{\substack{B^*,\theta^* \\ \theta^* \in S_{B^*}}} \mathbb{E}\left[\frac{1}{n} \left\| X(\theta^* - \hat{B}\hat{w} - \hat{\delta}) \right\|_2^2 \right] \gtrsim \frac{\sigma^2 d}{n_T}. \qquad \square$$

C.3 Simulations

In this section, we experimentally verify the linear hard case presented in Section 3.4. Since the empirical success of MAML and its variants has already been demonstrated extensively in practice and in existing work, it is not the focus of this section.

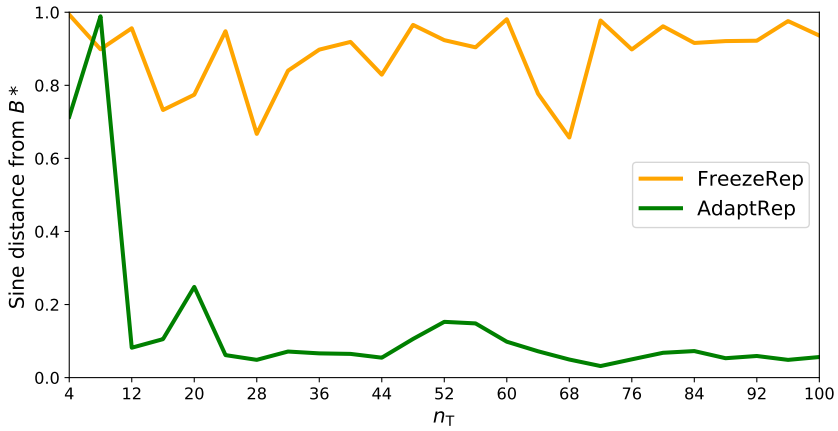

Figure 6: Sine distances of the representations learned by each method from the correct space B^*.

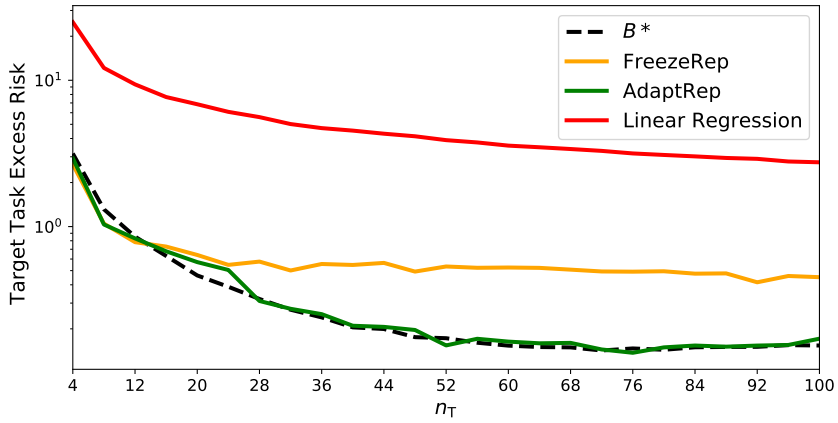

Figure 7: The worst-case average excess risk for several settings of n_{T}, plotted on a logarithmic y-axis.

We consider the high-dimensional setting where $d \asymp n_{\mathrm{T}}$, and $k = \Theta(1)$. We fix $k = 2$ throughout the experiment. The matrix B^* spans the first k coordinates, while the residuals δ_t^* lie in the span of the last k coordinates. Finally, we set the Gaussian noise variance to $\sigma^2 = 1$.

During source training, both FROZENREP and ADAPTREP are provided with 1000 tasks and $10d$ samples per task from the task distribution in Section 3.4 (both are provided with the same samples). During target time, we evaluate the methods on the worst-case regression task from the same family.

Nonconvexity in Source Procedure. Rather than optimizing (4) or (6) during source training, we use an additional Frobenius-norm regularizer on $B^\top B - WW^\top$ to balance B and W. For FROZENREP, the regularized objective was shown to have a favorable optimization landscape in Tripuraneni et al. (2020a). We used L-BFGS to optimize these regularized objectives. To further mitigate optimization issues, we evaluated both methods with 10 random restarts, and report the best of the 10 restarts (as measured by the worst-case performance on the target task) for both methods.

(Subspace Alignment). We plotted the alignment of the learned representation (using the best of the 10 restarts) with B^*. We measure this via the sine of the largest principal angle between the two spaces, i.e.

$$\sin\Theta_1(\hat{B}, B^*) = \sqrt{1 - \sigma_1^2\left(\hat{B}^\top B^*\right)}.$$

We plot the results in Figure 6. As predicted by Lemma C.1, FROZENREP does not learn B^*, in contrast to ADAPTREP.

(Target Task Performance). We evaluated how the methods fare on their corresponding worst-case target tasks. We do so by training with the representation over 1000 i.i.d. draws of the target dataset, and averaging the excess risk over all obtained representations. We provide the results in Figure 7, and include a comparison with standard linear regression. As predicted, ADAPTREP performs much better than FROZENREP, with a gap that grows with n_{T}.

D Proof of Theorem 4.1

In this section, we will prove the guarantee provided in Theorem 4.1, along with all related intermediate results. Along these lines, we proceed as we did for Section B. More explicitly, the overall outline of the proof follows the four major steps below. We have placed in parenthesis the corresponding intermediate steps in the linear representation case (Section B) as an additional illustration of the procedure:

(1) Provide a statistical rate for source training (Lemma B.3).

(2) Bound the difference in performance between the solution found by the optimization algorithm and the ERM solution (Lemma B.6+Theorem H.1).

(3) Prove uniform concentration over $\mathcal{A}_{\mathcal{C}_{\mathrm{T}}}(\theta_0)$ as a function of target sample size (Lemma B.5).

(4) Connect the best-case performance in $\mathcal{A}_{\mathcal{C}_{\mathrm{T}}}(\theta_0)$ to the performance of the learner on the source tasks (Lemmas B.4, B.5 and B.8).

Note that step (4) is provided by the (ν, ε)-diversity condition. We will now proceed to demonstrate the remaining steps.

D.1 Bounding Average Source Task Performance (1)

The (ν, ε)-diversity condition implies that we can bound the best-case performance during target time training via control over the average source task performance. We proceed to provide such a bound using a standard uniform convergence argument. Recall that for a set of vector-valued functions $\mathcal{H}$ mapping from $\mathbb{R}^m \to \mathbb{R}^n$, the Rademacher complexity of $\mathcal{H}$ on n_{S} samples, denoted $\mathcal{R}_{n_{\mathrm{S}}}(\mathcal{H})$, is given by

$$\mathcal{R}_{n_{\mathrm{S}}}(\mathcal{H}) \coloneqq \mathbb{E}\left[\sup_{h \in \mathcal{H}} \frac{1}{n_{\mathrm{S}}} \sum_{i \in [n_{\mathrm{S}}]} \sum_{j \in [n]} \varepsilon_{ij} h(x_i)_j\right],$$

where the expectation is over the samples (x_i) and i.i.d. Rademacher random variables ε_{ij}.

Lemma D.1 (Source Training Bound)**.** *Let $\theta \in \Theta_0$ be a minimizer of the training objective in (7). Then, with probability at least $1 - \delta$ over the random draw of inputs and labels,*

$$\frac{1}{T} \sum_{t \in [T]} \inf_{g_t \in \mathcal{A}_{\mathcal{C}_{\mathrm{S}}}(\theta)} \mathcal{L}^{\mathrm{ex}}_{\infty}(g_t, g^*_t) \lesssim \frac{1}{\delta T} \mathcal{R}_{n_{\mathrm{S}}}\left[\bigcup_{\theta \in \Theta_0} \left[\mathcal{A}_{\mathcal{C}_{\mathrm{S}}}(\theta)\right]^{\otimes T}\right] + \frac{B}{\delta \sqrt{n_{\mathrm{S}} T}}.$$

Proof. For any $t \in [T]$, we define the empirical risk minimizer $\bar{g}_t \coloneqq \operatorname{argmin}_{g \in \mathcal{A}_{\mathcal{C}_{\mathrm{S}}}(\theta)} \mathcal{L}(g, g^*_t)$. Then, since θ minimizes the training objective in (7),

$$\sum_{t \in [T]} \mathcal{L}(\bar{g}_t, g^*_t) - \mathcal{L}(g^*_t, g^*_t) \leq 0.$$

Following the canonical risk decomposition, we have that

$$\begin{aligned}
\sum_{t \in [T]} \mathcal{L}^{\mathrm{ex}}_{\infty}(\bar{g}_t, g^*_t) &\leq \sum_{t \in [T]} \mathcal{L}_{\infty}(\bar{g}_t, g^*_t) - \mathcal{L}(\bar{g}_t, g^*_t) + \sum_{t \in [T]} \mathcal{L}(\bar{g}_t, g^*_t) - \mathcal{L}(g^*_t, g^*_t) \\
&\quad + \sum_{t \in [T]} \mathcal{L}(g^*_t, g^*_t) - \mathcal{L}_{\infty}(g^*_t, g^*_t) \\
&\leq 2 \sup_{\substack{\theta \in \Theta_0 \\ g_t \in \mathcal{A}_{\mathcal{C}_{\mathrm{S}}}(\theta)}} \left| \sum_{t \in [T]} \mathcal{L}_{\infty}(g_t, g^*_t) - \mathcal{L}(g_t, g^*_t) \right|.
\end{aligned}$$

We give a high-probability bound on the right-hand side by first bounding its expectation, and then applying Markov's inequality to obtain the desired result. To bound the expectation, note that via

symmetrization,

$$\mathbb{E}\left[\sup_{\substack{\theta\in\Theta_0\\ g_t\in\mathcal{A}_{\mathcal{C}_\mathrm{S}}(\theta)}}\left|\sum_{t\in[T]}\mathcal{L}_\infty(g_t,g_t^*)-\mathcal{L}(g_t,g_t^*)\right|\right]$$
$$\leq 2\mathbb{E}\left[\sup_{\substack{\theta\in\Theta_0\\ g_t\in\mathcal{A}_{\mathcal{C}_\mathrm{S}}(\theta)}}\left|\frac{1}{n_\mathrm{S}}\sum_{t\in[T]}\sum_{i\in[n_\mathrm{S}]}\varepsilon_{ij}\ell(g_t(x_{i,t}),y_{i,t})\right|\right]$$

Recentering about $\ell(0,y_{i,t})$ and using the constant-shift property of the Rademacher complexity,

$$\mathbb{E}\left[\sup_{\substack{\theta\in\Theta_0\\ g_t\in\mathcal{A}_{\mathcal{C}_\mathrm{S}}(\theta)}}\left|\frac{1}{n_\mathrm{S}}\sum_{t\in[T]}\sum_{i\in[n_\mathrm{S}]}\varepsilon_{ij}\ell(g_t(x_{i,t}),y_{i,t})\right|\right]$$
$$\leq \mathbb{E}\left[\sup_{\substack{\theta\in\Theta_0\\ g_t\in\mathcal{A}_{\mathcal{C}_\mathrm{S}}(\theta)}}\left|\frac{1}{n_\mathrm{S}}\sum_{t\in[T]}\sum_{i\in[n_\mathrm{S}]}\varepsilon_{ij}[\ell(g_t(x_{i,t}),y_{i,t})-\ell(0,y_{i,t})]\right|\right]+B\sqrt{\frac{T}{n_\mathrm{S}}}$$

Finally, since the loss is 1-Lipschitz, we can apply the Rademacher contraction principle, from which we find that

$$\mathbb{E}\left[\sup_{\substack{\theta\in\Theta_0\\ g_t\in\mathcal{A}_{\mathcal{C}_\mathrm{S}}(\theta)}}\left|\frac{1}{n_\mathrm{S}}\sum_{t\in[T]}\sum_{i\in[n_\mathrm{S}]}\varepsilon_{ij}[\ell(g_t(x_{i,t}),y_{i,t})-\ell(0,y_{i,t})]\right|\right]$$
$$\lesssim \mathbb{E}\left[\sup_{\substack{\theta\in\Theta_0\\ g_t\in\mathcal{A}_{\mathcal{C}_\mathrm{S}}(\theta)}}\left|\frac{1}{n_\mathrm{S}}\sum_{t\in[T]}\sum_{i\in[n_\mathrm{S}]}\varepsilon_{ij}g_t(x_{i,t})\right|\right]=\mathcal{R}_{n_\mathrm{S}}\left[\bigcup_{\theta\in\Theta_0}\left[\mathcal{A}_{\mathcal{C}_\mathrm{S}}(\theta)\right]^{\otimes T}\right].$$

Putting everything together, we can thus bound the expectation of the maximum deviation as

$$\mathbb{E}\left[\sup_{\substack{\theta\in\Theta_0\\ g_t\in\mathcal{A}_{\mathcal{C}_\mathrm{S}}(\theta)}}\left|\sum_{t\in[T]}\mathcal{L}_\infty(g_t,g_t^*)-\mathcal{L}(g_t,g_t^*)\right|\right]\lesssim\mathcal{R}_{n_\mathrm{S}}\left[\bigcup_{\theta\in\Theta_0}\left[\mathcal{A}_{\mathcal{C}_\mathrm{S}}(\theta)\right]^{\otimes T}\right]+B\sqrt{\frac{T}{n_\mathrm{S}}}.$$

Therefore, by applying Markov's inequality, we have that with probability at least $1-\delta$,

$$\frac{1}{T}\sum_{t\in[T]}\mathcal{L}_\infty^{\mathrm{ex}}(\bar{g}_t,g_t^*)\lesssim\frac{1}{T}\sup_{\substack{\theta\in\Theta_0\\ g_t\in\mathcal{A}_{\mathcal{C}_\mathrm{S}}(\theta)}}\left|\sum_{t\in[T]}\mathcal{L}_\infty(g_t,g_t^*)-\mathcal{L}(g_t,g_t^*)\right|$$
$$\lesssim\frac{1}{\delta T}\mathcal{R}_{n_\mathrm{S}}\left[\bigcup_{\theta\in\Theta_0}\left[\mathcal{A}_{\mathcal{C}_\mathrm{S}}(\theta)\right]^{\otimes T}\right]+\frac{B}{\delta\sqrt{n_\mathrm{S}T}}.\qquad\square$$

D.2 Bounding Optimization Performance (2)

Having provided a bound on source task performance, we now proceed to analyze the objective being optimized during target time training. We first note that the approximate linearity assumption in Assumption 4.4 and the norm-boundedness of $\mathcal{C}_\mathrm{T}$ via Assumption 4.5 ensure that the results from Section H apply, as long as we can show that the empirical loss satisfies an $(\alpha/\sqrt{n_\mathrm{T}})$-Lipschitz condition. We proceed to show that this is indeed a simple consequence of the 1-Lipschitz assumption on the loss given by Assumption 4.2.

Lemma D.2. *Define the function* $\mathcal{L}:\mathbb{R}^{n_\mathrm{T}}\to\mathbb{R}$ *as*

$$\mathcal{L}(\hat{y})\coloneqq\frac{1}{n_\mathrm{T}}\sum_{i\in[n_\mathrm{T}]}\ell(\hat{y}_i,y_i).$$

for fixed $y_1,\ldots,y_{n_\mathrm{T}}\in\mathcal{Y}$*. Then, for any* $\hat{y}$*,* $\|\nabla_{\hat{y}}\mathcal{L}(\hat{y})\|_2^2\leq 1/n_\mathrm{T}$*.*

Proof. By direct computation,

$$\|\nabla_{\hat{y}}\mathcal{L}(\hat{y})\|_2^2 = \frac{1}{n_{\mathrm{T}}^2}\sum_{i\in n_{\mathrm{T}}}|\nabla_{\hat{y}_i}\ell(\hat{y}_i, y)|^2 \leq \frac{1}{n_{\mathrm{T}}},$$

where we have used the fact that $\ell(\cdot, y)$ is 1-Lipschitz for any $y \in \mathcal{Y}$ by Assumption 4.2. □

Having verified that the assumptions in Section H hold, it follows that we have the following performance bound on projected gradient descent during target time training:

Lemma D.3. *Assume that we run projected gradient descent during target training time for T_{PGD} iterations with step size η given by*

$$\eta = \frac{1}{\sqrt{T_{\mathrm{PGD}}}}\left(\frac{R}{\sqrt{L^2+\beta^2R^2}}\right).$$

Let $g_0, \ldots, g_{T_{\mathrm{PGD}}}$ denote the sequence of predictors obtained, where $g_0 = g_{\theta_0}$. Then, for any $g \in \mathcal{A}_{\mathcal{C}_{\mathrm{T}}}(\theta_0)$,

$$\min_t \mathcal{L}(g_t, g_t^*) - \mathcal{L}(g, g_t^*) \leq \beta R^2 + R\sqrt{\frac{L^2+\beta^2R^2}{T_{\mathrm{PGD}}}}.$$

D.3 Bounding the Performance of ERM during Target Training (3)

We now proceed to bound the performance of the ERM solution during target time training. Via following the canonical risk decomposition as in Lemma D.1, we can prove such a bound simply by bounding the maximum deviation between empirical and population losses over $\mathcal{A}_{\mathcal{C}_{\mathrm{T}}}(\theta_0)$.

Lemma D.4. *Let $\mathcal{S}$ be the support of ρ. With probability at least $1-\delta$ over the random draw of inputs and noise,*

$$\sup_{\substack{g^*\in\mathcal{S}\\ g\in\mathcal{A}_{\mathcal{C}_{\mathrm{T}}}(\theta_0)}} |\mathcal{L}(g, g^*) - \mathcal{L}_\infty(g, g^*)| \leq \frac{1}{\delta}\sup_{\theta\in\Theta_0}\mathcal{R}_{n_{\mathrm{T}}}\left[\mathcal{A}_{\mathcal{C}_{\mathrm{T}}}(\theta)\right] + \frac{B}{\delta\sqrt{n_{\mathrm{T}}}}.$$

Proof. The proof proceeds similarly to that of Lemma D.1. Note that the supremum over g^* does not affect the bound, since g^* only enters into the expression through the labels y_i, and no matter what choice of g^* is made, $|\ell(0, y_i)| \leq B$ for all i. The final result follows from taking a supremum over all possible initialization choices. □

D.4 Concluding: Proving Theorem 4.1

Having completed all the steps for the outline, we now proceed to compile the main result. Intuitively, since the diversity condition allows us to bound the performance of the best predictor, and projected gradient descent can perform as well as any predictor in $\mathcal{A}_{\mathcal{C}_{\mathrm{T}}}(\theta_0)$ (and thus, as well as the best predictor), we can obtain performance bounds on the iterates found while training on the target task.

Theorem 4.1 (General Performance Bound)**.** *Assume that all assumptions in Section 4.2 hold. Let (θ_t) be the set of iterates generated by PGD following the procedure in Section 4.1. Then, with probability at least $1-\delta$ over the random draw of samples,*

$$\mathbb{E}_{g^*\sim\rho}\left[\min_t \mathcal{L}_\infty^{\mathrm{ex}}(g_{\theta_t}, g^*)\right] \lesssim \underbrace{\beta R^2 + R\sqrt{\frac{L^2+\beta^2R^2}{T_{\mathrm{PGD}}}}}_{\varepsilon_{\mathrm{OPT}}} + \underbrace{\frac{1}{\delta}\sup_{\theta\in\Theta_0}\mathcal{R}_{n_{\mathrm{T}}}\left[\mathcal{A}_{\mathcal{C}_{\mathrm{T}}}(\theta)\right] + \frac{B}{\delta\sqrt{n_{\mathrm{T}}}}}_{\varepsilon_{\mathrm{EST}}}$$
$$+ \underbrace{\frac{1}{\nu}\left\{\frac{1}{\delta T}\mathcal{R}_{n_{\mathrm{S}}}\left[\bigcup_{\theta\in\Theta_0}\left[\mathcal{A}_{\mathcal{C}_{\mathrm{S}}}(\theta)\right]^{\otimes T}\right] + \frac{B}{\delta\sqrt{n_{\mathrm{S}}T}}\right\} + \varepsilon}_{\varepsilon_{\mathrm{REPR}}}.$$

Note that the $\mathcal{R}_{n_{\mathrm{T}}}$ complexity term samples from p. Meanwhile, the $\mathcal{R}_{n_{\mathrm{S}}}$ complexity term samples from $p^{\otimes T}$, which concatenates T i.i.d. samples from p (one for each task) for every draw.

Proof. We instantiate the high-probability events in Lemmas D.1 and D.4 both with failure probability $\delta/2$, which simultaneously occur with probability at least $1-\delta$ via a union-bound. Consequently, we have the bounds

$$\frac{1}{T}\sum_{t\in[T]}\inf_{g_t\in\mathcal{A}_{\mathcal{C}_\mathrm{S}}(\theta)}\mathcal{L}_\infty^{\mathrm{ex}}(g_t,g_t^*)\lesssim\frac{1}{\delta T}\mathcal{R}_{n_\mathrm{S}}\left[\bigcup_{\theta\in\Theta_0}\left[\mathcal{A}_{\mathcal{C}_\mathrm{S}}(\theta)\right]^{\otimes T}\right]+\frac{B}{\delta\sqrt{n_\mathrm{S}T}}$$

$$\sup_{\substack{g^*\in\mathcal{S}\\ g\in\mathcal{A}_{\mathcal{C}_\mathrm{T}}(\theta_0)}}\left|\mathcal{L}(g,g^*)-\mathcal{L}_\infty(g,g^*)\right|\lesssim\frac{1}{\delta}\sup_{\theta\in\Theta_0}\mathcal{R}_{n_\mathrm{T}}\left[\mathcal{A}_{\mathcal{C}_\mathrm{T}}(\theta)\right]+\frac{B}{\delta\sqrt{n_\mathrm{T}}}$$

Given g^*, let $\bar{g}$ be the population risk minimizer in $\mathcal{A}_{\mathcal{C}_\mathrm{T}}(\theta_0)$. Througout this proof, we omit g^* when writing $\mathcal{L}$ and $\mathcal{L}_\infty$ whenever it is understood. Then, by Lemma D.3, projected gradient descent always generates iterates satisfying

$$\min_t\mathcal{L}(g_t)-\mathcal{L}(\bar{g})\lesssim\beta R^2+R\sqrt{\frac{L^2+\beta^2R^2}{T_{\mathrm{PGD}}}},\tag{12}$$

independent of g^*. Let $\tilde{g}$ be the predictor achieving the minimum. We thus proceed to decompose the risk as

$$\begin{aligned}\min_t\mathcal{L}_\infty(g_t)-\mathcal{L}_\infty(g^*)\lesssim&\underbrace{\mathcal{L}_\infty(\tilde{g})-\mathcal{L}(\tilde{g})}_{=:T_1}+\underbrace{\mathcal{L}(\tilde{g})-\mathcal{L}(\bar{g})}_{=:T_2}\\&+\underbrace{\mathcal{L}(\bar{g})-\mathcal{L}_\infty(\bar{g})}_{=:T_3}+\underbrace{\mathcal{L}_\infty(\bar{g})-\mathcal{L}_\infty(g^*)}_{=:T_4}.\end{aligned}$$

We now bound the terms $T_1,\ldots,T_4$ individually. As a result of uniform convergence as guaranteed by Lemma D.4,

$$T_1,T_3\leq\sup_{\substack{g^*\in\mathcal{S}\\ g\in\mathcal{A}_{\mathcal{C}_\mathrm{T}}(\theta_0)}}\left|\mathcal{L}(g)-\mathcal{L}_\infty(g)\right|\lesssim\frac{1}{\delta}\mathcal{R}_{n_\mathrm{T}}\left[\mathcal{A}_{\mathcal{C}_\mathrm{T}}(\theta_0)\right]+\frac{B}{\delta\sqrt{n_\mathrm{T}}},$$

where $\mathcal{S}$ is the support of the target task distribution ρ. Furthermore, we have demonstrated a bound on T_2 in (12). Finally, by applying the (ν,ε)-diversity condition as guaranteed by Assumption 4.3,

$$\begin{aligned}\mathbb{E}_{g^*\sim\rho}\left[T_4\right]&=\mathbb{E}_{g^*\sim\rho}\left[\inf_{g\in\mathcal{A}_{\mathcal{C}_\mathrm{T}}(\theta_0)}\mathcal{L}_\infty^{\mathrm{ex}}(g)\right]\\&\leq\frac{1}{\nu}\left[\frac{1}{T}\sum_{t\in[T]}\inf_{g_t\in\mathcal{A}_{\mathcal{C}_\mathrm{S}}(\theta)}\mathcal{L}_\infty^{\mathrm{ex}}(g_t)\right]+\varepsilon\\&\leq\frac{1}{\nu}\left\{\frac{1}{\delta T}\mathcal{R}_{n_\mathrm{S}}\left[\bigcup_{\theta\in\Theta_0}\left[\mathcal{A}_{\mathcal{C}_\mathrm{S}}(\theta)\right]^{\otimes T}\right]+\frac{B}{\delta\sqrt{n_\mathrm{S}T}}\right\}+\varepsilon.\end{aligned}$$

Taking the expectation with respect to $g^*\sim\rho$ and putting all of these bounds together, we thus have that

$$\begin{aligned}\mathbb{E}_{g^*\sim\rho}\left[\mathcal{L}_\infty^{\mathrm{ex}}(g,g^*)\right]\lesssim&\,\beta R^2+R\sqrt{\frac{L^2+\beta^2R^2}{T_{\mathrm{PGD}}}}+\frac{1}{\delta}\sup_{\theta\in\Theta_0}\mathcal{R}_{n_\mathrm{T}}\left[\mathcal{A}_{\mathcal{C}_\mathrm{T}}(\theta)\right]+\frac{B}{\delta\sqrt{n_\mathrm{T}}}\\&+\frac{1}{\nu}\left\{\frac{1}{\delta T}\mathcal{R}_{n_\mathrm{S}}\left[\bigcup_{\theta\in\Theta_0}\left[\mathcal{A}_{\mathcal{C}_\mathrm{S}}(\theta)\right]^{\otimes T}\right]+\frac{B}{\delta\sqrt{n_\mathrm{S}T}}\right\}+\varepsilon.\end{aligned}$$

□

E Case Study: Two-Layer Neural Networks (Proofs)

E.1 Additional Assumptions on Initialization and Diversity

Before we proceed with the proofs for Section 5, we first formalize the additional assumptions that were referenced in the main text.

Assumption E.1 (Initialization Assumptions)**.** *The initialization (B^*, w_0^*) satisfies the following assumptions:*

(a) $B^* = [A^*, A^*]$ *for* $A^* \in \mathbb{R}^{d\times k}$*. Furthermore,* $w_0^* \in \{-1, 1\}^{2k}$*, with* $(w_0^*)_i = -(w_0^*)_{i+k}$ *for* $i \in [k]$*.*

(b) The columns of B^* *are norm-bounded by* 1*.*

(c) $c_1 I \preceq \mathbb{E}\left[\rho_{\theta_0^*}(x)\rho_{\theta_0^*}(x)^\top\right] \preceq c_2 I$ *for* $c_1 > 0$*. We define* $\kappa = c_2/c_1$*.*

We refer to initializations satisfying (a) as *antisymmetric initializations*. Note that under any antisymmetric initialization θ_0, $f^\beta_{\theta_0}(x) \equiv 0$. Additionally, the assumption on the representation covariance in (c) ensures that that representation is well-conditioned, with condition number κ.

Finally, we need to impose diversity conditions on the source tasks. Recall that for every $t \in [T]$, there exists unit-norm $w_t^* \in \mathbb{R}^{2k}$ and $\delta_t^* \in \mathbb{R}^k$ such that the corresponding source predictor is parameterized by

$$\theta_t^* = \left(B^* + \frac{1}{\beta}\sum_{i\in[k]}(\delta_t^*)_i\Delta_i^*\ ,\ w_0^* + \frac{1}{\beta}w_t^*\right).$$

We note that the fine-tuning step for $t \in [T]$ can be parametrized by $\omega_t^* = [w_t^*, \delta_t^*] \in \mathbb{R}^{3k}$ via a (natural) linear transformation. We assume that this parametrization has a matrix representation Γ/β, so that $\theta_t^* = \theta_0^* + \Gamma\omega_t^*/\beta$. Then, we impose the following condition on $(\omega_t^*)_t$:

Assumption E.2. *Let* $\Omega \in \mathbb{R}^{2k\times T}$ *be the matrix* $[\omega_1^*, \ldots, \omega_T^*]$*. Then,* $\sigma_{2k}(\Omega) \gtrsim T/k$*.*

The assumption above is analogous to the diversity conditions assumed in the previous sections.

E.2 Formalizing Approximate Linearity

In this section, we demonstrate that under Assumption E.1, $f^\beta_{\theta_t^*}$ behaves approximately like its linearization for large enough β.

Recall that the linearization of f^β_θ involves two feature vectors, $\phi_{B_0}(x)$ and $\psi_{B_0,w_0}(x)$. We now provide the following formal definitions for these vectors:

Definition E.1 (Feature vectors ϕ, ψ, ρ)**.** Let $\theta_0 = (B_0, w_0)$ be an antisymmetric parameter, i.e. satisfying Assumption E.1(c). Then, for every x, there exists feature vectors $\phi_{B_0}(x)$ and $\psi_{B_0,w_0}(x)$ such that

$$f^\beta_{\theta_0+(\Delta,w)}(x) = \beta w^\top\phi_{B_0}(x) + \beta\Delta^\top\psi_{B_0,w_0}(x) + \beta\zeta^{\Delta,w}_{B_0,w_0}(x).$$

We interpret the features $\phi_{B_0}(x)$ and $\psi_{B_0,w_0}(x)$ to be the gradients of $(B, w) \mapsto f^\beta_{(B,w)}(x)$ evaluated at θ_0, which have closed forms

$$\phi_{B_0}(x) = \sigma(B_0^\top x) \quad \text{and} \quad \psi_{B_0,w_0}(x) = w_0^{\mathrm{diag}}\sigma'(B_0^\top x)x^\top.$$

Additionally, $\zeta^{\Delta,w}_{B_0,w_0}(x)$ is the Taylor error. By Taylor's theorem, there exists $\bar{B} = B_0 + \alpha_1\Delta$ and $\bar{w} = w_0 + \alpha_2 w$ for some $\alpha_1, \alpha_2 \in [0, 1]$ such that $\zeta^{\Delta,w}_{(B_0,w_0)}(x)$ can be written as

$$\zeta^{\Delta,w}_{(B_0,w_0)}(x) = \sum_{i=1}^{2k}\bar{w}_i\sigma''(\bar{B}_i^\top x)(\Delta_i^\top x)^2 + 2w_i\sigma'(\bar{B}_i^\top x)(\Delta_i^\top x).$$

Finally, we define $\rho_{B_0,w_0}(x)$ to be the concatenation of $\phi_{B_0}(x)$ and $\psi_{B_0,w_0}(x)$. ◇

For $\theta_0 = (B_0, w_0)$ and $\delta = (\Delta, w)$, we will frequently abuse notation and let $\rho_{\theta_0}(x)\delta$ denote the linearization of $f_{\theta_0+\delta}$, *i.e.*

$$\rho_{\theta_0}(x)\delta = w^\top \phi_{B_0}(x) + \langle \psi_{\theta_0}(x), \Delta \rangle .$$

We will show that if $\|\Delta\|_F$ and $\|w\|_F$ are both $O(1/\beta)$, then the remainder term $\zeta^{\Delta,w}_{B_0,w_0}(x)$ is $O(1/\beta)$, and thus the function class is approximately linear in (Δ, w) with feature functions $\phi_{B_0}(\cdot)$ and $\psi_{B_0,w_0}(\cdot)$. We first bound the Hessian term, in order to control the remainder term $\zeta^{\Delta,w}_{(B_0,w_0)}(x)$, which depends on the Hessian.

Lemma E.1 (Hessian Bound)**.** *Fix a matrix $\bar{B} \in \mathbb{R}^{2k\times d}$ and a vector $\bar{w} \in \mathbb{R}^{2k}$. Assume that all rows of $\bar{B}$ are 2-norm-bounded, and that $\|\bar{w}\|_2 \le 2$. Then, for $x \sim p$ almost surely, $\left\|\nabla^2_\theta f^\beta_{(\bar{B},\bar{w})}(x)\right\|_2 \lesssim \beta(\mu + L)$.*

Proof. Throughout the proof, we will use the fact that $\|x\|_2 \le 1$ for $x \sim p$ almost surely. By definition,

$$\left\|\nabla^2_\theta f^\beta_{(\bar{B},\bar{w})}(x)\right\|_2 = \sup_{\|\Delta\|_F^2+\|w\|_2^2\le 1} \beta \left| \sum_{i\in[2k]} \bar{w}_i \sigma''(\bar{B}_i^\top x) \langle \Delta_i, x\rangle^2 + 2 \sum_{i\in[2k]} w_i \sigma'(\bar{B}_i^\top x) \langle \Delta_i, x\rangle \right|$$

To bound the first term, observe that by applying Hölder's inequality,

$$\begin{aligned}\sum_{i\in[2k]} \bar{w}_i \sigma''(\bar{B}_i^\top x) \langle \Delta_i, x\rangle^2 &\le \left(\sup_{i\in[2k]} \left|w_i \sigma''(\bar{B}_i{}^\top x)\right|\right) \sum_{i\in[2k]} \langle \Delta_i, x\rangle^2 \\ &= \left(\sup_{i\in[2k]} \left|\bar{w}_i \sigma''(\bar{B}_i{}^\top x)\right|\right) \|\Delta x\|_2^2 \le 2\mu.\end{aligned}$$

To bound the second term, we apply Cauchy-Schwarz, from which we see that

$$\begin{aligned}2 \sum_{i\in[2k]} w_i \sigma'(\bar{B}_i^\top x) \langle \Delta_i, x\rangle &\le 2 \left(\sum_{i\in 2k} \left[w_i \sigma'(\bar{B}_i{}^\top x)\right]^2\right)^{1/2} \|\Delta x\|_2 \\ &\le 2L \|w\|_2 \|\Delta x\|_2 \le 2L.\end{aligned}$$

Altogether, we thus have that almost surely for $x \sim p$,

$$\left\|\nabla^2_\theta f^\beta_{(\bar{B},\bar{w})}(x)\right\|_2 \lesssim \beta(\mu + L). \qquad \square$$

As an immediate corollary, since $\zeta^\delta_{\theta_0}(x)$ evaluates the Hessian at a point satisfying the preconditions of Lemma E.1 during both source and target trainng time, we have the following result:

Corollary E.1. *For any $\theta_0 \in \Theta_0$ and any (Δ, w) with $\|\Delta\|_F \le c_1/\beta, \|w\|_2 \le c_2/\beta$, we have that $\left|\beta\zeta^{(\Delta,w)}_{\theta_0}(x)\right| \le c_1c_2(\mu + L)/\beta$ for $x \sim p$ almost surely.*

Therefore, we see that for large enough β, the remainder term is close to 0, and thus the family is indeed approximately linear. Finally, we provide a norm bound on the combined representation vector $\rho_{\theta_0}(x)$.

Lemma E.2 (Representation Norm Bound)**.** *For $x \sim p$, $\|\rho_{\theta_0}(x)\|_2 \le 2L\sqrt{k}$ almost surely for any $\theta_0 = (B_0, w_0) \in \Theta_0$.*

Proof. Recall from definitions that $\|\rho_{\theta_0}(x)\|_2^2 = \|\phi_{B_0}(x)\|_2^2 + \|\psi_{\theta_0}(x)\|_F^2$. To bound the activation features, since $\sigma(0) = 0$ and $|\sigma'(x)| \le L$ for $x \in [-1, 1]$,

$$\|\phi_{B_0}(x)\|_2^2 = 2k \max_{i\in[2k]} \left|\sigma(B_{0,i}^\top x)\right|^2 \le 2k \max_{i\in[2k]} \left|\int_0^{B_{0,i}^\top x} \sigma'(z)\, \mathrm{d}z\right|^2 \le 2kL^2.$$

To bound the gradient features, we once again use the boundedness of σ' to obtain

$$\|\psi_{\theta_0}(x)\|_F^2 = \left\|w_0^{\mathrm{diag}} \sigma'(B_0^\top x) x^\top\right\|_F^2 = \left\|w_0^{\mathrm{diag}} \sigma'(B_0^\top x)\right\|_2^2 \|x\|_2^2 = \left\|\sigma'(B_0^\top x)\right\|_2^2 \|x\|_2^2 \le 2kL^2.$$

Putting the bounds together, we obtain the desired overall bound. $\square$

E.3 Verifying Assumptions of Section 4.2

Having shown that the function class is approximately linear as well as the norm bounds above, we now can proceed to verify the required assumptions. First, we verify the assumptions on the loss function; clearly, the squared error loss is convex, so we simply need to verify that the loss is Lipschitz over the prediction, and that $\ell(0, y)$ is bounded. Note that we have to prove separate bounds for source and target training, due to the change in the function class.

Lemma E.3 (Validating Loss Assumptions)**.** *Assume that $\gamma^2 \geq 2\kappa$. Define the quantities*

$$\alpha_1 \coloneqq L\sqrt{k} + \frac{\mu + L}{\beta} \quad \textit{and} \quad \alpha_2 \coloneqq \kappa^{1/2} L\sqrt{k} + \frac{\kappa}{\gamma}(\mu + L).$$

Then, during both source and target training time, $|\ell(0, y)| \leq \alpha_1^2$ for any y, and $\ell(\cdot, y)$ is α_1-Lipschitz during source training time and α_2-Lipschitz during target training time for any y, all up to universal constants.

Proof. For the squared error loss, $\ell(0, y) = y^2$ and $\nabla_{\hat{y}} \ell(\hat{y}, y) = 2(\hat{y} - y)$ for any y. Therefore, since the additive noise is $O(1)$ bounded, the claims hold as long as we can prove a bound on the predictions of any feasible predictor.

We first consider bounding the Lipschitz constant during source-time training. That is, we need to provide a uniform bound on $x \mapsto \beta \rho_{\theta_0}(x)^\top \delta + \beta \zeta^\delta_{\theta_0}(x)$ for $\theta_0 \in \Theta_0$, $\delta \in \mathcal{C}_{\mathrm{S}}$, and x in the support of p. To this end, we apply the bounds in Corollary E.1 and Lemma E.2, from which we obtain

$$\left|\beta \rho_{\theta_0}(x)^\top \delta + \beta \zeta^\delta_{\theta_0}(x)\right| \lesssim L\sqrt{k} + \frac{\mu + L}{\beta} =: \alpha_1.$$

By applying a similar argument for target-time training, we find that

$$\left|\gamma \rho_{\theta_0}(x)^\top \delta + \gamma \zeta^\delta_{\theta_0}(x)\right| \lesssim \kappa^{1/2} L\sqrt{k} + \frac{\kappa}{\gamma}(\mu + L) =: \alpha_2.$$

We thus can conclude that during both source and target training time, $|\ell(0, y)| \leq \alpha_1^2$ for any y, and that $\ell(\cdot, y)$ is α_1-Lipschitz during source training time and α_2-Lipschitz during target training time for any feasible y. □

Lemma E.4 (Neural network diversity)**.** *Assume that $\gamma > \max(\kappa\beta, \sqrt{2\kappa})$. Then, the source tasks satisfy a $(1, (L + \mu)^2/\beta^2)$-diversity condition with respect to the target task distribution ρ.*

Proof. Observe that we can write the excess risk of a predictor $f^\gamma_{(B_0, w_0)}$ as

$$\mathbb{E}\left[\left[(\gamma \rho_{B_0, w_0}(x)\delta - \beta \rho_{B^*, w_0^*}(x)\delta^*) + (\gamma \zeta^\delta_{B_0, w_0}(x) - \beta \zeta^{\delta^*}_{B^*, w_0^*}(x))\right]^2\right].$$

We proceed to upper bound the averaged best-case target performance, and lower bound the best-case performance averaged over source tasks.

Upper bounding the expected best-case target performance.

Fix a δ^*. Then, since $\gamma^2 > 2\kappa$ guarantees that Lemma E.1 holds,

$$\begin{aligned}
\mathbb{E}\left[(f^\gamma_{\theta_0 + \delta}(x) - f^\beta_{\theta_0^* + \delta^*}(x))^2\right] &\lesssim \mathbb{E}\left[(\gamma \rho_{\theta_0}(x)\delta - \rho_{\theta_0^*}(x)\delta^*)^2\right] + \mathbb{E}\left[(\gamma \zeta^\delta_{\theta_0}(x) - \beta \zeta^{\delta^*}_{\theta_0^*}(x))^2\right] \\
&\lesssim \mathbb{E}\left[\left(\gamma \rho_{\theta_0}(x)\delta - \beta \rho_{\theta_0^*}(x)\delta^*\right)^2\right] + \left[\frac{\kappa}{\gamma}(L + \mu)\right]^2 + \frac{(L + \mu)^2}{\beta^2} \\
&\lesssim \mathbb{E}\left[\left(\gamma \rho_{\theta_0}(x)\delta - \beta \rho_{\theta_0^*}(x)\delta^*\right)^2\right] + \frac{(L + \mu)^2}{\beta^2}.
\end{aligned}$$

Therefore, by taking the infimum over $\delta \in \mathcal{C}_{\mathrm{T}}$ and noting that the unconstrained infimum of the right-hand side is feasible by Proposition I.7 and the conditioning constraint imposed by Θ_0,

$$\inf_{\delta \in \mathcal{C}^\gamma_{\mathrm{T}}} \mathbb{E}\left[(f^\gamma_{\theta_0 + \delta}(x) - f^\beta_{\theta_0^* + \delta^*}(x))^2\right] \leq \beta^2 \left\|\Lambda(\rho_{\theta_0}, \rho_{\theta_0^*})^{1/2} \delta^*\right\|_2^2 + \left(\frac{L + \mu}{\beta}\right)^2$$

Now, by the choice of target task distribution, we can write δ^* as $\Gamma\omega^*/\beta$, where we note that $\mathbb{E}\left[\omega^*(\omega^*)^\top\right] = (1/k)I_{2k}$. Thus, by taking expectations, we have that

$$\mathbb{E}_{\delta^*\sim\rho}\left[\inf_{\delta\in\mathcal{C}_{\mathrm{T}}^\gamma}\mathbb{E}\left[(f^\gamma_{\theta_0+\delta}(x) - f^\beta_{\theta_0^*+\delta^*}(x))^2\right]\right] \leq \frac{1}{k}\left\|\Lambda(\rho_{\theta_0},\rho_{\theta_0^*})^{1/2}\Gamma\right\|_F^2 + \left(\frac{L+\mu}{\beta}\right)^2.$$

Lower bounding the source task-averaged best-case target performance.

Fix a source task $t \in [T]$. By Corollary I.1, we have the bound

$$\begin{aligned}\beta^2\mathbb{E}\left[(\rho_{\theta_0}(x)\delta - \rho_{\theta_0^*}(x)\delta_t^*)^2\right] &\lesssim \mathbb{E}\left[(f^\beta_{\theta_0+\delta}(x) - f^\beta_{\theta_t^*}(x))^2\right] + \mathbb{E}\left[(\zeta^\delta_{\theta_0}(x) - \zeta^{\delta^*}_{\theta_0^*}(x))^2\right]\\ &\lesssim \mathbb{E}\left[(f^\beta_{\theta_0+\delta}(x) - f^\beta_{\theta^*}(x))^2\right] + \left(\frac{L+\mu}{\beta}\right)^2.\end{aligned}$$

Therefore, by taking the infimum on both sides and applying Proposition I.7,

$$\left\|\Lambda(\rho_{\theta_0},\rho_{\theta^*})^{1/2}\Gamma\omega_t^*\right\|_2^2 \leq \inf_{\delta\in\mathcal{C}_{\mathrm{S}}}\mathbb{E}\left[(f^\beta_{\theta_0+\delta}(x) - f^\beta_{\theta^*}(x))^2\right] + \left(\frac{L+\mu}{\beta}\right)^2.$$

By averaging over the set of source tasks,

$$\frac{1}{T}\left\|\Lambda(\rho_{\theta_0},\rho_{\theta^*})^{1/2}\Gamma\Omega\right\|_F^2 \leq \frac{1}{T}\sum_{t\in[T]}\inf_{\delta\in\mathcal{C}_{\mathrm{S}}}\mathbb{E}\left[(f^\beta_{\theta_0+\delta}(x) - f^\beta_{\theta^*}(x))^2\right] + \left(\frac{L+\mu}{\beta}\right)^2.$$

Finally, by applying Assumption E.2 to lower bound the left-hand side, we find that

$$\frac{1}{k}\left\|\Lambda(\rho_{\theta_0},\rho_{\theta^*})^{1/2}\Gamma\right\|_F^2 \leq \frac{1}{T}\sum_{t\in[T]}\inf_{\delta\in\mathcal{C}_{\mathrm{S}}}\mathbb{E}\left[(f^\beta_{\theta_0+\delta}(x) - f^\beta_{\theta^*}(x))^2\right] + \left(\frac{L+\mu}{\beta}\right)^2.$$

Concluding.

Consequently, by putting the two parts together, we have that

$$\begin{aligned}&\mathbb{E}_{\delta^*\sim\rho}\left[\inf_{\delta\in\mathcal{C}_{\mathrm{T}}^\gamma}\mathbb{E}\left[(f^\gamma_{\theta_0+\delta}(x) - f^\beta_{\theta_0^*+\delta^*}(x))^2\right]\right]\\ &\quad\leq \frac{1}{T}\sum_{t\in[T]}\inf_{\delta\in\mathcal{C}_{\mathrm{S}}}\mathbb{E}\left[(f^\beta_{\theta_0+\delta}(x) - f^\beta_{\theta^*}(x))^2\right] + \left(\frac{L+\mu}{\beta}\right)^2.\end{aligned}$$

□

Lemma E.5 (Neural network approximate linearity)**.** *Assume that $\gamma^2 \geq 2\kappa$. For any $\theta_0 \in \Theta$,*

$$\sup_{\delta\in\mathcal{C}_{\mathrm{T}}}\frac{1}{n_{\mathrm{T}}}\sum_{i\in[n_{\mathrm{T}}]}\left\|\nabla^2_\theta f^\gamma_{\theta_0+\delta}(x_i)\right\|_2^2 \lesssim \gamma^2(L+\mu)^2 \quad \textit{and} \quad \frac{1}{n_{\mathrm{T}}}\sum_{i\in[n_{\mathrm{T}}]}\left\|\nabla_\theta f^\gamma_{\theta_0}(x_i)\right\|_2^2 \lesssim \gamma^2 kL^2.$$

Proof. The first bound is a trivial consequence of Lemma E.1. Meanwhile, we note that the second inequality, the quantity inside the norm is exactly $\beta\rho_{\theta_0}(x)$, and thus the bound follows by Lemma E.2. □

E.4 Computations

Lemma E.6 (Source Rademacher Bound)**.** *We have the following bound on the Rademacher complexity:*

$$\frac{1}{T}\mathcal{R}_{n_{\mathrm{S}}}\left[\bigcup_{\theta\in\Theta_0}\left[\mathcal{A}_{\mathcal{C}_{\mathrm{S}}}(\theta)\right]^{\otimes T}\right] \lesssim \frac{kL}{(n_{\mathrm{S}}T)^2} + \left[(L+\mu)\frac{k}{\sqrt{n_{\mathrm{S}}T}} + L\sqrt{\frac{k}{n_{\mathrm{S}}}}\right]\log(n_{\mathrm{S}}T) + \frac{\mu+L}{\beta}$$

Proof. We first consider bounding the empirical Rademacher complexity for a fixed set of inputs $(x_{i,t})$, after which we obtain the desired result by taking expectations over the sampled inputs. Expanding the definition, we have that

$$\begin{aligned}
&\frac{1}{T}\mathcal{R}_{n_\mathrm{S}}\left[\bigcup_{\theta\in\Theta_0}\left[\mathcal{A}_{\mathcal{C}_\mathrm{S}}(\theta)\right]^{\otimes T}\right] \\
&\quad= \frac{1}{nT}\mathbb{E}\left[\sup_{\substack{(B_0,w_0)\in\Theta_0\\(\Delta_t,w_t)\in\mathcal{C}_\mathrm{S}}}\left|\sum_{\substack{i\in[n_\mathrm{S}]\\t\in[T]}}\varepsilon_{i,t}\left[w_t^\top\phi(B^\top x_{i,t})+\langle\Delta_t,\psi_B(x_{i,t})\rangle+\zeta_{B,w_0}^{\Delta_t,w_t}(x_{i,t})\right]\right|\right] \\
&\quad\leq \frac{1}{T}\mathcal{R}_{n_\mathrm{S}}(\mathcal{F}_\Phi^{\otimes T}\circ\Phi)+\frac{1}{T}\mathcal{R}_{n_\mathrm{S}}(\mathcal{F}_\Psi^{\otimes T}\circ\Psi)+\frac{\mu+L}{\beta},
\end{aligned}$$

where

$$\begin{aligned}
\mathcal{F}_\Phi &\coloneqq \{z\mapsto\beta\langle w,z\rangle \mid \|w\|_2\leq 1/\beta\} \\
\mathcal{F}_\Psi &\coloneqq \{Z\mapsto\beta\langle\Delta,Z\rangle \mid \|\Delta\|_F\leq 1/\beta\} \\
\Phi &\coloneqq \{\phi_B \mid (\exists w)\,(B,w)\in\Theta_0\} \\
\Psi &\coloneqq \{\psi_{B,w} \mid (B,w)\in\Theta_0\}.
\end{aligned}$$

We proceed to bound the two complexity terms by noting that for any function class $\mathcal{F}$, $\mathcal{R}_{n_\mathrm{S}}(\mathcal{F})\lesssim\mathcal{G}_{n_\mathrm{S}}(\mathcal{F})$, and making use of the Gaussian complexity chain rule from Proposition I.6. First, note that for a fixed set of latent vectors $z_1,\ldots,z_{n_\mathrm{S}}$,

$$\mathcal{G}_Z(\mathcal{F}_\Phi)=\frac{1}{n_\mathrm{S}}\mathbb{E}\left[\sup_{\|w\|_2\leq 1}|\langle z,Zw\rangle|\right]\leq\frac{1}{n_\mathrm{S}}\sqrt{\operatorname{tr}ZZ^\top}=\frac{1}{n_\mathrm{S}}\sqrt{\sum_{i\in[n_\mathrm{S}]}\|z_i\|_2^2}.$$

Then, applying the representation norm bound Lemma E.2 to the latents,

$$\mathbb{E}\left[\max_{Z\in\mathcal{Z}}\mathcal{G}_Z(\mathcal{F}_\Phi)\right]\leq L\sqrt{\frac{k}{n_\mathrm{S}}}.$$

Similarly, for a fixed set of latent matrices $Z_1,\ldots,Z_{n_\mathrm{S}}$

$$\mathcal{G}_Z(\mathcal{F}_\Psi)\leq\frac{1}{n_\mathrm{S}}\sqrt{\sum_{i\in n_\mathrm{S}}\|Z_i\|_F^2}\implies\mathbb{E}\left[\max_{Z\in\mathcal{Z}}\mathcal{G}_Z(\mathcal{F}_\Psi)\right]\leq L\sqrt{\frac{k}{n_\mathrm{S}}},$$

where we once again used the representation bound from Lemma E.2.

Finally, we bound the complexity of the classes Φ and Ψ. Firstly, we have

$$\begin{aligned}
\frac{1}{T}\mathcal{G}_{n_\mathrm{S}}(\Phi)&=\frac{1}{n_\mathrm{S}T}\mathbb{E}\left[\sup_{(B_0,w_0)\in\Theta_0}\left|\sum_{r\in[2k]}\sum_{i\in[n_\mathrm{S}]}\sum_{t\in[T]}z_{r,i,t}\sigma(B_r^\top x_{i,t})\right|\right] \\
&\leq\frac{2k}{n_\mathrm{S}T}\mathbb{E}\left[\sup_{\|b\|_2\leq 1}\left|\sum_{i\in[n_\mathrm{S}]}\sum_{t\in[T]}z_{i,t}\sigma(b^\top x_{i,t})\right|\right] \\
&\lesssim\frac{kL}{n_\mathrm{S}T}\mathbb{E}\left[\left\|\sum_{i\in[n_\mathrm{S}T]}z_ix_i\right\|_2\right]\leq\frac{kL}{n_\mathrm{S}T}\sqrt{\sum_{i\in[n_\mathrm{S}T]}\|x_i\|_2^2}\leq\frac{kL}{\sqrt{n_\mathrm{S}T}}.
\end{aligned}$$

where the second inequality makes use of the Gaussian contraction result in Ledoux & Talagrand (1991, Corollary 3.17). Note that we reindex at the third line, which is made possible by the fact

that the source tasks have the same input distributions. Secondly,

$$\begin{aligned}
\frac{1}{T}\mathcal{G}_{n_{\mathrm{S}}}(\Psi) &= \frac{1}{n_{\mathrm{S}}T}\mathbb{E}\left[\sup_{(B_0,w_0)\in\Theta_0}\left|\sum_{r\in[2k]}\sum_{i\in[n_{\mathrm{S}}]}\sum_{t\in[T]}\sigma'(B_r^\top x_{i,t})\,\langle x_{i,t}, z_{r,i,t}\rangle\right|\right] \\
&= \frac{1}{n_{\mathrm{S}}T}\mathbb{E}\left[\sup_{(B_0,w_0)\in\Theta_0}\left|\sum_{r\in[2k]}\sum_{i\in[n_{\mathrm{S}}]}\sum_{t\in[T]}[\sigma'(B_r^\top x_{i,t})-\sigma'(0)]\,\langle x_{i,t}, z_{r,i,t}\rangle\right|\right] + \frac{L}{\sqrt{n_{\mathrm{S}}T}} \\
&= \frac{2k}{n_{\mathrm{S}}T}\mathbb{E}\left[\sup_{\|b\|_2\le 1}\left|\sum_{i\in[n_{\mathrm{S}}]}\sum_{t\in[T]}\left[\sigma'(b^\top x_{i,t})-\sigma'(0)\right]\langle x_{i,t}, z_{r,i,t}\rangle\right|\right] + \frac{L}{\sqrt{n_{\mathrm{S}}T}} \\
&\lesssim \frac{k\mu}{n_{\mathrm{S}}T}\mathbb{E}\left[\left\|\sum_{i\in[n_{\mathrm{S}}T]}\langle x_{i,t}, z_{i,t}\rangle\, x_{i,t}\right\|_2\right] + \frac{L}{\sqrt{n_{\mathrm{S}}T}} \le \frac{k\mu}{n_{\mathrm{S}}T}\sqrt{\sum_{i\in[n_{\mathrm{S}}T]}\|x_i\|_2^4} + \frac{L}{\sqrt{n_{\mathrm{S}}T}} \\
&\le \frac{k\mu}{\sqrt{n_{\mathrm{S}}T}} + \frac{L}{\sqrt{n_{\mathrm{S}}T}},
\end{aligned}$$

following the same reasoning as before. Therefore, by applying Proposition I.6, we obtain the bound

$$\frac{1}{T}\mathcal{R}_{n_{\mathrm{S}}}\left[\bigcup_{\theta\in\Theta_0}[\mathcal{A}_{\mathcal{C}_{\mathrm{S}}}(\theta)]^{\otimes T}\right] \lesssim \frac{kL}{(n_{\mathrm{S}}T)^2} + \left[(L+\mu)\frac{k}{\sqrt{n_{\mathrm{S}}T}} + L\sqrt{\frac{k}{n_{\mathrm{S}}}}\right]\log(n_{\mathrm{S}}T) + \frac{\mu+L}{\beta}. \quad \square$$

Lemma E.7 (Target Rademacher bound)**.** *We have the following Rademacher complexity bound:*

$$\sup_{\theta\in\Theta_0}\mathcal{R}_{n_{\mathrm{T}}}[\mathcal{A}_{\mathcal{C}_{\mathrm{T}}^\gamma}(\theta)] \lesssim L\sqrt{\kappa}\sqrt{\frac{k}{n_{\mathrm{T}}}} + \frac{\kappa}{\gamma}(L+\mu).$$

Proof. Following definitions, we have that

$$\begin{aligned}
&\sup_{\theta\in\Theta_0}\mathcal{R}_{n_{\mathrm{T}}}[\mathcal{A}_{\mathcal{C}_{\mathrm{T}}^\gamma}(\theta)] \\
&= \frac{\gamma}{n_{\mathrm{T}}}\mathbb{E}\left[\sup_{\delta\in\mathcal{C}_{\mathrm{T}}^\gamma}\left|\left\langle\varepsilon, \rho(X)\delta + \zeta_{\theta_0}^\delta(X)\right\rangle\right|\right] \le \frac{\gamma}{n_{\mathrm{T}}}\mathbb{E}\left[\sup_{\delta\in\mathcal{C}_{\mathrm{T}}^\gamma}|\langle\varepsilon, \rho(X)\delta\rangle|\right] + \frac{\kappa}{\gamma}(L+\mu) \\
&= \frac{\sqrt{\kappa}}{n_{\mathrm{T}}}\mathbb{E}\left[\left\|\rho(X)^\top\varepsilon\right\|_2\right] + \frac{\kappa}{\gamma}(L+\mu) \le \frac{\sqrt{\kappa}}{n_{\mathrm{T}}}\sqrt{\mathbb{E}\left[\left\|\rho(X)^\top\varepsilon\right\|_2^2\right]} + \frac{\kappa}{\gamma}(L+\mu) \\
&\le L\sqrt{\kappa}\sqrt{\frac{k}{n_{\mathrm{T}}}} + \frac{\kappa}{\gamma}(L+\mu),
\end{aligned}$$

where the final inequality uses the bound on the representation given in Lemma E.2. $\square$

E.5 Compiling the Bound

Theorem 5.1 (Neural net performance bound)**.** *Assume that Assumptions 5.1, E.1 and E.2 hold. Then, if $n_{\mathrm{S}} \ge n_{\mathrm{T}}$, there exists a setting of the training parameters (see Section E) such that with probability at least $1-\delta$, the iterates (θ_t) satisfy*

$$\mathbb{E}_{f^*\sim\rho}\left[\min_t \mathcal{L}_\infty^{\mathrm{ex}}(f_{\theta_t}^\gamma, f^*)\right] \lesssim \kappa L^2\frac{k}{\sqrt{n_{\mathrm{T}}}} + L(L+\mu)\frac{k^{3/2}}{\sqrt{n_{\mathrm{S}}T}} + \left(\frac{\mu+L}{\beta}\right)L\sqrt{k}.$$

Proof. By incorporating previous bounds and invoking Theorem 4.1, we have that for large enough γ,

$$\begin{aligned}
&\mathbb{E}_{f^*\sim\rho}\left[\min_t \mathcal{L}_\infty^{\mathrm{ex}}(f_{\theta_t}^\gamma, f^*)\right] \\
&\lesssim \sqrt{\frac{k\kappa(L+\mu)^2}{T_{\mathrm{PGD}}}} + \kappa L^2\frac{k}{\sqrt{n_{\mathrm{T}}}} + L(L+\mu)\frac{k^{3/2}}{\sqrt{n_{\mathrm{S}}T}} + \left(\frac{\mu+L}{\beta}\right)L\sqrt{k}.
\end{aligned}$$

Therefore, by running enough projected gradient descent iterations so that the first term matches $kL^2k/\sqrt{n_{\mathrm{T}}}$, we obtain the desired bound. $\square$

F Case Study: Logistic Regression

To further illustrate our framework, we analyze the performance of AdaptRep on logistic regression, as done by Tripuraneni et al. (2020b). In this setting, we let $\theta = (B, w)$ for $B \in \mathbb{R}^{d\times k}$ and $w \in \mathbb{R}^k$, and define the predictor corresponding to θ to be $g_\theta(x) = x^\top Bw$.

F.1 Statistical Assumptions

As in the linear setting, we consider an input distribution p with covariance Σ. We restrict the set of labels $\mathcal{Y}$ to $\{0, 1\}$, and consider the conditional distribution $q(y \mid g_\theta(x)) = \text{Ber}\left(\sigma(g_\theta(x))\right)$, where $\sigma(y) = 1/(1 + e^{-y})$ is the sigmoid function.

We define the optimal parameters for tasks $t \in [T]$ to be $(B^* + \Delta_t^*, w_t^*)$, where B^* is orthogonal and $\|\Delta_t^*\|_F \le \delta_0$. As before, we define $\delta_t^* \coloneqq \Delta_t^* w_t^*$ for any $t \in [T]$ and $W^* = [w_1, \ldots, w_T] \in \mathbb{R}^{k\times T}$. Having defined the prior quantities, we make use of the statistical assumptions presented in Section 3.1, reproduced below for convenience:

Assumption F.1 (Sub-Gaussian input)**.** *There exists $\rho > 0$ such that if $x \sim p_t$, then $\Sigma^{-1/2}x$ is ρ^2-sub-Gaussian.*

Assumption F.2 (Source task diversity)**.** *For any $t \in [T]$, $\|w_t^*\|_2 \le r$, and $\sigma_k^2(W^*) = \Omega(r^2T/k)$.*

Finally, we define the target task distribution ρ by sampling w^*, δ^* uniformly from the r– and δ_0–balls of $\mathbb{R}^k$, respectively, and letting $\theta^* = B^*w^* + \delta^*$.

F.2 Training Procedure

We use the standard logistic loss $\ell(\hat{y}, y) = -y\log[\sigma(\hat{y})] - (1 - y)\log[1 - \sigma(\hat{y})]$. During source training, we optimize over initializations $\Theta_0 \coloneqq \{(B, 0) \mid B \text{ is orthogonal}\}$. Let $B_0 \in \mathbb{R}^{d\times k}$ be the obtained representation. To adapt to the target task, we initialize the learner at $\theta_0 = ([B_0, B_0], [w_0, -w_0])$ for some unit-norm vector w_0, and scale the predictor by a fixed parameter β to be chosen later. Finally, we set the feasible sets for optimization to be

$$\mathcal{C}_{\text{S}} \coloneqq \{(\Delta, w) \mid \|\Delta\|_F \le \delta_0, \|w\|_2 \le r\} \quad \text{and} \quad \mathcal{C}_{\text{T}}^\beta \coloneqq \left\{(\Delta, w) \;\middle|\; \|\Delta\|_F \le \frac{\delta_0}{\beta}, \|w\|_2 \le \frac{r\kappa^{1/2}}{\beta}\right\},$$

where $\kappa \coloneqq \lambda_{\max}(\Sigma)/\lambda_{\min}(\Sigma)$. Note the similarity of this procedure to that of the linear setting.

F.3 Performance Guarantee

Having described the statistical assumptions and the training procedure, we now specialize the guarantee of Theorem 4.1 to this setting.

Theorem F.1 (Performance Guarantee for Logistic Regression)**.** *Assume that Assumption F.1 and Assumption F.2 both hold. Set the parameters for target time training to be*

$$\beta = (\kappa r^2 + \delta_0^2)\max\left(1, \frac{\sqrt{\text{tr}\,\Sigma}}{1/\delta\sqrt{n_{\text{T}}}}\right) \quad \textit{and} \quad T_{\text{PGD}} = \frac{(\kappa r^2 + \delta_0^2)\,\text{tr}\,\Sigma}{1/\delta^2 n_{\text{T}}}.$$

Then, for $n_{\text{S}}, n_{\text{T}} \gtrsim \rho^4 d$, we have that with probability at least $1 - \delta$ over the random draw of samples, the iterates (θ_t) satisfy

$$\begin{aligned}
&\mathbb{E}_{g^*\sim\rho}\left[\min_t \mathcal{L}_\infty^{\text{ex}}(g_{\theta_t}, g^*)\right] \\
&\quad\lesssim \frac{1}{\delta}\left\{r\kappa^{1/2}\|\Sigma\|_2^{1/2}\sqrt{\frac{k}{n_{\text{T}}}} + \frac{\delta_0}{\sqrt{n_{\text{T}}}}\sqrt{\text{tr}\,\Sigma}\right. \\
&\qquad\left. + \exp\left[\rho^3(r + \delta_0)\|\Sigma\|_2^{1/2}\right]\left[\frac{r\|\Sigma\|_2^{1/2}}{\sqrt{n_{\text{S}}T}}\sqrt{kT + kd\log n_{\text{S}}} + \frac{\delta_0}{\sqrt{n_{\text{S}}}}\sqrt{\text{tr}\,\Sigma}\right]\right\}.
\end{aligned}$$

F.4 Proofs

In this section, we prove the performance bound provided for logistic regression. First, we verify that the setting satisfies the assumptions of our general framework. Subsequently, we compute the quantities required to instantiate our bounds.

F.4.1 Verifying Assumptions of Section 4.2

In this section, we verify that the logistic regression setting, as described in Section F, satisfies the assumptions required by the general framework in Section 4.2.

It is easily verified that the logistic loss is 1-Lipschitz, convex, and that $|\ell(0, y)| \leq 1$ for $y \in \{0, 1\} = \mathcal{Y}$. Furthermore, as we have already characterized the approximate linearity of the function class in Lemma B.6, the approximate linearity assumption in Assumption 4.4 holds with high probability. Finally, $\mathcal{C}_{\mathrm{T}}$ is norm-bounded by $(\kappa r^2 + \delta_0^2)/\beta^2$.

Therefore, all that remains is verifying that a (ν, ε)-diversity condition holds in this setting. In what follows, we will do so by connecting the logistic loss to squared error loss via leveraging smoothness and local strong convexity, as was done by Tripuraneni et al. (2020b). Consequently, we can utilize the same argument as in the linear setting to obtain the desired diversity condition.

Lemma F.1 (Diversity condition, logistic regression)**.** *Under the assumptions above, the source tasks satisfy a* $(\Omega(\exp[-\rho^3(r + \delta_0) \|\Sigma\|_2^{1/2}]), 0)$*-diversity condition.*

Proof. We remark that under the choice of q and the logistic loss, we have that

$$\mathbb{E}_{x,y}\left[\ell(g_\theta(x), y) - \ell(g_{\theta'}(x), y)\right] = \mathbb{E}_x\left[\mathrm{KL}\left(\mathrm{Ber}\left(\sigma(g_\theta(x))\right) \| \mathrm{Ber}\left(\sigma(g_{\theta'}(x))\right)\right)\right]$$

Using the results in Tripuraneni et al. (2020b, Lemmas 2 and 3),

$$\frac{1}{8}\mathbb{E}_x\left[\exp(-\max(|g_\theta(x)|, |g_{\theta'}(x)|))(g_\theta(x) - g_{\theta'}(x))^2\right] \leq \mathbb{E}_{x,y}\left[\ell(g_\theta(x), y) - \ell(g_{\theta'}(x), y)\right] \quad (13)$$

$$\mathbb{E}_{x,y}\left[\ell(g_\theta(x), y) - \ell(g_{\theta'}(x), y)\right] \leq \frac{1}{8}\mathbb{E}_x\left[(g_\theta(x) - g_{\theta'}(x))^2\right]. \quad (14)$$

In what follows, we will make use of (13) to lower bound the task-averaged best-case performance, and (14) to upper bound the expected best-case target performance. Note that this results in bounds in terms of best-case mean squared error – we can then proceed to use arguments similar to that of the linear case to connect the two inequalities.

Upper bounding the expected best-case target performance.

Assume that the optimal predictor is given by $\theta^* = B^* w^* + \delta^*$. Then, defining

$$\hat{\Delta} = \frac{1}{\beta}\delta^* w_0^\top \quad \text{and} \quad \bar{w} = \frac{1}{\beta}(B_0^\top \Sigma B_0)^\dagger B_0^\top \Sigma B^* w^*,$$

we note that $\hat{\delta} \coloneqq ([\hat{\Delta}, 0], [0, \bar{w}^\top]^\top) \in \mathcal{C}_{\mathrm{T}}^\beta$, following the argument in Lemma B.8, and achieves a squared-error excess risk of

$$\mathbb{E}\left[(\beta g_{\theta+\hat{\delta}}(x) - g_{\theta^*}(x))^2\right] = \left\|\Sigma^{1/2}[(B_0\bar{w} + \delta^*) - (B^* w^* - \delta^*)]\right\|_2^2 = \left\|P^\perp_{\Sigma^{1/2}B_0}\Sigma^{1/2}B^* w^*\right\|_2^2.$$

Therefore, by taking the infimum of the excess risk over feasible δ and applying the excess risk upper bound in (14),

$$\begin{aligned}
\inf_{\delta \in \mathcal{C}_{\mathrm{T}}^\beta} \mathbb{E}_{x,y}\left[\ell(\beta g_{\theta_0+\delta}(x), y) - \ell(g_{\theta^*}(x), y)\right] &\leq \mathbb{E}_{x,y}\left[\ell(\beta g_{\theta_0+\hat{\delta}}(x), y) - \ell(g_{\theta^*}(x), y)\right] \\
&\leq \frac{1}{8}\mathbb{E}\left[(\beta g_{\theta_0+\hat{\delta}}(x) - g_{\theta^*}(x))^2\right] \\
&\leq \frac{1}{8}\left\|P^\perp_{\Sigma^{1/2}B_0}\Sigma^{1/2}B^* w^*\right\|_2^2.
\end{aligned}$$

Finally, recall that $\mathbb{E}_{\theta^* \sim \rho}\left[(w^*)(w^*)^\top\right] = (r^2/k)I$. Therefore, by taking expectations with respect to ρ,

$$\begin{aligned}
&\mathbb{E}_{\theta^* \sim \rho}\left[\inf_{\delta \in \mathcal{C}_{\mathrm{T}}^{\beta}} \mathbb{E}_{x,y}\left[\ell(\beta g_{\theta_0+\delta}(x), y) - \ell(g_{\theta^*}(x), y)\right]\right] \\
&\lesssim \mathbb{E}_{\theta^* \sim \rho}\left[\operatorname{tr} P^{\perp}_{\Sigma^{1/2}B_0}\Sigma^{1/2}B^*(w^*)(w^*)^\top\left(P^{\perp}_{\Sigma^{1/2}B_0}\Sigma^{1/2}B^*\right)^\top\right] \\
&= \frac{r^2}{k}\left\|P^{\perp}_{\Sigma^{1/2}B_0}\Sigma^{1/2}B^*\right\|_F^2.
\end{aligned}$$

Lower bounding the task-averaged best-case performance.

We proceed to lower bound the task-averaged best-case performance. To proof proceeds similarly to that of Tripuraneni et al. (2020b), lower bounding the excess risk by a constant multiple of the squared-error excess risk. Then, the result follows by an application of the transfer lemma in Lemma B.4.

Fix a task $t \in [T]$. Define $Z_1 \coloneqq x^\top(B^*w_t^* + \delta_t^*)$ and $Z_2 \coloneqq x^\top(B_0 w_t + \delta_t)$. We see that Z_1 is sub-Gaussian with parameter $\rho^2\left\|\Sigma^{1/2}(B^*w_t^* + \delta_t^*)\right\|_2^2 \le \rho^2(r^2 + \delta_0^2)\left\|\Sigma\right\|_2 =: \sigma^2$. Note that Z_2 is also sub-Gaussian with proxy σ^2. Then, by applying the lower bound in (15),

$$\mathbb{E}_{x,y}\left[\ell(g_{\hat{\theta}_t}(x), y) - \ell(g_{\theta_t^*}(x), y)\right] \ge \frac{1}{8}\mathbb{E}_x\left[e^{-\max(|Z_1|,|Z_2|)}(g_{\hat{\theta}_t}(x) - g_{\theta_t^*}(x))^2\right]. \tag{15}$$

We will now show that the right-hand side is lower bounded by a constant multiple of the squared error excess risk. Intuitively, since Z_1 and Z_2 concentrate around 0, $e^{-\max(|Z_1|,|Z_2|)}$ concentrates around 1.

Formally, consider the event $E_\alpha \coloneqq \{\max(|Z_1|, |Z_2|) \le \alpha\sigma\}$ for any α. Since the quantity inside the expectation of (15) is non-negative,

$$\begin{aligned}
&\frac{1}{8}\mathbb{E}_x\left[e^{-\max(|Z_1|,|Z_2|)}(g_{\hat{\theta}_t}(x) - g_{\theta_t^*}(x))^2\right] \\
&\quad\ge \frac{1}{8}\mathbb{E}\left[\mathbb{1}\left[E_\alpha\right]e^{-\alpha\sigma}(g_{\hat{\theta}_t}(x) - g_{\theta_t^*}(x))^2\right] \\
&\quad\ge \frac{e^{-\alpha\sigma}}{8}\left\{\mathbb{E}\left[(g_{\hat{\theta}_t}(x) - g_{\theta_t^*}(x))^2\right] - \mathbb{E}\left[\mathbb{1}\left[E_\alpha^C\right](g_{\hat{\theta}_t}(x) - g_{\theta_t^*}(x))^2\right]\right\}.
\end{aligned}$$

Now, we upper bound the last term. By Cauchy-Schwarz,

$$\mathbb{E}\left[\mathbb{1}\left[E_\alpha^C\right](g_{\hat{\theta}_t}(x) - g_{\theta_t^*}(x))^2\right] \le \sqrt{P(E_\alpha^C)}\sqrt{\mathbb{E}\left[(g_{\hat{\theta}_t}(x) - g_{\theta_t^*}(x))^4\right]},$$

where we have used the fact that $\operatorname{Var}[X] \le \sigma^2$ for a σ^2-sub-Gaussian random variable. We use properties of sub-Gaussian random variables to bound both factors. First, by Chebyshev's inequality,

$$P\left(E_\alpha^C\right) \le 2P(|Z_1| \ge \alpha\sigma) \le \frac{2\operatorname{Var}[Z_1]}{\alpha^2\sigma^2} \le \frac{2}{\alpha^2}.$$

For the second factor, $g_{\hat{\theta}_t}(x) - g_{\theta_t^*}(x)$ is sub-Gaussian with proxy $\rho^2\mathbb{E}\left[(g_{\hat{\theta}_t}(x) - g_{\theta_t^*}(x))^2\right]$, and so via an equivalent definition of sub-Gaussian random variables,

$$\sqrt{\mathbb{E}\left[(g_{\hat{\theta}_t}(x) - g_{\theta_t^*}(x))^4\right]} \lesssim \rho^2\mathbb{E}\left[(g_{\hat{\theta}_t}(x) - g_{\theta_t^*}(x))^2\right].$$

Putting these bounds together, we have that

$$\mathbb{E}\left[\mathbb{1}\left[E_\alpha^C\right](g_{\hat{\theta}_t}(x) - g_{\theta_t^*}(x))^2\right] \lesssim \frac{\rho^2}{\alpha}\mathbb{E}\left[(g_{\hat{\theta}_t}(x) - g_{\theta_t^*}(x))^2\right].$$

Therefore, by setting $\alpha \asymp \rho^2$, we can ensure that

$$\begin{aligned}
&\mathbb{E}_{x,y}\left[\ell(g_{\hat{\theta}_t}(x), y) - \ell(g_{\theta_t^*}(x), y)\right] \gtrsim e^{-\rho^3(r+\delta_0)\|\Sigma\|_2^{1/2}} \mathbb{E}\left[(x^\top \hat{\theta}_t - x^\top \theta_t^*)^2\right] \\
&\implies \frac{1}{T}\sum_{t\in[T]} \mathbb{E}_{x,y}\left[\ell(g_{\hat{\theta}_t}(x), y) - \ell(g_{\theta_t^*}(x), y)\right] \\
&\qquad \gtrsim \frac{e^{-\rho^3(r+\delta_0)\|\Sigma\|_2^{1/2}}}{T} \sum_{t\in[T]} \left\|\Sigma^{1/2}(B_0 w_t + \delta_t - B^* w_t^* + \delta_t^*)\right\|_2^2.
\end{aligned}$$

By applying Lemma B.4, we finally obtain the final lower bound

$$\frac{e^{\rho^3(r+\delta_0)\|\Sigma\|_2^{1/2}}}{T} \sum_{t\in[T]} \inf_{\delta_t \in \mathcal{C}_{\mathrm{S}}} \mathbb{E}_{x,y}\left[\ell(g_{\theta_0+\delta_t}(x), y) - \ell(g_{\theta_t^*}(x), y)\right] \geq \frac{r^2}{k}\left\|P^{\perp}_{\Sigma^{1/2}B_0}\Sigma^{1/2}B^*\right\|_F^2.$$

Putting everything together, we thus have the desired diversity condition

$$\begin{aligned}
&\inf_{\|\delta\|\leq\delta_0} \mathbb{E}_{x,y}\left[\ell(\beta g_{\theta_0+\delta}(x), y) - \ell(g_{\theta^*}(x), y)\right] \\
&\qquad \lesssim \frac{e^{\rho^3(r+\delta_0)\|\Sigma\|_2^{1/2}}}{T} \sum_{t\in[T]} \inf_{\|\delta_t\|\leq\delta_0} \mathbb{E}_{x,y}\left[\ell(g_{\theta_0+\hat{\delta}_t}(x), y) - \ell(g_{\theta_t^*}(x), y)\right]. \qquad \square
\end{aligned}$$

F.4.2 Computations

Having demonstrated that the assumptions in Section 4.2 hold, we proceed to calculate the relevant quantities required for establishing a performance bound in this setting. First, we compute the Rademacher complexity for source task training.

Lemma F.2 (Source Rademacher Bound)**.** *Assume that $n_{\mathrm{S}} \gtrsim \rho^4 d$. Then, we can bound the source Rademacher complexity as*

$$\frac{1}{T}\mathcal{R}_{n_{\mathrm{S}}}\left[\bigcup_{\theta\in\Theta_0}\left[\mathcal{A}_{\mathcal{C}_{\mathrm{S}}}(\theta)\right]^{\otimes T}\right] \lesssim \frac{r\,\|\Sigma\|_2^{1/2}}{\sqrt{n_{\mathrm{S}}T}}\sqrt{kT + kd\log n_{\mathrm{S}}} + \frac{\delta_0}{\sqrt{n_{\mathrm{S}}}}\sqrt{\operatorname{tr}\Sigma}.$$

Proof. Observe that if $(\Delta_t, w_t) \in \mathcal{C}_{\mathrm{S}}$, then $\delta_t \coloneqq \Delta_t w_t$ satifies $\|\delta_t\|_2 \leq \delta_0$. Therefore, we have that

$$\begin{aligned}
&\frac{1}{T}\mathcal{R}_{n_{\mathrm{S}}}\left[\bigcup_{\theta\in\Theta_0}\left[\mathcal{A}_{\mathcal{C}_{\mathrm{S}}}(\theta)\right]^{\otimes T}\right] \\
&\quad = \mathbb{E}\left[\sup_{\substack{B \\ (\Delta_t, w_t)\in\mathcal{C}_{\mathrm{S}}}} \frac{1}{n_{\mathrm{S}}T}\sum_{t\in[T]}\langle\varepsilon_t, X_t(B+\Delta_t)w_t\rangle\right] \\
&\quad = \underbrace{\mathbb{E}\left[\sup_{\substack{B \\ \|w_t\|_2\leq r}} \frac{1}{n_{\mathrm{S}}T}\sum_{t\in[T]}\langle\varepsilon_t, X_t B w_t\rangle\right]}_{=:\mathrm{(I)}} + \underbrace{\mathbb{E}\left[\sup_{\|\delta_t\|_2\leq\delta_0} \frac{1}{n_{\mathrm{S}}T}\sum_{t\in[T]}\langle\varepsilon_t, X\delta_t\rangle\right]}_{=:\mathrm{(II)}}
\end{aligned}$$

We proceed to bound these two quantities separately.

Bounding (I) via discretization.

In this section, we bound the complexity by discretizing the set $\mathcal{O}^{k\times d}$ of $(k\times d)$ orthogonal matrices. We remark that the argument is similar in form to that of Lemma B.3.

Let $\mathcal{S}$ be an ε-covering of $\mathcal{O}^{k\times d}$ in the Frobenius norm with at most $(6\sqrt{k}/\varepsilon)^{dk}$ elements guaranteed by Proposition I.1. Then,

$$\text{(I)} \le \underbrace{\mathbb{E}\left[\sup_{\substack{B\in\mathcal{S}\\ \|w_t\|_2\le r}} \frac{1}{n_{\mathrm{S}}T}\sum_{t\in[T]}\langle \varepsilon_t, X_tBw_t\rangle\right]}_{=:\text{(A)}} + \underbrace{\mathbb{E}\left[\sup_{\substack{B,\bar{B}\in\mathcal{S}\\ \|B-\bar{B}\|_F\le\varepsilon\\ \|w_t\|_2\le r}} \frac{1}{n_{\mathrm{S}}T}\sum_{t\in[T]}\langle \varepsilon_t, X_t(B-\bar{B})w_t\rangle\right]}_{=:\text{(B)}}$$

To bound (A), we bound the corresponding Gaussian complexity and use the fact that $\mathcal{R}(\cdot)\lesssim\mathcal{G}(\cdot)$. Consequently, via multiple applications of Cauchy-Schwarz,

$$\begin{aligned}\text{(A)} &\le \sqrt{\frac{\pi}{2}}\mathbb{E}\left[\sup_{\substack{B\in\mathcal{S}\\ \|w_t\|_2\le r}} \frac{1}{n_{\mathrm{S}}T}\sum_{t\in[T]}\langle P_{X_tB}z_t, X_tBw_t\rangle\right]\\ &\lesssim \sqrt{\mathbb{E}\left[\sup_{B\in\mathcal{S}}\frac{1}{n_{\mathrm{S}}T}\sum_{t\in[T]}\|P_{X_tB}z_t\|_2^2\right]}\sqrt{\mathbb{E}\left[\sup_{\substack{B\in\mathcal{S}\\ \|w_t\|_2\le r}}\frac{1}{n_{\mathrm{S}}T}\sum_{t\in[T]}\|X_tBw_t\|_2^2\right]}.\end{aligned}$$

Conditioned on X_t, $\sum_{t\in[T]}\|P_{X_tB}z_t\|_2^2$ is distributed as a chi-squared random variable with kT degrees of freedom, with mean kT. Therefore, using known bounds on expectations of finite maxima of subexponential random variables,

$$\begin{aligned}&\mathbb{E}\left[\sup_{B\in S}\sum_{t\in[T]}\|P_{X_tB}z_t\|_2^2 - kT\right] \lesssim \sqrt{kT\log|\mathcal{S}|}+\log|S|\\ &\implies \mathbb{E}\left[\sup_{B\in S}\sum_{t\in[T]}\|P_{X_tB}z_t\|_2^2\right]\lesssim kT+\log|S|\,.\end{aligned} \tag{16}$$

Furthermore, by applying the expectation bound on the empirical spectral norm in Proposition I.5,

$$\begin{aligned}\sqrt{\mathbb{E}\left[\sup_{\substack{B\in\mathcal{S}\\ \|w_t\|_2\le r}}\frac{1}{n_{\mathrm{S}}T}\sum_{t\in[T]}\|X_tBw_t\|_2^2\right]} &= r\sqrt{\mathbb{E}\left[\sup_{B\in\mathcal{S}}\frac{1}{T}\sum_{t\in[T]}\lambda_{\max}\left(\frac{B^\top X_t^\top X_tB}{n_{\mathrm{S}}}\right)\right]} &(17)\\ &\le r\sqrt{\mathbb{E}\left[\lambda_{\max}\left(\frac{X_t^\top X_t}{n_{\mathrm{S}}}\right)\right]}\\ &\lesssim r\,\|\Sigma\|_2^{1/2}\,. &(18)\end{aligned}$$

Therefore, by combining the inequalities from (16) and (18),

$$\text{(A)}\lesssim\frac{\|\Sigma\|_2^{1/2}}{\sqrt{n_{\mathrm{S}}T}}\sqrt{kT+\log|S|}.$$

We proceed to bound (B), the error arising from discretization. We have that

$$\text{(B)}\le\mathbb{E}\left[\sup_{\substack{B,\bar{B}\in\mathcal{S}\\ \|B-\bar{B}\|_F\le\varepsilon\\ \|w_t\|_2\le r}}\frac{1}{n_{\mathrm{S}}T}\sum_{t\in[T]}\langle\varepsilon_t, X_t(B-\bar{B})w_t\rangle\right]\lesssim r\varepsilon\left\|\Sigma^{1/2}\right\|_2.$$

Finally, by setting $\varepsilon = \sqrt{k/n_{\mathrm{S}}}$, we have an overall bound of

$$\text{(I)} \leq \frac{r \left\|\Sigma\right\|_2^{1/2}}{\sqrt{n_{\mathrm{S}} T}} \sqrt{kT + kd \log n_{\mathrm{S}}}.$$

Bounding (II).

By bounding by the Gaussian complexity,

$$\text{(II)} \leq \frac{1}{n_{\mathrm{S}} T} \mathbb{E}\left[\sup_{\left\|\delta_t\right\|_2 \leq \delta_0} \sum_{t \in [T]} \left\langle z_t, X_t \delta_t \right\rangle\right] \lesssim \frac{\delta_0}{\sqrt{n_{\mathrm{S}}}} \mathbb{E}\left[\left\| \frac{1}{\sqrt{n}} X_1^\top z \right\|_2\right] \leq \frac{\delta_0}{\sqrt{n_{\mathrm{S}}}} \sqrt{\operatorname{tr} \Sigma}.$$

Putting the bounds on (I) and (II) together, we thus have that

$$\frac{1}{T} \mathcal{R}_{n_{\mathrm{S}}}\left[\bigcup_{\theta \in \Theta_0} \left[\mathcal{A}_{\mathcal{C}_{\mathrm{S}}}(\theta)\right]^{\otimes T}\right] \lesssim \frac{r \left\|\Sigma\right\|_2^{1/2}}{\sqrt{n_{\mathrm{S}} T}} \sqrt{kT + kd \log n_{\mathrm{S}}} + \frac{\delta_0}{\sqrt{n_{\mathrm{S}}}} \sqrt{\operatorname{tr} \Sigma}. \qquad \square$$

Now, we compute the Rademacher complexity term associated with target task training.

Lemma F.3 (Target Rademacher Bound)**.** *For $\beta \gtrsim r\kappa^{1/2}$, the Rademacher complexity of the feasible set during target time training is bounded by*

$$\sup_{\theta \in \Theta_0} \mathcal{R}_{n_{\mathrm{T}}}\left[\mathcal{A}_{\mathcal{C}_{\mathrm{T}}}(\theta_0)\right] \lesssim r\kappa^{1/2} \left\|\Sigma\right\|_2^{1/2} \sqrt{\frac{k}{n_{\mathrm{T}}}} + \frac{\delta_0}{\sqrt{n_{\mathrm{T}}}} \sqrt{\operatorname{tr} \Sigma}.$$

Proof. We can write the Rademacher complexity for a fixed $\theta_0 \in \Theta_0$ as

$$\mathcal{R}_{n_{\mathrm{T}}}\left[\mathcal{A}_{\mathcal{C}_{\mathrm{T}}}(\theta_0)\right] = \beta \mathbb{E}\left[\sup_{(\Delta, w) \in \mathcal{C}_{\mathrm{T}}^{\beta}} \frac{1}{n_{\mathrm{T}}} \left\langle \varepsilon, X[B_0, B_0] w + X \Delta w_0 + X \Delta w \right\rangle\right].$$

Firstly, by converting to Gaussian complexity and following standard arguments, we can bound the first term as

$$\begin{aligned}
\beta \mathbb{E}\left[\sup_{\left\|w\right\|_2 \leq r\kappa^{1/2}/\beta} \frac{1}{n_{\mathrm{T}}} \left\langle z, X[B_0, B_0] w \right\rangle\right] &\leq \beta \mathbb{E}\left[\sup_{\left\|w\right\|_2 \leq \kappa/\beta} \frac{1}{n_{\mathrm{T}}} \left\langle [B_0, B_0]^\top X^\top z, w \right\rangle\right] \\
&\leq \frac{r\kappa^{1/2}}{\sqrt{n_{\mathrm{T}}}} \mathbb{E}\left[\left\| B_0^\top \frac{X^\top}{\sqrt{n_{\mathrm{T}}}} z \right\|_2\right] \\
&\leq r\kappa^{1/2} \left\|\Sigma\right\|_2^{1/2} \sqrt{\frac{k}{n_{\mathrm{T}}}}.
\end{aligned}$$

For the second term, note that we only need to consider the set of rank-1 matrices of the form $\Delta = v[w_0, -w_0]^\top$, where $\left\|v\right\|_2 \leq \delta_0/\beta\sqrt{2}$. As such,

$$\beta \mathbb{E}\left[\sup_{\left\|v\right\|_2 \leq \delta_0/\beta\sqrt{2}} \frac{1}{n_{\mathrm{T}}} \left\langle z, Xv \right\rangle\right] \leq \frac{\delta_0}{\sqrt{n_{\mathrm{T}}}} \sqrt{\mathbb{E}\left[\left\| \frac{X}{\sqrt{n_{\mathrm{T}}}} z \right\|_2^2\right]} \lesssim \frac{\delta_0}{\sqrt{n_{\mathrm{T}}}} \sqrt{\operatorname{tr} \Sigma}.$$

Finally, to bound the Hessian term, note that $\left\|\Delta w\right\|_2 \leq r\kappa^{1/2}\delta_0/\beta^2$. Therefore, by a similar argument to the previous term,

$$\beta \mathbb{E}\left[\sup_{(\Delta, w) \in \mathcal{C}_{\mathrm{T}}} \frac{1}{n_{\mathrm{T}}} \left\langle \varepsilon, X \Delta w \right\rangle\right] \leq \frac{\kappa \delta_0}{\beta \sqrt{n_{\mathrm{T}}}} \sqrt{\operatorname{tr} \Sigma}.$$

Thus, as long as $\beta \geq r\kappa^{1/2}$, we can take suprema and obtain the overall bound of

$$\sup_{\theta \in \Theta_0} \mathcal{R}_{n_{\mathrm{T}}}\left[\mathcal{A}_{\mathcal{C}_{\mathrm{T}}}(\theta)\right] \lesssim r\kappa^{1/2} \left\|\Sigma\right\|_2^{1/2} \sqrt{\frac{k}{n_{\mathrm{T}}}} + \frac{\delta_0}{\sqrt{n_{\mathrm{T}}}} \sqrt{\operatorname{tr} \Sigma}. \qquad \square$$

F.4.3 Compiling the Bound

Having performed all required computations, we now prove the provided performance guarantee.

Theorem F.1 (Performance Guarantee for Logistic Regression)**.** *Assume that Assumption F.1 and Assumption F.2 both hold. Set the parameters for target time training to be*

$$\beta = (\kappa r^2 + \delta_0^2) \max\left(1, \frac{\sqrt{\operatorname{tr}\Sigma}}{1/\delta\sqrt{n_{\mathrm{T}}}}\right) \quad \textit{and} \quad T_{\mathrm{PGD}} = \frac{(\kappa r^2 + \delta_0^2)\operatorname{tr}\Sigma}{1/\delta^2 n_{\mathrm{T}}}.$$

Then, for $n_{\mathrm{S}}, n_{\mathrm{T}} \gtrsim \rho^4 d$, we have that with probability at least $1-\delta$ over the random draw of samples, the iterates (θ_t) satisfy

$$\begin{aligned}
&\mathbb{E}_{g^* \sim \rho}\left[\min_t \mathcal{L}_\infty^{\mathrm{ex}}(g_{\theta_t}, g^*)\right] \\
&\quad \lesssim \frac{1}{\delta}\Bigg\{ r\kappa^{1/2}\left\|\Sigma\right\|_2^{1/2}\sqrt{\frac{k}{n_{\mathrm{T}}}} + \frac{\delta_0}{\sqrt{n_{\mathrm{T}}}}\sqrt{\operatorname{tr}\Sigma} \\
&\qquad\qquad + \exp\left[\rho^3(r+\delta_0)\left\|\Sigma\right\|_2^{1/2}\right]\left[\frac{r\left\|\Sigma\right\|_2^{1/2}}{\sqrt{n_{\mathrm{S}}T}}\sqrt{kT + kd\log n_{\mathrm{S}}} + \frac{\delta_0}{\sqrt{n_{\mathrm{S}}}}\sqrt{\operatorname{tr}\Sigma}\right]\Bigg\}.
\end{aligned}$$

Proof. With probability at least $1-\delta/2$, the approximate linearity property of the function class holds via Lemma B.2 and Lemma B.6. Therefore, by combining all prior calculations and instantiating Theorem 4.1 with failure probability $\delta/2$, projected gradient descent finds a predictor g satisfying the desired bound in the theorem statement. □

G The Nonlinear Hard Case

G.1 Construction

In what follows, we establish the existence of a nonlinear setting where there exists a sample complexity separation between ADAPTREP and FROZENREP, similar to Section 3.4. As before, fix $k, d \in \mathbb{N}$ with $2k < d$. The construction relies on the observation that linear predictors lying in a rank-k space are representable as linear functions of $2k$ appropriately chosen ReLU neurons.

Following the discussion in Section E, note that when we take $\beta \to \infty$, the resulting function class can be expressed as

$$f_{(B+\Delta,w)} = w^\top \sigma(B^\top x) + \left\langle x\sigma'(x^\top B), \Delta \right\rangle,$$

where $B, \Delta \in \mathbb{R}^{d\times 2k}$, and $w \in \mathbb{R}^{2k}$. We further constrain B so that the first k columns are equal to the negation of the last k columns. Finally, we choose $\sigma(x) = \max(x, 0)$, which we will also write as x_+ for convenience[8]. For convenience, we follow the convention in Section E of writing ϕ_B, ψ_B and ρ_B for the activation, gradient, and concatenated features corresponding to B, respectively, as defined in Definition E.1.

We briefly review the construction in Section 3.4. Let the input distribution p be a Gaussian distribution on $\mathbb{R}^d$ with covariance

$$\Sigma = \begin{bmatrix} \varepsilon I_{d-k} & 0 \\ 0 & I_k \end{bmatrix}$$

for a fixed $\varepsilon \in (0, 1)$. Furthermore, we define $E^*, E_k \subset \mathbb{R}^d$ to be the two eigenspaces of Σ determined by the two blocks, *i.e.*

$$E^* = \operatorname{Col}\begin{bmatrix} \varepsilon I_{d-k} \\ 0 \end{bmatrix} \quad \text{and} \quad E_k = \operatorname{Col}\begin{bmatrix} 0 \\ I_k \end{bmatrix}.$$

Then, for any orthogonal matrix $A \in \mathbb{R}^{d\times k}$ with $\operatorname{Col} A \subseteq E^*$, define a distribution over θ given by

$$\theta = \frac{1}{\sqrt{2\varepsilon}} Av + \delta, \tag{19}$$

where v and δ are sampled uniformly at random from the unit spheres in $\mathbb{R}^k$ and E_k, respectively.

Now, we lift this linear task distribution setting into the ReLU setting. In particular, we sample the source tasks by fixing orthogonal $A \in \mathbb{R}^{d\times k}$, sampling v and δ as before, and letting the optimal predictor be $f_{(B+\Delta,w)}$, where

$$B \coloneqq [A, -A],\ w \coloneqq \frac{1}{\sqrt{2\varepsilon}}[v, -v],\ \Delta \coloneqq \frac{1}{k}\mathbb{1}\delta^\top. \tag{20}$$

One can easily verify algebraically that

$$f_{(B+\Delta,w)}(x) = x^\top \left(\frac{1}{\sqrt{2\varepsilon}} Av + \delta \right),$$

as desired. As before, we consider the family of task distributions induced by any A^* with $\operatorname{Col} A^* \subseteq E^*$.

With the above task distribution, we can then prove the following hardness result on FROZENREP:

Theorem G.1 (FROZENREP Minimax Bound, ReLU)**.** *For an orthogonal matrix $A^* \in \mathbb{R}^{d\times k}$ such that $\operatorname{Col} A^* \subset E^*$, let $B^* = [A^*, -A^*]$, and define S_{A^*} to be the set*

$$S_{A^*} = \left\{ \frac{1}{\sqrt{2\varepsilon}} A^* v + \delta \;\middle|\; \|v\|_2 \le 1, \|\delta\|_2 \le 1, \delta \in E_k \right\}.$$

Furthermore, let $\hat{B}$ be the output of FROZENREP *with access to infinite per-task samples and tasks, with the task distribution deterined by B^*. Then, with high probability over the draw of $n_T \gg d$ samples during target training, we have that*

$$\min_{\bar{B},\hat{w},\hat{\Delta}} \max_{\substack{A^* \\ \theta^* \in S_{A^*}}} \mathbb{E}\left[\frac{1}{n_{\mathrm{T}}} \left\| X\theta^* - f_{(\bar{B}+\hat{\Delta},\hat{w})}(X) \right\|_2^2 \right] \gtrsim \frac{\sigma^2 d}{n_{\mathrm{T}}},$$

[8] Note that $\sigma'(x) = \mathbb{1}[x > 0]$.

where $\bar{B}$, $\hat{w}$, and $\hat{\Delta}$ are all measurable functions of (X, y) to $\mathbb{R}^{d\times 2k}$, $\mathbb{R}^{2k}$, and $\mathbb{R}^{d\times 2k}$, respectively, and $\operatorname{Col}\bar{B} = \operatorname{Col}\hat{B}$. Furthermore, the expectation is over the randomness in the labels $y \sim \mathcal{N}\left(X\theta^*, \sigma^2 I_{n_{\mathrm{T}}}\right)$.

In contrast, we have the following upper bound on the performance of ADAPTREP:

Lemma G.1 (Adaptation target performance, ReLU)**.** *Let $B_0 = [A_0, -A_0]$, and fix a $\theta^* \in S_{A^*}$. Consider a learner which solves*

$$\min_{B} \min_{\substack{(w_t,\Delta_t)\\ \|\Delta_t\|_F\le 1}} \frac{1}{n_{\mathrm{S}}} \sum_{t\in[T]} \|\rho_B(X)[w,\Delta] - y\|_2^2 .$$

during source training, and

$$\min_{\substack{w,\Delta\\ \|\Delta\|_F\le 1}} \frac{1}{n_{\mathrm{T}}} \|\rho_{B_0}(X)[w,\Delta] - y\|_2^2$$

during target training, where B_0 is the representation obtained from source training. Finally, we set $k = \Theta(1)$ and $\varepsilon = k/d$. Then, with access to infinite per-task samples and tasks during source training, the learner achieves target loss bounded as

$$\frac{1}{n_{\mathrm{T}}} \|\rho_{B_0}(X)[w,\Delta] - X\theta^*\|_2^2 \lesssim \frac{\sigma^2}{n_{\mathrm{T}}}\left(1 + \log\frac{1}{\delta}\right) + \frac{\sigma}{\sqrt{n_{\mathrm{T}}}}\sqrt{1+\log\frac{1}{\delta}}$$

with probability at least $1-\delta$ over the draw of target samples.

Proofs of the above results are provided in the following section. As before, we compare the two methods when $n_{\mathrm{T}} = \Theta(d)$. Then, from the results above, the lower bound on the loss of FROZENREP is $\Omega(1)$, while the upper bound on the loss of ADAPTREP is $O(1/\sqrt{n_{\mathrm{T}}})$. Therefore, we also see a strict separation between the two methods within this setting as well, which grows with $n_{\mathrm{T}} \to \infty$.

G.2 Proofs

Throughout this section, we write $\theta_t^* = A^* v_t^* + \delta_t^*$ for the linear predictor parameter for source task $t \in [T]$. Furthermore, we assume that this linear predictor corresponds to ReLU predictor $f_{(B^*+\Delta^*, w^*)}$. First, we prove the following intermediate technical result which will be used throughout this section.

Lemma G.2 (Optimal ReLU Predictor is Linear)**.** *Let $\theta \in \mathbb{R}^d$, and assume that $B = [A, -A]$. Then, if we set*

$$w = \operatorname*{argmin}_{w\in\mathbb{R}^{2k}} \mathbb{E}\left[(w^\top (B^\top x)_+ - x^\top\theta_t^*)^2\right],$$

then the predictor $x \mapsto w^\top(B^\top x)$ is in fact a linear function of x, and takes the form of $x^\top A v$ for some $v \in \mathbb{R}^k$.

Proof. Write $w = [w_+, w_-]$, and thus the objective defining w can be written as

$$\min_{w_+, w_- \in \mathbb{R}^k} \mathbb{E}\left[(w_+^\top (A^\top x)_+ + w_-^\top (-A^\top x)_+ - x^\top\theta_t^*)^2\right].$$

Now, define the matrices

$$\Omega \coloneqq \mathbb{E}\left[(A^\top x)_+ (x^\top A)_+\right] = \mathbb{E}\left[(-A^\top x)_+ (-x^\top A)_+\right]$$
$$\Gamma \coloneqq \mathbb{E}\left[(-A^\top x)_+ (x^\top A)_+\right] = \mathbb{E}\left[(A^\top x)_+ (-x^\top A)_+\right],$$

where the equalities follow from the fact that Ax and $-Ax$ are equal in distribution. Additionally,,

$$\mathbb{E}\left[(A^\top x)_+ x^\top\right] = \frac{1}{2}\mathbb{E}\left[(A^\top x)_+ x^\top\right] - \frac{1}{2}\mathbb{E}\left[(-A^\top x)_+ x^\top\right] = \frac{1}{2}A^\top \mathbb{E}\left[xx^\top\right] = \frac{1}{2}A^\top\Sigma,$$

and thus once again since Ax and $-Ax$ are equal in distribution,

$$\mathbb{E}\left[(-A^\top x)_+ x^\top\right] = -\mathbb{E}\left[(A^\top x)_+ x^\top\right] = -\frac{1}{2}A^\top\Sigma.$$

Thus, by explicitly solving for the optimum of the convex objective,

$$\begin{aligned}\begin{bmatrix} w_+ \\ w_- \end{bmatrix} &= \frac{1}{2}\begin{bmatrix} \Omega & \Gamma \\ \Gamma & \Omega \end{bmatrix}^{-1}\begin{bmatrix} A^\top \Sigma \theta_t^* \\ -A^\top \Sigma \theta_t^* \end{bmatrix} \\ &= \begin{bmatrix} (\Omega - \Gamma\Omega\Gamma)^{-1} & 0 \\ 0 & (\Omega - \Gamma\Omega\Gamma)^{-1} \end{bmatrix}\begin{bmatrix} I & -\Gamma\Omega^{-1} \\ -\Gamma\Omega^{-1} & I \end{bmatrix}\begin{bmatrix} A^\top \Sigma \theta_t^* \\ -A^\top \Sigma \theta_t^* \end{bmatrix},\end{aligned}$$

where we have applied blockwise matrix inversion to obtain the second inequality. From this, we see that $w_+ = w_-$, and thus the corresponding optimal predictor is linear, and can be shown to be $x \mapsto x^\top A w_+$. □

G.2.1 Hardness Result for FROZENREP

Lemma G.3 (FROZENREP learns incorrect neurons)**.** *Fix an orthogonal matrix $A^* \in \mathbb{R}^{d\times k}$ with $\operatorname{Col} A^* \subseteq E^*$, and assume that we sample tasks from the distribution in (20). Assume that $\hat{B} = [\hat{A}, -\hat{A}]$ is the representation found by* FROZENREP*. Then, with infinitely many tasks and per-task samples (i.e. $n_{\mathrm{S}}, T \to \infty$), $\hat{A} = \operatorname{Col} E_k$.*

Proof. Intuitively, the result follows from the fact that the optimal predictor is equivalent to a linear predictor, and thus the result follows by Lemma C.1. More formally, by Lemma G.2,

$$\begin{aligned}\mathcal{L}(B) &= \frac{1}{T}\sum_{t\in[T]} \min_w \mathbb{E}\left[(w^\top (Bx)_+ - x^\top \theta_t^*)^2\right] = \frac{1}{T}\sum_{t\in[T]} \min_v \mathbb{E}\left[(x^\top A v - x^\top \theta_t^*)^2\right] \\ &= \frac{1}{T}\sum_{t\in[T]} \min_v \left\|\Sigma^{1/2}(Av - \theta_t^*)\right\|_2^2 = \frac{1}{T}\sum_{t\in[T]} \left\|P^{\perp}_{\Sigma^{1/2}A}\Sigma^{1/2}\theta_t^*\right\|_2^2.\end{aligned}$$

At this point, we recognize that the objective is equivalent to the one analyzed in Lemma C.1, and thus the same characterization of global optima holds. That is, the solution $\hat{B}$ found by FROZENREP can be expressed as $\hat{B} = [\hat{A}, -\hat{A}]$, where $\operatorname{Col}\hat{A} = E_k$. □

Theorem G.1 (FROZENREP Minimax Bound, ReLU)**.** *For an orthogonal matrix $A^* \in \mathbb{R}^{d\times k}$ such that $\operatorname{Col} A^* \subset E^*$, let S_{A^*} be the set*

$$S_{A^*} = \left\{ \frac{1}{\sqrt{2\varepsilon}} A^* v + \delta \;\middle|\; \|v\|_2 \le 1, \|\delta\|_2 \le 1, \delta \in E_k \right\}.$$

We consider the following procedure:

1. *We draw n_{T} samples for target time training, which are collected into a matrix X.*
2. *Player chooses target-time estimator $x \mapsto \hat{w}^\top(\bar{B}^\top x) + \mathbb{1}\left[x^\top \bar{B} \ge 0\right]\hat{\Delta}^\top x$, where $\bar{B}$, $\hat{w}$ and $\hat{\Delta}$ are measurable functions of (X, y), and $\bar{B}$ is an orthogonal matrix with $\operatorname{Col}\bar{B} = \operatorname{Col}\hat{B}$.*
3. *Adversary chooses an orthogonal matrix $A^* \in \mathbb{R}^{d\times k}$ satisfying $\operatorname{Col} A^* \subset E_k$, and a target time predictor $\theta^* \in S_{A^*}$.*
4. *Compute the representation $\hat{B}$ returned by* FROZENREP *under the setting of Lemma G.3 with the task distribution determined by A^*.*
5. *Target time samples are generated using $y \sim \mathcal{N}\left(X\theta^*, \sigma^2 I_n\right)$, and the player estimator is evaluated.*

Then, with probability at least $1-\delta$ over the draw of X, we have that

$$\min_{\bar{B},\hat{w},\hat{\Delta}} \max_{\substack{A^* \\ \theta^* \in S_{A^*}}} \mathbb{E}\left[\frac{1}{n_{\mathrm{T}}}\sum_{i=1}^{n_{\mathrm{T}}}\left(x_i^\top\theta^* - \hat{w}^\top(\bar{B}^\top x_i) - \mathbb{1}\left[x_i^\top\bar{B} > 0\right]\hat{\Delta}^\top x\right)^2\right] \gtrsim \frac{\sigma^2 d}{n_{\mathrm{T}}},$$

where the expectation is over the randomness in the labels $y \sim \mathcal{N}\left(X\theta^*, \sigma^2 I_{n_{\mathrm{T}}}\right)$.

Proof. Throughout this proof, we assume the high-probability event in Lemma B.2, which guarantees that with probability at least $1 - \delta$,

$$0.9\Sigma \preceq \frac{1}{n_{\mathrm{T}}} X^\top X \preceq 1.1\Sigma.$$

Let S denote the intersection of the d-dimensional unit sphere with E_k. For any infinite-dimensional vector f indexed by S and a measure μ on S, we define $\mu^\top f$ to denote integration with respect to μ, i.e.

$$\mu^\top f = \int_S f \, \mathrm{d}\mu.$$

We then define the infinite-dimensional vectors η and $\zeta^{(i)}$ for $i = 1, \ldots, d$, both indexed by S, as

$$\eta(x) \coloneqq (v^\top x)_+ \quad \text{and} \quad \zeta^{(i)}(x) \coloneqq \mathbb{1}\left[v^\top x > 0\right] x_i.$$

Recall that by Lemma G.3, $\operatorname{Col} \hat{B} = E_k$. With the preceding discussion in mind, we can equivalently think of the player as choosing $d + 1$ signed measures α, $\beta_1, \ldots, \beta_d$ over S, all with a common support of $2k$ elements in S, as a function of (X, y). The player then plays the predictor

$$x \mapsto \alpha^\top \eta(x) + \sum_{i=1}^{d} \beta_i^\top \zeta^{(i)}(x).$$

Then, if we let $T = \left\{\theta \;\middle|\; \|P_{E^*}\theta\|_2^2 \le 1/2\varepsilon, \|P_{E_k}\theta\|_2^2 \le 1\right\} \subseteq S_{A^*}$, then we have the inequality

$$\begin{aligned}
&\min_{\hat{w}, \hat{\Delta}} \max_{\substack{A^* \\ \theta^* \in S_{A^*}}} \mathbb{E}\left[\frac{1}{n_{\mathrm{T}}} \sum_{i=1}^{n_{\mathrm{T}}} \left(x_i^\top \theta^* - \hat{w}^\top (\hat{B}^\top x_i)_+ - \mathbb{1}\left[x_i^\top \hat{B} > 0\right] \hat{\Delta}^\top x\right)^2\right] \\
&\quad = \min_{\alpha, \beta_1, \ldots, \beta_d} \max_{\substack{A^* \\ \theta^* \in S_{A^*}}} \mathbb{E}\left[\frac{1}{n_{\mathrm{T}}} \sum_{i=1}^{n_{\mathrm{T}}} \left(x_i^\top \theta^* - \alpha^\top \eta(x) - \sum_{j=1}^{d} \beta_j^\top \zeta^{(j)}(x)\right)^2\right] \\
&\quad = \min_{\alpha, \beta_1, \ldots, \beta_d} \max_{\theta^* \in T} \mathbb{E}\left[\frac{1}{n_{\mathrm{T}}} \sum_{i=1}^{n_{\mathrm{T}}} \left(x_i^\top \theta^* - \alpha^\top \eta(x) - \sum_{j=1}^{d} \beta_j^\top \zeta^{(j)}(x)\right)^2\right],
\end{aligned}$$

where the second equality follows from the fact that the expression has no direct dependence on A^*.

The rest of the argument follows makes use of Fano's minimax bound, applied similarly as in the linear setting. First, note that for any $\theta \in T$, there exists $d + 1$ signed measures on S with common support size $2k$, which we denote with α^θ and $\beta_1^\theta, \ldots, \beta_d^\theta$, such that $x^\top \theta = (\alpha^\theta)^\top \eta(x) + \sum_j (\beta_j^\theta)^\top \zeta^{(j)}(x)$. Using this construction, we can lift the local packing set from the linear setting into this setting, where the associated seminorm (defined over vector measures) is given by

$$\|(\alpha, \beta_1, \ldots, \beta_d)\|^2 = \frac{1}{n_{\mathrm{T}}} \sum_{i=1}^{n_{\mathrm{T}}} \left(\alpha^\top \eta(x) + \sum_{j=1}^{d} \beta_j^\top \zeta^{(j)}(x)\right)^2.$$

More formally, let B be the unit ball under the Σ-norm, which we observe satisfies $(1/\sqrt{2})B \subseteq T$. Recall that there exists a $(1/2)$-packing of the unit Σ-ball in the Σ-norm with at least 2^d elements, via a standard volumetric argument - let this set be P. Equivalently, there exists a $(2\delta/\sqrt{0.9})$-packing of $(4\delta/\sqrt{0.9})B$ with at least 2^d elements, which we denote as P. Note that for any $\theta, \theta' \in P$,

$$\left\|(\alpha^\theta, \beta_1^\theta, \ldots, \beta_d^\theta) - (\alpha^{\theta'}, \beta_1^{\theta'}, \ldots, \beta_d^{\theta'})\right\|^2 = \frac{1}{n_{\mathrm{T}}} \|X(\theta - \theta')\|_2^2 \ge 0.9 \|\theta - \theta'\|_\Sigma^2 > 4\delta^2,$$

which implies that the vector measures corresponding to the elements of P are 2δ-separated in the associated seminorm. Furthermore, for any $\theta, \theta' \in P$,

$$\mathrm{KL}\left(\mathcal{N}\left(X\theta, \sigma^2 I_{n_{\mathrm{T}}}\right) \middle\| \mathcal{N}\left(X\theta', \sigma^2 I_{n_{\mathrm{T}}}\right)\right) = \frac{1}{2\sigma^2} \|X(\theta - \theta')\|_2^2 \lesssim \frac{n_{\mathrm{T}}}{2\sigma^2} \|\theta - \theta'\|_\Sigma^2 \le \frac{32 n_{\mathrm{T}}}{0.9\sigma^2} \delta^2.$$

Therefore, for any $\delta^2 \le 1/32$, which ensures that $P \subseteq 4\delta B \subseteq (1/\sqrt{2})B \subseteq T$, Fano's inequality implies that

$$\min_{\alpha,\beta_1,\ldots,\beta_d} \max_{\theta^* \in T} \mathbb{E}\left[\left\|(\alpha^{\theta^*}, \beta_1^{\theta^*}, \ldots, \beta_d^{\theta^*}) - (\alpha, \beta_1, \ldots, \beta_d)\right\|^2\right] \ge \delta^2\left(1 - \frac{32 n_{\mathrm{T}}}{0.9\sigma^2 \log 2}\delta^2 - \frac{1}{d}\right)$$
$$\ge \delta^2\left(\frac{3}{4} - \frac{32 n_{\mathrm{T}}}{0.9\sigma^2 d \log 2}\delta^2\right).$$

Therefore, as long as $n_{\mathrm{T}} \ge \left(\frac{0.9\log 2}{2}\right)\sigma^2 d$, we can set $\delta^2 = \left(\frac{0.9\log 2}{64}\right)\frac{\sigma^2 d}{n_{\mathrm{T}}}$, which implies that

$$\min_{\hat{w},\hat{\Delta}} \max_{\substack{A^* \\ \theta^* \in S_{A^*}}} \mathbb{E}\left[\frac{1}{n_{\mathrm{T}}}\sum_{i=1}^{n_{\mathrm{T}}}\left(x_i^\top \theta^* - \hat{w}^\top(\bar{B}^\top x_i) - \mathbb{1}\left[x_i^\top \bar{B} > 0\right]\hat{\Delta}^\top x\right)^2\right] \gtrsim \frac{\sigma^2 d}{n_{\mathrm{T}}}. \qquad \square$$

G.2.2 Adaptation Upper Bound

Having proven the minimax result for FROZENREP, we now proceed to prove a corresponding upper bound on the performance of ADAPTREP. To do so, we need to prove a result analogous to Lemma B.4 for the ReLU setting.

Before we proceed to the proof, we need to find proper generalizations for relevant objects in the proof of Lemma B.4. In particular, we recall the prominent use of the projector $P_{\Sigma B_0}$, which, loosely speaking, can be thought of as representing the "average" component of the signal that can be represented by a parameter in $\operatorname{Col} B_0$.

Recall that the inputs (which are element of $\mathbb{R}^d$) are sampled from a distribution p. This input distribution induces the $L_2(p)$-norm[9] on vector-valued functions of $\mathbb{R}^d$ and its associated inner product via

$$\|\zeta\|_{L_2} = \mathbb{E}\left[\|\zeta(x)\|_2^2\right] \quad \text{and} \quad \langle \zeta, \xi\rangle_{L_2} = \mathbb{E}\left[\zeta(x)^\top \xi(x)\right].$$

Now, let $\zeta : \mathbb{R}^d \to \mathbb{R}^p$ and $\xi : \mathbb{R}^d \to \mathbb{R}^q$ be two representation functions on $\mathbb{R}^d$. Then, we can define the *linear projector onto* ζ as the linear operator P_ζ taking representations $\mathbb{R}^d \to \mathbb{R}^q$ onto itself via

$$[P_\zeta \xi](x) \coloneqq \mathbb{E}\left[\xi(x)\zeta(x)^\top\right]\mathbb{E}\left[\zeta(x)\zeta(x)^\top\right]^\dagger \zeta(x).$$

We denote the corresponding orthogonal projection as $P_\zeta^\perp \xi \coloneqq \xi - P_\zeta \xi$. We note that these operators satisfy the orthogonality property

$$\mathbb{E}\left[[P_\zeta \xi](x)^\top [P_\zeta^\perp \xi](x)\right] = 0,$$

which can be easily verified algebraically.

To understand the operator P_ζ, consider the problem of approximating a linear function of ξ via a linear function of ζ, as measured via the input distribution p. More formally, for $v \in \mathbb{R}^q$, we want to find

$$w^* = \operatorname*{argmin}_{w \in \mathbb{R}^p} \left\|\xi(\cdot)^\top v - \zeta(\cdot)^\top w\right\|_{L_2}^2 \quad \text{and} \quad f^* = \zeta(\cdot)^\top w^*.$$

As the problem is differentiable and convex in w, we can simply use standard optimality conditions to find that

$$w^* = \mathbb{E}\left[\zeta(x)\zeta(x)^\top\right]^\dagger \mathbb{E}\left[\zeta(x)\xi(x)^\top\right] v \quad \text{and} \quad f^* = \zeta(\cdot)^\top w^*.$$

That is, $P_\zeta \xi$ is performing exactly the transformation required on ζ such that $[P_\zeta \xi(\cdot)]^\top v$ is the best approximation to $\xi(\cdot)^\top v$ via linear functions of ζ in L_2-norm. To further connect this construction to the linear setting, observe that if $\zeta(x) = B_0^\top x$ and $\xi(x) = x$, then for any $\theta, v \in \mathbb{R}^d$,

$$\left\langle \xi(\cdot)^\top v, [P_\zeta^\perp \xi](\cdot)^\top \theta\right\rangle_{L_2} = v^\top \Sigma^{1/2} P_{\Sigma B_0}^\perp \Sigma^{1/2} \theta,$$

and thus $P_\zeta \xi$ is indeed the desired generalization of the projection operators used in the proof of Lemma B.4. Having introduced the required mathematical tools, we now prove the corresponding transfer lemma for this setting.

[9] we write L_2 throughout as a shorthand for $L_2(p)$.

Lemma G.4 (Transfer Lemma, ReLU)**.** *Assume that $B_0 = [A_0, -A_0]$. Then, we have that*

$$\frac{1}{T}\left\|P^{\perp}_{\Sigma^{1/2}A_0}\Sigma^{1/2}A^*W^*\right\|_F^2 \lesssim \frac{1}{T}\sum_{i=1}^{T}\min_{w,\Delta}\mathbb{E}\left[\left(x^\top\theta_t^* - \rho_{B_0}(x)[w,\Delta]\right)^2\right].$$

Proof. This proof follows the outline of Lemma B.4, with all inner products computed with respect to $\langle\cdot,\cdot\rangle_{L_2}$. Throughout the proof, we will write P and $P^\perp$ as shorthand for the operators $P_{\phi_{B_0}}$ and $P^\perp_{\phi_{B_0}}$, respectively. First, we decompose the task-averaged population risk as for any choices of $(w_t), (\Delta_t)$ as

$$\begin{aligned}
&\frac{1}{T}\sum_{i=1}^{T}\left\|(\cdot)^\top\theta_t^* - \rho_{B_0}(\cdot)[w_t,\Delta_t]\right\|_{L_2}^2 \\
&\quad\gtrsim \frac{1}{T}\sum_{t\in[T]}\left\|\phi_{B^*}(\cdot)^\top w_t^* - [P\rho_{B_0}](\cdot)[w_t,\Delta_t]\right\|_{L_2}^2 \\
&\qquad + \frac{1}{T}\sum_{t\in[T]}\left\|\psi_{B^*}(\cdot)^\top\Delta_t^* - [P^\perp\rho_{B_0}](\cdot)[w_t,\Delta_t]\right\|_{L_2}^2 \\
&\qquad - \frac{1}{T}\sum_{t\in[T]}\left|\left\langle\phi_{B^*}(\cdot)^\top w_t^* - [P\rho_{B_0}](\cdot)[w_t,\Delta_t], \psi_{B^*}(\cdot)^\top\Delta_t^* - [P^\perp\rho_{B_0}](\cdot)[w_t,\Delta_t]\right\rangle_{L_2}\right| \\
&\quad\gtrsim \frac{1}{T}\sum_{t\in[T]}\left\|P^\perp_{\Sigma^{1/2}A_0}\Sigma^{1/2}A^*v_t^*\right\|_2^2 + \frac{1}{T}\sum_{t\in[T]}\left\|\psi_{B^*}(\cdot)^\top\Delta_t^* - [P^\perp\rho_{B_0}](\cdot)[w_t,\Delta_t]\right\|_{L_2}^2 \\
&\qquad - \frac{1}{T}\sum_{t\in[T]}\left|\left\langle\phi_{B^*}(\cdot)^\top w_t^* - [P\rho_{B_0}](\cdot)[w_t,\Delta_t], \psi_{B^*}(\cdot)^\top\Delta_t^* - [P^\perp\rho_{B_0}](\cdot)[w_t,\Delta_t]\right\rangle_{L_2}\right|.
\end{aligned}$$

The second inequality follows from noting that $\phi_{B^*}(x)^\top w_t^*$ is a linear function of x, and thus by Lemma G.2, the best linear predictor on ϕ is a linear function of x. Since the L_2-norm on linear functions of x is equivalent to the Σ-norm on the parameters, we obtain the inequality above[10].

Now, we proceed to prove a bound on the inner product above. For any fixed $t \in [T]$,

$$\begin{aligned}
&\left|\left\langle\phi_{B^*}(\cdot)^\top w_t^* - [P\rho_{B_0}](\cdot)[w,\Delta], \psi_{B^*}(\cdot)^\top\Delta_t^* - [P^\perp\rho_{B_0}](\cdot)[w,\Delta]\right\rangle_{L_2}\right| \\
&\quad\le \left|\left\langle\phi_{B^*}(\cdot)^\top w_t^* - [P\rho_{B^*}](\cdot)[w_t^*,\Delta_t^*], \psi_{B^*}(\cdot)^\top\Delta_t^* - [P^\perp\rho_{B^*}](\cdot)[w_t^*,\Delta_t^*]\right\rangle_{L_2}\right| \\
&\qquad + \left|\left\langle\phi_{B^*}(\cdot)^\top w_t^* - [P\rho_{B^*}](\cdot)[w_t^*,\Delta_t^*], [P^\perp\rho_{B^*}](\cdot)[w_t^*,\Delta_t^*] - [P^\perp\rho_{B_0}](\cdot)[w_t,\Delta_t]\right\rangle_{L_2}\right| \\
&\qquad + \left|\left\langle[P\rho_{B^*}](\cdot)[w_t^*,\Delta_t^*] - [P\rho_{B_0}](\cdot)[w_t,\Delta_t], \psi_{B^*}(\cdot)^\top\Delta_t^* - [P^\perp\rho_{B_0}](\cdot)[w_t,\Delta_t]\right\rangle_{L_2}\right| \\
&\quad\le \left|\left\langle\phi_{B^*}(\cdot)^\top w_t^* - [P\rho_{B^*}](\cdot)[w_t^*,\Delta_t^*], \psi_{B^*}(\cdot)^\top\Delta_t^* - [P^\perp\rho_{B^*}](\cdot)[w_t^*,\Delta_t^*]\right\rangle_{L_2}\right| \\
&\qquad + \left|\left\langle[P^\perp\phi_{B^*}](\cdot)^\top w_t^*, [P^\perp\rho_{B^*}](\cdot)[w_t^*,\Delta_t^*] - [P^\perp\rho_{B_0}](\cdot)[w_t,\Delta_t]\right\rangle_{L_2}\right| \\
&\qquad + \left|\left\langle[P\rho_{B^*}](\cdot)[w_t^*,\Delta_t^*] - [P\rho_{B_0}](\cdot)[w_t,\Delta_t], \psi_{B^*}(\cdot)^\top\Delta_t^* - [P^\perp\rho_{B_0}](\cdot)[w_t,\Delta_t]\right\rangle_{L_2}\right|,
\end{aligned}$$

where adding and subtracting $[P\rho_{B^*}](\cdot)[w_t^*,\Delta_t^*]$ and $[P^\perp\rho_{B^*}](\cdot)[w_t^*,\Delta_t^*]$ in the first and second arguments of the inner product, respectively, results in the first inequality. The second inequality then uses the orthogonality properties of P and $P^\perp$ in the second term. Furthermore, observe that

$$\begin{aligned}
&\left|\left\langle\phi_{B^*}(x)^\top w_t^* - [P\rho_{B^*}](x)[w_t^*,\Delta_t^*], \psi_{B^*}(x)^\top\Delta_t^* - [P^\perp\rho_{B^*}](x)[w_t^*,\Delta_t^*]\right\rangle_{L_2}\right| \\
&\quad= \left|\left\langle[P^\perp\phi_{B^*}](x)^\top w_t^* - [P\psi_{B^*}](x)^\top\Delta_t^*, [P^\perp\psi_{B^*}](x)^\top\Delta_t^* - [P\phi_{B^*}](x)[w_t^*,\Delta_t^*]\right\rangle_{L_2}\right| \\
&\quad= \left|\left\langle\phi_{B^*}(x)^\top w_t^*, \psi_{B^*}(x)^\top\Delta_t^*\right\rangle_{L_2}\right| = |w_t^*A^*\Sigma\delta_t^*| \\
&\quad= 0,
\end{aligned}$$

[10]This is exactly the argument used in Lemma B.4.

where we have used the algebraically-verifiable fact that $[P\rho_{B^*}](\cdot)[w,\Delta] = [P\phi_{B^*}](x)^\top + [P\psi_{B^*}](x)[\Delta]$, and that $[P^\perp\rho_{B^*}][w,\Delta]$ decomposes in a similar fashion. Finally, since $\phi_B^*(x)^\top w_t^*$ is linear, we can apply Lemma G.2, and thus

$$\left\|[P^\perp\phi_{B^*}](x)^\top w_t^*\right\|_{L_2} = \left\|P^\perp_{\Sigma^{1/2}A_0}\Sigma^{1/2}A^*v_t^*\right\|_2.$$

From here, we note that we have an analogous quadratic inequality to that of Lemma B.4 in the terms

$$\frac{1}{T}\sum_{t\in[T]}\left\|P^\perp_{\Sigma^{1/2}A_0}\Sigma^{1/2}A^*v_t^*\right\|_2^2 \quad \text{and} \quad \frac{1}{T}\sum_{t\in[T]}\left\|\psi_{B^*}(\cdot)^\top\Delta_t^* - [P^\perp\rho_{B_0}](\cdot)[w_t,\Delta_t]\right\|_{L_2}^2$$

upon applying Cauchy-Schwarz as before. Furthermore, note that via orthogonality, we have the Pythagorean identity

$$\|P\rho\|_{L_2}^2 + \left\|P^\perp\rho\right\|_{L_2}^2 \le \|\rho\|_{L_2}^2,$$

and thus by following the exact same algebraic argument as in Lemma B.4 of using Proposition I.2,

$$\begin{aligned}\frac{1}{T}\left\|P^\perp_{\Sigma^{1/2}A_0}\Sigma^{1/2}A^*V^*\right\|_F^2 &= \frac{1}{T}\sum_{t\in[T]}\left\|P^\perp_{\Sigma^{1/2}A_0}\Sigma^{1/2}A^*v_t^*\right\|_2^2 \\ &\le \frac{1}{T}\sum_{t\in[T]}\mathbb{E}\left[(x^\top\theta_t^* - \rho_{B_0}(x)[w_t,\Delta_t])^2\right].\end{aligned}$$

Since the result holds for any (w_t) and (Δ_t), it holds for the minimizers. □

Lemma G.1 (Adaptation target performance, ReLU)**.** *Let $B_0 = [A_0, -A_0]$, and fix a $\theta^* \in S_{A^*}$. Consider a learner which solves*

$$\min_B \min_{\substack{(w_t,\Delta_t)\\ \|\Delta_t\|_F\le 1}} \frac{1}{n_{\mathrm{S}}}\sum_{t\in[T]}\|\rho_B(X)[w,\Delta] - y\|_2^2.$$

during source training, and

$$\min_{\substack{w,\Delta\\ \|\Delta\|_F\le 1}} \frac{1}{n_{\mathrm{T}}}\|\rho_{B_0}(X)[w,\Delta] - y\|_2^2$$

during target training, where B_0 is the representation obtained from source training. Then, with probability at least $1-\delta$,

$$\begin{aligned}\frac{1}{n_{\mathrm{T}}}\|\rho_{B_0}(X)[w,\Delta] - X\theta^*\|_2^2 \lesssim \left\|P^\perp_{\Sigma^{1/2}A_0}\Sigma^{1/2}A^*\right\|_2^2 &+ \frac{\sigma^2 k}{n_{\mathrm{T}}}\left(1+\log\frac{1}{\delta}\right) \\ &+ \frac{\sigma}{\sqrt{n_{\mathrm{T}}}}\sqrt{k\,\mathrm{tr}\,\Sigma\left(1+\log\frac{1}{\delta}\right)}.\end{aligned}$$

In particular, with $k = \Theta(1)$ and $\varepsilon = k/d$, and assuming access to infinite per-task samples and tasks during source training, the learner achieves target loss bounded as

$$\frac{1}{n_{\mathrm{T}}}\|\rho_{B_0}(X)[w,\Delta] - X\theta^*\|_2^2 \lesssim \frac{\sigma^2}{n_{\mathrm{T}}}\left(1+\log\frac{1}{\delta}\right) + \frac{\sigma}{\sqrt{n_{\mathrm{T}}}}\sqrt{1+\log\frac{1}{\delta}}$$

with probability at least $1-\delta$ over the draw of target samples.

Proof. Throughout the proof, we instantiate the high-probability event in Lemma B.2, which guarantees that with probability at least $1-\delta/9$,

$$0.9\Sigma \preceq \frac{1}{n_{\mathrm{T}}}X^\top X \preceq 1.1\Sigma.$$

By forming the least-squares basic inequality, we have that

$$\begin{aligned}
&\frac{1}{n_{\mathrm{T}}} \left\| \rho_{B_0}(X)[w, \Delta] - X(A^* v^* + \delta^*) \right\|_2^2 \\
&\quad \lesssim \underbrace{\frac{1}{n_{\mathrm{T}}} \left\| P_{X A_0}^{\perp} X A^* v^* \right\|_2^2}_{=:\zeta} + \frac{1}{n_{\mathrm{T}}} \left| \langle z, \rho_{B_0}(X)[w, \Delta] - P_{X A_0} X A^* v^* - X \delta^* \rangle \right|.
\end{aligned}$$

Then, since $P_{X A_0} X A^* v^*$ is in the span of $\phi_{B_0}(X)$, we can bound the right-hand side of the basic inequality by

$$\begin{aligned}
&\zeta + \frac{1}{n_{\mathrm{T}}} \left| \left\langle z, \underbrace{P_{\phi_{B_0}(X)} \left(\rho_{B_0}(X)[w, \Delta] - P_{X A_0} X A^* v^* - X \delta^* \right)}_{=:T_{\text{low-rank}}} \right\rangle \right| \\
&\qquad + \frac{1}{n_{\mathrm{T}}} \left| \left\langle z, P_{\phi_{B_0}(X)}^{\perp} \left(\psi_{B_0}(X)[\Delta] - X \delta^* \right) \right\rangle \right|.
\end{aligned}$$

Furthermore, we can lower bound the left-hand side by

$$\frac{1}{n_{\mathrm{T}}} \left\| \rho_{B_0}(X)[w, \Delta] - X(A^* v^* + \delta^*) \right\|_2^2 \geq \frac{1}{n_{\mathrm{T}}} \left\| T_{\text{low-rank}} \right\|_2^2 + \frac{2}{n_{\mathrm{T}}} \left\langle P_{X A_0}^{\perp} X A^* v^*, T_{\text{low-rank}} \right\rangle.$$

Note that $\phi_{B_0}(X)$ is a matrix with rank $\leq 2k$, and therefore, $P_{\phi_{B_0}}(X) z$ is a chi-squared random variable, with at most $2k$ degrees of freedom. Subsequently, by applying known bounds on chi-squared random variables, we have that with probability at least $1 - (4/9)\delta$,

$$\begin{aligned}
\frac{1}{n_{\mathrm{T}}} \left\| T_{\text{low-rank}} \right\|_2^2 \leq \zeta &+ \left(\sqrt{\zeta} + \frac{2 \sigma k}{\sqrt{n_{\mathrm{T}}}} \sqrt{1 + \log \frac{1}{\delta}} \right) \frac{1}{\sqrt{n_{\mathrm{T}}}} \left\| T_{\text{low-rank}} \right\|_2 \\
&+ \frac{1}{n_{\mathrm{T}}} \left| \left\langle z, P_{\phi_{B_0}(X)}^{\perp} \left(\psi_{B_0}(X)[\Delta] - X \delta^* \right) \right\rangle \right|,
\end{aligned}$$

which thus implies via Proposition I.2 that

$$\begin{aligned}
&\frac{1}{n_{\mathrm{T}}} \left\| \rho_{B_0}(X)[w, \Delta] - X(A^* v^* + \delta^*) \right\|_2^2 \\
&\quad \lesssim \zeta + \frac{\sigma^2 k}{n_{\mathrm{T}}} \left(1 + \log \frac{1}{\delta} \right) + \frac{1}{n_{\mathrm{T}}} \left| \left\langle z, P_{\phi_{B_0}(X)}^{\perp} \left(\psi_{B_0}(X)[\Delta] - X \delta^* \right) \right\rangle \right|.
\end{aligned}$$

To bound the final term, note that

$$\begin{aligned}
&\frac{1}{n_{\mathrm{T}}} \left| \left\langle z, P_{\phi_{B_0}(X)}^{\perp} \left(\psi_{B_0}(X)[\Delta] - X \delta^* \right) \right\rangle \right| \\
&\quad \leq \frac{1}{\sqrt{n_{\mathrm{T}}}} \left[\left\| \frac{1}{\sqrt{n_{\mathrm{T}}}} \psi_{B_0}(X)^{\top} P_{\phi_{B_0}(X)}^{\perp} z \right\|_2 + \left\| \frac{1}{\sqrt{n_{\mathrm{T}}}} X^{\top} P_{\phi_{B_0}(X)}^{\perp} z \right\|_2 \right],
\end{aligned}$$

by applying the norm bounds on Δ and δ^*, and thus by applying the Hanson-Wright inequality, with probability at least $1 - (4/9)\delta$,

$$\begin{aligned}
&\frac{1}{n_{\mathrm{T}}} \left| \left\langle z, P_{\phi_{B_0}(X)}^{\perp} \left(\psi_{B_0}(X)[\Delta] - X \delta^* \right) \right\rangle \right| \\
&\quad \leq \frac{\sigma}{\sqrt{n_{\mathrm{T}}}} \left[\sqrt{\frac{1}{n_{\mathrm{T}}} \operatorname{tr} \psi_{B_0}(X)^{\top} \psi_{B_0}(X)} + \sqrt{\frac{1}{n_{\mathrm{T}}} \operatorname{tr} X^{\top} X} \right] \sqrt{1 + \log \frac{1}{\delta}} \\
&\quad \lesssim \frac{\sigma}{\sqrt{n_{\mathrm{T}}}} \sqrt{k \operatorname{tr} \Sigma \left(1 + \log \frac{1}{\delta} \right)}.
\end{aligned}$$

Therefore, since $\left\| v^* \right\|_2 \leq 1$, we have an overall bound of

$$\begin{aligned}
\frac{1}{n_{\mathrm{T}}} \left\| \rho_{B_0}(X)[w, \Delta] - X \theta^* \right\|_2^2 \lesssim &\left\| P_{\Sigma^{1/2} A_0}^{\perp} \Sigma^{1/2} A^* \right\|_2^2 + \frac{\sigma^2 k}{n_{\mathrm{T}}} \left(1 + \log \frac{1}{\delta} \right) \\
&+ \frac{\sigma}{\sqrt{n_{\mathrm{T}}}} \sqrt{k \operatorname{tr} \Sigma \left(1 + \log \frac{1}{\delta} \right)}.
\end{aligned}$$

To prove the second part of the statement, note that with infinite source tasks and samples, global optimality with respect to the source loss together with Lemma G.4 implies that

$$\frac{1}{T}\left\|P^{\perp}_{\Sigma^{1/2}A_0}\Sigma^{1/2}A^*V^*\right\|_F^2 = 0 \implies \left\|P^{\perp}_{\Sigma^{1/2}A_0}\Sigma^{1/2}A^*\right\|_2 = 0,$$

since $(k/T)\left[{V^*}^\top V^*\right] = I_k$. Furthermore, with the choice of d and ε, $\operatorname{tr}\Sigma = \Theta(1)$. Therefore, we obtain the second claim. □

H A Performance Bound for Projected Gradient Descent

In this section, we provide a performance bound for projected gradient descent on the objective

$$\mathcal{L}(\theta) \coloneqq \frac{1}{n} \sum_{i \in [n]} \ell(g_\theta(x_i), y_i)$$

for $\theta \in \Theta$, where $\Theta \subseteq \mathbb{R}^q$. We assume that Θ is norm-bounded by D, and that Θ is convex and contains 0. Furthermore, we assume that g_θ is twice-differentiable as a function of θ.

Key to the performance bound that we will demonstrate is that g_θ is "approximately linear" in the parameter θ, which we formally define below. Under this assumption, we demonstrate $\mathcal{L}$ is approximately convex over Θ if $\mathcal{L}$ is Lipschitz as a function of the vector of predictions and ℓ is convex in the first argument. Therefore, with slight modifications to the online analysis of projected gradient descent, we obtain the desired performance bound.

Following the discussion above, we make the following assumptions:

Assumption H.1 (Approximate linearity)**.** *There exists β and L such that*

$$\sup_{\theta \in \Theta} \frac{1}{n} \sum_{i \in [n]} \left\| \nabla_\theta^2 g_\theta(x_i) \right\|_2^2 \leq \beta^2 \quad \textit{and} \quad \frac{1}{n} \sum_{i \in [n]} \left\| \nabla_\theta g_0(x) \right\|_2^2 \leq L^2.$$

Assumption H.2 (Assumptions on ℓ)**.** *We assume that ℓ is convex in the first argument. Furthermore, if $\mathcal{L}$ is viewed as a function of the vector of predictions $g_\theta(X)$, then we have that*

$$\left\| \nabla_g \mathcal{L}(g_\theta(X)) \right\|_2^2 \leq \frac{\alpha^2}{n}$$

for any $\theta \in \Theta$, i.e. $\mathcal{L}$ is $(\alpha/\sqrt{n})$-Lipschitz as a function of the vector of predictions.

Note that we abuse notation in Assumption H.2, using $\mathcal{L}$ to reference both the function of the parameter, and of the vector of predictions. Given these two assumptions, we now proceed to demonstrate that the loss landscape of $\mathcal{L}$ has several desirable properties.

Lemma H.1 (Approximate convexity in Θ)**.** *Let $\theta_1, \theta_2 \in \Theta$. Then,*

$$\langle \nabla_\theta \mathcal{L}(\theta_1), \theta_2 - \theta_1 \rangle \leq \mathcal{L}(\theta_2) - \mathcal{L}(\theta_1) + 4\alpha\beta D^2.$$

Proof. Note that $\mathcal{L}$ is a convex function of the vector of predictions $g_\theta(X)$, as it is a sum of convex functions by Assumption H.2. Therefore,

$$\mathcal{L}(\theta_2) - \mathcal{L}(\theta_1) \geq \langle \nabla_g \mathcal{L}(g_{\theta_1}(X)), g_{\theta_2}(X) - g_{\theta_1}(X) \rangle .$$

Furthermore, by the chain rule,

$$\langle \nabla_\theta \mathcal{L}(\theta_1), \theta_2 - \theta_1 \rangle = \langle \nabla_g \mathcal{L}(g_{\theta_1}(X)), [\nabla_\theta g_{\theta_1}(X)](\theta_2 - \theta_1) \rangle .$$

Putting the two statements together and applying Cauchy-Schwarz,

$$\begin{aligned}
&\mathcal{L}(\theta_2) - \mathcal{L}(\theta_1) - \langle \nabla_\theta \mathcal{L}(\theta_1), \theta_2 - \theta_1 \rangle \\
&\quad \geq \langle \nabla_g \mathcal{L}(g_{\theta_1}(X)), g_{\theta_2}(X) - g_{\theta_1}(X) - [\nabla_\theta g_{\theta_1}(X)](\theta_2 - \theta_1) \rangle \\
&\quad \geq - \left\| \nabla_g \mathcal{L}(g_{\theta_1}(X)) \right\|_2 \left\| g_{\theta_2}(X) - g_{\theta_1}(X) - [\nabla_\theta g_{\theta_1}(X)](\theta_2 - \theta_1) \right\|_2 \\
&\quad \geq - \frac{\alpha}{\sqrt{n}} \left\| g_{\theta_2}(X) - g_{\theta_1}(X) - [\nabla_\theta g_{\theta_1}(X)](\theta_2 - \theta_1) \right\|_2
\end{aligned}$$

Now, by Taylor's theorem, there exists $\bar{\theta} = \lambda\theta_1 + (1-\lambda)\theta_2$ such that

$$g_{\theta_2}(X) - g_{\theta_1}(X) - \nabla_\theta g_{\theta_1}(X)(\theta_2 - \theta_1) = [(\theta_2 - \theta_1)^\top \nabla_\theta^2 g_{\bar{\theta}}(x_i)(\theta_2 - \theta_1)]_{i \in [n]}.$$

Consequently, by rearranging,

$$\begin{aligned}
\langle \nabla_\theta \mathcal{L}(\theta_1), \theta_2 - \theta_1 \rangle &\leq \mathcal{L}(\theta_2) - \mathcal{L}(\theta_1) + \frac{\alpha}{\sqrt{n}} \sqrt{\sum_{i \in [n]} [(\theta_2 - \theta_1)^\top \nabla_\theta^2 g_{\bar{\theta}}(x_i)(\theta_2 - \theta_1)]^2} \\
&\leq \mathcal{L}(\theta_2) - \mathcal{L}(\theta_1) + 4\alpha\beta D^2.
\end{aligned}$$

□

Lemma H.2 (Gradient bound in Θ)**.** *For any $\theta \in \Theta$, we have that*

$$\|\nabla_\theta \mathcal{L}(\theta)\|_2^2 \lesssim \alpha^2(L^2 + \beta^2 D^2)$$

Proof. Using the Lipschitz assumption in Assumption H.2,

$$\begin{aligned}\|\nabla_\theta \mathcal{L}(\theta)\|_2^2 &= \left\|[\nabla_\theta g_\theta(X)]^\top \nabla_g \mathcal{L}(g_\theta(X))\right\|_2^2 \leq \|\nabla_\theta g_\theta(X)\|_F^2 \, \|\nabla_g \mathcal{L}(g_\theta(X))\|_2^2 \\ &\leq \frac{\alpha^2}{n} \sum_{i\in[n]} \|\nabla_\theta g_\theta(x_i)\|_2^2\end{aligned}$$

Now, by integrating, we have that for any fixed $i \in [n]$,

$$\nabla_\theta g_\theta(x_i) = \nabla_\theta g_0(x_i) + \left[\int_0^1 \nabla_\theta^2 g_{\alpha\theta}(x_i) \, \mathrm{d}\alpha\right] \theta.$$

Therefore,

$$\begin{aligned}\|\nabla_\theta \mathcal{L}(\theta)\|_2^2 &\lesssim \frac{\alpha^2}{n} \sum_{i\in[n]} \|\nabla_\theta g_0(x_i)\|_2^2 + \frac{\alpha^2}{n} \left\|\int_0^1 \nabla_\theta^2 g_{\alpha\theta}(x_i) \, \mathrm{d}\alpha\right\|_2^2 \|\theta\|_2^2 \\ &\leq \frac{\alpha^2}{n} \sum_{i\in[n]} \|\nabla_\theta g_{\theta_0}(x_i)\|_2^2 + \frac{\alpha^2 D^2}{n} \int_0^1 \frac{1}{n} \sum_{i\in[n]} \left\|\nabla_\theta^2 g_{\alpha\theta}(x_i)\right\|_2^2 \, \mathrm{d}\alpha \\ &\leq \alpha^2(L^2 + \beta^2 D^2).\end{aligned}$$

□

Having proven the results above, we now proceed to the main claim of this section.

Theorem H.1 (Bound on PGD performance)**.** *Assume we run projected gradient descent on $\mathcal{L}$ with constraint set Θ for T_{PGD} iterations with step size η given by*

$$\eta = \frac{1}{\sqrt{T_{\mathrm{PGD}}}} \left(\frac{D}{\alpha\sqrt{L^2 + \beta^2 D^2}}\right).$$

Let $(\theta_t)_{t\in[T_{\mathrm{PGD}}]}$ denote the sequence of PGD iterates obtained, where $\theta_0 = 0$. Then, for any $\theta \in \Theta$,

$$\min_t \mathcal{L}(\theta_t) - \mathcal{L}(\theta) \lesssim \alpha\beta D^2 + \alpha D \sqrt{\frac{L^2 + \beta^2 D^2}{T_{\mathrm{PGD}}}}.$$

Proof. For any $t \in [T_{\mathrm{PGD}}]$,

$$\begin{aligned}r_{t+1}^2 &= \|\theta_{t+1} - \theta\|_2^2 \\ &\leq \|\theta_t - \eta\nabla_\theta \mathcal{L}(\theta_t) - \theta\|_2^2 \\ &= r_t^2 + 2\eta \left\langle \nabla_\theta \mathcal{L}(\theta_t), \theta - \theta_t \right\rangle + \eta^2 \|\nabla_\theta \mathcal{L}(\theta_t)\|_2^2 \\ &\leq r_t^2 + 2\eta \left[\mathcal{L}(\theta_0 + \delta) - \mathcal{L}(\theta_0 + \delta_t) + 4\alpha\beta D^2\right] + \eta^2\alpha^2 \left[L^2 + \beta^2 D^2\right],\end{aligned}$$

where the first inequality follows from the nonexpansive property of projections onto convex sets. Furthermore, the last inequality makes use of the approximate convexity property from Lemma H.1 and the gradient bound over Θ from Lemma H.2. Therefore, via telescoping,

$$r_{T_{\mathrm{PGD}}}^2 \leq r_0^2 + 2\eta \left[\sum_{t=0}^{T_{\mathrm{PGD}}-1} \mathcal{L}(\theta) - \mathcal{L}(\theta_t)\right] + 8\eta T_{\mathrm{PGD}} \alpha\beta D^2 + \eta^2 T_{\mathrm{PGD}} \alpha^2 [L^2 + \beta^2 D^2],$$

or by rearranging,

$$\begin{aligned}\frac{r_0^2 - r_{T_{\mathrm{PGD}}}^2}{2\eta T_{\mathrm{PGD}}} + 4\alpha\beta D^2 + \frac{\eta\alpha^2}{2} \left[L^2 + \beta^2 D^2\right] &\geq \frac{1}{T_{\mathrm{PGD}}} \sum_{t=0}^{T_{\mathrm{PGD}}-1} \mathcal{L}(\theta_0 + \delta_t) - \mathcal{L}(\theta_0 + \delta) \\ &\geq \min_{t=0,\ldots,T_{\mathrm{PGD}}-1} \mathcal{L}(\theta_0 + \delta_t) - \mathcal{L}(\theta_0 + \delta).\end{aligned}$$

Now, using the choice of step size, observe that

$$\begin{aligned}\frac{r_0^2 - r_{T_{\mathrm{PGD}}}^2}{2\eta T_{\mathrm{PGD}}} + \frac{\eta}{2}\left[\alpha^2 L^2 + \beta^2 D^2\right] &\leq \frac{D^2}{2\eta T_{\mathrm{PGD}}} + \frac{\eta\alpha^2}{2}[L^2 + \beta^2 D^2] \\ &= \alpha D\sqrt{\frac{L^2 + \beta^2 D^2}{T_{\mathrm{PGD}}}},\end{aligned}$$

from which the desired claim easily follows. □

I Technical Lemmas

Proposition I.1 (Du et al. (2020), Lemma A.5)**.** *Let $\mathcal{O}^{d_1 \times d_2}$ be the set of matrices in $\mathbb{R}^{d_1 \times d_2}$ with orthonormal columns, $d_1 \geq d_2$. Then, there exists an ε-covering of $\mathcal{O}^{d_1 \times d_2}$ with at most $(6\sqrt{d_2}/\varepsilon)^{d_1 d_2}$ elements.*

Proposition I.2 (Solving quadratic inequalities)**.** *Assume that $ax^2 \leq bx + c$ for $a, b, c > 0$. Then, $bx + c \lesssim (b^2/a) + c$.*

Proof. Since $a > 0$, the solution set to the inequality is given by the interval $[r_1, r_2]$, where r_1 and r_2 are the roots of $ax^2 - bx - c$. By the quadratic formula, the larger root r_2 is given by

$$r_2 = \frac{b + \sqrt{b^2 + 4ac}}{2a} \leq \frac{b}{a} + \sqrt{\frac{c}{a}}.$$

Therefore,

$$x \leq r_2 \leq \frac{b}{a} + \sqrt{\frac{c}{a}} \implies bx + c \leq \frac{b^2}{a} + \left(\frac{b}{\sqrt{a}}\right)\sqrt{c} + c \lesssim \frac{b^2}{a} + c,$$

where the last inequality makes use of the Cauchy-Schwarz inequality. □

Corollary I.1. *Let X and Y be random variables.*

$$\mathbb{E}\left[X^2\right] \lesssim \mathbb{E}\left[(X+Y)^2\right] + \mathbb{E}\left[Y^2\right]$$

Proof. We have that

$$\mathbb{E}\left[(X+Y)^2\right] = \mathbb{E}\left[X^2\right] + 2\mathbb{E}\left[XY\right] + \mathbb{E}\left[Y^2\right] \geq \mathbb{E}\left[X^2\right] - 2\mathbb{E}\left[|XY|\right].$$

Therefore, by applying Cauchy-Schwarz,

$$\mathbb{E}\left[X^2\right] \leq \mathbb{E}\left[(X+Y)^2\right] + 2\mathbb{E}\left[|XY|\right] \leq \mathbb{E}\left[(X+Y)^2\right] + 2\sqrt{\mathbb{E}\left[X^2\right]\mathbb{E}\left[Y^2\right]}.$$

Finally, by applying Proposition I.2,

$$\mathbb{E}\left[X^2\right] \lesssim \mathbb{E}\left[(X+Y)^2\right] + \mathbb{E}\left[Y^2\right]. \qquad \square$$

Proposition I.3. *Let A, B be matrices with compatible dimensions, and assume that $\operatorname{rank} A = r > 0$. Then,*

$$\|P_A B\|_F \leq \frac{1}{\sigma_r(A)} \left\|A^\top B\right\|_F.$$

Proof. Let $(U, \Sigma, V^\top)$ be the compact singular value decomposition of A, *i.e.* we only retain positive singular values in Σ. By rotational invariance,

$$\left\|A^\top B\right\|_F^2 = \left\|V\Sigma U^\top B\right\|_F^2 = \left\|\Sigma U^\top B\right\|_F^2.$$

Furthermore, by definition,

$$\left\|\Sigma U^\top B\right\|_F^2 = \sum_i \left\|\Sigma U^\top B e_i\right\|_2^2 \geq \sigma_r^2(\Sigma) \sum_i \left\|U^\top B e_i\right\|_2^2 = \sigma_r^2(A) \left\|U^\top B\right\|_F^2.$$

Finally, by applying rotational invariance once more,

$$\left\|A^\top B\right\|_F^2 \geq \sigma_r^2(A) \left\|UU^\top B\right\|_F^2 = \sigma_r^2(A) \left\|P_A B\right\|_F^2,$$

from which the desired claim follows. □

Proposition I.4. *Let $\lambda, \gamma > 0$, and fix a vector y. Then,*

$$\min_{\substack{A,x \\ Ax=y}} \frac{\lambda}{2}\|A\|_F^2 + \frac{\gamma}{2}\|x\|_2^2 = \sqrt{\lambda\gamma}\,\|y\|_2.$$

Proof. We proceed by cases. If $y = 0$, then the result is trivial.

Otherwise, if $y \neq 0$, note that $x^* \neq 0$. Now, for any fixed $x \neq 0$, the minimizing choice for A is $zx^\top$ for some z. To see this, observe that if A is not rank-1, then we can achieve a lower Frobenius norm by reducing its rank. Consequently, for a given x, the minimizing choice for A is $yx^\top / \|x\|_2^2$ necessarily. Therefore,

$$\min_{\substack{A,x \\ Ax=y}} \frac{\lambda}{2}\|A\|_F^2 + \frac{\gamma}{2}\|x\|_2^2 = \min_x \frac{\lambda}{2}\left(\frac{\|y\|_2^2}{\|x\|_2^2}\right) + \frac{\gamma}{2}\|x\|_2^2 = \min_{z>0} \frac{\lambda}{2}\left(\frac{\|y\|_2^2}{z}\right) + \frac{\gamma z}{2}.$$

This final optimization problem is convex in z – using first-order optimality conditions, we can thus easily see that $z^* = \sqrt{\lambda/\gamma}\,\|y\|_2$, and therefore

$$\min_{\substack{A,x \\ Ax=y}} \frac{\lambda}{2}\|A\|_F^2 + \frac{\gamma}{2}\|x\|_2^2 = \sqrt{\lambda\gamma}\,\|y\|_2 . \qquad \square$$

Proposition I.5 (Expectation bound on empirical spectral norm)**.** *Let $X \in \mathbb{R}^{n\times d}$ be a matrix with rows drawn i.i.d. from a zero-mean distribution with covariance Σ. Furthermore, assume that the whitened distribution is ρ^2-sub-Gaussian. Then, whenever $n \gtrsim \rho^4 d$,*

$$\mathbb{E}\left[\lambda_{\max}\left(\frac{X^\top X}{n}\right)\right] \lesssim \|\Sigma\|_2 .$$

Proof. By Weyl's inequality,

$$\mathbb{E}\left[\lambda_{\max}\left(\frac{X^\top X}{n}\right)\right] \leq \|\Sigma\|_2 + \mathbb{E}\left[\left|\lambda_{\max}\left(\frac{X^\top X}{n}\right) - \Sigma\right|\right] \leq \|\Sigma\|_2 + \mathbb{E}\left[\left\|\frac{X^\top X}{n} - \Sigma\right\|_2\right].$$

Thus, by applying the result in Vershynin (2017, Theorem 4.4.1), we have that as long as $n_{\mathrm{S}} \gtrsim \rho^4 d$,

$$\mathbb{E}\left[\lambda_{\max}\left(\frac{X^\top X}{n}\right)\right] \lesssim \|\Sigma\|_2 . \qquad \square$$

Proposition I.6 (Gaussian complexity chain rule, Tripuraneni et al. (2020b), Theorem 7)**.** *Assume that $\mathcal{F}$ is a class of functions $\mathbb{R}^k \to \mathbb{R}$ such that every $f \in \mathcal{F}$ is L-Lipschitz in the L_2-norm. Furthermore, assume that Φ is a class of functions $\mathbb{R}^d \to \mathbb{R}^k$ such that for any $\phi \in \Phi$, $\phi(x)$ is norm-bounded by D for any x in the support of the input distribution. Then, we have the bound*

$$\frac{1}{T}\mathcal{G}_n(\mathcal{F}^{\otimes T} \circ \Phi) \leq \frac{8D}{(nT)^2} + 128\left(\frac{L}{T}\mathcal{G}_n(\Phi) + \mathbb{E}\left[\sup_{Z\in\mathcal{Z}} \mathcal{G}_Z(\mathcal{F})\right]\right)\log(nT),$$

where $\mathcal{Z}$ is the random set $\{(\phi(x_{i_1}), \ldots, \phi(x_{i_n})) \mid i_1, \ldots, i_n \in [nT]\}$ and $\mathcal{G}_Z(\mathcal{F})$ is the empirical Gaussian complexity on samples Z. Note that the inner expectation is over the nT input samples, and that we have assumed that all input samples come from a single distribution.

Proposition I.7 (Tripuraneni et al. (2020b), Lemma 6)**.** *Let $h, h^* : \mathbb{R}^d \to \mathbb{R}^k$ be representation functions, and define*

$$\Lambda(h, h^*) \coloneqq \mathbb{E}\left[h^*(x)h^*(x)^\top\right] - \mathbb{E}\left[h^*(x)h(x)^\top\right]\left(\mathbb{E}\left[h(x)h(x)^\top\right]\right)^\dagger \mathbb{E}\left[h(x)h^*(x)^\top\right].$$

Then, $\inf_v \mathbb{E}\left[(h(x)^\top v - h^*(x)^\top v^*)^2\right] = (v^*)^\top \Lambda(h, h^*) v^*$. *Furthermore, if*

$$\sigma_{\min}(\mathbb{E}\left[h(x)h(x)^\top\right]) \geq c_1 > 0 \quad \textit{and} \quad \sigma_{\max}(\mathbb{E}\left[h^*(x)h^*(x)^\top\right]) \leq c_2,$$

then this infimum is achieved within the ball of radius $\|v^*\|_2 \sqrt{c_2/c_1}$.

Proof. The calculation of the infimum is provided in Tripuraneni et al. (2020b), and is thus omitted. However, we prove the sharper radius bound below.

Define $F_{h,h'} \coloneqq \mathbb{E}\left[h(x)h'(x)^\top\right]$, so that $\Lambda(h, h^*) = F_{h^*,h^*} - F_{h^*,h}F_{h,h}^\dagger F_{h,h^*}$. Then, since $\Lambda(h, h^*) \succeq 0$, and recalling that the infimum is achieved at $v = F_{h,h}^\dagger F_{h,h^*} v^*$,

$$\left\|F_{h,h}^\dagger F_{h,h^*} v^*\right\|_2^2 \leq \frac{1}{c_1}\left\|F_{h,h}^{1/2} F_{h,h}^\dagger F_{h,h^*} v^*\right\|_2^2 \leq \frac{1}{c_1}\left\|F_{h^*,h^*}^{1/2} v\right\|_2^2 \leq \frac{c_1}{c_2}\|v^*\|_2^2 ,$$

as desired. $\square$