A Missing Proofs

A.1 Connection to discrepancy measure

In this section, we discuss how our assumption relates to discrepancy assumptions. Consider $\mathcal{Y}$-discrepancy that measures the maximum absolute distance between the loss function: $\operatorname{dist}(\mathcal{D}_1, \mathcal{D}_2) \coloneqq \sup_{h \in \mathcal{H}} |\mathcal{L}_{\mathcal{D}_1}(h) - \mathcal{L}_{\mathcal{D}_2}(h)|$, where $\mathcal{D}_1$ and $\mathcal{D}_1$ represents the source domain and target domain and $\mathcal{L}_{\mathcal{D}_1}$ and $\mathcal{L}_{\mathcal{D}_2}$ are expected loss for two domains.

Note that under the GLM assumption, the L_2 distance in unknown parameters resembles the discrepancy using square loss. Consider a funnel with two layers and $\|\theta_1 - \theta_2\|_2 = q$. Lemma 3 indicates that $q \approx \operatorname{dist}(\mathcal{D}_1, \mathcal{D}_2)$.

Lemma 3. *We have under square loss function,* $\operatorname{dist}(\mathcal{D}_1, \mathcal{D}_2) \le 4\kappa d_x q$.

Proof.

We first show the second inequality.

$$
\begin{aligned}
&\operatorname{dist}(\mathcal{D}_1, \mathcal{D}_2) \\
&= \sup_\theta |\mathbb{E}_x(\mu(x^T\theta)) - \mu(x^T\theta_1^*))^2 - \mathbb{E}_x(\mu(x^T\theta) - \mu(x^T\theta_2^*))^2| \\
&\le \sup_\theta |\mathbb{E}_x \mu(x^T\theta)(\mu(x^T\theta_1^*) - \mu(x^T\theta_2^*))| + |\mathbb{E}_x(\mu^2(x^T\theta_1^*) - \mu^2(x^T\theta_2^*))| \\
&\le 4|\mathbb{E}_x(\mu(x^T\theta_1^*) - \mu(x^T\theta_2^*))| \\
&\le 4\mathbb{E}_x|\mu(x^T\theta_1^*) - \mu(x^T\theta_2^*)| \\
&\le 4\kappa\mathbb{E}_x|x^T(\theta_1^* - \theta_2^*)| \\
&\le 4\kappa d_x q
\end{aligned}
$$

On the other hand, an lower bound of $\operatorname{dist}(\mathcal{D}_1, \mathcal{D}_2)$ is also closely related to q.

$$
\begin{aligned}
&\operatorname{dist}(\mathcal{D}_1, \mathcal{D}_2) \\
&= \sup_\theta |\mathbb{E}_x(\mu(x^T\theta)) - \mu(x^T\theta_1^*))^2 - \mathbb{E}_x(\mu(x^T\theta) - \mu(x^T\theta_2^*))^2| \\
&= \sup_\theta |\mathbb{E}_x(\mu(x^T\theta_1^*) - \mu(x^T\theta_2^*))(\mu(x^T\theta_1^*) + \mu(x^T\theta_2^*) + \mu(x^T\theta))| \\
&= \sup_\theta |\mathbb{E}_x \int_t \mu'\left(tx^T\theta_1^* + (1-t)x^T\theta_2^*\right) dt(x^T(\theta_1^* - \theta_2^*))(\mu(x^T\theta_1^*) + \mu(x^T\theta_2^*) + \mu(x^T\theta))| \\
&= \sup_\theta |(\theta_1^* - \theta_2^*)^T[\mathbb{E}_x x \int_t \mu'\left(tx^T\theta_1^* + (1-t)x^T\theta_2^*\right) dt(\mu(x^T\theta_1^*) + \mu(x^T\theta_2^*) + \mu(x^T\theta))]| \\
&\text{(Let } \theta \to -\infty) \\
&\ge |(\theta_1^* - \theta_2^*)^T \nu_{\theta_1^*,\theta_2^*}| \text{ (letting } \nu_{\theta_1^*,\theta_2^*} = [\mathbb{E}_x x \int_t \mu'\left(tx^T\theta_1^* + (1-t)x^T\theta_2^*\right) dt(\mu(x^T\theta_1^*) + \mu(x^T\theta_2^*))]).
\end{aligned}
$$

Let $\theta_2^* = \theta_1^* + \|\theta_1^* - \theta_2^*\|_2 \mu$, where μ is a unit vector. For sufficient small $\|\theta_1^* - \theta_2^*\|_2, \nu_{\theta_1^*,\theta_2^*} \to 2\mathbb{E}_x[x\mu'(x^T\theta_1^*)\mu(x^T\theta_1^*)] =: \nu_{\theta_1^*}$, which is a constant vector. Thus

$$
\lim_{\|\theta_1^* - \theta_2^*\|_2 \to 0} \frac{\operatorname{dist}(\mathcal{D}_1, \mathcal{D}_2)}{\|\theta_1^* - \theta_2^*\|_2} = |\mu^T \nu_{\theta_1^*}|.
$$

For sufficient small $\|\theta_1^* - \theta_2^*\|$, discrepancy scales with $\|\theta_1^* - \theta_2^*\|$.

A.2 Proof of Lemma 1

In this subsection, we introduce the proof of Lemma 1. Many proofs could achieve a very similar bound. Here we use the idea of local Rademacher complexity.

Proof.

We discuss two cases: 1) $\hat{\theta} \in \mathrm{int}(\Theta_0)$. 2) $\hat{\theta} \notin \mathrm{int}(\Theta_0)$.

In both cases, one simply has

$$|\mu(x^T\hat{\theta}) - \mu(x^T\theta^*)| \leq \kappa|x^T(\hat{\theta} - \theta^*)| \leq \kappa \sup_{\theta_1,\theta_2\in\Theta_0} |x^T(\theta_1 - \theta_2)|,$$

which completes the first term in the minimum.

Now we prove the parametric bound. We first assume that case 1 holds. In this case, the constraint does not come into effects and $\hat{\theta}$ is the global minimal. By Theorem 26.5 in Shalev-Shwartz and Ben-David (2014), we have under an event, whose probability is at least $1 - \delta$,

$$L(\hat{\theta}) - L(\theta^*) \leq 2R_n(\boldsymbol{z}) + 5\sqrt{\frac{2\ln(8/\delta)}{n}}, \tag{8}$$

where $R(\boldsymbol{z})$ is the Rademacher complexity defined by

$$R_n(\boldsymbol{z}) = \mathbb{E}_{\boldsymbol{\sigma}} \frac{1}{n} \sup_{\theta\in\Theta} \sum_{i=1}^{n} \|z_i - \mu(x_i^T\theta)\|_{M_n}^2 \sigma_i,$$

and the variables in $\boldsymbol{\sigma}$ are distributed i.i.d. from Rademacher distribution. Let us call the event E_A.

As for any $i \in [n]$, let $\phi_i(t) \coloneqq (z_i - \mu(t))^2$, which satisfies $|\phi_i'(t)| = |2(z_i - \mu(t))\mu'(t)| \leq \kappa$, using Contraction lemma (Shalev-Shwartz and Ben-David, 2014), we have

$$\begin{aligned}
R_n(\boldsymbol{z}) &\leq \mathbb{E}_{\boldsymbol{\sigma}} \frac{1}{n} \sup_{\theta\in\Theta} \sum_i \kappa(x_i^T\theta)\sigma_i \\
&= \kappa\mathbb{E}_{\boldsymbol{\sigma}} \frac{1}{n} \sup_{\theta\in\Theta} \sum_{i=1}^{n} x_i^T(\theta - \theta^*)\sigma_i. \\
&\leq \kappa\mathbb{E}_{\boldsymbol{\sigma}} \frac{1}{n} \sup_{\theta\in\Theta} \|\sum_i x_i\sigma_i\|_{M_n^{-1}} \|\theta - \theta^*\|_{M_n} \\
&\leq \kappa\mathbb{E}_{\boldsymbol{\sigma}} \frac{1}{n} \|\sum_i x_i\sigma_i\|_{M_n^{-1}} \sup_{\theta\in\Theta_0} \|\theta - \theta^*\|_{M_n}.
\end{aligned} \tag{9}$$

Next, using Jensen's inequality we have that

$$\begin{aligned}
&\mathbb{E}_{\boldsymbol{\sigma}} \frac{1}{n} \|\sum_i x_i\sigma_i\|_{M_n^{-1}} \\
&\leq \frac{1}{n} \left(\mathbb{E}_{\boldsymbol{\sigma}} \|\sum_i x_i\sigma_i\|_{M_n^{-1}}^2 \right)^{1/2} \\
&= \frac{1}{n} \left(\mathbb{E}_{\boldsymbol{\sigma}} tr[M_n^{-1} (\sum_i x_i\sigma_i)(\sum_i x_i\sigma_i)^T] \right)^{1/2} \\
&= \frac{1}{n} \left(tr[M_n^{-1} \mathbb{E}_{\boldsymbol{\sigma}} (\sum_i x_i\sigma_i)(\sum_i x_i\sigma_i)^T] \right)^{1/2}
\end{aligned} \tag{10}$$

Finally, since the variables $\sigma_1, \dots, \sigma_m$ are independent we have

$$\begin{aligned}
&\mathbb{E}_{\boldsymbol{\sigma}}(\sum_i x_i\sigma_i)(\sum_i x_i\sigma_i)^T \\
&= \mathbb{E}_{\boldsymbol{\sigma}} \sum_{k,l\in[n]} \sigma_k\sigma_l x_k x_l^T \\
&= \mathbb{E}_{\boldsymbol{\sigma}} \sum_{i\in[n]} \sigma_i^2 x_i x_i^T \\
&= \sum_{i\in[n]} x_i x_i^T = nM_n.
\end{aligned}$$

Plugging this into (10), assuming M_n is full rank, we have

$$(9) \leq \sqrt{d/n} \sup_{\theta\in\Theta_0} \|\theta - \theta^*\|_{M_n}. \tag{11}$$

Lemma 4. *Under the notation in Lemma 1 and Assumption 2, if an estimate $\hat{\theta}$ satisfies $L(\hat{\theta}) \leq L(\theta^*) + b_n$, then*

$$\|\hat{\theta} \ - \theta^*\|_{M_n}^2 \leq \frac{d_x b_n}{c_\mu}.$$

Proof. Let $g_n(\theta) = \sum_i x_i(\mu(x_i^T\theta) - \mu(x_i^T\theta^*))$. For any θ, $\nabla g_n(\theta) = \sum_i x_i x_i^T \mu'(x_i^T\theta)$. By simple calculus,

$$g_n(\theta^*) - g_n(\hat{\theta}) = \int_0^1 \nabla g_n\left(s\theta^* + (1-s)\hat{\theta}\right) ds(\theta^* - \hat{\theta}).$$

As $\mu(t) \geq c_\mu$, we have $\int_0^1 \nabla g_n\left(s\theta^* + (1-s)\hat{\theta}\right) ds \succ c_\mu M_n$. Plugging this into the inequality above we have

$$\|\theta^* - \hat{\theta}\|_{M_n}^2 \leq \frac{1}{c_\mu}(\sum_i x_i(\mu(x_i^T\theta) - \mu(x_i^T\theta^*)))^2 = \frac{1}{c_\mu}\epsilon^T M_n \epsilon \leq \frac{d_x}{c_\mu}\epsilon^T\epsilon = \frac{d_x}{c_\mu}(L(\hat{\theta}) - L(\theta^*)),$$

where $\epsilon \coloneqq (\mu(x_i^T\hat{\theta}) - \mu(x_i^T\theta^*))_{i=1}^n$.

Applying (8) and Lemma 4, we complete the proof by

$$\begin{aligned}
\|\hat{\theta} - \theta^*\|_{M_n} &\leq \sqrt{\frac{2d_x\sqrt{d}}{c_\mu\sqrt{n}} \sup_{\theta\in\Theta} \|\theta - \theta^*\|_{M_n} + 5\sqrt{\frac{2\ln(8/\delta)}{n}}} \\
&\leq \sqrt{\frac{20\sqrt{2\ln(8/\delta)}d_x\sqrt{d}\sup_{\theta\in\Theta_0}\|\theta - \theta^*\|_{M_n}}{c_\mu\sqrt{n}}}.
\end{aligned} \tag{12}$$

□

We apply (12) iteratively [1]. Let $\Theta_{(1)} \coloneqq \Theta_0$. For any $t > 1$, let $\Theta_{(t)} = \{\theta \in \mathbb{R}^d : \|\theta - \hat{\theta}\|_{M_n} \leq \sqrt{\frac{20\sqrt{2d\ln(8/\delta)}}{c_\mu\sqrt{n}} \sup_{\theta\in\Theta_{(t-1)}} \|\theta - \theta^*\|_{M_n}}\}$. When $t \to \infty$, we have

$$\Theta_{(\infty)} = \frac{20d_x\sqrt{2d\ln(8/\delta)}}{c_\mu\sqrt{n}}.$$

By (12), we have $\theta^* \in \cap_{t\geq 1}\Theta_{(\infty)}$ and $\|\hat{\theta} - \theta^*\|_{M_n} \leq \frac{40d_x\sqrt{2d\ln(8/\delta)}}{c_\mu\sqrt{n}}$, which completes the second part of Lemma 1.

[1] Note that (12) holds under the same event E_A as the estimates $\hat{\theta}$ keeps the same each round as it is the global minimizer.

For any $x \in \mathcal{X}$, we have

$$|\mu(x^T\hat{\theta}) - \mu(x^T\theta^*)| \le \kappa \|x\|_{M_n^{-1}} \frac{40 d_x \sqrt{2d\ln(8/\delta)}}{c_\mu \sqrt{n}}. \tag{13}$$

When case 2 holds, let $\hat{\theta}'$ be the global minimizer. Using the analysis above, we have

$$\|\hat{\theta}' - \theta^*\|_{M_n} \le \frac{40 d_x \sqrt{2d\ln(8/\delta)}}{c_\mu \sqrt{n}}.$$

Then by triangle inequality

$$\|\hat{\theta} - \theta^*\|_{M_n} \le \|\hat{\theta} - \hat{\theta}'\|_{M_n} + \|\hat{\theta}' - \theta^*\|_{M_n} \le \frac{80 d_x \sqrt{2d\ln(8/\delta)}}{c_\mu \sqrt{n}}.$$

A.3 Tightness of Lemma 1

We use an example to show the tightness of Lemma 1. Assume a linear predictor, i.e. $\mu(t) = t$. Consider the following distribution, let X be uniform over the d-standard basis vector e_m, for $m = 1, \dots, d$. Let $Z \mid (X = e_i) \sim Bern(r_i)$, where $r_i \in [0,1]$ is pre-determined and unknown. The optimal parameter $\theta^* = (r_1, \dots, r_d)^T$. Let n_m be the number of samples collected for dimension m. Let $\Theta_0 \coloneqq \{\theta : \|\theta\|_2 \le q\}$.

When n is sufficiently large $n > 1/q^2$, $\hat{\theta}$ is the regularized minimizer. It can be shown that for any $\hat{\theta}$, there exists θ^* such that $\mathbb{E}[\hat{\theta}_i - \theta_i^*]^2 \ge (r_m(1-r_m))/n_m$. Then $\mathbb{E}\|\hat{\theta} - \theta^*\|_2^2 \ge \sum_{m=1}^d \frac{r_i(1-r_i)}{n_m} \ge \frac{d^2(r_i(1-r_i))}{n} = \Omega(\frac{d^2}{n})$.

Then we also see that when n is small ($\le \frac{1}{q^2}$), the estimation error is $\Omega(q)$. We use the same example as above. This time, we assume $\|\theta^*\| \le \frac{q}{2}$. If we have a $\|\hat{\theta}\| = q$, then $\|\theta^* - \hat{\theta}\| \ge q/2 = \Omega(q)$. Otherwise, we use the lower bound above: $\|\theta^* - \hat{\theta}\| \ge \Omega(\frac{d}{\sqrt{n}}) = \Omega(dq)$.

The above argument corresponds to the upper bound in Lemma 1, where we use prior knowledge when n is small and use the parametric bound when n is large.

A.4 Proof of Theorem 1

In this subsection, we show the missing proof for Theorem 1.

Theorem 4 (Prediction error under sequential dependency)**.** *For any funnel with a sequential dependency of parameters $q_1, \dots, q_J$, let $\hat{\theta}_1, \dots, \hat{\theta}_J$ be the estimates from Algorithm 1. If $n_{j+1} \le n_j/4$, $q_1 \ge, \dots, \ge q_J$ and Assumption 5 is satisfied, then with a probability at least $1-\delta$, for any $j_0 \in [J]$, we have*

$$PE_j \le \begin{cases} \kappa\|x\|_2 \frac{c_\delta}{c_\mu \lambda}\sqrt{\frac{d}{n_j}}, & \text{if } j < j_0, \\ \kappa\|x\|_2 (\frac{c_\delta}{c_\mu \lambda}\sqrt{\frac{d}{n_{j_0}}} + \sum_{i=j_0+1}^{j} q_j), & \text{if } j \ge j_0, \end{cases} \tag{14}$$

where we let $n_0 = \infty$. The bound is smallest when j_0 is the smallest $j \in [J]$, such that

$$\frac{4 c_\delta \sqrt{d}}{c_\mu \lambda}\left(\frac{1}{\sqrt{n_j}} - \frac{1}{\sqrt{n_{j-1}}}\right) \ge q_j, \tag{15}$$

if none of j's in $[J]$ satisfies (15), $j_0 = J + 1$.

Proof. First we reshape the ellipsoid in (3) to a ball.

Lemma 5 (Reshape)**.** *For any vector $x \in \mathbb{R}^d$ and any matrix $M \succ 0 \in \mathbb{R}^{d\times d}$, $\|x\|_2 \le \frac{1}{\lambda}\|x\|_M$, where λ is the minimum eigenvalue of M.*

Proof. We directly use the definition of positive definite matrix: $\lambda^2\|x\|_2^2 - \|x\|_M^2 = x^T(\lambda^2 I - M)x \le 0$. Thus, $\|x\|_2 \le \frac{1}{\lambda^2}\|x\|_M$. #

Using Lemma 5 and Assumption 5, we have $\|\bar{\theta}_j - \theta_j^*\|_2 \leq \frac{1}{\lambda}\|\bar{\theta}_j - \theta_j^*\|_{M_n} \leq \frac{4c_\delta}{c_\mu\lambda}\sqrt{\frac{d}{n}}$. Thus the set $\hat{\Theta}_j \subset \{\theta : \|\theta - \bar{\theta}_j\|_2 \leq \frac{4c_\delta}{c_\mu\lambda}\sqrt{\frac{d}{n}}\} =: \hat{\Theta}_j^{ball}$.

For every j, one can derive two bounds. First we can directly apply Corollary 1 and get $PE_j \leq \kappa\|x\|_2 \frac{4c_\delta}{c_\mu\lambda}\sqrt{\frac{d}{n_j}}$. Second, for any j_0, we have $\theta_j^* \in \Theta_1[j] \subset \{\theta : \|\bar{\theta}_{j_0} - \theta\|_2 \leq \frac{4c_\delta}{c_\mu\lambda}\sqrt{\frac{d}{n_{j_0}}} + \sum_{j_0+1\leq i\leq j} q_i\}$ and get $PE_j \leq \kappa\|x\|_2(\frac{c_\delta}{c_\mu\lambda}\sqrt{\frac{d}{n_{j_0}}} + \sum_{i=j_0+1}^{j} q_j)$.

Now we show the second argument: of all those bounds the one defined in (14) with j_0 defined in (15) is the smallest. For any $j \leq j_0$ and $j_1 \leq j$, we have

$$\frac{4c_\delta}{c_\mu\lambda}\sqrt{\frac{d}{n_j}} = \frac{4c_\delta\sqrt{d}}{c_\mu\lambda}\left(\sum_{i=j_1+1}^{j}(\frac{1}{\sqrt{n_i}} - \frac{1}{\sqrt{n_{i-1}}}) + \frac{1}{\sqrt{n_{j_1}}}\right) \leq \frac{4c_\delta}{c_\mu\lambda}\sqrt{\frac{d}{n_{j_1}}} + \sum_{i=j_1+1}^{j} q_i. \tag{16}$$

The second inequality is given by $(\frac{1}{\sqrt{n_i}} - \frac{1}{\sqrt{n_{i-1}}}) \leq q_i$ for all $i < j_0$. For any $j \geq j_0$ and $j_1 \leq j_0$, by (16), we have

$$\frac{4c_\delta}{c_\mu\lambda}\sqrt{\frac{d}{n_{j_0}}} + \sum_{i=j_0+1}^{j} q_i \leq \frac{4c_\delta}{c_\mu\lambda}\sqrt{\frac{d}{n_{j_1}}} + \sum_{i=j_1+1}^{j} q_i.$$

Now we prove that for all $i \geq j_0$,

$$\frac{4c_\delta\sqrt{d}}{c_\mu\lambda}(\frac{1}{\sqrt{n_i}} - \frac{1}{\sqrt{n_{i-1}}}) \geq q_i. \tag{17}$$

We use induction. Assume for some i_1, (17) is satisfied. Under the assumption that $n_{i_1-1} \leq n_{i_1}/4$ and $q_{i_1} \geq q_{i_1+1}$, we have

$$\begin{aligned}
\frac{4c_\delta\sqrt{d}}{c_\mu\lambda}(\frac{1}{\sqrt{n_{i_1+1}}} - \frac{1}{\sqrt{n_{i_1}}}) &= \frac{4c_\delta\sqrt{d}}{c_\mu\lambda}(\frac{1}{\sqrt{n_{i_1+1}}} + \frac{1}{\sqrt{n_{i_1-1}}} - \frac{2}{\sqrt{n_{i_1}}} + \frac{1}{\sqrt{n_{i_1}}} - \frac{1}{\sqrt{n_{i_1-1}}}) \\
&\geq \frac{4c_\delta\sqrt{d}}{c_\mu\lambda}(\frac{1}{\sqrt{n_{i_1+1}}} + \frac{2}{\sqrt{n_{i_1}}} - \frac{2}{\sqrt{n_{i_1}}} + \frac{1}{\sqrt{n_{i_1}}} - \frac{1}{\sqrt{n_{i_1-1}}}) \\
&\geq \frac{4c_\delta\sqrt{d}}{c_\mu\lambda}(+\frac{1}{\sqrt{n_{i_1}}} - \frac{1}{\sqrt{n_{i_1-1}}}) \\
&\geq q_{i_1} \geq q_{i_1+1}.
\end{aligned}$$

Using 17, for any $j \geq j_1 > j_0$,

$$\frac{4c_\delta}{c_\mu\lambda}\sqrt{\frac{d}{n_{j_0}}} + \sum_{i=j_0+1}^{j} q_i = \frac{4c_\delta\sqrt{d}}{c_\mu\lambda}\left(\sum_{i=j_0+1}^{j_1}(\frac{1}{\sqrt{n_{i-1}}} - \frac{1}{\sqrt{n_i}}) + \frac{1}{\sqrt{n_{j_1}}}\right) + \sum_{i=j_0+1}^{j} q_i \leq \frac{4c_\delta}{c_\mu\lambda}\sqrt{\frac{d}{n_{j_1}}} + \sum_{i=j_1+1}^{j} q_i.$$

Finally, we conclude that j_0 gives the smallest bound. # □

Similar argument can be used to show Theorem 2. For any $j_0 \in [J]$, we have

$$PE_j \leq \kappa\|x\|_2 \min\left\{\frac{c_{\delta/J}}{c_\mu\lambda}\sqrt{\frac{d}{n_{j0}}} + q_j, \frac{c_{\delta/J}}{c_\mu\lambda}\sqrt{\frac{d}{n_j}}\right\}.$$

Out of all the choices of j_0, the best one is achieved by $j_0 = \arg\min_{j\in[J]} \frac{c_\delta}{c_\mu\lambda}\sqrt{\frac{d}{n_j}} + q_j$.

A.5 Proof of Theorem 3

Theorem 5. *Using Algorithm 2, under the Assumptions 1-4, with a probability at least $1-\delta$, the total regret*

$$\begin{aligned}&\sum_{t=1}^{T}\left[P(x_t,\boldsymbol{\theta}^*_{a_t^*})-P(x_t,\boldsymbol{\theta}^*_{a_t})\right]\\ &\le 2\sqrt{2}c_0\sum_{a,j}\sqrt{n_{a,j}^T}+\sum_{a,j}\frac{8c_0^2Jd_x^4\log(6AJT/\delta)}{\bar{p}_{a,j}^2}-\sum_{a,j}\Delta_{a,j}.\end{aligned} \tag{18}$$

where $\mathcal{O}$ ignores all the constant terms and logarithmic terms for better demonstrations, $c_0=(\kappa d_x c_{\delta/AJT}\sqrt{d})/(c_\mu\bar{\lambda})$, $\bar{p}_a \coloneqq \mathbb{E}_x P_{J-1}(x^T\boldsymbol{\theta}^*_a)$ and

$$\Delta_{a,j}=\sum_{t=1;a_t=a}^{T}P_j(x_t^T\hat{\theta}^t_{a_t})\left[c_0\frac{1}{\sqrt{n_{a,j}^t\vee 1}}-\Delta\mu_{a,j}^t\right].$$

represents the benefits of transfer learning.

Let $\bar{p}_{a,j}\coloneqq\mathbb{E}_xP_{j-1}\left(x^T\theta_a^*\right)$. We first show that upper bound the number of steps t with $\lambda^t_{a_t,j}\le\bar{\lambda}/2$ or $n^t_{a,j}\le\frac{1}{2}n^t_{a,1}\bar{p}_{a,j}$. These steps are considered bad events.

Lemma 6 shows that with high probability, the number of observations for each layer is close to its expectation.

Lemma 6. *With a probability at least $1-\delta$, we have $n^t_{a,j}\ge n^t_{a,1}\bar{p}_{a,j}-\sqrt{2n^t_{a,1}\log(1/\delta)}$. Especially, when $n^t_{a,1}>8\log(1/\delta)/\bar{p}^2_{a,j}=:c_{n,a}$, we have $n^t_{a,j}\ge\frac{1}{2}n^t_{a,1}\bar{p}_{a,j}$.*

Proof. This is a direct application of Hoeffding inequality. □

Lemma 7. *For any $x_1,\ldots,x_n$ i.i.d, $\|x_i\|\le d_x$, let λ_n be the minimum eigenvalue of $\sum_i x_ix_i^T/n$ and $\bar{\lambda}$ be the minimum eigenvalue of its expectation. We have $\lambda_n\ge\bar{\lambda}/2$, when $n>d_x^4\log(1/\delta)/\bar{\lambda}^2$.*

Proof. For all $x_1,\ldots,x_n$, write $x_i=\sum_{s=1}^d\nu_{s,i}\tilde{x}_s$, where $\tilde{x}_1,\ldots,\tilde{x}_d$ are any basis of $\mathbb{R}^d$. We have $\mathbb{E}\nu^2_{s,i}\ge\bar{\lambda}$. For Hoeffding's inequality, since $\nu_{s,i}\le d_x$, with a probability $1-\delta$, we have

$$\frac{1}{n}\sum_i\nu^2_{s,i}\ge\mathbb{E}\nu^2_{s,1}-d_x^2\sqrt{\frac{\log(1/\delta)}{n}}\ge\bar{\lambda}-d_x^2\sqrt{\frac{\log(1/\delta)}{n}}.$$

For $n>d_x^4\log(1/\delta)/\bar{\lambda}^2$, we have $\frac{1}{n}\sum_i\nu^2_{s,i}\ge\bar{\lambda}/2$. There exists a choice of $\tilde{x}_1,\ldots,\tilde{x}_d$ such that $\lambda_n=\frac{1}{n}\sum_i\nu^2_{s,i}$. □

Combining Lemma 6 and Lemma 7, we have with a probability at least $1-\delta/3$, $\#\{t:\exists j,\lambda^t_{a_t,j}\le\bar{\lambda}/2$ or $n^t_{a,j}\le\frac{1}{2}n^t_{a,1}\bar{p}_{a,j}\}$ can be upper bounded by

$$\sum_{a,j}\max\left\{8\log(6AJT/\delta)/\bar{p}^2_{a,j},2d_x^4\log(6AJT/\delta)/(\bar{\lambda}^2\bar{p}_{a,j})\right\}. \tag{19}$$

In the following proof, we assume for all t, $\lambda^t_{a,j}\ge\bar{\lambda}/2$ and $n^t_{a,j}\ge\frac{1}{2}n^t_{a,1}\bar{p}_{a,j}$. We also assume the event in Lemma 1 happens for all $a\in[A],j\in[J]$ and $t<T$. The probability is at least $1-\delta/3$ as each probability is at least $1-\delta/(3AJT)$.

The total regret is

$$
\begin{aligned}
&\sum_{t=1}^{T}\Big[P(x_t,\boldsymbol{\theta}^*_{a_t^*})-P(x_t,\boldsymbol{\theta}^*_{a_t})\Big]\\
&\le \sum_{t=1}^{T}\Big[P(x_t,\boldsymbol{\theta}^*_{a_t^*})-P^+(x_t,\hat{\boldsymbol{\theta}}_{a_t})+P^+(x_t,\hat{\boldsymbol{\theta}}_{a_t})-P(x_t,\boldsymbol{\theta}^*_{a_t})\Big]\\
&\text{(Using } P(x_t,\boldsymbol{\theta}^*_{a_t^*})-P^+(x_t,\hat{\boldsymbol{\theta}}_{a_t})\le 0)\\
&\le \sum_{t=1}^{T}\Big[P^+(x_t,\hat{\boldsymbol{\theta}}^t_{a_t})-P(x_t,\boldsymbol{\theta}^*_{a_t})\Big]\\
&\text{(Using Lemma 2)}\\
&\le \sum_{t=1}^{T}\left[\sum_j \frac{P_J\left(x,\hat{\theta}^t_{a_t}\right)}{\mu\left(x^T\hat{\theta}^t_{a_t,j}\right)}\Delta\mu^t_{a_t,j}+\sum_{i\ne j}\Delta\mu^t_{a_t,j}\Delta\mu^t_{a_t,i}\right]\\
&\le \sum_{t=1}^{T}\left[\sum_j P_j(x_t,\hat{\theta}^t_{a_t})\Delta\mu^t_{a_t,j}+\sum_{i\ne j}\Delta\mu^t_{a_t,j}\Delta\mu^t_{a_t,i}\right]\\
&= \sum_{t=1}^{T}\left[\sum_j (P_j(x_t,\theta^*_{a_t})+P_j(x_t,\hat{\theta}^t_{a_t})-P_j(x_t,\theta^*_{a_t}))\Delta\mu^t_{a_t,j}+\sum_{i\ne j}\Delta\mu^t_{a_t,j}\Delta\mu^t_{a_t,i}\right]\\
&\le \underbrace{\sum_{t=1}^{T}\sum_j P_j\left(x_t,\theta^*_{a_t}\right)\frac{c_0}{\sqrt{n^t_{a_t,j}}}}_{①}-\underbrace{\sum_{t=1}^{T}\sum_j P_j\left(x_t,\theta^*_{a_t}\right)\Big(\frac{c_0}{\sqrt{n^t_{a_t,j}}}-\Delta\mu^t_{a_t,i}\Big)}_{②}+\underbrace{\sum_{t=1}^{T}\sum_{i\ne j}\Delta\mu^t_{a_t,j}\Delta\mu^t_{a_t,i}}_{③}\\
&\quad+\underbrace{\sum_{t=1}^{T}\left[\sum_j (P_j(x_t,\hat{\theta}^t_{a_t})-P_j(x_t,\theta^*_{a_t}))\Delta\mu^t_{a_t,j}\right]}_{④}.
\end{aligned}
$$

We further bound the terms separately. The first term ① represents the bound one could have without multi-task learning.

$$
\begin{aligned}
&\sum_{t=1}^{T}\sum_j P_j(x_t,\theta^*_{a_t})\frac{c_0}{\sqrt{n^t_{a_t,j}}}\\
&\le \sum_{t=1}^{T}\sum_j \mathbf{1}(r_{t,j-1}=1)\frac{c_0}{\sqrt{n^t_{a_t,j}}}+\sum_{t=1}^{T}\sum_j (P_j(x_t,\theta^*_{a_t})-\mathbf{1}(r_{t,j-1}=1))\frac{c_0}{\sqrt{n^t_{a_t,j}}}\\
&\quad\text{(Using Lemma 19 in Jaksch et al. (2010))}\\
&\le c_0 2\sqrt{2}\sum_{a,j}\sqrt{n^T_{a,j}}+\sum_{t=1}^{T}\sum_j (P_j(x_t,\theta^*_{a_t})-\mathbf{1}(r_{t,j-1}=1))\frac{c_0}{\sqrt{n^t_{a_t,j}}} \qquad (20)
\end{aligned}
$$

As $\mathbb{E}[P_j(x_t,\theta^*_{a_t})-\mathbf{1}(r_{t,j-1}=1)]=0$, the second term in (20) is a martingale. Using Azuma-Hoeffding inequality, with a probability at least $1-\delta/3$, for all T,

$$
\sum_{t=1}^{T}\sum_j (P_j(x_t,\theta^*_{a_t})-\mathbf{1}(r_{t,j-1}=1))\frac{c_0}{\sqrt{n^t_{a_t,j}}}\le c_0\sqrt{2\log(3TJ/\delta)}. \qquad (21)
$$

Combined with (20),

$$① \leq 2\sqrt{2}c_0 \sum_{a,j} \sqrt{n^T_{a,j}} + c_0\sqrt{2\log(3TJ/\delta)}. \tag{22}$$

Next we bound ③. We notice that this is a quadratic term. We first show Lemma 6 that lower bounds the number of observations for each layer. Lemma 6 is a direct application of Hoeffding's inequality.

For any pair i, j, we have

$$\begin{aligned}
&\sum_{t=1}^{T} \Delta\mu^t_{a_t,j}\Delta\mu^t_{a_t,i} \\
&\leq c_0^2 \sum_{t=1}^{T} \frac{1}{\sqrt{n^t_{a_t,i}}}\frac{1}{\sqrt{n^t_{a_t,j}}} \\
&\leq c_0^2 \sum_{t=1}^{T} \left[\mathbf{1}(n^t_{a_t,1} \leq c_{n,a_t})\frac{1}{\sqrt{n^t_{a_t,i}}}\frac{1}{\sqrt{n^t_{a_t,j}}} + \mathbf{1}(n^t_{a_t,1} > c_{n,a_t})\frac{1}{\sqrt{n^t_{a_t,i}}}\frac{1}{\sqrt{n^t_{a_t,j}}}\right] \\
&\leq c_0^2 \sum_{a} c_{n,a} + c_0^2 \sum_{t} \frac{4}{\bar{p}_a^2 n^t_{a_t,1}} \\
&\leq c_0^2 \sum_{a} c_{n,a} + c_0^2 \sum_{a} \frac{4\log(n^T_{a,1})}{\bar{p}_a^2} \\
&\leq 4c_0^2 \sum_{a} \frac{\log(n^T_{a,1}A/\delta)}{\bar{p}_a^2}.
\end{aligned} \tag{23}$$

where we let $\bar{p}_a := \mathbb{E}_x P_J\left(x^T\theta^*_a\right)$.

Thus, ③ is upper bounded by $4c_0^2J^2\sum_a \frac{\log(n^T_{a,1}A/(3\delta))}{\bar{p}_a^2}$.

Finally we bound term ④. Using Lemma 2 on only first j layers, we have

$$④ \leq \sum_t \sum_j [\sum_i \Delta\mu^t_{a_t,i} + \sum_{i,k} \Delta\mu^t_{a_t,k}\Delta\mu^t_{a_t,i}]\Delta\mu^t_{a_t,j} \leq (J+1) \times ③. \tag{24}$$

The proof is completed by combining Equations (19), (21), (22), (23) and (24).

B Experiments

B.1 Practical algorithm

Algorithm 3 Practical Algorithm for Contextual Bandit with a Funnel Structure

$t \rightarrow 1$, total number of steps T, memory $\mathcal{H}_a = \{\}$ for all $a \in [A]$. Initialize $\hat{\theta}_{a,\star}$ with zero vectors.
$\hat{\theta}_{a,0} \rightarrow 0$.
for $t = 1$ to T **do**
  Receive context x_t.
  Choose $a_t = \arg\max_{a \in \mathcal{A}} \hat{P}_J(x_t, \hat{\theta}_{a,j})$.
  Set $a_t = \mathrm{Unif}([A])$ with probability ϵ.
  Receive $r_{t,1}, \dots, r_{t,J}$ from funnel F_{a_t}.
  Set $\mathcal{H}_{a_t} \rightarrow \mathcal{H}_{a_t} \cup \{(x_t, (r_{t,1}, \dots, r_{t,J}))\}$.
  for $j = 1, \dots, J$ **do**
    # *For sequential dependency*

$$\hat{\theta}_{a_t,j} \rightarrow \arg\min_{\theta} l(\theta, \mathcal{H}_{a_t}) + \lambda_j \|\theta - \hat{\theta}_{a_t,j-1}\|_2$$

    # *For clustered dependency*

$$\hat{\theta}_{a_t,j} \rightarrow \arg\min_{\theta} l(\theta, \mathcal{H}_{a_t}) + \lambda_j \|\theta - \frac{1}{J} \sum_i \hat{\theta}_{a_t,i}\|_2$$

  end for
end for

B.2 Tuned hyper-parameters

Simulated environment.

1. Target: units 16
2. Mix: units 32
3. Sequential: units 32
4. Multi-layer Clustered: units 4; λ 0.001
5. Multi-layer Sequential: units 8; λ 0.001

Data-based environment.

1. Target: units 64
2. Mix: units 64
3. Sequential: units 64
4. Multi-layer Clustered: units 64; λ 0.005
5. Multi-layer Sequential: units 16; λ 0.001