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: $\text{dist}(\mathcal{D}_1, \mathcal{D}_2) := \sup_{h \in \mathcal{H}} |\mathcal{L}_{\mathcal{D}_1}(h) - \mathcal{L}_{\mathcal{D}_2}(h)|$, where $\mathcal{D}_1$ and $\mathcal{D}_2$ 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 $q \approx \text{dist}(\mathcal{D}_1, \mathcal{D}_2)$.

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

**Proof.**

We first show the second inequality.

\[
\text{dist}(\mathcal{D}_1, \mathcal{D}_2) = \sup_{\theta} |\mathbb{E}_x(\mu(x^T \theta)) - \mu(x^T \theta^*_1))^2 - \mathbb{E}_x(\mu(x^T \theta) - \mu(x^T \theta^*_2))^2| \\
\leq \sup_{\theta} |\mathbb{E}_x(\mu(x^T \theta)(\mu(x^T \theta^*_1) - \mu(x^T \theta^*_2)))| + |\mathbb{E}_x(\mu^2(x^T \theta^*_1) - \mu^2(x^T \theta^*_2))| \\
\leq 4|\mathbb{E}_x(\mu(x^T \theta^*_1) - \mu(x^T \theta^*_2))| \\
\leq 4\kappa \mathbb{E}_x|x^T(\theta^*_1 - \theta^*_2)| \\
\leq 4\kappa d_x q
\]

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

\[
\text{dist}(\mathcal{D}_1, \mathcal{D}_2) = \sup_{\theta} |\mathbb{E}_x(\mu(x^T \theta)) - \mu(x^T \theta^*_1))^2 - \mathbb{E}_x(\mu(x^T \theta) - \mu(x^T \theta^*_2))^2| \\
= \sup_{\theta} |\mathbb{E}_x(\mu(x^T \theta)^2) - \mathbb{E}_x(\mu(x^T \theta)\mu(x^T \theta^*_1) + \mu(x^T \theta)\mu(x^T \theta^*_2))| \\
= \sup_{\theta} |\mathbb{E}_x \int_\mu' (tx^T \theta^*_1 + (1-t)x^T \theta^*_2) dt(x^T \theta^*_1 - \theta^*_2)(\mu(x^T \theta^*_1) + \mu(x^T \theta^*_2))| \\
\geq |(\theta^*_1 - \theta^*_2)^T \nu_{\theta^*_1, \theta^*_2}| (\text{letting } \nu_{\theta^*_1, \theta^*_2} = |\mathbb{E}_x \int_\mu' (tx^T \theta^*_1 + (1-t)x^T \theta^*_2) dt(\mu(x^T \theta^*_1) + \mu(x^T \theta^*_2)))|)
\]

Let $\theta^*_2 = \theta^*_1 + \|\theta^*_1 - \theta^*_2\|_2 \mu$, where $\mu$ is a unit vector. For sufficient small $\|\theta^*_1 - \theta^*_2\|_2, \nu_{\theta^*_1, \theta^*_2} \rightarrow 2\mathbb{E}_x[x \mu'(x^T \theta^*_1)\mu(x^T \theta^*_1)] =: \nu_{\theta^*_1}$, which is a constant vector. Thus

\[
\lim_{\|\theta^*_1 - \theta^*_2\|_2 \rightarrow 0} \frac{\text{dist}(\mathcal{D}_1, \mathcal{D}_2)}{\|\theta^*_1 - \theta^*_2\|_2} = |\mu^T \nu_{\theta^*_1}|
\]

For sufficient small $\|\theta^*_1 - \theta^*_2\|$, discrepancy scales with $\|\theta^*_1 - \theta^*_2\|_2$.

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) \( \hat{\theta} \in \text{int}(\Theta_0) \). 2) \( \hat{\theta} \notin \text{int}(\Theta_0) \).

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 \( 1 - \delta \),

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

where \( R(z) \) is the Rademacher complexity defined by

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

and the variables in \( \sigma \) are distributed i.i.d. from Rademacher distribution. Let us call the event \( E_A \).

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

\[
R_n(z) \leq \mathbb{E}_{\sigma} \frac{1}{n} \sup_{\theta \in \Theta} x_i^T (\theta - \theta^*) \sigma_i
= \kappa \mathbb{E}_{\sigma} \frac{1}{n} \sup_{\theta \in \Theta} x_i^T (\theta - \theta^*) \sigma_i.
\]

\[
\leq \kappa \mathbb{E}_{\sigma} \frac{1}{n} \sup_{\theta \in \Theta} \| x_i \sigma_i \| M_n^{-1} \| \theta - \theta^* \| M_n
\]

\[
\leq \kappa \mathbb{E}_{\sigma} \frac{1}{n} \| x_i \sigma_i \| M_n^{-1} \sup_{\theta \in \Theta_0} \| \theta - \theta^* \| M_n.
\]

(9)

Next, using Jensen’s inequality we have that

\[
\mathbb{E}_{\sigma} \frac{1}{n} \| x_i \sigma_i \| M_n^{-1}
\]

\[
\leq \frac{1}{n} \left( \mathbb{E}_{\sigma} \| x_i \sigma_i \| M_n^{-1} \right)^{1/2}
\]

\[
= \frac{1}{n} \left( \mathbb{E}_{\sigma} \text{tr}[M_n^{-1}(x_i \sigma_i)(x_i \sigma_i)^T] \right)^{1/2}
\]

\[
= \frac{1}{n} \left( \text{tr}[M_n^{-1} \mathbb{E}_{\sigma}(x_i \sigma_i)(x_i \sigma_i)^T] \right)^{1/2}.
\]

(10)
Finally, since the variables $\sigma_1, \ldots, \sigma_m$ are independent we have

$$
\mathbb{E}_\sigma \left( \sum_i x_i \sigma_i \right) \left( \sum_i x_i \sigma_i \right)^T
= \mathbb{E}_\sigma \left( \sum_{k,l} \sigma_k \sigma_l x_k x_l^T \right)
= \mathbb{E}_\sigma \left( \sum_{i \in [n]} \sigma_i^2 x_i x_i^T \right)
= \sum_{i \in [n]} x_i x_i^T = n M_n.
$$

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

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

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

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

Proof. Let $g_n(\theta) = \sum_i x_i (\mu(x_i^T \theta) - \mu(x_i^T \theta^*))$. For any $\theta$, $\nabla g_n(\theta) = \sum_i x_i x_i^T \mu'(x_i^T \theta)$. 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 $\int_0^1 \nabla g_n \left( s \theta^* + (1-s) \hat{\theta} \right) ds \geq c_\mu M_n$. Plugging this into the inequality above we have

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

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

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

$$
\| \hat{\theta} - \theta^* \|_{M_n} \leq \sqrt{\frac{2 d \sqrt{d}}{c_\mu \sqrt{n}} \sup_{\theta \in \Theta} \| \theta - \theta^* \|_{M_n} + 5 \sqrt{\frac{2 \ln(8/\delta)}{n}}}
\leq \sqrt{\frac{20 \sqrt{2d \ln(8/\delta)}}{c_\mu \sqrt{n}} \sup_{\theta \in \Theta_{(t-1)}} \| \theta - \theta^* \|_{M_n}}.
$$

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

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

By (12), we have $\theta^* \in \cap_{t \geq 1} \Theta_{(t)}$ and $\| \hat{\theta} - \theta^* \|_{M_n} \leq \frac{40 d \sqrt{2d \ln(8/\delta)}}{c_\mu \sqrt{n}}$, which completes the second part of Lemma 4.

\[1\] 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
\[
|\mu(x^T \hat{\theta}) - \mu(x^T \theta^*)| \leq \kappa \|x\|_{M_n^{-1}} \frac{40d_\mathcal{X} \sqrt{2d \ln(8/\delta)}}{c_\mu \sqrt{n}}. \tag{13}
\]

When case 2 holds, let \( \hat{\theta}' \) be the global minimizer. Using the analysis above, we have
\[
\|\hat{\theta}' - \theta^*\|_{M_n} \leq \frac{40d_\mathcal{X} \sqrt{2d \ln(8/\delta)}}{c_\mu \sqrt{n}}. \tag{14}
\]

Then by triangle inequality
\[
\|\hat{\theta} - \theta^*\|_{M_n} \leq \|\hat{\theta} - \hat{\theta}'\|_{M_n} + \|\hat{\theta}' - \theta^*\|_{M_n} \leq \frac{80d_\mathcal{X} \sqrt{2d \ln(8/\delta)}}{c_\mu \sqrt{n}}.
\]

### 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 \( m = 1, \ldots, d \). Let \( Z \mid (X = e_i) \sim Bern(r_i) \), where \( r_i \in [0, 1] \) is pre-determined and unknown. The optimal parameter \( \theta^* = (r_1, \ldots, r_d)^T \). Let \( n_m \) be the number of samples collected for dimension \( m \). Let \( \Theta_0 := \{ \theta : \|\theta\|_2 \leq q \} \).

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 \( \theta^* \) such that \( \mathbb{E}[\hat{\theta} - \theta^*]^2 \geq (r_m(1 - r_m))/n_m \). Then \( \mathbb{E} \|\hat{\theta} - \theta^*\|^2 \geq \sum_{m=1}^d \frac{r_m(1 - r_m)}{n_m} \geq d^2(r_1(1 - r_1)) = \Omega(d^2/n) \).

Then we also see that when \( n \) is small \((\leq \frac{1}{4})\), the estimation error is \( \Omega(q) \). We use the same example as above. This time, we assume \( \|\theta^*\| \leq \frac{q}{2} \). If we have a \( \|\hat{\theta}\| = q \), then \( \|\theta^* - \hat{\theta}\| \geq q/2 = \Omega(q) \). Otherwise, we use the lower bound above: \( \|\theta^* - \hat{\theta}\| \geq \Omega(dq/n) = \Omega(dq) \).

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 \( q_1, \ldots, q_J \), let \( \hat{\theta}_1, \ldots, \hat{\theta}_J \) be the estimates from Algorithm 1. If \( n_j + 1 \geq n_j/4, q_1 \geq \ldots \geq q_J \) and Assumption 5 is satisfied, then with a probability at least \( 1 - \delta \), for any \( j_0 \in [J] \), we have
\[
PE_j \leq \left\{ \begin{array}{ll}
\kappa \|x\|_2 \frac{c_\nu}{\sqrt{n_{j}}} \sqrt{\frac{M_j}{\sqrt{n_j}}} & \text{if } j < j_0, \\
\kappa \|x\|_2 \frac{c_\nu}{\sqrt{n_{j_0}}} \sqrt{\frac{M_j}{\sqrt{n_{j_0}}}} + \sum_{i=j_0+1}^J q_i & \text{if } j \geq j_0. 
\end{array} \right. \tag{15}
\]
where we let \( n_0 = \infty \). The bound is smallest when \( j_0 \) is the smallest \( j \in [J] \), such that
\[
\frac{4c_\nu \sqrt{d}}{c_\mu \lambda} \left( \frac{1}{\sqrt{n_j}} - \frac{1}{\sqrt{n_{j-1}}} \right) \geq q_j. \tag{16}
\]

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 \( M > 0 \in \mathbb{R}^{d \times d} \), \( \|x\|_2 \leq \frac{1}{\lambda} \|x\|_M \), where \( \lambda \) is the minimum eigenvalue of \( M \).

**Proof.** We directly use the definition of positive definite matrix: \( \lambda^2 \|x\|_2^2 - \|x\|_M^2 = x^T (\lambda^2 I - M)x \leq 0 \). Thus, \( \|x\|_2 \leq \frac{1}{\lambda} \|x\|_M \).
Using Lemma 3 and Assumption 5, we have $\|\hat{\theta}_j - \theta^*_j\|_2 \leq \frac{1}{L} \|\hat{\theta}_j - \theta^*_j\|_n \leq \frac{4c_\delta}{c\mu \lambda} \sqrt{\frac{d}{n}}$. Thus the set $\hat{\Theta}_j \subset \{\theta : \|\theta - \hat{\theta}_j\|_2 \leq \frac{4c_\delta}{c\mu \lambda} \sqrt{\frac{d}{n}}\} = \hat{\Theta}_{ball}$.

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 $\theta^*_j \in \Theta_1[j] \subset \{\theta : \|\hat{\theta}_{j_0} - \theta\|_2 \leq \frac{4c_\delta}{c\mu \lambda} \sqrt{\frac{d}{n_{j_0}}} + \sum_{j_0+1 \leq j \leq j} q_i\}$ and get $PE_j \leq \kappa \|x\|_2 (\frac{4c_\delta}{c\mu \lambda} \sqrt{\frac{d}{n_{j_0}}} + \sum_{i=j_0+1}^j q_i)$.

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

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

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 $i \geq j_0$,

$$\frac{4c_\delta}{c\mu \lambda} \sqrt{\frac{d}{n_{j_0}}} \left(\frac{1}{\sqrt{n_{i+1}}} - \frac{1}{\sqrt{n_i}}\right) \geq q_i.$$

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

$$\frac{4c_\delta}{c\mu \lambda} \sqrt{\frac{d}{n_{j_0}}} \left(\frac{1}{\sqrt{n_{i_1+1}}} - \frac{1}{\sqrt{n_{i_1}}}ight) \geq \frac{4c_\delta}{c\mu \lambda} \sqrt{\frac{d}{n_{j_1}}} \left(\sum_{i=j_0+1}^{j_1} \left(\frac{1}{\sqrt{n_i}} - \frac{1}{\sqrt{n_{i-1}}}ight)\right) + \sum_{i=j_0+1}^{j_1} q_i \leq \frac{4c_\delta}{c\mu \lambda} \sqrt{\frac{d}{n_{j_1}}} + \sum_{i=j_1+1}^j q_i.$$

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

$$PE_j \leq \kappa \|x\|_2 \min \left\{ \frac{c_{\delta/j}}{c\mu \lambda} \sqrt{\frac{d}{n_{j_0}}} + q_j, \frac{c\mu \lambda}{c_{\delta/j}} \sqrt{\frac{d}{n_{j}}} \right\}.$$

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

A.5 Proof of Theorem 5

Theorem 5. Using Algorithm 3 under the Assumptions 1-4, with a probability at least 1 − δ, the total regret

$$
\sum_{t=1}^{T} \left[ P(x_t, \theta^*_t) - P(x_t, \theta^*_t) \right] 
\leq 2\sqrt{2c_0} \sum_{a,j} \sqrt{n_{a,j}^T} + \sum_{a,j} \frac{8c_0 J_2^1 \log(6AJT/\delta)}{\bar{p}_{a,j}} - \sum_{a,j} \Delta_{a,j}. \tag{18}
$$

where $\mathcal{O}$ ignores all the constant terms and logarithmic terms for better demonstrations, $c_0 = (ndx^3/\lambda_{\max}^2/\sqrt{d})/(\mu\lambda)$, $\bar{p}_a := E_x P_{T-1}(x^T \theta^*_t)$ and

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

represents the benefits of transfer learning.

Let $\bar{p}_{a,j} := E_x P_{T-1}(x^T \theta^*_a)$. We first show that upper bound the number of steps $t$ with $\lambda_{a,j}^t \leq \bar{\lambda}/2$ or $n_{a,j}^t \leq \frac{1}{2}n_{a,j}^1\bar{p}_{a,j}$. 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 − δ, we have $n_{a,j}^t \geq n_{a,1}^1\bar{p}_{a,j} - \sqrt{2n_{a,1} \log(1/\delta)}$. Especially, when $n_{a,1} > 8 \log(1/\delta)/\bar{p}_{a,j}^2 =: c_{n,a}$, we have $n_{a,j}^t \geq \frac{1}{2}n_{a,1} \bar{p}_{a,j}$.

Proof. This is a direct application of Hoeffding inequality.

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

Proof. For all $x_1, \ldots, x_n$, write $x_i = \sum_{s=1}^{d} \nu_{s,i} \hat{x}_s$, where $\hat{x}_1, \ldots, \hat{x}_d$ are any basis of $\mathbb{R}^d$. We have $E_{s,i}^2 \geq \bar{\lambda}$. For Hoeffding’s inequality, since $\nu_{s,i} \leq d_x$, with a probability 1 − δ, we have

$$
\frac{1}{n} \sum_i \nu_{s,i}^2 \geq E_{s,i}^2 - d_x^2 \sqrt{\frac{\log(1/\delta)}{n}} \geq \bar{\lambda} - d_x^2 \sqrt{\frac{\log(1/\delta)}{n}}.
$$

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

Combining Lemma 6 and Lemma 7, we have with a probability at least 1 − δ/3, $\#\{t : \exists j, \lambda_{a,j}^t \leq \bar{\lambda}/2 \text{ or } n_{a,j}^t \leq \frac{1}{2}n_{a,1} \bar{p}_{a,j}\}$ can be upper bounded by

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

In the following proof, we assume for all $t$, $\lambda_{a,j}^t \geq \bar{\lambda}/2$ and $n_{a,j}^t \geq \frac{1}{2}n_{a,1} \bar{p}_{a,j}$. 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 − δ/3 as each probability is at least 1 − δ/(3AJT).
The total regret is
\[
\sum_{t=1}^{T} \left[ P(x_t, \theta^*_a) - P(x_t, \hat{\theta}_a) \right]
\]
\[
\leq \sum_{t=1}^{T} \left[ P(x_t, \theta^*_a) - P^+(x_t, \hat{\theta}_a) + P^+(x_t, \hat{\theta}_a) - P(x_t, \theta^*_a) \right]
\]
(Using \(P(x_t, \theta^*_a) - P^+(x_t, \hat{\theta}_a) \leq 0\))
\[
\leq \sum_{t=1}^{T} \left[ P^+(x_t, \hat{\theta}_a^t) - P(x_t, \theta^*_a) \right]
\]
(Using Lemma 2)
\[
\leq \sum_{t=1}^{T} \left[ \sum_j P_j(x_t, \hat{\theta}_a^t) \Delta \mu_{a,j}^t + \sum_{i \neq j} \Delta \mu_{a,j}^t \Delta \mu_{a,i}^t \right]
\]
\[
\leq \sum_{t=1}^{T} \left[ \sum_j (P_j(x_t, \theta^*_a) + P_j(x_t, \hat{\theta}_a^t) - P_j(x_t, \theta^*_a)) \Delta \mu_{a,j}^t + \sum_{i \neq j} \Delta \mu_{a,j}^t \Delta \mu_{a,i}^t \right]
\]
\[
\leq \sum_{t=1}^{T} \left[ \sum_j P_j(x_t, \theta^*_a) \frac{c_0}{\sqrt{n^t_{a,j}}} - \sum_{t=1}^{T} \sum_j P_j(x_t, \theta^*_a) (\frac{c_0}{\sqrt{n^t_{a,j}}} - \Delta \mu_{a,i}^t) + \sum_{t=1}^{T} \sum_{i \neq j} \Delta \mu_{a,j}^t \Delta \mu_{a,i}^t \right]
\]
\[
+ \sum_{t=1}^{T} \left[ \sum_j (P_j(x_t, \hat{\theta}_a^t) - P_j(x_t, \theta^*_a)) \Delta \mu_{a,j}^t \right]
\]

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

\[
\sum_{t=1}^{T} \sum_j P_j(x_t, \theta^*_a) \frac{c_0}{\sqrt{n^t_{a,j}}}
\]
\[
\leq \sum_{t=1}^{T} \sum_j 1 (r_{t,j-1} = 1) \frac{c_0}{\sqrt{n^t_{a,j}}} + \sum_{t=1}^{T} \sum_j (P_j(x_t, \theta^*_a) - 1 (r_{t,j-1} = 1)) \frac{c_0}{\sqrt{n^t_{a,j}}}
\]
(Using Lemma 19 in Jaksch et al. [2010])
\[
\leq c_0 2 \sqrt{2} \sum_{a,j} \sqrt{n^T_{a,j}} + \sum_{t=1}^{T} \sum_j (P_j(x_t, \theta^*_a) - 1 (r_{t,j-1} = 1)) \frac{c_0}{\sqrt{n^t_{a,j}}}
\]
(20)

As \( E[P_j(x_t, \theta^*_a) - 1 (r_{t,j-1} = 1)] = 0 \), the second term in (20) is a martingale. Using Azuma-Hoeffding inequality, with a probability at least \( 1 - \delta/3 \), for all \( T \),
\[
\sum_{t=1}^{T} \sum_j (P_j(x_t, \theta^*_a) - 1 (r_{t,j-1} = 1)) \frac{c_0}{\sqrt{n^t_{a,j}}} \leq c_0 \sqrt{2 \log(3TJ/\delta)}.
\]
(21)
Next we bound (3). We notice that this is a quadratic term. We first show Lemma [2] that lower bounds the number of observations for each layer. Lemma [2] is a direct application of Hoeffding’s inequality.

For any pair \( i, j \), we have

\[
\sum_{t=1}^{T} \Delta \mu_{i,j}^t \Delta \mu_{i,i}^t \\
\leq c_0^2 \sum_{t=1}^{T} \frac{1}{\sqrt{n_{i,i}^t}} \frac{1}{\sqrt{n_{j,j}^t}} \\
\leq c_0^2 \sum_{t=1}^{T} \left[ 1(n_{i,1}^t \leq c_{n,i}) \frac{1}{\sqrt{n_{i,i}^t}} \frac{1}{\sqrt{n_{i,j}^t}} + 1(n_{i,1}^t > c_{n,i}) \frac{1}{\sqrt{n_{i,i}^t}} \frac{1}{\sqrt{n_{i,j}^t}} \right] \\
\leq c_0^2 \sum_{a} c_{n,a} + c_0^2 \sum_{t} \frac{4}{\bar{p}_a n_{i,1}^t} \\
\leq c_0^2 \sum_{a} c_{n,a} + c_0^2 \sum_{a} \frac{4 \log(n_{a,1}^T)}{\bar{p}_a^2} \\
\leq 4c_0^2 \sum_{a} \frac{\log(n_{a,1}^T A/\delta)}{\bar{p}_a^2}. \tag{23}
\]

where we let \( \bar{p}_a := \mathbb{E}_x P_j (x^T \theta_a^*) \).

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

Finally we bound term (4). Using Lemma [2] on only first \( j \) layers, we have

\[
4 \leq \sum_{t} \sum_{j} \left( \sum_{i} \Delta \mu_{i,i}^t + \sum_{i,k} \Delta \mu_{i,k}^t \Delta \mu_{i,i}^t \right) \Delta \mu_{i,j}^t \leq (J + 1) \times (3). \tag{24}
\]

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 $\hat{\theta}_{a,*}$ with zero vectors.  

$\hat{\theta}_{a,0} \rightarrow 0$.  

for $t = 1$ to $T$ do 

Receive context $x_t$.  

Choose $a_t = \arg \max_{a \in A} \hat{P}_J(x_t, \hat{\theta}_{a,j})$.  

Set $a_t = \text{Unif}([A])$ with probability $\epsilon$.  

Receive $r_{t,1}, \ldots, r_{t,J}$ from funnel $F_{a_t}$.  

Set $\mathcal{H}_{a_t} \rightarrow \mathcal{H}_{a_t} \cup \{(x_t, (r_{t,1}, \ldots, r_{t,J}))\}$.  

for $j = 1, \ldots, J$ 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; $\lambda$ 0.001  
5. Multi-layer Sequential: units 8; $\lambda$ 0.001

**Data-based environment.**

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