

A APPENDIX

Proof of Lemma 2. Denote W^* as the ground truth diagonal matrix for the training instances, i.e., $W_{ii}^* = 1$ if the i -th instance is a clean instance, otherwise $W_{ii}^* = 0$. Accordingly, define D_c as the ground truth set of clean instances. For clearness of the presentation, we may drop the subscript t when there is no ambiguity. For mislabeled instances, the output is written in the form of $y_i = r_i + n_i$ where n_i represents the observation noise, and y_i depends on the specific setting we consider. Under this general representation, we can re-write the term w_{t+1} as

$$\begin{aligned} w_{t+1} &= (\Phi(X)^\top W \Phi(X))^{-1} \Phi(X)^\top W (W^* \Phi(X) w_c^* + (I - W^*)r + e) \\ &= w_c^* + (\Phi(X)^\top W \Phi(X))^{-1} (\Phi(X)^\top W W^* \Phi(X) w_c^* \\ &\quad + \Phi(X)^\top W r - \Phi(X)^\top W W^* r - \Phi(X)^\top W \Phi(X) w_c^* + \Phi(X)^\top W e) \\ &= w_c^* + (\Phi(X)^\top W \Phi(X))^{-1} \Phi(X)^\top (W W^* - W) (\Phi(X) w_c^* - r - e) \\ &\quad + (\Phi(X)^\top W \Phi(X))^{-1} \Phi(X)^\top W W^* e \end{aligned}$$

Therefore, the l_2 distance between the learned parameter and ground truth parameter can be bounded by:

$$\begin{aligned} \|w_{t+1} - w_c^*\| &= \|(\Phi(X)^\top W \Phi(X))^{-1} \Phi(X)^\top (W W^* - W) \\ &\quad (\Phi(X) w_c^* - r - e) + (\Phi(X)^\top W \Phi(X))^{-1} \Phi(X)^\top W W^* e\|_2 \\ &\leq \underbrace{\|(\Phi(X)^\top W \Phi(X))^{-1}\|_2}_{v_1} \cdot \\ &\quad \left(\underbrace{\|\Phi(X)^\top (W W^* - W) (\Phi(X) w_c^* - r - e)\|_2}_{v_2} \right. \\ &\quad \left. + \underbrace{\|\Phi(X)^\top W W^* e\|_2}_{v_3} \right) \end{aligned}$$

where basic spectral norm inequalities and triangle inequalities. For the term v_1 , notice that W selects αN rows of $\Phi(X)$, i.e., $\text{Tr}(W) = \alpha N$. Therefore, $v_1 \leq \frac{1}{\Psi^-(\alpha N)}$.

Next, the term v_2 can be bounded as:

$$\begin{aligned} v_2^2 &= \|(\Phi(X)^\top (W - W W^*) (\Phi(X) w_c^* - r - e))\|_2^2 \\ &= (\Phi(X) w_c^* - r - e)^\top \\ &\quad [(W - W W^*) \Phi(X) \Phi(X)^\top (W - W W^*)] (\Phi(X) w_c^* - r - e) \\ &\leq 2(\Phi(X) w_c^* - \Phi(X) w_t)^\top \\ &\quad [(W - W W^*) \Phi(X) \Phi(X)^\top (W - W W^*)] (\Phi(X) w_c^* - \Phi(X) w_t) \\ &\quad + 2(\Phi(X) w_t - r - e)^\top \\ &\quad [(W - W W^*) \Phi(X) \Phi(X)^\top (W - W W^*)] (\Phi(X) w_t - r - e) \\ &\leq 2\sigma_{\max}(\Phi(X)^\top (W - W W^*) \Phi(X))^2 \|w_c^* - w_t\|_2^2 \\ &\quad + 2(\Phi(X) w_t - r - e)^\top \\ &\quad [(W - W W^*) \Phi(X) \Phi(X)^\top (W - W W^*)] (\Phi(X) w_t - r - e) \end{aligned}$$

For the first term $2\sigma_{\max}(\Phi(X)^\top (W - W W^*) \Phi(X))^2 \|w_c^* - w_t\|_2^2$, let $|D_t \setminus D_c|$ be the number of mislabeled instances in D_t . Then, the eigenvalue is bounded by $\Psi^+(|D_t \setminus D_c|)$. And the last term $2(\Phi(X) w_t - r - e)^\top [(W - W W^*) \Phi(X) \Phi(X)^\top (W - W W^*)] (\Phi(X) w_t - r - e)$ is defined as $\varphi_t := \varphi(D_t, D_c, \|w_c^* - w_t\|_2) = \|\sum_{i \in D \setminus D_c} (\delta(x_i)^\top w_t - r_i - e_i) \delta(x_i)\|_2$. The term v_3 can be bounded as:

$$\begin{aligned}
v_3^2 &= \|\Phi(X)^\top W W^* e\|_2 \\
&\leq e^\top \Phi(X) \Phi(X)^\top e \\
&= \sum_{i=1}^d \left(\sum_{j=1}^N e_j \delta(x_j)_i \right)^2 \\
&\leq c \sum_{i=1}^N \|\delta(x_i)\|_2^2 \log N \zeta^2
\end{aligned}$$

where the last inequality holds with high probability by the sub-exponential concentration property, and all randomness comes from the measurement noise e . Then, as a summary, combining the results for all three terms, we have:

$$\begin{aligned}
\|w_{t+1} - w_c^*\|_2 &\leq \frac{\sqrt{2}\Psi^+(|D_t \setminus D_c|)}{\Psi^-(\alpha N)} \|w_t - w_c^*\|_2 \\
&\quad + \frac{\sqrt{2}\varphi(D_t, D_c, \|w_c^* - w_t\|_2)}{\Psi^-(\alpha N)} \\
&\quad + c \frac{\sqrt{\sum_{i=1}^N \|\delta(x_i)\|_2^2 \log N \zeta}}{\Psi^-(\alpha N)}
\end{aligned}$$

The following does the same for the more general non-linear case.

Non-Linear Case. Assume $\pi : \mathbb{R} \rightarrow \mathbb{R}$ monotone and differentiable. Assume $\pi'(u) \in [a, b]$ for all $u \in \mathbb{R}$, where a, b are positive constants. Define $F : \mathbb{R}^n \rightarrow \mathbb{R}^n$ as an entry-wise $f(\cdot)$ -operation. Denote learning rate as η .

$$\begin{aligned}
w_{t+1} &= w_t - \frac{\eta}{\alpha N} \sum_{i \in D_t} (f(\delta(x_i)^\top w_t) - y_i) \cdot f'(\delta(x_i)^\top w_t) \cdot \delta(x_i) \\
&= w_t - \frac{\eta}{\alpha N} \Phi(X)^\top \text{Diag}(F'(\Phi(X)w_t)) W_t (F(\Phi(X)w_t) - y) \\
&= w_t - \frac{\eta}{\alpha N} \Phi(X)^\top \text{Diag}(F'(\Phi(X)w_t)) W_t (F(\Phi(X)w_t) - W^* F(\Phi(X)w_c^*) - (I - W^*)(r + e) - W^* e) \\
&= w_t - \frac{\eta}{\alpha N} \Phi(X)^\top \text{Diag}(F'(\Phi(X)w_t)) W_t (F(\Phi(X)w_t) - W^* F(\Phi(X)w_c^*) - (I - W^*) F(\Phi(X)w_c^*)) \\
&\quad - \frac{\eta}{\alpha N} \Phi(X)^\top \text{Diag}(F'(\Phi(X)w_t)) W_t ((I - W^*) F(\Phi(X)w_c^*) - (I - W^*)(r + e) - W^* e) \\
&= w_t - \frac{\eta}{\alpha N} \Phi(X)^\top \text{Diag}(F'(\Phi(X)w_t)) W_t (F(\Phi(X)w_t) - F(\Phi(X)w_c^*)) \\
&\quad - \frac{\eta}{\alpha N} \Phi(X)^\top \text{Diag}(F'(\Phi(X)w_t)) (W_t - W_t W^*) (F(\Phi(X)w_c^*) - r - e) \\
&\quad + \frac{\eta}{\alpha N} \Phi(X)^\top \text{Diag}(F'(\Phi(X)w_t)) W_t W^* e
\end{aligned}$$

We simplify the notation using $H_t \triangleq \text{Diag}(F'(\Phi(X)w_t))$. Also, by mean value theorem, for any a, b , there exists some $c \in [a, b]$, such that $\frac{f(b) - f(a)}{b - a} = f'(c)$. Therefore, for the term $F(\Phi(X)w_t) - F(\Phi(X)w_c^*)$, there exists a diagonal matrix C_t , such that $F(\Phi(X)w_t) - F(\Phi(X)w_c^*) = C_t \Phi(X)(w_t - w_c^*)$. Therefore, we have

$$\begin{aligned}
\|w_{t+1} - w_c^*\|_2 &\leq \underbrace{\left(1 - \frac{\eta}{\alpha N} \Phi(X)^\top H_t W_t C_t \Phi(X) \right)}_{U_1} \|w_t - w_c^*\|_2 \\
&\quad + \underbrace{\frac{\eta}{\alpha N} \|\Phi(X)^\top H_t (W_t - W_t W^*) (F(\Phi(X)w_c^*) - r - e)\|_2}_{U_2} \\
&\quad + \underbrace{\frac{\eta}{\alpha N} \|\Phi(X) H_t W_t W^* e\|_2}_{U_3}
\end{aligned}$$

Here,

$$U_1 \leq 1 - \eta a^2 \frac{\Psi^-(\alpha N)}{\alpha N}, U_3 \leq b \xi_t \zeta$$

For U_2 , define $\hat{\delta}_t$ similar to δ_t :

$$\hat{\phi}_t = \left\| \sum_{i \in D_t \setminus D_c} (\pi(\delta(x_i)^\top \mathbf{w}_c^*) - r_i - e_i) \pi'(\delta(x_i)^\top \mathbf{w}_c^*) \delta(x_i) \right\|.$$

As a result, we have:

$$\|\mathbf{w}_{t+1} - \mathbf{w}_c^*\|_2 \leq \left(1 - \frac{\eta}{\alpha N} a^2 \Psi^-(\alpha N)\right) \|\mathbf{w}_t - \mathbf{w}_c^*\|_2 + \eta \frac{\hat{\phi}_t + \xi_t b \zeta}{\alpha N}$$