Proof of Theory 2 and Theory 3

January 31, 2018

1 Proof of Theory 2

The algorithm updates the model whenever adaptive margin is less than 0 ($Z_t = 1$). If there is no update, $\mathbf{U}_t = \mathbf{U}_{t-1}$ yields $\inf_{\mathbf{U}} G_t(\mathbf{U}) = \inf_{\mathbf{U}} G_{t-1}(\mathbf{U})$. Given $\ell_t(alg) = \|\mathbf{y}_t - \mathbf{f}_t\|^2$, we have

$$\inf_{\mathbf{U}} G_t(\mathbf{U}) - \inf_{\mathbf{U}} G_{t-1}(\mathbf{U}) = Z_t \left(\ell_t(alg) - a_t^2 \mathbf{x}_t^\top \mathbf{A}_t^{-1} \mathbf{x}_t + a_t - 1\right),$$

holds for all trial t, which is similar to the proof of [2]. Summing over $t = 1, \ldots, T$ with $\|\mathbf{y}_t\|^2 = 1$, we obtain with expanding the square,

$$\begin{aligned}
&\sum_{t \in \mathcal{Z}} (a_t \|\mathbf{y}_t\|^2 - 2\mathbf{y}_t \cdot \mathbf{f}_t - a_t^2 \mathbf{x}_t^\top \mathbf{A}_t^{-1} \mathbf{x}_t + \|\mathbf{f}_t\|^2) \\
=& \inf_{\mathbf{U}} (b\|\mathbf{U}\|^2 + \sum_t a_t \|\mathbf{y}_t - \mathbf{U}^\top \mathbf{x}_t\|^2) - (\inf_{\mathbf{U}} (b\|\mathbf{U}\|^2 + L_0^{\mathbf{a}}(\mathbf{U}))) \\
\leq& \sum_{t \in \mathcal{Z}} a_t (\|\mathbf{y}_t\|^2 - 2\mathbf{y}_t \cdot \mathbf{U}^\top \mathbf{x}_t) + \mathrm{tr}(\mathbf{U}^\top (b\mathbf{I} + \sum_{t \in \mathcal{Z}} a_t \mathbf{x}_t \mathbf{x}_t^\top) \mathbf{U}).
\end{aligned}$$

Assume that $\mathbf{A}_{\mathcal{Z}} = b\mathbf{I} + \sum_{t \in \mathcal{Z}} a_t \mathbf{x}_t \mathbf{x}_t^\top$, and $\sigma_t = \frac{1}{2} a_t^2 \mathbf{x}_t^\top \mathbf{A}_t^{-1} \mathbf{x}_t$ we obtain,

$$\sum_{t \in \mathcal{Z}} (-\mathbf{f}_t \mathbf{y}_t - \sigma_t) \leq -\sum_{t \in \mathcal{Z}} a_t \mathbf{y}_t \cdot \mathbf{U}^\top \mathbf{x}_t + \frac{1}{2} \mathrm{tr}(\mathbf{U}^\top \mathbf{A}_{\mathcal{Z}} \mathbf{U}),$$

where we omit $\|\mathbf{f}_t\|^2$ since it does not affect the upper bound. We add $\sum_t a_t$ on the both sides with $a_t = \frac{1}{1 - \mathbf{x}_t^\top \mathbf{A}_{t-1}^{-1} \mathbf{x}_t} \geq 1$,

$$\begin{aligned}
&\sum_t (1 - \mathbf{f}_t \mathbf{y}_t - \sigma_t) \leq \sum_t (a_t - \mathbf{f}_t \mathbf{y}_t - \sigma_t) \\
\leq& \sum_t a_t (1 - \mathbf{y}_t \cdot \mathbf{U}^\top \mathbf{x}_t) + \frac{1}{2} \mathrm{tr}(\mathbf{U}^\top \mathbf{A}_{\mathcal{Z}} \mathbf{U}) \leq \sum_t a_t \tilde{\mathcal{L}}(\mathbf{y}_t \cdot \mathbf{U}^\top \mathbf{x}_t) + \frac{1}{2} \mathrm{tr}(\mathbf{U}^\top \mathbf{A}_{\mathcal{Z}} \mathbf{U}),
\end{aligned} \tag{1}$$

where the last inequality holds due to hinge loss $\tilde{\mathcal{L}}(x) = \max(0, 1 - x) \geq 1 - x$. There are two types of update trials: (I) when an error occurs, i.e., $t \in \mathcal{M}$ and $-\mathbf{f}_t \mathbf{y}_t \geq 0$,

$$\sum_t (1 - \mathbf{f}_t \mathbf{y}_t - \sigma_t) \geq M - \sum_{t \in \mathcal{M}} \sigma_t;$$

and (II) when no error occurs, i.e., $t \in \mathcal{D}$ and $0 \le \mathbf{f}_t \mathbf{y}_t \le \sigma_t \Rightarrow -\mathbf{f}_t \mathbf{y}_t + \sigma_t \ge 0$,

$$\sum_t (1 - \mathbf{f}_t \mathbf{y}_t + \sigma_t - 2\sigma_t) \ge D - 2 \sum_{t \in \mathcal{D}} \sigma_t.$$

Combining two cases with the upper bound (1), and substituting the inequality $\sum_{t \in \mathcal{Z}} \sigma_t \le \frac{b}{2(b-1)} \log(\frac{1}{b} \mathbf{A}_{\mathcal{Z}})$ inspired by [3], we finish the proof. $\square$

2 Proof of Theory 3

Note that the update rule is

$$\mathbf{B}_t = \mathbf{B}_{t-1} + a_t \mathbf{x}_t \mathbf{y}_t^\top \quad \mathbf{A}_t = \mathbf{A}_{t-1} + a_t \mathbf{x}_t \mathbf{x}_t^\top,$$

or $\mathbf{A}_t^{-1} = \mathbf{A}_{t-1}^{-1} - \mathbf{A}_{t-1}^{-1} \mathbf{x}_t \mathbf{x}_t^\top \mathbf{A}_{t-1}^{-1}$ according to Woodbury identity. Given an annotation $\mathcal{D}_t(\mathbf{U}, \mathbf{V}) = \|\mathbf{U} - \mathbf{V}\|_{\mathbf{A}_t}^2$, the following equations can be derived,

$$a_t \left(\|\mathbf{y}_t - \mathbf{W}_{t-1}^\top \mathbf{x}_t\|^2 - \|\mathbf{y}_t - \mathbf{U}^\top \mathbf{x}_t\|^2 \right) = \mathcal{D}_{t-1}(\mathbf{U}, \mathbf{W}_{t-1}) - \mathcal{D}_t(\mathbf{U}, \mathbf{W}_t) + \mathcal{D}_t(\mathbf{W}_{t-1}, \mathbf{W}_t), \tag{2}$$

$$a_t^2 \|\mathbf{y}_t - \mathbf{W}_{t-1}^\top \mathbf{x}_t\|^2 \mathbf{x}_t^\top \mathbf{A}_t^{-1} \mathbf{x}_t = \mathcal{D}_t(\mathbf{W}_{t-1}, \mathbf{W}_t). \tag{3}$$

Assume that $\ell_t(\mathbf{U}) = \|\mathbf{y}_t - \mathbf{U}^\top \mathbf{x}_t\|^2 \le r, (r > 1)$ for any $\mathbf{U} \in \mathbb{R}^{n \times K}$ and $a_t = \frac{1}{1 - \mathbf{x}_t^\top \mathbf{A}_{t-1}^{-1} \mathbf{x}_t} \le \frac{b}{b-1}$ for any $t \in [T]$, the cumulative sum of Eq. (3) can be bounded,

$$\begin{aligned}
\sum_{t=1}^T \mathcal{D}_t(\mathbf{W}_{t-1}, \mathbf{W}_t) &= \sum_{t=1}^T a_t^2 \|\mathbf{y}_t - \mathbf{W}_{t-1}^\top \mathbf{x}_t\|^2 \mathbf{x}_t^\top \mathbf{A}_t^{-1} \mathbf{x}_t \\
&\le \frac{rb}{b-1} \sum_{t=1}^T a_t \mathbf{x}_t^\top \mathbf{A}_t^{-1} \mathbf{x}_t \le \frac{rb}{b-1} \sum_{t=1}^T \log \frac{|\mathbf{A}_t|}{|\mathbf{A}_t - a_t \mathbf{x}_t \mathbf{x}_t^\top|} \\
&= \frac{rb}{b-1} \log |\frac{1}{b} \mathbf{A}_T|,
\end{aligned} \tag{4}$$

where the last inequality is similar to the proof of Theorem 5 in [3]. Equipped with the bound (4), the cumulative sum of Eq. (2) can be bounded

$$\sum_{s=1}^{t-1} a_s \left(\|\mathbf{y}_s - \mathbf{W}_{s-1}^\top \mathbf{x}_s\|^2 - \|\mathbf{y}_s - \mathbf{U}^\top \mathbf{x}_s\|^2 \right) \le \mathcal{D}_0(\mathbf{U}, \mathbf{0}) - \mathcal{D}_{t-1}(\mathbf{U}, \mathbf{W}_{t-1}) + \frac{rb}{b-1} \log |\frac{1}{b} \mathbf{A}_{t-1}|. \tag{5}$$

According to the Cauchy-Schwarz inequality (dual norms), we have

$$\|\hat{\Delta}_t - \Delta_t\|^2 = \|(\mathbf{W}_{t-1}^\top - \mathbf{U}^\top) \mathbf{x}_t\|^2 \le 2 \mathbf{x}_t^\top \mathbf{A}_{t-1}^{-1} \mathbf{x}_t \mathcal{D}_{t-1}(\mathbf{U}, \mathbf{W}_{t-1}). \tag{6}$$

Due to $a_t = \frac{1}{1 - \mathbf{x}_t^\top \mathbf{A}_{t-1}^{-1} \mathbf{x}_t} > 1$ for all t, we can infer according to the proof of Lemma 2 to Lemma 5 in [1],

$$\sum_{s=1}^{t-1} a_s \left(\|\mathbf{y}_s - \mathbf{W}_{s-1}^\top \mathbf{x}_s\|^2 - \|\mathbf{y}_s - \mathbf{U}^\top \mathbf{x}_s\|^2 \right) \ge -36 \log \frac{t+4}{\delta} \tag{7}$$

holds with probability at least $1-\delta$ over the t rounds. Substituting Eq. (5)-Eq. (7), we obtain for any $\mathbf{U} \in \mathbb{R}^{n\times K}$,

$$\|\hat{\Delta}_t - \Delta_t\|^2 \leq 2\mathbf{x}^\top \mathbf{A}_{t-1}^{-1}\mathbf{x}_t \left(b\|\mathbf{U}\|_F^2 + \frac{rb}{b-1}\log|\frac{1}{b}\mathbf{A}_{t-1}| + 36\log\frac{t+4}{\delta} \right) \quad (8)$$

hold with probability at least $1-\delta$ over the $t-1$ rounds.

Since $\mathbf{A}_t^{-1} \preceq \mathbf{A}_{t-1}^{-1}$, we have $\mathbf{x}_t^\top \mathbf{A}_{t-1}^{-1}\mathbf{x}_t \leq \ldots \leq \mathbf{x}_t^\top \mathbf{A}_0^{-1}\mathbf{x}_t = \frac{1}{b}\|\mathbf{x}_t\|^2$. Assume that $\|\mathbf{x}_t\| \leq 1$, we infer that $0 \leq \mathbf{x}_t^\top \mathbf{A}_{t-1}^{-1}\mathbf{x}_t \leq \frac{1}{b}$ where we let $b > 1$. Thus,

$$1 - \mathbf{x}_t^\top \mathbf{A}_{t-1}^{-1}\mathbf{x}_t \geq 1 - \frac{1}{b} \;\Rightarrow\; \frac{b}{b-1}(1 - \mathbf{x}_t^\top \mathbf{A}_{t-1}^{-1}\mathbf{x}_t) \geq 1.$$

Multiplying $\mathbf{x}_t^\top \mathbf{A}_{t-1}^{-1}\mathbf{x}_t$ on both sides, we obtain

$$\mathbf{x}_t^\top \mathbf{A}_{t-1}^{-1}\mathbf{x}_t \leq \frac{b}{b-1}\mathbf{x}_t^\top(\mathbf{A}_{t-1}^{-1} - \mathbf{A}_{t-1}^{-1}\mathbf{x}_t\mathbf{x}_t^\top\mathbf{A}_{t-1}^{-1})\mathbf{x}_t = \frac{b}{b-1}\mathbf{x}_t^\top \mathbf{A}_t^{-1}\mathbf{x}_t. \quad (9)$$

Substituting Eq. (9) into Eq. (8), we obtain,

$$\begin{aligned}
\|\hat{\Delta}_t - \Delta_t\|^2 &\leq \frac{1}{2}(\frac{b}{b-1})^2\mathbf{x}_t^\top \mathbf{A}_t^{-1}\mathbf{x}_t \left(4(b-1)\|\mathbf{U}\|_F^2 + 4r\log|\frac{1}{b}\mathbf{A}_{t-1}| + \frac{b-1}{b}144\log\frac{t+4}{\delta} \right) \\
&\leq \sigma_t(\frac{b}{b-1})^2 \left(4(b-1)\|\mathbf{U}\|_F^2 + 4r\log|\frac{1}{b}\mathbf{A}_{t-1}| + 144\log\frac{t+4}{\delta} \right),
\end{aligned}$$

where $\sigma_t = \frac{1}{2}a_t^2\mathbf{x}_t^\top \mathbf{A}_t^{-1}\mathbf{x}_t$ and the last inequality holds due to $a_t \geq 1$. We assume that

$$\varphi_t^2 = (\frac{b}{b-1})^2 \left(4(b-1)\|\mathbf{U}\|_F^2 + 4r\log|\frac{1}{b}\mathbf{A}_{t-1}| + 144\log\frac{t+4}{\delta} \right), \quad (10)$$

and bound the cumulative sum of σ_t for $t \in [T]$,

$$\sum_{s=1}^T \sigma_s = \frac{1}{2}\sum_{s=1}^T a_s^2\mathbf{x}_s^\top \mathbf{A}_s^{-1}\mathbf{x}_s \leq \frac{b}{2(b-1)}\log|\frac{1}{b}\mathbf{A}_T| \leq \frac{b}{2(b-1)}Kn\log(1+\frac{T}{Knb}).$$

Assume that

$$H_1 = 2(b-1)\|\mathbf{U}\|_F^2 + 72\log\frac{t+4}{\delta}, \quad H_2 = 2Knr\log(1+\frac{T}{Knb})$$

we have that

$$\sum_{t=1}^T \|\hat{\Delta}_t - \Delta_t\|^2 \leq \sum_{t=1}^T \varphi_t^2\sigma_t \leq 2(H_1+H_2)(\frac{b}{b-1})^2\sum_{t=1}^T \sigma_t \leq (\frac{b}{b-1})^3(H_1+H_2)H_2. \quad (11)$$

with probability at least $1-\delta$ over T rounds. Since $\sum_{t=1}^T A_t^2 \le M$ implies $\sum_{t=1}^T A_t \le \sqrt{TM}$, we obtain

$$\begin{aligned}\sum_{t=1}^{T}(\mathbb{P}_t(y_t \neq \hat{y}_t) - \mathbb{P}_t(y_t \neq y_t^*)) \le \sum_{t=1}^{T} |\Delta_t - \hat{\Delta}_t| \\ \le \sqrt{(\frac{b}{b-1})^3 T}\sqrt{H_1 H_2 + H_2^2} \le \sqrt{(\frac{b}{b-1})^3 T}(\sqrt{H_1 H_2} + H_2),\end{aligned} \tag{12}$$

where the last inequality holds due to $\sqrt{A+B} \le \sqrt{A} + \sqrt{B}$.

References

[1] Koby Crammer and Claudio Gentile. Multiclass classification with bandit feedback using adaptive regularization. *Machine learning*, 90(3):347–383, 2013.

[2] Jürgen Forster. On relative loss bounds in generalized linear regression. In *Fundamentals of Computation Theory*, pages 269–280, 1999.

[3] Edward Moroshko and Koby Crammer. Weighted last-step min–max algorithm with improved sub-logarithmic regret. *Theoretical Computer Science*, 558:107–124, 2014.