January 29, 2018

1 Proof of Theory 2

The algorithm updates model whenever adaptive margin is less than 0 ($Z_t = 1$). If there is no update, $\mathbf{U}_t = \mathbf{U}_{t-1}$ yields $\inf_{\mathbf{U}} G_t(\mathbf{U}) = \inf_{\mathbf{U}} G_{t-1}(\mathbf{U})$. Inspired by the proof in [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. Summing over $t = 1, \dots, T$, we obtain with expanding the square,

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

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

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

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

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

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 $0 \le \mathbf{f}_t \mathbf{y}_t \le \sigma_t \Rightarrow -\mathbf{f}_t \mathbf{y}_t + \sigma_t \ge 0$,

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

Combine two cases with $\sum_{t=1}^{T} \sigma_t \le \frac{b}{2(b-1)} \ln(\frac{1}{b}\mathbf{A}_T)$, we finish the proof. $\square$

2 Proof of Theory 3

From the update rule that

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

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, we can infer the following two equations with an annotation $\mathcal{D}_t(\mathbf{U}, \mathbf{V}) = \|\mathbf{U} - \mathbf{V}\|_{\mathbf{A}_t}^2$,

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

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

Assume that $\ell_t = \|\mathbf{y}_t - \mathbf{W}_{t-1}^\top \mathbf{x}_t\|^2 \le r$ 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. (2) can be bounded,

$$\begin{aligned} \sum_t \mathcal{D}_t(\mathbf{W}_{t-1}, \mathbf{W}_t) &= \sum_t a_t^2 \|\mathbf{y}_t - \mathbf{W}_{t-1}^\top \mathbf{x}_t\|^2 \mathbf{x}_t^\top \mathbf{A}_t^{-1} \mathbf{x}_t \\ &\le \frac{rb}{b-1} \sum_t a_t \mathbf{x}_t^\top \mathbf{A}_t^{-1} \mathbf{x}_t \le \frac{rb}{b-1} \sum_t \ln \frac{|\mathbf{A}_t|}{|\mathbf{A}_{t-1}|} \\ &= \frac{rb}{b-1} \ln |\frac{1}{b} \mathbf{A}_T|, \end{aligned} \tag{3}$$

which is similar to the proof of Theorem 5 in [3]. Equipped with the bound (3), the cumulative sum of Eq. (1) can be bounded

$$\sum_{t=1}^{t-1} a_t \left(\|\mathbf{y}_t - \mathbf{W}_{t-1}^\top \mathbf{x}_t\|^2 - \|\mathbf{y}_t - \mathbf{U}^\top \mathbf{x}_t\|^2 \right) \le \mathcal{D}_0(\mathbf{U}, \mathbf{0}) - \mathcal{D}_{t-1}(\mathbf{U}, \mathbf{W}_{t-1}) + \frac{rb}{b-1} \ln |\frac{1}{b} \mathbf{A}_T|. \tag{4}$$

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

$$\|\hat{\Delta}_t - \Delta_t\|^2 = \|(\mathbf{W}_{t-1}^\top - \mathbf{U}^\top) \mathbf{x}_t\|^2 \le 2 \mathbf{x}^\top \mathbf{A}_{t-1}^{-1} \mathbf{x}_t \mathcal{D}_{t-1}(\mathbf{U}, \mathbf{W}_{t-1}). \tag{5}$$

According to the proof of Lemma 2 to Lemma 5 in [1], we can infer that

$$\sum_{t=1}^{t-1} a_t \left(\|\mathbf{y}_t - \mathbf{W}_{t-1}^\top \mathbf{x}_t\|^2 - \|\mathbf{y}_t - \mathbf{U}^\top \mathbf{x}_t\|^2 \right) \ge -36 \log \frac{t+4}{\delta} \tag{6}$$

holds with probability at least $1-\delta$ over the t rounds. From Eq. (4) to Eq. (6), we can infer that for any $\mathbf{U}$,

$$\|\hat{\Delta}_t - \Delta_t\|^2 \leq 2\mathbf{x}^\top \mathbf{A}_{t-1}^{-1}\mathbf{x}_t \left(b\|\mathbf{U}\|_F^2 + \frac{rb}{b-1} \ln |\frac{1}{b}\mathbf{A}_T| + 36 \log \frac{t+4}{\delta} \right) \tag{7}$$

hold with probability at least $1-\delta$ over the t 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 $0 \leq \mathbf{x}_t^\top \mathbf{A}_{t-1}^{-1}\mathbf{x}_t \leq 1/b$. Thus, we infer that

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

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

$$\mathbf{x}_t^\top \mathbf{A}_{t-1}^{-1}\mathbf{x}_t \leq \frac{b}{b-1}\mathbf{x}_t^\top(\mathbf{A}_{t-1}^{-1} - \mathbf{A}_{t-1}^{-1}\mathbf{x}_t\mathbf{x}_t^\top\mathbf{A}_{t-1}^{-1})\mathbf{x}_t = \frac{b}{b-1}\mathbf{x}_t^\top \mathbf{A}_t^{-1}\mathbf{x}_t. \tag{8}$$

Substituting Eq. (8) into Eq. (7), we obtain with $\sigma_t = \frac{1}{2}a_t^2\mathbf{x}_t^\top \mathbf{A}_t^{-1}\mathbf{x}_t$,

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

where the second inequality is due to $a_t \geq 1$. We assume that

$$\varphi_t^2 = (\frac{b}{b-1})^2 \left(4(b-1)\|\mathbf{U}\|_F^2 + 4r\ln|\frac{1}{b}\mathbf{A}_T| + 144\log\frac{t+4}{\delta} \right), \tag{9}$$

and we bound

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

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

$$\sum_{t=1}^{T}\varphi_t^2\sigma_t \leq 2(H_1+H_2)\sum_{t=1}^{T}\sigma_t \leq (\frac{b}{b-1})^3(H_1+H_2)H_2. \tag{10}$$

with probability at least $1-\delta$ over T rounds. Finally, since $\sum_{t=1}^{T}\varphi_t^2\sigma_t \leq$

$(\frac{b}{b-1})^3(H_1+H_2)H_2$ implies that

$$\begin{aligned}
&\sum_{t=1}^{T}(\mathbb{P}_t(y_t \neq \hat{y}_t) - \mathbb{P}_t(y_t \neq y_t^*)) = \sum_{t=1}^{T}\frac{|\Delta_t - \hat{\Delta}_t|}{2} \\
\leq &\sum_{t=1}^{T}\varphi_t\sqrt{\sigma_t} \leq \sqrt{T(\frac{b}{b-1})^3(H_1+H_2)H_2} \qquad (11)\\
= &\sqrt{(\frac{b}{b-1})^3 T}\sqrt{H_1H_2+H_2^2} \leq \sqrt{(\frac{b}{b-1})^3 T}(\sqrt{H_1H_2}+H_2),
\end{aligned}$$

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

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.