Fix-Margin and Adaptive-Margin Theoretical Comparison

March 21, 2018

1 Fix-Margin Algorithm Theorem

Lemma 1. *Given an arbitrary node sequence* $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_T, y_T)$*, an online algorithm predicts with* $\hat{y}_T = \arg\max_{i\in[K]}(\mathbf{B}_{T-1}^\top \mathbf{A}_{T-1}^{-1}\mathbf{x}_T)_i$*, where* $\mathbf{A}_T$ *and* $\mathbf{B}_T$ *are defined in Eq. (6), and updates model when an error occurs, i.e.,* $\hat{\Delta}_t = \mathbf{f}_t \cdot \mathbf{y}_t \leq 0$ *(called Fixed Margin), then the following inequality holds,*

$$M \leq \sum_{t\in\mathcal{M}} a_t \tilde{\mathcal{L}}(\mathbf{y}_t \cdot \mathbf{U}^\top \mathbf{x}_t) + \frac{1}{2} tr(\mathbf{U}^\top \mathbf{A}_{\mathcal{M}} \mathbf{U}) + \frac{b}{2(b-1)} \log |\frac{1}{b}\mathbf{A}_{\mathcal{M}}|.$$

Proof: Since a update is issued when an error occurs, the update trials are defined as $\mathcal{M} = \{t : \hat{\Delta}_t \leq 0, \hat{y}_t \neq y_t\}$ with $M = |\mathcal{M}|$ includes the indices on which an error occurs. Given $\ell_t(alg) = \|\mathbf{y}_t - \mathbf{f}_t\|^2$, we derive when $t \in \mathcal{M}$,

$$\begin{aligned}
&\ell_t(alg) + \inf_{\mathbf{U}}(b\|\mathbf{U}\|^2 + L_{t-1}^{\mathbf{a}}(\mathbf{U})) - \inf_{\mathbf{U}}(b\|\mathbf{U}\|^2 + L_t^{\mathbf{a}}(\mathbf{U}))\\
=&\|\mathbf{f}_t - \mathbf{y}_t\|^2 - a_t\|\mathbf{y}_t\|^2 - \mathrm{tr}(\mathbf{B}_{t-1}^\top \mathbf{A}_{t-1}^{-1}\mathbf{B}_{t-1}) + \mathrm{tr}(\mathbf{B}_t^\top \mathbf{A}_t^{-1}\mathbf{B}_t)\\
=&(1-a_t)\|\mathbf{y}_t\|^2 - 2\mathbf{y}_t \cdot \mathbf{f}_t + \|\mathbf{f}_t\|^2 - \mathrm{tr}(\mathbf{B}_{t-1}^\top \mathbf{A}_{t-1}^{-1}\mathbf{B}_{t-1}) + \mathrm{tr}(\mathbf{B}_t^\top \mathbf{A}_t^{-1}\mathbf{B}_t)\\
=&(1-a_t)\|\mathbf{y}_t\|^2 - 2\mathbf{y}_t \cdot (a_t\mathbf{B}_{t-1}^\top \mathbf{A}_t^{-1}\mathbf{x}_t) + \mathrm{tr}((\mathbf{B}_{t-1} + a_t\mathbf{x}_t\mathbf{y}_t^\top)^\top \mathbf{A}_t^{-1}(\mathbf{B}_{t-1} + a_t\mathbf{x}_t\mathbf{y}_t^\top))\\
&+ \mathrm{tr}(\mathbf{B}_{t-1}^\top \mathbf{A}_{t-1}^{-1}\mathbf{x}_t\mathbf{x}_t^\top \mathbf{A}_{t-1}^{-1}\mathbf{B}_{t-1}) - \mathrm{tr}(\mathbf{B}_{t-1}^\top \mathbf{A}_{t-1}^{-1}\mathbf{B}_{t-1})\\
=&\mathrm{tr}(\mathbf{B}_{t-1}^\top(\mathbf{A}_{t-1}^{-1}\mathbf{x}_t\mathbf{x}_t^\top \mathbf{A}_{t-1}^{-1} - \mathbf{A}_{t-1}^{-1} + \mathbf{A}_t^{-1})\mathbf{B}_{t-1}) + \mathrm{tr}(a_t^2\mathbf{y}_t\mathbf{x}_t^\top \mathbf{A}_t^{-1}\mathbf{x}_t\mathbf{y}_t^\top) + (1-a_t)\|\mathbf{y}_t\|^2\\
=&(a_t^2\mathbf{x}_t^\top \mathbf{A}_t^{-1}\mathbf{x}_t + 1 - a_t)\mathrm{tr}(\mathbf{y}_t\mathbf{y}_t^\top) = a_t^2\mathbf{x}_t^\top \mathbf{A}_t^{-1}\mathbf{x}_t - a_t + 1.
\end{aligned}$$

When no error occurs, $\mathbf{U}_t = \mathbf{U}_{t-1}$ yields $\inf_{\mathbf{U}} G_t(\mathbf{U}) = \inf_{\mathbf{U}} G_{t-1}(\mathbf{U})$. When an error occurs, we conclude that

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

holds for all trial $t \in \mathcal{M}$, which is similar to the proof of [1]. Summing over

$t = 1, \ldots, T$ with $\|\mathbf{y}_t\|^2 = 1$, we obtain with expanding the square,

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

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

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

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 $a_t = \frac{1}{1-\mathbf{x}_t^\top\mathbf{A}_{t-1}^{-1}\mathbf{x}_t} \geq 1$,

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

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

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

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

2 Adaptive-Margin Algorithm Theorem

Theorem 1. *Algorithm 1 runs on an arbitrary node sequence* $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_T, y_T)$ *and update model when* $\Theta_t = \hat{\Delta}_t - \sigma_t \leq 0$. *Let* $\tilde{\mathcal{L}}(x) = \max(0, 1-x)$ *be hinge loss, for any* $\mathbf{U} \in \mathbb{R}^{n\times K}$, *the following inequality holds,*

$$M \leq \sum_{t\in\mathcal{Z}} a_t\tilde{\mathcal{L}}(\mathbf{y}_t \cdot \mathbf{U}^\top\mathbf{x}_t) + \frac{1}{2}tr(\mathbf{U}^\top\mathbf{A}_\mathcal{Z}\mathbf{U}) + \frac{b}{b-1}\log|\frac{1}{b}\mathbf{A}_\mathcal{Z}| - D.$$

Proof: In Algorithm 1, the update trials are partitioned into two disjoint sets, $\mathcal{M} = \{t : \hat{\Delta}_t \leq 0, \hat{y}_t \neq y_t\}$ with $M = |\mathcal{M}|$ includes the indices on which an update is issued when an error occurs, and $\mathcal{D} = \{t : 0 < \hat{\Delta}_t < \sigma_t, \hat{y}_t = y_t\}$ with $D = |\mathcal{D}|$ includes the indices on which an aggressive update is issued for low-confident prediction, even if the prediction is correct. Let $\mathcal{Z} = \{t : Z_t = 1\}$

with $Z = |\mathcal{Z}|$ be the update trials containing $Z = M + D$. Similar with Eq. (1) in lemma 1, we derive for $t \in \mathcal{Z}$,

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

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 $0 \leq \mathbf{f}_t \mathbf{y}_t \leq \sigma_t \Rightarrow -\mathbf{f}_t \mathbf{y}_t + \sigma_t \geq 0$,

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

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

Conclusion: Due to the deduction of the low-confident update trials $|\mathcal{D}|$, the error bound of Algorithm 1 can be lower than that of the weighted min-max algorithm using error-driven update rules (Shown in Lemma 1).

References

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

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