Supplementary Material for "Faster Online Learning of Optimal Threshold for Consistent F-measure Optimization"

Mingrui Liu[*†], Xiaoxuan Zhang[*†], Xun Zhou[‡], Tianbao Yang[†]
[†]Department of Computer Science, The University of Iowa, Iowa City, IA 52242, USA
[‡]Department of Management Sciences, The University of Iowa, Iowa City, IA 52242, USA
`mingrui-liu, xiaoxuan-zhang, xun-zhou, tianbao-yang@uiowa.edu`

1 Extension to Other Metrics

In this section, we consider the extension of the proposed method to other metrics, in particular Jaccard similarity coefficient, and F_β measure.

Let us first consider Jaccard similarity coefficient (JAC) [2]:

$$\mathrm{JAC}(f) = \frac{\int_{\mathcal{X}} \eta(\mathbf{x}) f(\mathbf{x}) d\mu(\mathbf{x})}{\pi + \int_{\mathcal{X}} f(\mathbf{x}) d\mu(\mathbf{x}) - \int_{\mathcal{X}} \eta(\mathbf{x}) f(\mathbf{x}) d\mu(\mathbf{x})}.$$

Then we have

$$\frac{1}{\mathrm{JAC}(f)} = \frac{\pi + \int_{\mathcal{X}} f(\mathbf{x}) d\mu(\mathbf{x})}{\int_{\mathcal{X}} \eta(\mathbf{x}) f(\mathbf{x}) d\mu(\mathbf{x})} - 1 = \frac{2}{F(f)} - 1.$$

Therefore

$$\mathrm{JAC}(f) = \frac{F(f)}{2 - F(f)}.$$

If $F(f)$ is maximized so is $\mathrm{JAC}(f)$. According to [2], the optimal threshold $\theta_{\mathrm{JAC},*}$ for maximizing $\mathrm{JCA}(\eta_\theta(\mathbf{x}))$ is given by $\theta_{\mathrm{JAC},*} = \frac{\mathrm{JAC}_*}{1+\mathrm{JAC}_*} = F_*/2 = \theta_{F,*}$, where $\theta_{F,*}$ is the optimal threshold for F-measure maximization. Given an estimate of θ_F for F-measure optimization, we can set the threshold for JAC maximization as $\theta_{\mathrm{JAC}} = \theta_F$, and when $\theta_F \to \theta_{F,*}$, we have $\theta_{\mathrm{JAC}} \to \theta_{\mathrm{JAC},*}$. As a result, the proposed algorithm FOFO is still applicable.

Next, let us consider F_β-measure:

$$F_\beta(f) = \frac{(1+\beta^2) \int_{\mathcal{X}} \eta(\mathbf{x}) f(\mathbf{x}) d\mu(\mathbf{x})}{\beta^2 \pi + \int_{\mathcal{X}} f(\mathbf{x}) d\mu(\mathbf{x})}.$$

Following the same analysis as in [3][Lemma 13, Lemma 14], we can have that $F_\beta(\eta_\theta)$ is maximized at a point $\theta_{\beta,*}$ that is the root of the following equation:

$$\pi\beta^2\theta - \mathbb{E}_{\mathbf{x}}\left[(\eta(\mathbf{x}) - \theta)_+\right] = 0,$$

which is the optimal solution of the following strongly convex function

$$Q(\theta) \triangleq \frac{1}{2}\mathbb{E}_{\mathbf{x}}\left[(\eta(\mathbf{x}) - \theta)_+^2\right] + \frac{1}{2}\pi\beta^2\theta^2.$$

[*]equal contribution

32nd Conference on Neural Information Processing Systems (NIPS 2018), Montréal, Canada.

For the optimal threshold $\theta_{\beta,*}$ and optimal $F_{\beta,*}$, we have $\theta_{\beta,*} = \frac{F_{\beta,*}}{1+\beta^2}$. Hence, we can search for $\theta_{\beta,*}$ by solving the following problem:

$$\min_{\theta \in [0, 1/(1+\beta^2)]} Q(\theta) \triangleq \frac{1}{2}\mathbb{E}_{\mathbf{x}}\left[(\eta(\mathbf{x}) - \theta)_+^2\right] + \frac{1}{2}\pi\beta\theta^2.$$

We can modify FOFO a little to account for this change.

2 Missing Proofs

2.1 $\mathbf{w}_*$ minimizes the expected logistic loss

Under the assumption that

$$\eta(\mathbf{x}) = \Pr(y = 1|\mathbf{x}, \mathbf{w}_*) = \frac{1}{1 + \exp(-\mathbf{w}_*^\top \phi(\mathbf{x}))},$$

we prove that $\mathbf{w}_*$ is the minimizer of the following problem:

$$\min_{\mathbf{w} \in \mathbb{R}^d} L(\mathbf{w}) \triangleq \mathbb{E}_{\mathbf{x},y} \log(1 + \exp(-(2y-1)\mathbf{w}^\top \phi(\mathbf{x}))). \tag{7}$$

Using variable change $\tilde{y} = 2y - 1$, $\Pr(\tilde{y}|\mathbf{x}, \mathbf{w}_*) = \frac{1}{1+\exp(-\tilde{y}\mathbf{w}_*^\top \phi(\mathbf{x}))}$, and $L(\mathbf{w}) = \mathbb{E}_{\mathbf{x},\tilde{y}} \log(1 + \exp(-\tilde{y}\mathbf{w}^\top \phi(\mathbf{x})))$. Then,

$$\begin{aligned} L(\mathbf{w}) = \mathbb{E}_{\mathbf{x},\tilde{y}} \log(1 + \exp(-\tilde{y}\mathbf{w}^\top \phi(\mathbf{x}))) &= -\int_{\mathbf{x}} \mathbb{E}_{\tilde{y}|\mathbf{x}}[\log \Pr(\tilde{y}|\mathbf{x}, \mathbf{w})] d\mu(\mathbf{x}) \\ &= \int_{\mathbf{x}} \left[-\sum_{\tilde{y}} \Pr(\tilde{y}|\mathbf{x}, \mathbf{w}_*) \log \Pr(\tilde{y}|\mathbf{x}, \mathbf{w}) \right] d\mu(\mathbf{x}) \end{aligned}$$

Note that the term in the square brackets is the KL divergence between two distributions $\Pr(\tilde{y}|\mathbf{x}, \mathbf{w}_*)$ and $\Pr(\tilde{y}|\mathbf{x}, \mathbf{w})$ plus a constant independent of $\mathbf{w}$. Therefore $\mathbf{w} = \mathbf{w}_*$ minimizes this term and hence minimizes $L(\mathbf{w})$.

2.2 Proof of Lemma 1

We prove the strong convexity parameter here.

$$\begin{aligned} Q(\theta) &= \frac{1}{2}\int_{\eta(\mathbf{x})\geq\theta} (\theta^2 - 2\eta(\mathbf{x})\theta + \eta(\mathbf{x})^2) d\mu(\mathbf{x}) + \frac{1}{2}\pi\theta^2 \\ &= \frac{1}{2}\theta^2(\rho_\theta + \pi) - \theta \int_{\eta(\mathbf{x})\geq\theta} \eta(\mathbf{x}) d\mu(\mathbf{x}) + c \end{aligned}$$

where $\rho_\theta = \int_{\eta(\mathbf{x})\geq\theta} d\mu(\mathbf{x})$, c is a constant independent of θ. Then we can see the strong convexity parameter of $Q(\theta)$ over $[0, 0.5]$ is $\pi + \min_{\theta\in[0,0.5]} \rho_\theta$.

2.3 Proof of Lemma 2

Proof. For $A \subseteq \mathcal{X}$, define $\rho(A) = \int_{\mathbf{x}\in\mathcal{A}} 1 \cdot d\mu(\mathbf{x}) = \Pr(\mathbf{x} \in A)$. Let $\mathcal{X}_* = \{\mathbf{x} \in \mathcal{X} | \eta(\mathbf{x}) \geq \theta_*\}$ and $\mathcal{X}' = \{\mathbf{x} \in \mathcal{X} | \eta(\mathbf{x}) \geq \theta\}$, and note that $\eta_\theta(\mathbf{x}) = \mathbb{I}(\eta(\mathbf{x}) \geq \theta)$, we have

$$\frac{1}{2}F(\eta_\theta) = \frac{\int_{\mathcal{X}'} \eta(\mathbf{x}) d\mu(\mathbf{x})}{\pi + \rho(\mathcal{X}')} \tag{8}$$

According to [3], $F(\eta_{\theta_*}) = 2\theta_*$. Thus,

$$\theta_* = \frac{1}{2}F(\eta_{\theta_*}) = \frac{\int_{\mathcal{X}_*} \eta(\mathbf{x}) d\mu(\mathbf{x})}{\pi + \rho(\mathcal{X}_*)}$$

$$\int_{\mathcal{X}_*} \eta(\mathbf{x}) d\mu(\mathbf{x}) = \theta_*(\pi + \rho(\mathcal{X}_*)) \tag{9}$$

Then we consider two cases based on the relation between θ and θ_*.

Case 1. $0 \le \theta \le \theta_*$

Since $\mathcal{X}_* \subseteq \mathcal{X}'$, let $A = \mathcal{X}' - \mathcal{X}_* = \{\mathbf{x} \in \mathcal{X} | \theta \le \eta(\mathbf{x}) < \theta_*\}$. From (8),

$$\frac{1}{2}F(\eta_\theta) = \frac{\int_{\mathcal{X}_*} \eta(\mathbf{x}) d\mu(\mathbf{x}) + \int_A \eta(\mathbf{x}) d\mu(\mathbf{x})}{\pi + \rho(\mathcal{X}_*) + \rho(A)}$$

On A, we have $\eta(\mathbf{x}) \ge \theta$, thus $\int_A \eta(\mathbf{x}) d\mu(\mathbf{x}) \ge \theta\rho(A)$. From (9), we have

$$\begin{aligned}\frac{1}{2}F(\eta_\theta) \ge & \frac{\theta_*(\pi + \rho(\mathcal{X}_*)) + \theta\rho(A)}{\pi + \rho(\mathcal{X}_*) + \rho(A)} = \frac{\theta_*(\pi + \rho(\mathcal{X}_*)) + \theta_*\rho(A) - \theta_*\rho(A) + \theta\rho(A)}{\pi + \rho(\mathcal{X}_*) + \rho(A)} \\ = & \theta_* - \frac{(\theta_* - \theta)\rho(A)}{\pi + \rho(\mathcal{X}_*) + \rho(A)} \ge \theta_* - (\theta_* - \theta) = \theta\end{aligned}$$

Thus $F(\eta_{\theta_*}) - F(\eta_\theta) \le 2(\theta_* - \theta) \le \frac{2}{\pi}|\theta_* - \theta|$.

Case 2. $\theta_* < \theta \le 0.5$

Since $\mathcal{X}' \subseteq \mathcal{X}_*$, let $A = \mathcal{X}_* - \mathcal{X}' = \{\mathbf{x} \in \mathcal{X} | \theta_* \le \eta(\mathbf{x}) < \theta\}$. From (8),

$$\frac{1}{2}F(\eta_\theta) = \frac{\int_{\mathcal{X}_*} \eta(\mathbf{x}) d\mu(\mathbf{x}) - \int_A \eta(\mathbf{x}) d\mu(\mathbf{x})}{\pi + \rho(\mathcal{X}_*) - \rho(A)}$$

On A, we have $\eta(\mathbf{x}) < \theta$, thus $\int_A \eta(\mathbf{x}) d\mu(\mathbf{x}) \le \theta\rho(A)$. From (9), we have

$$\begin{aligned}\frac{1}{2}F(\eta_\theta) \ge & \frac{\theta_*(\pi + \rho(\mathcal{X}_*)) - \theta\rho(A)}{\pi + \rho(\mathcal{X}_*) - \rho(A)} = \frac{\theta_*(\pi + \rho(\mathcal{X}_*)) - \theta_*\rho(A) + \theta_*\rho(A) - \theta\rho(A)}{\pi + \rho(\mathcal{X}_*) - \rho(A)} \\ = & \theta_* - \frac{(\theta - \theta_*)\rho(A)}{\pi + \rho(\mathcal{X}_*) - \rho(A)} \ge \theta_* - \frac{1}{\pi}(\theta - \theta_*).\end{aligned}$$

The last step holds because $\rho(A) < 1$ and $\rho(\mathcal{X}_*) - \rho(A) = \rho(\mathcal{X}') \ge 0$. Then we have $F(\eta_{\theta_*}) - F(\eta_\theta) \le 2\left(\theta_* - \theta_* + \frac{1}{\pi}(\theta - \theta_*)\right) = \frac{2}{\pi}(\theta - \theta_*) = \frac{2}{\pi}|\theta_* - \theta|$.

We combine both cases and get the final result. □

2.4 Proof of Theorem 2

Proof. Here we consider any stage k. Let τ denote the iteration index of SFO and $t = T_0 + \tau$ denote the global index. Define $g(\theta) = q(\theta) = \partial Q(\theta)$, $\mathbf{z} = (\mathbf{x}, y)$, $G(\theta, \mathbf{z}) = \pi\theta - (\eta(\mathbf{x}) - \theta)_+$, $\hat{G}_t(\theta, \mathbf{z}) = \widehat{\pi}_t\theta - (\hat{\eta}_t(\mathbf{x}) - \theta)_+$. It is clear that $\mathbb{E}[G(\theta, \mathbf{z})] = g(\theta)$, and $\max\left(|g(\theta_\tau)|, |G(\theta_\tau, \mathbf{z}_t)|, |\hat{G}_t(\theta_\tau, \mathbf{z}_t)|\right) \le 2$ for any τ. Following standard analysis of gradient descent, we have

$$\frac{1}{T}\sum_{\tau=1}^{T}(\theta_\tau - \theta_*)\hat{G}_t(\theta_\tau, \mathbf{z}_t) \le \frac{|\theta_1 - \theta_*|^2}{2\gamma T} + \frac{\gamma \max(\hat{G}_t(\theta_\tau, \mathbf{z}_t))^2}{2}$$

Then by the convexity of $Q(\theta)$, we have

$$\begin{aligned}Q(\bar{\theta}_T) - Q(\theta_*) \le & \frac{\|\theta_1 - \theta_*\|_2^2}{2\gamma T} + \frac{4\gamma}{2} + \frac{\sum_{\tau=1}^{T}(\theta_\tau - \theta_*)(g(\theta_\tau) - G(\theta_\tau, \mathbf{z}_t))}{T} \\ & + \frac{\sum_{\tau=1}^{T}(\theta_\tau - \theta_*)(G(\theta_\tau, \mathbf{z}_t) - \hat{G}_t(\theta_\tau, \mathbf{z}_t))}{T} \\ = & \mathbf{I} + \mathbf{II} + \mathbf{III} + \mathbf{IV}\end{aligned}$$

Now we try to bound the four terms respectively. Note that $\mathbf{I} \le \frac{R^2}{2\gamma T}$, $\mathbf{II} \le 2\gamma$. To bound the third term, we utilize the similar analysis of SGD (e.g. [4]). Define

$$\begin{aligned}&\tilde{\theta}_1 = \theta_1 \in [0, 0.5] \cap \mathcal{B}(\theta_1, R), \\ &\tilde{\theta}_{\tau+1} = \Pi_{[0,0.5]\cap\mathcal{B}(\theta_1,R)}(\tilde{\theta}_\tau - \gamma(g(\theta_\tau) - G(\theta_\tau, \mathbf{z}_t))).\end{aligned}$$

Then we have

$$\begin{aligned}\sum_{\tau=1}^{T}\gamma(\tilde{\theta}_\tau-\theta_*)(g(\theta_\tau)-G(\theta_\tau,\mathbf{z}_t))\leq&\frac{\|\tilde{\theta}_1-\theta_*\|_2^2}{2}+\frac{1}{2}\sum_{\tau=1}^{T}\gamma^2\|g(\theta_\tau)-G(\theta_t,\mathbf{z}_t)\|_2^2\\ \leq&\frac{R^2}{2}+8\gamma^2T.\end{aligned}\tag{10}$$

Note that both θ_τ and $\tilde{\theta}_\tau$ are measurable with respect to $\mathcal{F}_{t-1}=\{\mathbf{z}_1,\ldots,\mathbf{z}_{t-1}\}$, and $\{S_\tau:\gamma(\theta_\tau-\tilde{\theta}_\tau)(g(\theta_\tau)-G(\theta_\tau,\mathbf{z}_t)),\tau=1,\ldots,T\}$ is a martingale difference sequence, and for any τ we have $|\gamma(\theta_\tau-\tilde{\theta}_\tau)(g(\theta_\tau)-G(\theta_\tau,\mathbf{z}_t))|\leq 4\gamma\|\theta_\tau-\tilde{\theta}_\tau\|_2\leq 4\gamma\times 2R=8\gamma R$. Then by Azuma-Hoeffding's inequality, we have with probability at least $1-\frac{\delta}{3}$,

$$\sum_{\tau=1}^{T}\gamma(\theta_\tau-\tilde{\theta}_\tau)(g(\theta_\tau)-G(\theta_\tau,\mathbf{z}_t))\leq 8\gamma R\sqrt{2T\ln(\frac{3}{\delta})}.\tag{11}$$

Adding (10) and (11) together suffices to show that with probability at least $1-\frac{\delta}{3}$, we have

$$\mathbf{III}\leq\frac{R^2}{2\gamma T}+8\gamma+\frac{8R\sqrt{2\ln(\frac{3}{\delta})}}{\sqrt{T}}.$$

Next we bound **IV** according to the Lemma 3 introduced later. By union bound, we have with probability at least $1-\frac{\delta}{3}$, we have

$$\begin{aligned}\mathbf{IV}\leq&\frac{1}{T}\sum_{\tau=1}^{T}\left(\sup_{\tau}(\|\theta_\tau-\theta_1\|_2+\|\theta_1-\theta_*\|_2)\cdot\sup_{\theta\in[0,0.5],\mathbf{z}\in\mathcal{Z}}\|\widehat{G}_t(\theta,\mathbf{z})-G(\theta,\mathbf{z})\|_2\right)\\ \leq&\frac{2R\cdot(1+C\kappa)\times\sum_{t=1}^{T}\sqrt{\frac{\ln(12T/\delta)}{t}}}{T}\leq\frac{4R(1+C\kappa)\sqrt{\ln\left(\frac{12T}{\delta}\right)}}{\sqrt{T}},\end{aligned}$$

where the last inequality holds since $\sum_{t=1}^{T}\frac{1}{\sqrt{t}}\leq 2\sqrt{T}$. Combining these inequalities together, we have with probability at least $1-\delta$, we have

$$Q(\bar{\theta}_T)-Q(\theta_*)\leq\frac{R^2}{\gamma T}+10\gamma+\frac{R(20+4C\kappa)\sqrt{\ln(12T/\delta)}}{\sqrt{T}}.$$

Choosing $\gamma=\frac{R}{\sqrt{10T}}$, we have

$$Q(\bar{\theta}_T)-Q(\theta_*)\leq\frac{\left(2\sqrt{10}+(20+4C\kappa)\sqrt{\ln(12T/\delta)}\right)R}{\sqrt{T}}.$$

□

Lemma 3. *With probability at least* $1-\delta$*,*

$$\sup_{\theta\in[0,0.5],\mathbf{z}\in\mathcal{Z}}\|\widehat{G}_t(\theta,\mathbf{z})-G(\theta,\mathbf{z})\|_2\leq(1+C\kappa)\sqrt{\frac{\ln(4/\delta)}{t}}.$$

Proof. For any θ and any $\mathbf{z}$, the following argument holds. By Hoeffding's inequality, we have with probability at least $1-\frac{\delta}{2}$,

$$|\widehat{\pi}_t-\pi|\leq\sqrt{\frac{\ln(4/\delta)}{2t}}.$$

By the Assumption 1, we have with probability at least $1-\frac{\delta}{2}$,

$$|\hat{\eta}_t(\mathbf{x})-\eta(\mathbf{x})|\leq C\kappa\sqrt{\frac{\ln(4/\delta)}{t}}.$$

Note that $0\leq\theta\leq 0.5$, and hence we know that with probability at least $1-\delta$,

$$\text{LHS}\leq|\hat{\pi}_t-\pi|\cdot\theta+|\hat{\eta}_t(\mathbf{x}_t)-\eta(\mathbf{x}_t)|\leq(1+C\kappa)\sqrt{\frac{\ln(4/\delta)}{t}}.$$

□

2.5 Proof of Theorem 3

Given Theorem 2, the proof of Theorem 3 follows similar as the analysis [1] by noting that the objective function $Q(\theta)$ is strongly convex which is a special case of uniformly convex. For completeness, we give a proof here.

Proof. Define

$$\bar{\delta} = \frac{2\delta}{\log_2 n}, \quad a(n, \bar{\delta}) = \frac{2\sqrt{10} + (20 + 4C\kappa)\sqrt{\ln(12n/\bar{\delta})}}{\sqrt{n}}.$$
$$\mu_0 = \frac{2a(n_0, \bar{\delta})}{R_0}, \quad \mu_k = 2^k \mu_0, \quad R_k = R_0/2^k$$

where $k = 1, \ldots, m$. Then we have $\mu_k R_k^2 = 2^{-k}\mu_0 R_0^2$.

By definition of m in Algorithm 1 (FOFO), when $n \geq 100$,

$$0 < \frac{1}{2}\log_2 \frac{2n}{\log_2 n} - 2 \leq m \leq \frac{1}{2}\log_2 \frac{2n}{\log_2 n} - 1 \leq \frac{1}{2}\log_2 n, \tag{12}$$

so we have

$$2^m \geq \frac{1}{4}\sqrt{\frac{2n}{\log_2 n}}. \tag{13}$$

Define $c = \sqrt{\frac{2}{\sigma}}$, and note that $Q(\theta)$ is σ-strongly convex, and hence $\|\theta - \theta^*\|_2 \leq c(Q(\theta) - Q(\theta^*))^{\frac{1}{2}}$, where θ^* is the closest point to θ in $[0, 0.5]$.

Without loss of generality, we assume $c^2 \geq \frac{R_0}{2}$, i.e., $\frac{1}{c^2} \leq \frac{2}{R_0}$. Now we prove that $\frac{2}{R_0} \leq \mu_m$. When $n \geq 100$, we have

$$\begin{aligned}
\mu_m &= 2^m \mu_0 \\
&\geq \frac{1}{4}\sqrt{\frac{2n}{\log_2 n}}\frac{4}{R_0}\left(\frac{\sqrt{10}}{\sqrt{n_0}} + \frac{(10 + 2C\kappa)\sqrt{\ln(12n_0/\bar{\delta})}}{\sqrt{n_0}}\right) \\
&\geq \frac{2}{R_0}\cdot\frac{1}{2}\sqrt{\frac{2n}{\log_2 n}}\left(\frac{\sqrt{10}}{\sqrt{n_0}} + \frac{8\sqrt{\ln(6\log_2 n)}}{\sqrt{n_0}}\right) \\
&\geq \frac{2}{R_0}\sqrt{\frac{2n}{\log_2 n}}\sqrt{\frac{(8\sqrt{10})\sqrt{\ln(6\log_2 n)}}{n_0}} \\
&\geq \frac{2}{R_0}\cdot\sqrt{\frac{2n}{\log_2 n}}\sqrt{\frac{(8\sqrt{10})\sqrt{\ln(3\log_2 n)}}{\frac{n}{m}}} \\
&\geq \frac{2}{R_0}\cdot\sqrt{\frac{2n}{\log_2 n}}\sqrt{\frac{(8\sqrt{10})\sqrt{\ln(3\log_2 n)}}{\frac{n}{\frac{1}{2}\log_2\frac{2n}{\log_2 n} - 2}}} \\
&= \frac{2}{R_0}\sqrt{(8\sqrt{10})\sqrt{\ln(3\log_2 n)}\left(1 - \frac{\log_2\log_2 n + 3}{\log_2 n}\right)} \\
&\geq \frac{2}{R_0}.
\end{aligned}$$

where the first inequality holds because of (13), the second inequality stems from the fact that $10 + 2C\kappa > 8$, $0 < \delta < 1$, $n_0 \geq 1$, and the definition of $\bar{\delta}$, the third inequality holds by employing $a + b \geq 2\sqrt{ab}$, the fourth inequality holds because $0 < n_0 = \lfloor n/m \rfloor \leq n/m$, the fifth inequality holds because of the lower bound of m in (12), and the last inequality holds since when $n \geq 100$, the function $(8\sqrt{10})\sqrt{\ln(3\log_2 n)}\left(1 - \frac{\log_2\log_2 n+3}{\log_2 n}\right)$ is monotonically increasing with respect to n, and hence is greater than 1. So $\frac{2}{R_0} \leq \mu_m$. Recall that $\frac{1}{c^2} \leq \frac{2}{R_0}$, and thus, $\frac{1}{c^2} \leq \mu_m$.

Given $\widehat{\theta}_k$, denote $\widehat{\theta}_k^*$ by the closest optimal solution to $\widehat{\theta}_k$. We consider two cases.

Case 1. If $\frac{1}{c^2} \geq \mu_0$, then $\mu_0 \leq \frac{1}{c^2} \leq \mu_m$. So there exists a k^* such that $\mu_{k^*} \leq \frac{1}{c^2} \leq \mu_{k^*+1} = 2\mu_{k^*}$, where $0 \leq k^* < m$. To utilize this fact, we have the following lemma.

Lemma 4. *Let k^* satisfy $\mu_{k^*} \leq \frac{1}{c^2} \leq 2\mu_{k^*}$. Then for any $1 \leq k \leq k^*$, there exists a Borel set $\mathcal{A}_k \subset \Omega$ of probability at least $1 - k\bar{\delta}$, such that for $\omega \in \mathcal{A}_k$, the points $\{\widehat{\theta}_k\}_{k=1}^m$ generated by the Algorithm 1 satisfy*

$$\|\widehat{\theta}_{k-1} - \widehat{\theta}^*_{k-1}\|_2 \leq R_{k-1} = 2^{-k+1}R_0, \tag{14}$$

$$Q(\widehat{\theta}_k) - Q_* \leq \mu_k R_k^2 = 2^{-k}\mu_0 R_0^2. \tag{15}$$

Moreover, for $k > k^*$ there is a Borel set $\mathcal{C}_k \subset \Omega$ of probability at least $1 - (k - k^*)\bar{\delta}$ such that on $\mathcal{C}_k$, we have

$$Q(\widehat{\theta}_k) - Q(\widehat{\theta}_{k^*}) \leq \mu_{k^*} R_{k^*}^2. \tag{16}$$

Proof. We prove (14) and (15) by induction. Note that (14) holds for $k = 1$. Assume it is true for some $k > 1$ on $\mathcal{A}_{k-1}$. According to the Theorem 2, there exists a Borel set $\mathcal{B}_k$ with $\Pr(\mathcal{B}_k) \geq 1 - \bar{\delta}$ such that

$$Q(\widehat{\theta}_k) - Q_* \leq R_{k-1}a(n_0, \bar{\delta}) = \frac{1}{2}\mu_k 2^{-k} R_0 R_{k-1} = \mu_k R_k^2,$$

which is (15). By the inductive hypothesis, $\|\widehat{\theta}_{k-1} - \widehat{\theta}^*_{k-1}\|_2 \leq R_{k-1}$ on the set $\mathcal{A}_{k-1}$. Define $\mathcal{A}_k = \mathcal{A}_{k-1} \cap \mathcal{B}_k$. Note that

$$\Pr(\mathcal{A}_k) \geq \Pr(\mathcal{A}_{k-1}) + \Pr(\mathcal{B}_k) - 1 \geq 1 - k\bar{\delta},$$

and on $\mathcal{A}_k$, by the strong-convexity of $Q(\theta)$ and the definition of k^*, we have

$$\|\widehat{\theta}_k - \widehat{\theta}^*_k\|_2^2 \leq c^2(Q(\widehat{\theta}_k) - Q_*) \leq \frac{Q(\widehat{\theta}_k) - Q_*}{\mu_{k^*}} \leq \frac{\mu_k R_k^2}{\mu_{k^*}} \leq R_k^2,$$

which is (14) for $k + 1$.

Now we prove (16). For $k > k^*$, one can apply the similar strategy as in Theorem 2. Specifically, at the k-th stage with $k > k^*$, employing the similar proof of Theorem 2 by substituting all θ_* to $\widehat{\theta}_{k-1}$, the first term of RHS becomes zero and hence we get a tighter bound of $Q(\widehat{\theta}_k) - Q(\widehat{\theta}_{k-1})$, we here relax the bound to be $R_{k-1}a(n_0, \bar{\delta})$.

So there exists a Borel set $\mathcal{B}_k$ with $\Pr(\mathcal{B}_k) \geq 1 - \bar{\delta}$ such that

$$Q(\widehat{\theta}_k) - Q(\widehat{\theta}_{k-1}) \leq R_{k-1}a(n_0, \bar{\delta}) = 2^{k^*-k} R_{k^*-1}a(n_0, \bar{\delta}) = 2^{k^*-k}\mu_{k^*} R_{k^*}^2 = \mu_k R_k^2,$$

which implies that on $\mathcal{C}_k = \cap_{j=k^*+1}^k \mathcal{B}_j$, we have

$$Q(\widehat{\theta}_k) - Q(\widehat{\theta}_{k^*}) = \sum_{j=k^*+1}^{k} \Big(Q(\widehat{\theta}_j) - Q(\widehat{\theta}_{j-1})\Big) \leq \sum_{j=k^*+1}^{k} 2^{k^*-j}\mu_{k^*} R_{k^*}^2 \leq \mu_{k^*} R_{k^*}^2.$$

By union bound, we have $\Pr(\mathcal{C}_k) = \Pr(\cap_{j=k^*+1}^k \mathcal{B}_j) \geq 1 - (k - k^*)\bar{\delta}$. Here completes the proof. □

Now we proceed the proof as follows. Note that $\mu_0 \leq \frac{1}{c^2} \leq \mu_m$. At the end of k^*-th stage, on the Borel set $\mathcal{A}_{k^*}$ of probability at least $1 - k^*\bar{\delta}$, we have

$$Q(\widehat{\theta}_{k^*}) - Q_* \leq \mu_{k^*} R_{k^*}^2.$$

Then on the Borel set $\mathcal{D}_m = \mathcal{C}_m \cap \mathcal{A}_{k^*} = (\cap_{j=k^*+1}^m \mathcal{B}_j) \cap A_{k^*}$ with $\Pr(\mathcal{D}_m) \geq 1 - m\bar{\delta}$, we have

$$\begin{aligned} Q(\widehat{\theta}_m) - Q_* &= Q(\widehat{\theta}_m) - Q(\widehat{\theta}_{k^*}) + (Q(\widehat{\theta}_{k^*}) - Q_*) \leq 2\mu_{k^*} R_{k^*}^2 \leq 4(\frac{\mu_{k^*}}{c^{-2}})\mu_{k^*} R_{k^*}^2 \\ &= (4c \cdot a(n_0, \bar{\delta}))^2. \end{aligned}$$

By the definition of m and $\bar{\delta}$, and the fact that $m \leq \frac{1}{2}\log_2 n$, we have $m\bar{\delta} \leq \delta$. So $\Pr(\mathcal{D}_m) \geq 1 - \delta$.

Table 1: Offline Testing F-measure (bold numbers represent the best performance)

Datasets	FOFO	OFO	LR	STAMP	OMCSL
webspam	**.9348 ± .0003**	.9348 ± .0004	.9347 ± .0005	.9312 ± .0014	.9282 ± .0046
a9a	**.6789 ± .0015**	.6755 ± .0020	.6518 ± .0026	.6735 ± .0034	.6704 ± .0096
ijcnn1	**.6412 ± .0020**	.5776 ± .0039	.4441 ± .0040	.5987 ± .0328	.6050 ± .0225
w8a	**.7159 ± .0118**	.6695 ± .0134	.6621 ± .0222	.6706 ± .0289	.6627 ± .0370
covtype (2 vs o)	**.7627 ± .0005**	.7625 ± .0005	.7557 ± .0004	.7568 ± .0055	.7557 ± .0081
covtype (1 vs o)	**.7090 ± .0004**	.7082 ± .0002	.6770 ± .0010	.7039 ± .0047	.7000 ± .0093
cov (3 vs o)	**.7277 ± .0009**	.7257 ± .0005	.6914 ± .0039	.7213 ± .0050	.7210 ± .0050
covtype (7 vs o)	**.6723 ± .0022**	.6521 ± .0025	.6140 ± .0037	.6417 ± .0197	.6513 ± .0150
covtype (6 vs o)	**.4468 ± .0015**	.4251 ± .0014	.1258 ± .0072	.3971 ± .0516	.4237 ± .0142
covtype (5 vs o)	**.2648 ± .0036**	.2488 ± .0027	.0000 ± .0000	.2218 ± .0246	.2362 ± .0304
covtype (4 vs o)	**.5512 ± .0035**	.5228 ± .0083	.4123 ± .0130	.3682 ± .0724	.5139 ± .0256
Sensorless (1 vs o)	**.7549 ± .0047**	.6732 ± .0022	.4774 ± .0156	.6243 ± .1394	.5401 ± .2360
Sensorless (2 vs o)	**.4698 ± .0178**	.2388 ± .0083	.1667 ± .0000	.3284 ± .1485	.4689 ± .0330
Sensorless (3 vs o)	.2138 ± .0047	**.2254 ± .0048**	.1345 ± .0709	.1819 ± .0812	.1804 ± .0413
Sensorless (4 vs o)	**.5895 ± .0055**	.3117 ± .0102	.1360 ± .0717	.3778 ± .2152	.4530 ± .0813
Sensorless (5 vs o)	**.3089 ± .0049**	.2343 ± .0047	.1009 ± .0868	.2264 ± .1186	.1782 ± .1228
Sensorless (6 vs o)	**.3607 ± .0062**	.2789 ± .0078	.0993 ± .0854	.2772 ± .0702	.2266 ± .1503
Sensorless (7 vs o)	.9994 ± .0002	**.9996 ± .0001**	.9986 ± .0010	.9988 ± .0009	.9982 ± .0017
Sensorless (8 vs o)	**.4085 ± .0017**	.3158 ± .0047	.0496 ± .0799	.3185 ± .1159	.3484 ± .0583
Sensorless (9 vs o)	**.2783 ± .0037**	.2069 ± .0039	.1346 ± .0710	.1749 ± .1352	.1902 ± .1251
Sensorless (10 vs o)	**.6025 ± .0080**	.4897 ± .0113	.1659 ± .0000	.4089 ± .2345	.5170 ± .0566
Sensorless (11 vs o)	.9997 ± .0000	.9997 ± .0002	.9998 ± .0002	.9997 ± .0001	**.9998 ± .0002**
protein (1 vs o)	.5008 ± .0026	**.5037 ± .0059**	.4643 ± .0114	.4914 ± .0163	.4930 ± .0116
protein (2 vs o)	**.6849 ± .0035**	.6835 ± .0040	.6390 ± .0053	.6787 ± .0069	.6735 ± .0144
protein (0 vs o)	.7479 ± .0017	**.7483 ± .0014**	.7183 ± .0023	.7430 ± .0071	.7423 ± .0052

Case 2. If $\frac{1}{c^2} < \mu_0$, then on $\mathcal{A}_1 = \mathcal{B}_1$,

$$Q(\widehat{\theta}_1) - Q_* \le R_0 \cdot a(n_0, \bar{\delta}) = \frac{R_0}{a(n_0, \bar{\delta})} \cdot a(n_0, \bar{\delta})^2 = \frac{2}{\mu_0} a(n_0, \bar{\delta})^2 \le 2 \left(c \cdot a(n_0, \bar{\delta})\right)^2.$$

Hence on $\mathcal{A}_1 \cap \mathcal{C}_m$, by using Lemma 4 and a similar argument as in case 1, we have

$$Q(\widehat{\theta}_m) - Q_* = Q(\widehat{\theta}_m) - Q(\widehat{\theta}_1) + Q(\widehat{\theta}_1) - Q_* \le 2R_0 \cdot a(n_0, \bar{\delta}) \le (2c \cdot a(n_0, \bar{\delta}))^2,$$

where $\Pr(\mathcal{A}_1 \cap \mathcal{C}_m) \ge 1 - \delta$.

Combining the two cases, we have with probability at least $1 - \delta$,

$$Q(\widehat{\theta}_m) - Q_* \le (4c \vee 2c)^2 \left(a(n_0, \bar{\delta})\right)^2 = \widetilde{O}\left(\frac{\ln(\frac{1}{\delta})}{\sigma n}\right).$$

□

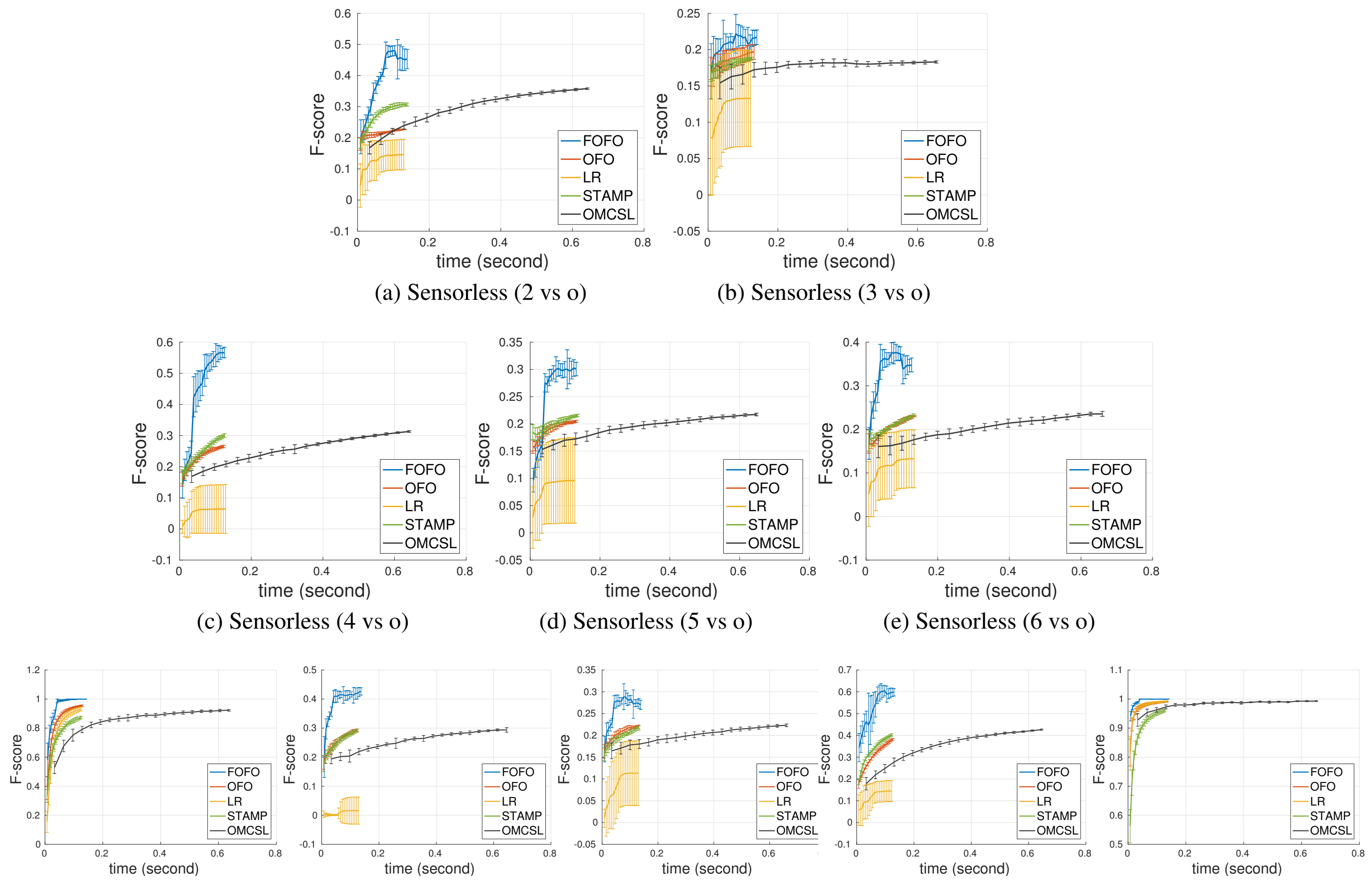
3 More Experimental Results

More experimental results are reported in Table 1 (offline testing results) and Figure 1 (online F-measure vs running time).

References

[1] Anatoli Juditsky and Yuri Nesterov. Deterministic and stochastic primal-dual subgradient algorithms for uniformly convex minimization. *Stoch. Syst.*, 2014.

[2] Nagarajan Natarajan, Oluwasanmi Koyejo, Pradeep Ravikumar, and Inderjit S. Dhillon. Consistent binary classification with generalized performance metrics. In *Neural Information Processing Systems (NIPS)*, 2014.



(a) Sensorless (2 vs o) (b) Sensorless (3 vs o)

(c) Sensorless (4 vs o) (d) Sensorless (5 vs o) (e) Sensorless (6 vs o)

(f) Sensorless (7 vs o) (g) Sensorless (8 vs o) (h) Sensorless (9 vs o) (i) Sensorless (10 vs o) (j) Sensorless (11 vs o)

(k) protein (1 vs o) (l) protein (2 vs o)

Figure 1: Online F-measure vs Running Time for other datasets

[3] Ming-Jie Zhao, Narayanan Edakunni, Adam Pocock, and Gavin Brown. Beyond fano's inequality: bounds on the optimal f-score, ber, and cost-sensitive risk and their implications. *Journal of Machine Learning Research*, 14(Apr):1033–1090, 2013.

[4] Martin Zinkevich. Online convex programming and generalized infinitesimal gradient ascent. 2003.