Supplementary Material for “Fast Stochastic AUC Maximization with $O(1/n)$-Convergence Rate”

Mingrui Liu [1] **Xiaoxuan Zhang** [1] **Zaiyi Chen** [2] **Xiaoyu Wang** [3] **Tianbao Yang** [1]

1. Technical Lemmas

We introduce two concentration inequalities in Lemma 4, which are used frequently in the proofs.

Lemma 4.

- ***(Randomized version of Hoeffding's inequality)*** *Suppose T is a random variable taking value on $\mathbb{N}^+$, and let $X_1, \ldots, X_T$ be independent random variables. Define $\bar{X}_T = \frac{1}{T}(X_1 + \ldots + X_T)$. If every X_i is strictly bounded by the intervals $[a_i, b_i]$, then we have with probability at least $1 - \delta$,*

$$\bar{X}_T - \mathbb{E}(\bar{X}_T) \leq \sqrt{\frac{\ln(1/\delta)\sum_{i=1}^T (b_i - a_i)^2}{2T^2}}. \tag{7}$$

Similarly, with probability at least $1 - \delta$,

$$\mathbb{E}(\bar{X}_T) - \bar{X}_T \leq \sqrt{\frac{\ln(1/\delta)\sum_{i=1}^T (b_i - a_i)^2}{2T^2}} \tag{8}$$

- ***(Randomized version of vector concentration inequality)*** *Suppose T is a random variable taking value on $\mathbb{N}^+$, and let $X_1, \ldots, X_T \in \mathbb{R}^d$ be i.i.d. random variables. If $\phi : \mathbb{R}^d \to \mathcal{H}$, where $\mathcal{H}$ is a Hilbert space endowed with norm $\|\cdot\|$ (actually we can take $\mathcal{H}$ to be $\mathbb{R}^d$ endowed with infinity norm). Suppose $B = \sup_{\mathbf{x}\in\mathbb{R}^d} \|\phi(\mathbf{x})\| < \infty$. Then we have with probability at least $1 - \delta$,*

$$\left\| \frac{1}{T}\sum_{i=1}^{T} \phi(X_i) - \mathbb{E}\left(\phi(X_1)\right) \right\| \leq \frac{B}{\sqrt{n}}\left[2 + \sqrt{2\log(1/\delta)}\right]. \tag{9}$$

Proof. The proof is quite straightforward. For the randomized version of Hoeffding's inequality, note that

$$\begin{aligned}
&\Pr\left(\frac{1}{T}\left(X_1 + \ldots + X_T\right) - \mathbb{E}(X_1) \geq \sqrt{\frac{\ln(1/\delta)\sum_{i=1}^T (b_i - a_i)^2}{2T^2}}\right) \\
&= \sum_{t=1}^{\infty} \Pr\left(\bar{X}_T - \mathbb{E}(\bar{X}_T) \geq \sqrt{\frac{\ln(1/\delta)\sum_{i=1}^T (b_i - a_i)^2}{2T^2}} \middle| T = t\right) \\
&\quad \cdot \Pr\left(T = t\right) \\
&\leq \sum_{t=1}^{\infty} \delta \cdot \Pr(T = t) = \delta,
\end{aligned}$$

where the first inequality follows from the deterministic version of Hoeffding's inequality.

It is easy to show the correctness of the randomized version of vector concentration inequality by employing the same technique. The deterministic version can be derived via McDiarmid's inequality (McDiarmid, 1989). A standard proof can be found in the section 4.1 of (Shawe-Taylor & Cristianini, 2004). For completeness, we include the proof here. To derive the deterministic version, define $S_1 = (X_1, \ldots, X_T)$, $S_1' = (X_1', \ldots, X_n')$ to be two collections of independent samples, $S_2 = (X_1, \ldots, X_{i-1}, X_i', X_{i+1}, \ldots, X_T)$, and $f(S) = \|\frac{1}{T}\sum_{i=1}^T \phi(X_i) - \mathbb{E}(\phi(X_1))\|$, we have $|f(S_1) - f(S_2)| \leq 2B/n$. By McDiarmid's inequality, we have

$$\Pr\left(f(S_1) - \mathbb{E}(f(S_1)) > \epsilon\right) \leq \exp\left(-\frac{2T\epsilon^2}{4B^2}\right) \tag{10}$$

Define $\sigma = (\sigma_1, \ldots, \sigma_n)$ to be Rademacher variables, i.e. $\Pr(\sigma_i = 1) = \Pr(\sigma_i = -1) = 1/2$, and σ_i's are i.i.d.

[1] Department of Computer Science, The University of Iowa, IA 52242, USA [2] University of Science and Technology of China [3] Intellifusion. Correspondence to: Mingrui Liu <mingrui-liu@uiowa.edu>, Tianbao Yang <tianbao-yang@uiowa.edu>.

Proceedings of the 35^{th} International Conference on Machine Learning, Stockholm, Sweden, PMLR 80, 2018. Copyright 2018 by the author(s).

Define $\bar{\phi}_{S_1} = \frac{1}{T}\sum_{i=1}^{T}\phi(Z_i)$, then

$$\begin{aligned}
&\mathbb{E}\left(f(S_1)\right) = \mathbb{E}\left(\left\|\bar{\phi}_{S_1} - \mathbb{E}(\phi_{S_1})\right\|\right) = \mathbb{E}\left(\|\bar{\phi}_{S_1} - \mathbb{E}(\phi_{S_1'})\|\right) \\
&= \mathbb{E}\left(\left\|\mathbb{E}(\bar{\phi}_{S_1} - \phi_{S_1'})\right\|\right) \le \mathbb{E}\left(\left\|\phi_{S_1} - \phi_{S_1'}\right\|\right) \\
&= \mathbb{E}\left(\frac{1}{T}\left\|\sum_{i=1}^{T}\sigma_i(\phi(X_i) - \phi(X_i'))\right\|\right) \\
&\le 2\mathbb{E}\left(\frac{1}{T}\left\|\sum_{i=1}^{T}\sigma_i\phi(X_i)\right\|\right) \\
&= 2\mathbb{E}\left(\frac{1}{T}\sqrt{\sum_{i=1}^{T}\sigma_i^2\phi^2(X_i) + \sum_{i\neq j}\sigma_i\sigma_j\phi(X_i)\phi(X_j)}\right) \\
&\le \frac{2}{T}\sqrt{\mathbb{E}\left(\sum_{i=1}^{T}\sigma_i^2\phi^2(X_i) + \sum_{i\neq j}\sigma_i\sigma_j\phi(X_i)\phi(X_j)\right)} \\
&= \frac{2}{T}\sqrt{\sum_{i=1}^{T}\mathbb{E}(\phi^2(X_i))} \le \frac{2B}{\sqrt{T}}.
\end{aligned}$$

Combing this result with (10), and taking $\epsilon = B\sqrt{\frac{2}{T}\log(\frac{1}{\delta})}$ suffice to get the result. □

2. Proof of Lemma 2

Proof. According to the equation (6) in (Ying et al., 2016), we have

$$\begin{aligned}
&f(\mathbf{v},\alpha) = f(\mathbf{w},a,b,\alpha) = p(1-p)\Bigg\{ \\
&\int_{\mathbf{x}}\left((\mathbf{w}^\top\mathbf{x} - a)^2 - 2(1+\alpha)\mathbf{w}^\top\mathbf{x}\right)P(\mathbf{x}|y=1)dx + \\
&\int_{\mathbf{x}}\left((\mathbf{w}^\top\mathbf{x} - b)^2 + 2(1+\alpha)\mathbf{w}^\top\mathbf{x}\right)P(\mathbf{x}|y=-1) - \alpha^2\Bigg\} \\
&= p(1-p)\Bigg\{\mathbf{w}^\top\left(\mathbb{E}(\mathbf{x}\mathbf{x}^\top|y=1) + \mathbb{E}(\mathbf{x}\mathbf{x}^\top|y=-1)\right)\mathbf{w} \\
&- 2\mathbf{w}^\top\left(a\mathbb{E}(\mathbf{x}|y=1) + b\mathbb{E}(\mathbf{x}|y=-1)\right) + a^2 + b^2 \\
&+ 2(1+\alpha)\mathbf{w}^\top\left(\mathbb{E}(\mathbf{x}|y=-1) - \mathbb{E}(\mathbf{x}|y=1)\right) - \alpha^2\Bigg\}
\end{aligned}$$

When $\alpha = \mathbf{w}^\top\left(\mathbb{E}(\mathbf{x}|y=-1) - \mathbb{E}(\mathbf{x}|y=1)\right) \in \Omega_2$, (it is easy to see that $\alpha \in \Omega_2$ by employing Cauchy-Schwarz inequality, i.e. $|\alpha| \le \|\mathbf{w}\|_1 \cdot \|\mathbb{E}(\mathbf{x}|y=-1) - \mathbb{E}(\mathbf{x}|y=1)\|_\infty \le 2R\kappa$), $f(v,\alpha)$ achieves its maximum with respect to α, so we get

$$\begin{aligned}
&f_1(\mathbf{v}) = f\left(\mathbf{w},a,b,\mathbf{w}^\top\left(\mathbb{E}(\mathbf{x}|y=-1) - \mathbb{E}(\mathbf{x}|y=1)\right)\right) \\
&= p(1-p)\Bigg\{\mathbf{w}^\top\mathbb{E}(\mathbf{x}\mathbf{x}^\top|y=1)\mathbf{w} - 2a\mathbf{w}^\top\mathbb{E}(\mathbf{x}|y=1) + a^2 \\
&+ \mathbf{w}^\top\mathbb{E}(\mathbf{x}\mathbf{x}^\top|y=-1)\mathbf{w} - 2b\mathbf{w}^\top\mathbb{E}(\mathbf{x}|y=-1) + b^2 \\
&+ [\left(\mathbf{w}^\top\left(\mathbb{E}(\mathbf{x}|y=1) - \mathbb{E}(\mathbf{x}|y=-1)\right)\right)^2 \\
&+ 2\mathbf{w}^\top\left(\mathbb{E}(\mathbf{x}|y=-1) - \mathbb{E}(\mathbf{x}|y=1)\right)]\Bigg\} \\
&= p(1-p)\left[(\mathbf{w},a,b)^\top \cdot M \cdot (\mathbf{w},a,b) + \text{affine function of } \mathbf{v}\right],
\end{aligned}$$

where $M = M_1 + M_2 + M_3$,

$$M_1 = \begin{bmatrix}\mathbb{E}(\mathbf{x}\mathbf{x}^\top|y=1) & -\mathbb{E}(\mathbf{x}|y=1) & 0\\ -\mathbb{E}(\mathbf{x}|y=1) & 1 & 0\\ 0 & 0 & 0\end{bmatrix}$$

$$M_2 = \begin{bmatrix}\mathbb{E}(\mathbf{x}\mathbf{x}^\top|y=-1) & 0 & -\mathbb{E}(\mathbf{x}|y=-1)\\ 0 & 0 & 0\\ -\mathbb{E}(\mathbf{x}|y=-1) & 0 & 1\end{bmatrix}$$

$$M_3 = \begin{bmatrix}\mathbf{q}\mathbf{q}^\top & 0 & 0\\ 0 & 0 & 0\\ 0 & 0 & 0\end{bmatrix}$$

$\mathbf{q} = \mathbb{E}(\mathbf{x}|y=1) - \mathbb{E}(\mathbf{x}|y=-1)$. Note that M_1, M_2, M_3 are positive semidefinite matrices, and hence M is positive semidefinite. So $f_1(\mathbf{v})$ is convex and piecewise quadratic. Since Ω_1 is a polyhedron, according to Corollary 3.1 of (Li, 2013), we can know that $f_1(\mathbf{v})$ restricted on Ω_1 satisfies the quadratic growth condition.

□

3. Proof of Lemma 3

Proof. By applying the inequality (9) in Lemma 4, the triangle inequality, and the union bound, we have with probability at least $1 - \frac{\delta}{6}$,

$$\|\widehat{A} - A\|_2 \le \frac{2\kappa}{\sqrt{T_-}}\left(2 + \sqrt{2\ln(\frac{12}{\delta})}\right) + \frac{2\kappa}{\sqrt{T_+}}\left(2 + \sqrt{2\ln(\frac{12}{\delta})}\right), \tag{11}$$

Note that both T_- and T_+ follow the Bernoulli distribution, and denote $p = \Pr(y=1)$. By applying deterministic version of Hoeffding's inequality in Lemma 4 (i.e., inequality 8) to indicator functions of random variables $\mathbb{I}_{[y_i=-1]}$ and $\mathbb{I}_{[y_i=1]}$ respectively and the union bound, we have with probability at least $1 - \frac{\delta}{6}$, the following two equations hold simultaneously:

$$T_- \ge (1-p)T - \sqrt{\frac{\ln(\frac{12}{\delta})T}{2}},\ T_+ \ge pT - \sqrt{\frac{\ln(\frac{12}{\delta})T}{2}}. \tag{12}$$

According to (11) and (12), by union bound, we know that with probability at least $1-\frac{\delta}{3}$, we have

$$\begin{aligned}&\|\widehat{A}-A\|_2\\ &\leq \frac{2\kappa\left(2+\sqrt{2\ln(\frac{12}{\delta})}\right)}{\sqrt{(1-p)T-\sqrt{\frac{\ln(\frac{12}{\delta})T}{2}}}}+\frac{2\kappa\left(2+\sqrt{2\ln(\frac{12}{\delta})}\right)}{\sqrt{pT-\sqrt{\frac{\ln(\frac{12}{\delta})T}{2}}}}.\end{aligned} \tag{13}$$

Note that (12) is equivalent to

$$\begin{aligned}&pT \geq \widehat{p}_T T-\sqrt{\frac{T\ln(\frac{12}{\delta})}{2}},\\ &(1-p)T \geq (1-\widehat{p}_T)T-\sqrt{\frac{T\ln(\frac{12}{\delta})}{2}}.\end{aligned} \tag{14}$$

Utilizing (14) and plugging it into (13), we know that with probability at least $1-\frac{\delta}{3}$,

$$\begin{aligned}\|\widehat{A}-A\|_2 &\leq \frac{2\kappa\left(2+\sqrt{2\ln(\frac{12}{\delta})}\right)}{\sqrt{(1-\widehat{p}_T)T-\sqrt{\frac{T\ln(\frac{12}{\delta})}{2}}-\sqrt{\frac{T\ln(\frac{12}{\delta})}{2}}}}\\ &+\frac{2\kappa\left(2+\sqrt{2\ln(\frac{12}{\delta})}\right)}{\sqrt{\widehat{p}_T T-\sqrt{\frac{T\ln(\frac{12}{\delta})}{2}}-\sqrt{\frac{T\ln(\frac{12}{\delta})}{2}}}}\\ &=\frac{2\kappa\left(2+\sqrt{2\ln(\frac{12}{\delta})}\right)}{\sqrt{(1-\widehat{p}_T)T-\sqrt{2T\ln(\frac{12}{\delta})}}}+\frac{2\kappa\left(2+\sqrt{2\ln(\frac{12}{\delta})}\right)}{\sqrt{\widehat{p}_T T-\sqrt{2T\ln(\frac{12}{\delta})}}}\\ &\leq \frac{4\kappa\left(2+\sqrt{2\ln(\frac{12}{\delta})}\right)}{\sqrt{\xi T}},\end{aligned}$$

where

$$\xi \equiv \min(\widehat{p}_T, 1-\widehat{p}_T)-\sqrt{\frac{2\ln(\frac{12}{\delta})}{T}}.$$

□

4. Proof of Theorem 1

Proof. Define $\bar{\delta}=\frac{2\delta}{\log_2 n}$, and $a(n,\bar{\delta})=G(\frac{2\sqrt{3\gamma_1}}{\sqrt{n}}+\frac{\gamma_2\sqrt{\ln(\frac{6n}{\bar{\delta}})}}{\sqrt{n}})$, $\mu_0=2R_0^{-1}a(n_0,\bar{\delta})$, $\mu_k=2^k\mu_0$, $R_k=R_0/2^k$, where $k=1,\ldots,m$. Then we have $\mu_k R_k^2=2^{-k}\mu_0 R_0^2$.

By definition of m, when $n \geq 100$,

$$0<\frac{1}{2}\log_2\frac{2n}{\log_2 n}-2 \leq m \leq \frac{1}{2}\log_2\frac{2n}{\log_2 n}-1 \leq \frac{1}{2}\log_2 n, \tag{15}$$

so we have

$$2^m \geq \frac{1}{4}\sqrt{\frac{2n}{\log_2 n}}. \tag{16}$$

To employ the result of Lemma 2, at the i-th stage, we need to satisfy $T \geq \frac{R^2}{R_i^2}=\frac{\frac{R_0^2}{2+4\kappa^2}}{4^{-i}R_0^2}=4^i/(2+4\kappa^2)$, which should hold for any $1 \leq i \leq m$. So $n_0 \geq 4^m$ suffices to achieve this requirement. Now we argue that this condition can be implied by $n \geq 100$. Note that

$$\begin{aligned}n_0=\lfloor n/m\rfloor \geq \frac{n}{m}-1 &\geq \frac{n}{\frac{1}{2}\log_2\frac{2n}{\log_2 n}-1}-1\\ &\geq n\left(\frac{1}{\frac{1}{2}\log_2\frac{2n}{\log_2 n}-1}-\frac{1}{n}\right)\\ &\geq n\left(\frac{1}{\frac{1}{2}\log_2\frac{2n}{\log_2 n}-1}-\frac{1}{\log_2 n}\right),\end{aligned}$$

and $4^m \leq 4^{\frac{1}{2}\log_2\frac{2n}{\log_2 n}-1}=\frac{n}{2\log_2 n}$. To show the implication from $n \geq 100$ to $n_0 \geq 4^m$, it suffices to prove that when $n \geq 100$, $\frac{1}{2}\log_2\frac{2n}{\log_2 n}-1 \leq \frac{2}{3}\log_2 n$, i.e., $\frac{\sqrt{\frac{2n}{\log_2 n}}}{2} \leq n^{\frac{2}{3}}$, which obviously holds.

According to Lemma 2, we know that $P(\mathbf{v}) \equiv f_1(\mathbf{v})$ satisfies the quadratic growth condition, which implies that there exists some $c>0$, such that $\|\mathbf{v}-\mathbf{v}^*\|_2 \leq c(P(\mathbf{v})-P(\mathbf{v}^*))^{\frac{1}{2}}$, where $\mathbf{v}^*$ is the closest point to $\mathbf{v}$ in Ω_*.

We can assume $c^2 \geq \frac{R_0}{G}$, i.e., $\frac{1}{c^2} \leq \frac{G}{R_0}$. Otherwise we can set c^2 to be $\frac{R_0}{G}$ such that the quadratic growth property in Lemma 2 still holds.

When $n \geq 100$, we have

$$\begin{aligned}\mu_m &= 2^m\mu_0\\ &\geq \frac{1}{4}\sqrt{\frac{2n}{\log_2 n}}2R_0^{-1}G\left(\frac{2\sqrt{3\gamma_1}}{\sqrt{n_0}}+\frac{\gamma_2\sqrt{\ln(6n_0/\bar{\delta})}}{\sqrt{n_0}}\right)\\ &\geq \frac{G}{R_0}\cdot\frac{1}{2}\sqrt{\frac{2n}{\log_2 n}}\left(\frac{2\sqrt{3}}{\sqrt{n_0}}+\frac{2\sqrt{\ln(3n_0\log_2 n)}}{\sqrt{n_0}}\right)\\ &\geq \frac{G}{R_0}\sqrt{\frac{2n}{\log_2 n}}\sqrt{\frac{(2\sqrt{3})2\sqrt{\ln(3n_0\log_2 n)}}{n_0}}\\ &\geq \frac{G}{R_0}\cdot\sqrt{\frac{2n}{\log_2 n}}\sqrt{\frac{(2\sqrt{3})2\sqrt{\ln(3\log_2 n)}}{\frac{n}{m}+1}}\\ &\geq \frac{G}{R_0}\cdot\sqrt{\frac{2n}{\log_2 n}}\sqrt{\frac{(2\sqrt{3})2\sqrt{\ln(3\log_2 n)}}{\frac{n}{\frac{1}{2}\log_2\frac{2n}{\log_2 n}-2}+1}}\\ &= \frac{G}{R_0}\sqrt{\frac{(2\sqrt{3})2\sqrt{\ln(3\log_2 n)}}{\frac{1}{1-\frac{\log_2\log_2 n+3}{\log_2 n}}+\frac{\log_2 n}{2n}}} \geq \frac{G}{R_0}.\end{aligned}$$

where the first inequality holds because of (16), the second inequality stems from the fact that $\gamma_1 \geq 1$, $\gamma_2 \geq 2$, $0 < \delta < 1$, and the definition of $\bar{\delta}$, the third inequality holds by employing $a + b \geq 2\sqrt{ab}$, the fourth inequality holds because $1 \leq n_0 \leq \frac{n}{m} + 1$, the fifth inequality holds because of the lower bound of m in (15), and the last inequality holds since $n \geq 100$ and the function is monotonically increasing with respect to n. So $\frac{G}{R_0} \leq \mu_m$. Recall that $\frac{1}{c^2} \leq \frac{G}{R_0}$, and thus $\frac{1}{c^2} \leq \mu_m$.

Given $\widehat{\mathbf{v}}_k$, denote $\widehat{\mathbf{v}}_k^*$ by the closest optimal solution to $\widehat{\mathbf{v}}_k$. 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 $\mu_{k^*} \leq \frac{1}{c^2} \leq 2\mu_{k^*}$, where $0 \leq k^* < m$. To utilize this fact, we have the following lemma.

Lemma 5. *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 $\{\widehat{\mathbf{v}}_k\}_{k=1}^m$ generated by the Algorithm 1 satisfy*

$$\|\widehat{\mathbf{v}}_{k-1} - \widehat{\mathbf{v}}_{k-1}^*\|_2 \leq R_{k-1} = 2^{-k+1}R_0, \tag{17}$$

$$P(\widehat{\mathbf{v}}_k) - P_* \leq \mu_k R_k^2 = 2^{-k}\mu_0 R_0^2. \tag{18}$$

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

$$P(\widehat{\mathbf{v}}_k) - P(\widehat{\mathbf{v}}_{k^*}) \leq \mu_{k^*} R_{k^*}^2. \tag{19}$$

Proof. We prove (17) and (18) by induction. Note that (17) holds for $k = 1$. Assume it is true for some $k > 1$ on $\mathcal{A}_{k-1}$. According to the Lemma 1, there exists a Borel set $\mathcal{B}_k$ with $\Pr(\mathcal{B}_k) \geq 1 - \bar{\delta}$ such that

$$P(\widehat{\mathbf{v}}_k) - P_* \leq R_{k-1}a(n_0, \bar{\delta}) = \frac{1}{2}\mu_k 2^{-k} R_0 R_{k-1} = \mu_k R_k^2,$$

which is (18). By the inductive hypothesis, $\|\widehat{\mathbf{v}}_{k-1} - \widehat{\mathbf{v}}_{k-1}^*\|_2 \leq R_{k-1}$ 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 quadratic growth condition and the definition of k^*, we have

$$\begin{aligned}\|\widehat{\mathbf{v}}_k - \widehat{\mathbf{v}}_k^*\|_2^2 &\leq c^2(P(\widehat{\mathbf{v}}_k) - P_*)\\ &\leq \frac{P(\widehat{\mathbf{v}}_k) - P_*}{\mu_{k^*}} \leq \frac{\mu_k R_k^2}{\mu_{k^*}} \leq R_k^2,\end{aligned}$$

which is (17) for $k + 1$.

Now we prove (19). For $k > k^*$, it is easy to show a similar conclusion as in Lemma 1 (Remark: At k-th stage with $k > k^*$, one can use the similar proof of Lemma 1 by substituting all $\mathbf{v}_*$ to $\widehat{\mathbf{v}}_{k-1}$, the first term (**I**) on the RHS becomes zero and hence we can get a tighter bound of $P(\widehat{\mathbf{v}}_k) - P(\widehat{\mathbf{v}}_{k-1})$, we here relax the bound to be $R_{k-1}a(n_0, \bar{\delta})$)), which is, there exists a Borel set $\mathcal{B}_k$ with $\Pr(\mathcal{B}_k) \geq 1 - \bar{\delta}$ such that

$$\begin{aligned}P(\widehat{\mathbf{v}}_k) - P(\widehat{\mathbf{v}}_{k-1}) &\leq R_{k-1}a(n_0, \bar{\delta})\\ &= 2^{k^*-k}R_{k^*-1}a(n_0, \bar{\delta}) = 2^{k^*-k}\mu_{k^*}R_{k^*}^2 = \mu_k R_k^2,\end{aligned}$$

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

$$\begin{aligned}P(\widehat{\mathbf{v}}_k) - P(\widehat{\mathbf{v}}_{k^*}) &= \sum_{j=k^*+1}^{k} (P(\widehat{\mathbf{v}}_j) - P(\widehat{\mathbf{v}}_{j-1}))\\ &\leq \sum_{j=k^*+1}^{k} 2^{k^*-j}\mu_{k^*}R_{k^*}^2 \leq \mu_{k^*}R_{k^*}^2.\end{aligned}$$

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

$$P(\widehat{\mathbf{v}}_{k^*}) - P_* \leq \mu_{k^*}R_{k^*}^2.$$

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

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

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 $\Pr(\mathcal{D}_m) \geq 1 - \delta$.

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

$$\begin{aligned}P(\widehat{\mathbf{v}}_1) - P_* &\leq R_0 \cdot a(n_0, \bar{\delta}) = \frac{R_0}{a(n_0, \bar{\delta})} \cdot a(n_0, \bar{\delta})^2\\ &= \frac{2}{\mu_0}a(n_0, \bar{\delta})^2 \leq 2\left(c \cdot a(n_0, \bar{\delta})\right)^2.\end{aligned}$$

Hence on $\mathcal{A}_1 \cap \mathcal{C}_m$, by using Lemma 5 and a similar argument as in case 1, we have

$$\begin{aligned}P(\widehat{\mathbf{v}}_m) - P_* =& P(\widehat{\mathbf{v}}_m) - P(\widehat{\mathbf{v}}_1) + P(\widehat{\mathbf{v}}_1) - P_*\\ \leq& 2R_0 \cdot a(n_0, \bar{\delta}) \leq (2c \cdot a(n_0, \bar{\delta}))^2,\end{aligned}$$

where $\Pr(\mathcal{A}_1 \cap \mathcal{C}_m) \geq 1 - \delta$. Combining the two cases, we have with probability at least $1 - \delta$,

$$
\begin{aligned}
&P(\widehat{\mathbf{v}}_m) - P_* \\
&\leq (4c \vee 2c)^2 \left(G \left(\frac{2\sqrt{3\gamma_1}}{\sqrt{n_0}} + \frac{\gamma_2 \sqrt{\ln\left(\frac{6n_0 \log_2 n}{2\delta}\right)}}{\sqrt{n_0}} \right) \right)^2 \\
&= \tilde{O}\left(\frac{\ln(\frac{1}{\delta})}{n} \right).
\end{aligned}
$$

□

References

Li, G. Global error bounds for piecewise convex polynomials. *Mathematical Programming*, pp. 1–28, 2013.

McDiarmid, C. On the method of bounded differences. *Surveys in combinatorics*, 141(1):148–188, 1989.

Shawe-Taylor, J. and Cristianini, N. *Kernel methods for pattern analysis*. Cambridge university press, 2004.

Ying, Y., Wen, L., and Lyu, S. Stochastic online auc maximization. In *Advances in Neural Information Processing Systems*, pp. 451–459, 2016.