CS861: Theoretical Foundations of Machine Learning Lecture 21 - 10/23/2023
University of Wisconsin–Madison, Fall 2023

Lecture 2 ncentration and structured bandits

Lecturer: Kirthe *Scribed by: Zhifeng Chen, Xinyan Wang*

Disclaimer: *These n jected to the usual scrutiny reserved for formal publications. They may be distributed outside this class only with the permission of the instructor.*

In the previous lectures, we have shown that $R_T \in \tilde{O}(d\sqrt{T})$ under the following good event $G = \left\{ \left| f(\theta_*^T a) - f(\hat{\theta}_t^T a) \right| \le \rho \|a\|_{V_{t-1}^{-1}}, \forall a \in \mathcal{A}, \forall t \in \{d+1, \dots, T\} \right\}$. In this lecture, we will show $\mathbb{P}(G^c) \le 1/T$ using martingale concentration inequalities.

1 Martingale Concentration Inequality

Theorem 1. *Let $\mathcal{F} = \{\mathcal{F}_t\}_{t\ge0}$ be a filtration. Let $\{A_t\}_{t\ge1}$ be an $\mathbb{R}_d$-valued stochastic process predictable w.r.t $\mathcal{F}$, and let $\{\epsilon_t\}_{t\ge1}$ be a real-valued martingale difference sequence adapted to $\{\mathcal{F}_t\}_{t\ge1}$. Assume ϵ_t is σ-subGaussian. Let $V_t = \sum_{s=1}^t A_s A_s^T$, $\xi_t = \sum_{s=1}^t A_s \epsilon_s$ and say $A_s^T A_s \le C^2, \forall s \in [T]$. Suppose $V_t \ge I, \forall t > t_0$. Then for all $\delta \ge e^{-\frac{1}{\sqrt{2}}}$, with probability at least $1-\delta$,*

$$\|\xi_t\|_{V_t^{-1}} \le \gamma\sigma\sqrt{2d\log(t)\log(d/\delta)}$$

Where $\gamma = \sqrt{3 + 2\log(1+2C)}$

To prove this theorem, we need the following lemma.

Lemma 1. *If A and B are random variables s.t $\mathbb{E}[e^{\lambda A - \frac{\lambda^2 B^2}{2}}] \le 1$, then $\forall \tau \ge \sqrt{2}$ and $y > 0$,*

$$\mathbb{P}\left(|A| > \tau\sqrt{(B^2+y)\left(1 + \frac{1}{2}\log\left(1 + \frac{B^2}{y}\right)\right)} \right) \le e^{-\frac{\tau^2}{2}}$$

Remark If we don't think of B as a random variable, but as a constant, then A is B-subGaussian by $\mathbb{E}[e^{\lambda A}] \le e^{\frac{\lambda^2 B^2}{2}}$. So we have $\mathbb{P}(|A| \ge B\tau) \le 2e^{\tau^2/2}$. This lemma gives a similar result when B is a random variable.

Now we can start to prove Theorem 1.

Proof Let $x \in \mathbb{R}^d$ be given. We will apply the lemma with $A = \frac{x^T\xi_t}{\sigma}$ and $B = \|x\|_{V_t} = \sqrt{x^T V_t x}$. First we should check the condition $\mathbb{E}[e^{\lambda A - \frac{\lambda^2 B^2}{2}}] \le 1 \quad \forall\lambda$.

$$\begin{aligned} \lambda A - \frac{\lambda^2 B^2}{2} &= \lambda\frac{x^T\xi_t}{\sigma} - \lambda^2\frac{x^T V_t x}{2} \\ &= \sum_{s=1}^{t} \underbrace{\left(\frac{\lambda}{\sigma} x^T A_s \epsilon_s - \frac{\lambda^2}{2} x^T A_s A_s^T x \right)}_{Q_s} \end{aligned}$$

As A_s is $\mathcal{F}_{s-1}$ measurable, it is a non-stochastic quantity given $\mathcal{F}_{s-1}$,

$$\begin{aligned}\mathbb{E}[e^{Q_s}|\mathcal{F}_{s-1}] &= [illegible] \left\|x^T A_s\right\|^2\right)\Big|\mathcal{F}_{s-1}\Big] \\ &= [illegible]_{s-1}\Big] \exp\left(-\frac{\lambda^2}{2}\left\|x^T A_s\right\|^2\right) \\ &\leq [illegible]\Big) \exp\left(-\frac{\lambda^2}{2}\left\|x^T A_s\right\|^2\right) \quad \text{(as } \epsilon_s \text{ is } \sigma\text{-sub-Gaussian)} \\ &= 1\end{aligned}$$

Therefore,

$$\begin{aligned}\mathbb{E}[e^{\lambda A-\frac{\lambda^2 B^2}{2}}] &= \mathbb{E}[e^{\sum_{s=1}^{t} Q_s}] \\ &= \mathbb{E}\left[e^{\sum_{s=1}^{t-1} Q_s}\,\mathbb{E}[e^{Q_t}|\mathcal{F}_{t-1}]\right] \\ &\leq \mathbb{E}[e^{\sum_{s=1}^{t-1} Q_s}] \\ &\leq \dots \leq 1\end{aligned}$$

We will now apply the lemma with $y = \|x\|_2^2$ and $\tau = 2\log(1/\delta')$. We will choose δ' later in terms of δ on. We require $\tau \geq \sqrt{2}$, which is satisfied if $\delta' \leq e^{-\frac{1}{\sqrt{2}}}$. Then with probability at least $1-\delta'$

$$|A| = \left|\frac{x^T \xi_t}{\sigma}\right| \leq \underbrace{\sqrt{(\|x\|_{V_t}^2 + \|x\|_2^2)\left(1 + \frac{1}{2}\log\left(1 + \frac{\|x\|_{V_t}^2}{\|x\|_2^2}\right)\right)}}_{(*)} \cdot \sqrt{2\log\frac{1}{\delta'}}$$

Next, we will show that $(*) \sim \|x\|_{V_t}^2$. For $t > t_0$, since $I \leq V_t = \sum_{s=1}^{t} A_s A_s^T \leq tC^2 I$, we have

$$\|x\|_2^2 \leq \|x\|_{V_t}^2 \leq tC^2 \|x\|_2^2$$

$$\|x\|_2^2 + \|x\|_{V_t}^2 \leq 2\|x\|_{V_t}^2$$

We can also show that $1 + \frac{1}{2}\log\left(1 + \frac{\|x\|_{V_t}^2}{\|x\|_2^2}\right) \leq \frac{\gamma^2 \log(t)}{2}$, where $\gamma = \sqrt{3 + 2\log(1 + 2C)}$ as given in the theorem. Therefore, with probability at least $1 - \delta'$ $\forall x \in \mathbb{R}^d$,

$$|x^T \xi_t| \leq \sigma\gamma \|x\|_{V_t} \sqrt{2\log(t)\log\frac{1}{\delta'}} \tag{1}$$

We can decompose $\|\xi_t\|_{V_t^{-1}}^2$ in the following way.

$$\begin{aligned}\|\xi_t\|_{V_t^{-1}}^2 &= \xi_t^T V_t^{-1} \xi_t \\ &= \xi_t^T V_t^{-\frac{1}{2}} I V_t^{-\frac{1}{2}} \xi_t \\ &= \sum_{j=1}^{d} \xi_t^T V_t^{-\frac{1}{2}} e_j e_j^T V_t^{-\frac{1}{2}} \xi_t\end{aligned}$$

Now for any $s > 0$,

$$
\begin{aligned}
&\mathbb{P}\left(\sum_{j=1}^{d} \xi_t^T V_t^{-\frac{1}{2}} e_j e_j^T V_t^{-\frac{1}{2}} \xi_t > ds^2\right) \\
&\sum_{j=1}^{d} \mathbb{P}\left(\xi_t^T V_t^{-\frac{1}{2}} e_j e_j^T V_t^{-\frac{1}{2}} \xi_t > s^2\right) \\
&\sum_{j=1}^{d} \mathbb{P}\left(\left|\xi_t^T V_t^{-\frac{1}{2}} e_j\right| > s\right)
\end{aligned}
$$

We will apply (1) with $x = V_t^{-\frac{1}{2}} e_j$, $\delta' = \delta/d$ and let $s = \sigma\gamma \left\|V_t^{-1/2} e_j\right\|_{V_t} \sqrt{\log(t)\log(d/s)} = \sigma\gamma\sqrt{\log(t)\log(d/s)}$

Finally we get

$$\mathbb{P}\left(\|\xi_t\|^2_{V_t^{-1}} \geq d\gamma^2\sigma^2 \log(t) \log\frac{d}{\delta}\right) \leq \delta$$

□

2 Bounding $\mathbb{P}(G^c)$

We have proved the martingale concentration inequality so we now proceed to prove $\mathbb{P}(G^c) \leq 1/T$. We can also use the following fact about generalized linear models.

Define $g_t(\theta) := \sum_{s=1}^{t} A_s f(A_s^T\theta)$, so we can write

$$
\begin{aligned}
\hat{\theta}_t &= \arg\min_{\theta\in\Theta} \left\|\sum_{s=1}^{t} A_s \left(f(A_s^T\theta) - X_s\right)\right\|_{V_t^{-1}} \\
&= \arg\min_{\theta\in\Theta} \left\|g_t(\theta) - \sum_{s=1}^{t} A_s X_s\right\|_{V_t^{-1}}
\end{aligned}
$$

Fact: By using quasi-maximum likelihood estimators in the exponential family, $\exists$ a unique $\tilde{\theta}_t \in \mathbb{R}^d$ s.t

$$g_t(\tilde{\theta}_t) - \sum_{s=1}^{t} A_s X_s = \sum_{s=1}^{t} A_s \left(f(A_s^T\tilde{\theta}_t) - X_s\right) = 0$$

Therefore we can write

$$\hat{\theta}_t = \arg\min_{\theta\in\Theta} \left\|g_t(\theta) - g_t(\tilde{\theta}_t)\right\|_{V_t^{-1}}$$

Consider

$$
\begin{aligned}
\left\|g_t(\theta_*) - g_t(\hat{\theta}_t)\right\|_{V_t^{-1}} &\leq \left\|g_t(\theta_*) - g_t(\tilde{\theta}_t)\right\|_{V_t^{-1}} + \left\|g_t(\tilde{\theta}_t) - g_t(\hat{\theta}_t)\right\|_{V_t^{-1}} \\
&\leq 2\left\|g_t(\theta_*) - g_t(\tilde{\theta}_t)\right\|_{V_t^{-1}} \\
&= 2\left\|\sum_{s=1}^{t} A_s\epsilon_s\right\|_{V_t^{-1}}
\end{aligned}
$$

We now prove the claim $\mathbb{P}(G^c) \leq 1/T$, where $G = \left\{ \left| f(\theta_*^T a) - f(\hat{\theta}_t^T a) \right| \leq \rho \|a\|_{V_{t-1}^{-1}}, \forall a \in \mathcal{A}, \forall t \in \{d+1, \dots, T\} \right\}$

Proof Pick a round ny $a \in \mathcal{A}$. By the L-Lipschitz property of f, We know

$$f(\hat{\theta}_t^T a) \Big| \leq L \left| (\theta_* - \hat{\theta}_t)^T a \right|$$

Now we bound $\theta_* - \hat{\theta}_t$.

$$\sum_{s=1} {}_s A_s^T f'(A_s^T \theta) \geq c \sum_{s=1}^{t-1} A_s A_s^T \geq cI$$

As f' is continuous, by the fundamental theorem of calculus,

$$g_{t-1}(\theta_*) - g_{t-1}(\hat{\theta}_{t-1}) = G_{t-1} (\theta_* - \hat{\theta}_{t-1})$$

Where $G_{t-1} = \int_0^1 \nabla g_{t-1} \left(s\theta_* + (1-s)\hat{\theta}_{t-1} \right) ds$

(proof to be continued in the next class) □