CS861: Theoretical Foundations of Machine Learning Lecture 1 - 10/20/2023

University of Wisconsin–Madison, Fall 2023

Lecture 20: Structured Bandits, Martingales

Lecturer: Kirthevasan Kandasamy *Scribed by: Alex Clinton, Chenghui Zheng*

Disclaimer: *These notes have not been subjected to the usual scrutiny reserved for formal publications. They may be distributed outside this class only with the permission of the instructor.*

In the last lecture we introduced a more general bandit framework and proposed an analogous UCB algorithm. In this lecture we will analyze the algorithm and introduce the formalism of martingales so that we can utilize their popular results.

1 Structured Bandits

Theorem 1. *Consider the algorithm introduced at the end of the previous lecture. For sufficiently large T we have*

$$R_T = Tf(\theta^T a_*) - \mathbb{E}[\sum_{t=1}^{T} x_i] \in \tilde{O}(d\sqrt{T})$$

where $a_* = \underset{a \in \mathcal{A}}{\arg\max} f(\theta_*^T a)$

Proof We start by defining a good event analogous to the one we define in the analysis of UCB. Let $G = \{|f(\theta_t^T a) - f(\hat{\theta}_{t-1}^T a)| \leq \rho ||a||_{V_{t-1}^{-1}}, \forall a \in \mathcal{A}, \forall t \in \{d+1, \dots, T\}\}$. Observe that here $\rho||a||_{V_{t-1}^{-1}}$ is playing the role of an upper confidence bound. We will use the following two claims to aid in the proof.

Claim 1. $\mathbb{P}(G^c) \leq \frac{1}{T}$. *We will prove this later using some martingale concentration results.*

Claim 2. *Under G, $f(\theta_*^T a_*) - (\theta_*^T A_t) \leq 2\rho(T)||A_t||_{V_{t-1}^{-1}}$ for all $t > d$.*

Claim 2 can be verified via the following simple calculation.

$$\begin{aligned} f(\theta_*^T a_i) - f(\theta_*^T A_t) &\leq f(\hat{\theta}_{t-1} a_*) + \rho(t)||a_*||_{V_{t-1}^{-1}} - (f(\hat{\theta}_{t-1} A_t) + \rho(t)||A_t||_{V_{t-1}^{-1}}) \\ &\leq f(\hat{\theta}_{t-1} A_t) - \rho(t)||A_t||_{V_{t-1}^{-1}} - (f(\hat{\theta}_{t-1} A_t) + \rho(t)||A_t||_{V_{t-1}^{-1}}) \\ &\leq 2\rho(t)||A_t||_{V_{t-1}^{-1}} \leq 2\rho(T)||A_t||_{V_{t-1}^{-1}} \end{aligned}$$

Next, to bound the regret, write the pseudo-regret $\bar{R}_T = Tf(\theta_*^T a_*) - \sum_{i=1}^{T} f(\theta_*^T A_t)$ so that $R_T = \mathbb{E}[\bar{R}_T]$. Using the tower property, we have:

$$R_T = \mathbb{E}(\bar{R}_T|G)\underbrace{\mathbb{P}(G)}_{\leq 1} + \underbrace{\mathbb{E}(\bar{R}_T|G^c)}_{\leq Tf_{\max}}\underbrace{\mathbb{P}(G^c)}_{\leq \frac{1}{T}}$$

where $f_{\max} = \max\limits_{a,a' \in \mathcal{A}} (f(\theta_*^T a) - f(\theta_*^T a'))$. Under the good event G,

$$\begin{aligned}
\bar{R}_t &= (f(\theta_*^T a_*) - f(\theta_*^T A_t)) \\
&\leq df_{\max} + \sum_{t=d+1}^{T} \min(f_{\max}, f(\theta_*^T a_*) - f(\theta_*^T A_t)) \\
&\leq df_{\max} + \sum_{t=d+1}^{T} \min(f_{\max}, 2\rho(t)||A_t||_{V_{t-1}^{-1}}) \\
&\leq df_{\max} + 2\rho(t) \sum_{t=d+1}^{T} \min(1, ||A_t||^2_{V_{t-1}^{-1}}) \quad \text{, as } f_{\max} \leq 2\rho(t) \text{ for sufficiently large T} \\
&\leq df_{\max} + 2\rho(t) \sqrt{T \sum_{t=d+1}^{T} \min(1, ||A_t||^2_{V_{t-1}^{-1}})} \text{ , by the Cauchy-Schwarz inequality,} \\
&\text{where } a_i = 1, b_i = \min(1, ||A_t||^2_{V_{t-1}^{-1}})
\end{aligned} \tag{1}$$

Next, we will bound $\sum_{t=d+1}^{T} min(1, ||A_t||^2_{V_{t-1}^{-1}})$. Consider for $t > d$,

$$\begin{aligned}
\det(V_t) &= \det(V_{t-1} + A_t A_t^T), \text{ where } V_t = \sum_{s=1}^{t} A_s A_s^T \\
&= \det(V_{t-1}^{\frac{1}{2}}(I + V_{t-1}^{-\frac{1}{2}} A_t A_t^T V_{t-1}^{-\frac{1}{2}}) V_{t-1}^{\frac{1}{2}}) \\
&= \det(V_{t-1}) \det(I + (V_{t-1}^{-\frac{1}{2}} A_t)(V_{t-1}^{-\frac{1}{2}} A_t)^T) \text{ , since } \det(AB) = \det(A)\det(B) \\
&= \det(V_{t-1})(1 + ||A_t||^2_{V_{t-1}^{-1}}), \text{ since } \det(I + UV^T) = 1 + U^T V
\end{aligned} \tag{2}$$

Therefore, $\det(V_T) = \det(V_d) \prod_{t=d+1}^{T} (1 + ||A_t||^2_{V_{t-1}^{-1}}) = \prod_{t=d+1}^{T} (1 + ||A_t||^2_{V_{t-1}^{-1}})$.

This means that

$$\sum_{t=d+1}^{T} \log\left(1 + ||A_t||^2_{V_{t-1}^{-1}}\right) = \log\left(\det(V_T)\right) \leq d\log(T)$$

where the last inequality follows from the fact

$$\det(V_T) \leq \left(\frac{\text{Trace}(V_t)}{d}\right)^d = \left(\frac{\sum_{s=1}^{T} ||A_s||_2^2}{d}\right)^d = \left(\frac{dT}{d}\right)^d = T^d$$

We will now use the following inequality: $x \leq 2\log(1+x)$, $\forall x \in [0, 2\log(2)] \supseteq [0,1]$ to get

$$\begin{aligned}
\sum_{t=d+1}^{T} \min(1, ||A_t||_2^2) &\leq 2 \sum_{t=d+1}^{T} \log\left(1 + \min\left(1, ||A_t||^2_{V_{t-1}^{-1}}\right)\right) \\
&\leq 2 \sum_{t=d+1}^{T} \log\left(1 + ||A_t||^2_{V_{t-1}^{-1}}\right) \\
&\leq 2d\log(T)
\end{aligned}$$

Therefore, under G $\overline{R}_T \leq df_{\max} + 2\rho(T)\sqrt{2Td\log(T)}$, and so using $\overline{R}_T = \mathbb{E}[\overline{R}_T | \, G]\mathbb{P}(G) + \underbrace{\mathbb{E}[\overline{R}_T | \, G^c]}_{\leq T f_{max}} \underbrace{\mathbb{P}(G^c)}_{\leq \frac{1}{T}}$, we can conclude that

$$\overline{R}_T \leq (d+1)f_{max} + 2 \underbrace{\rho(T)}_{\in \tilde{O}(\sqrt{d})} \sqrt{dT} \in \tilde{O}(d\sqrt{T})$$

□

Next, we need to prove Claim 1. In order to prove this result, we will begin with a review of martingales.

2 Review of martingales

In sequential decision making, information is revealed to the learner sequentially, and the learner makes decisions based on the information available. Filtrations are a construct used to formalize the amount of information available to the learner at a given time.

Definition 1. *$\mathcal{F} = \{\mathcal{F}_t\}_{t\in\mathbb{N}}$ is a filtration if $\forall t,\ \mathcal{F}_t$ is a σ-algebra and $\mathcal{F}_t \subseteq \mathcal{F}_{t+1}$*

In the context of stochastic bandits, $\mathcal{F}_t = \sigma\left(\{A_s, X_s\}_{s=1}^{t-1}\right)$ is the σ-algebra generated by actions and rewards up to round t.

Definition 2. *Predictable processes and adapted processes:*

1. *A stochastic process $\{X_t\}_{t\in\mathbb{N}}$ is predictable with respect to a filtration $\{\mathcal{F}\}_{t\in\mathbb{N}}$ if X_{t+1} is measurable (predictable).*
2. *A stochastic process $\{X_t\}_{t\in\mathbb{N}}$ is adapted to a filtration $\{\mathcal{F}\}_{t\in\mathbb{N}}$ if X_t is $\mathcal{F}_t$-measurable.*

Example 2. In stochastic bandits, the actions A_t are predictable as A_t is determined based on actions up to round $t-1$.

Definition 3. *Martingales and martingale difference sequences*

1. *An $\mathcal{F}$-adapted sequence of random variables $\{X_t\}_{t\in\mathbb{N}}$ is a martingale if*
   - *(i)* $\mathbb{E}[X_t|\mathcal{F}_{t-1}] = X_{t-1}$
   - *(ii)* $\mathbb{E}[|X_t|] < \infty$
2. *An $\mathcal{F}$-adapted sequence of random variables $\{Y_t\}_{t\in\mathbb{N}}$ is a martingale difference sequence if*
   - *(i)* $\mathbb{E}[Y_t|X_t] = 0$
   - *(ii)* $\mathbb{E}[|Y_t|] < \infty$

Example 3. If $\{X_t\}_{t\in\mathbb{N}}$ is a martingale, then $Y_t = X_t - X_{t-1}$ is a martingale difference sequence.

2.1 Martingale contraction

There are many popular martingale concentration results that we can use, such as the Hoeffding-Azuma inequality, and a martingale version of the Bernstein inequality (e.g Freedman 2009). Often however, in sequential feedback settings, we may need to develop a customized result suited to our problem setting. To that end, we will introduce and prove the following result.

Lemma 1. *Let $\mathcal{F} = \{\mathcal{F}_t\}_{t\geq 0}$ be a filtration. Let $\{A_t\}_{t\geq 0}$ be an $\mathbb{R}^d$-valued stochastic process predictable with respect to $\mathcal{F}$, and let $\{\varepsilon\}_{t\geq 1}$ be a real-valued martingale difference sequence adapted to $\{\mathbb{F}_t\}_{t\geq 2}$. Assume ε_t is σ-sub Gaussian, i.e. $\forall\lambda,\ \mathbb{E}[e^{\lambda\varepsilon t}] \leq e^{\frac{\lambda^2\sigma^2}{2}}$. Let $V_t = \sum_{s=1}^{t} A_s A_s^T$ and $\xi_t = \sum_{s=1}^{t} A_s \xi_s$ where $A_s^T A_s \leq c,\ \forall x$. Suppose $v_t \succcurlyeq I, \forall t > t_0$. Then for all $\delta \leq e^{-\frac{1}{\sqrt{2}}}$, with probability at least $1-\delta$*

$$||\xi_t||_{V_t^{-1}} \leq \gamma\sigma\sqrt{2d\log(t)\log\left(\frac{d}{\delta}\right)} \qquad \textit{where} \qquad \gamma = \sqrt{3 + 2\log(1+2c)}$$