Note for Bandit Algorithms

Chenmi'en Tan
E-mail: chenmientan@outlook.com
Actively Updating (Last update: June 26, 2021)

Abstract

This is a personal note covers solution for some exercises proposed in *bandit algorithms* written by Tor Lattimore and Csaba Szepesvári.

Theorem 1 (Exercise 5.14, Bernstein's inequality). *Suppose that $X_1, \dots, X_n$ are independent random variables where $X_i - \mathbb{E}[X_i] \leq b, \forall i = 1, \dots, n$ almost surely, then by denoting $S_n = X_1 + \cdots + X_n$, for any $\varepsilon \geq 0$ there is*

$$\mathbb{P}(S_n - \mathbb{E}[S_n] \geq \varepsilon) \leq \exp(-\frac{\frac{1}{2}\varepsilon^2}{\mathbb{V}[S_n] + \frac{\varepsilon b}{3}})$$

Proof. Without the loss of generality, assume $\mathbb{E}[X_i] = 0, \forall i = 1, \dots, n$. For any $\lambda > 0$, there is

$$\begin{aligned}\mathbb{P}(S_n \geq \varepsilon) &= \mathbb{P}(\exp(\lambda S_n) \geq \exp(\lambda\varepsilon)) \leq \exp(-\lambda\varepsilon)\mathbb{E}[\exp(\lambda S_n)] \\ &= \exp(-\lambda\varepsilon)\prod_{i=1}^{n}\mathbb{E}[\exp(\lambda X_i)]\end{aligned}$$

Since $g(x) = \frac{\exp(x)-x-1}{x^2}$ is increasing, there is $g(\lambda X_i) \leq g(\lambda b), \forall i = 1, \dots, n$ almost surely, which is equivalent to

$$b^2(\exp(\lambda X_i) - \lambda X_i - 1) \leq X_i^2(\exp(\lambda b) - \lambda b - 1)$$

almost surely. Through assigning expectation to the both sides we can obtain $b^2(\mathbb{E}[\exp(\lambda X_i)] - 1) \leq \mathbb{V}[X_i](\exp(\lambda b) - \lambda b - 1)$. Thus,

$$\begin{aligned}\mathbb{P}(S_n \geq \varepsilon) &\leq \exp(-\lambda\varepsilon)\prod_{i=1}^{n}(1 + \frac{\exp(\lambda b) - \lambda b - 1}{b^2}\mathbb{V}[X_i]) \\ &\leq \exp(\frac{\exp(\lambda b) - \lambda b - 1}{b^2}\mathbb{V}[S_n] - \lambda\varepsilon)\end{aligned}$$

By letting $\lambda = \frac{1}{b}\log(1 + \frac{\varepsilon b}{\mathbb{V}[S_n]})$, we obtain

$$\mathbb{P}(S_n \geq \varepsilon) \leq \exp(-\frac{\mathbb{V}[S_n]}{b^2}[(1 + \frac{\varepsilon b}{\mathbb{V}[S_n]})\log(1 + \frac{\varepsilon b}{\mathbb{V}[S_n]}) - \frac{\varepsilon b}{\mathbb{V}[S_n]}])$$

Combined with $(1+x)\log(1+x) - x \geq \frac{3x^2}{6+2x}, \forall x \in \mathbb{R}_+$, we conclude the desired conclusion. □

1 Stochastic Bandit

Algorithm 1: ε-Greedy

Data: number of arms k and parameter sequence $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$

for $t = 1, \ldots, n$ **do**

Choose the action $A_t \leftarrow \arg\max_{i \in \{1,\ldots,k\}} \hat{\mu}_i(t-1)$ with probability $1 - \varepsilon_t$, otherwise choose an arm uniformly at rondom

Observe the reward and update the expectation estimation

end

Theorem 2 (Exercise 6.7)**.** *Let $\Delta_{\min} = \min\{\Delta_i : \Delta_i > 0\}$ and $\varepsilon_t = \min\{1, \frac{Ck}{t\Delta_{\min}^2}\}$, where $C > 0$ is sufficiently large. Suppose that the bandit $\nu \in \mathcal{E}_{\mathrm{SG}}^k(1)$, then for ε-Greedy depends on parameter sequence $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$, there exists a universal $C' > 0$ such that*

$$R_n \leq C' \sum_{i=1}^{k} (\Delta_i + \frac{\Delta_i}{\Delta_{\min}^2} \log \max\{e, \frac{n\Delta_{\min}^2}{k}\})$$

Proof. Let $T = \sum_{t=1}^{n} \frac{\varepsilon_t}{2k}$. Denote $X_i(t)$ as the indicator that represents arm i is selected at random in episode t and $T_i^R(n) = X_i(1) + \cdots + X_i(n)$ as the number of actions where arm i is selected at random in n rounds. It is clear that for any i, t, $X_i(t) - \frac{\varepsilon_t}{k}$ is zero-mean random variable upper bounded by 1. Combined with $\mathbb{E}[T_i^R(n)] = \sum_{t=1}^{n} \frac{\varepsilon_t}{k} = 2T$ and $\mathbb{V}[T_i^R(n)] = \sum_{t=1}^{n} \frac{\varepsilon_t}{k}(1 - \frac{\varepsilon_t}{k}) \leq \sum_{t=1}^{n} \frac{\varepsilon_t}{k} = 2T$, through theorem 1, we conclude that $\mathbb{P}(T_i^R(t) \leq T) \leq \exp(-\frac{3T}{14})$. Meanwhile, for any i, t, there is

$$\begin{aligned}
\mathbb{P}(\hat{\mu}_i(t) - \mu_i \geq \frac{\Delta_i}{2}) &= \sum_{s=1}^{t} \mathbb{P}(T_i(t) = s | \hat{\mu}_{is}(t) - \mu_i \geq \frac{\Delta_i}{2}) \mathbb{P}(\hat{\mu}_{is} - \mu_i \geq \frac{\Delta_i}{2}) \\
&\leq \sum_{s=1}^{t} \mathbb{P}(T_i(t) = s | \hat{\mu}_{is}(t) - \mu_i \geq \frac{\Delta_i}{2}) \exp(-\frac{s\Delta_i^2}{8})
\end{aligned}$$

Since $\sum_{s=T+1}^{\infty} \exp(-ks) \leq \exp(-kT)/k$, we can further obtain

$$\begin{aligned}
\mathbb{P}(\hat{\mu}_i(t) - \mu_i \geq \frac{\Delta_i}{2}) &\leq \frac{8\exp(-\frac{\Delta_i^2 \lfloor T \rfloor}{8})}{\Delta_i^2} + \sum_{s=1}^{\lfloor T \rfloor} \mathbb{P}(T_i(t) = s | \hat{\mu}_{is}(t) - \mu_i \geq \frac{\Delta_i}{2}) \\
&\leq \frac{8\exp(-\frac{\Delta_i^2 \lfloor T \rfloor}{8})}{\Delta_i^2} + \sum_{s=1}^{\lfloor T \rfloor} \mathbb{P}(T_i^R(t) \leq s) \\
&\leq \frac{8\exp(-\frac{\Delta_i^2 \lfloor T \rfloor}{8})}{\Delta_i^2} + T\exp(-\frac{3T}{14})
\end{aligned}$$

Similarily we can derive $\mathbb{P}(\hat{\mu}_1(t) - \mu_1 \le -\frac{\Delta_i}{2}) \le \frac{8\exp(-\frac{\Delta_i^2 \lfloor T \rfloor}{8})}{\Delta_i^2} + T\exp(-\frac{3T}{14})$. Hence

$$\begin{aligned}\mathbb{P}(A_t = i) &\le \frac{\varepsilon_t}{k} + \mathbb{P}(\hat{\mu}_i(t-1) \ge \hat{\mu}_1(t-1)) \\ &\le \frac{\varepsilon_t}{k} + \mathbb{P}(\{\hat{\mu}_i(t-1) - \mu_i \ge \frac{\Delta_i}{2}\} \cup \{\hat{\mu}_1(t-1) - \mu_1 \le -\frac{\Delta_i}{2}\}) \\ &\le \frac{\varepsilon_t}{k} + \frac{16\exp(-\frac{\Delta_i^2 \lfloor T \rfloor}{8})}{\Delta_i^2} + 2T\exp(-\frac{3T}{14})\end{aligned}$$

□

Algorithm 2: Upper Confidence Bound (UCB)

Data: number of arms k and confidence parameter δ
for $t = 1, \ldots, n$ **do**
  Choose the action $A_t \leftarrow \arg\max_{i \in \{1,\ldots,k\}} \hat{\mu}_i(t-1) + \sqrt{\frac{2\log(1/\delta)}{T_i(t-1)}}$
  Observe the reward and update the upper confidence bound
end

Theorem 3. *Suppose that the bandit $\nu \in \mathcal{E}_{\mathrm{SG}}^k(1)$, then for UCB depends on confidence parameter $\delta \in (0, 1]$, there is*

$$\mathbb{P}(\overline{R}_n \ge (4k-4)\sqrt{2n\log\frac{1}{\delta}} + 2\sum_{i=1}^{k}\Delta_i) \le (n+k-1)\delta$$

Proof. Without loss of the gernerality, assume the first arm is optimal. For any $(u_2, \ldots, u_k) \in \mathbb{N}_+^{k-1}$, define

$$G = \{\mu_1 < \min_{t \in \{1,\ldots,n\}} \mathrm{UCB}_1(t, \delta)\} \cap \bigcap_{i=2}^{k} \{\hat{\mu}_{iu_i} + \sqrt{\frac{2}{u_i}\log\frac{1}{\delta}} < \mu_1\}$$

When G occurs, there is $T_i(n) \le u_i, \forall i = 2, \ldots, k$. Hence $T_i(n) \ge u_i + 1, \exists i = 2, \ldots, k$ implies G^c. Assume that $u_i, i = 2, \ldots, k$ are large sufficiently to satisfy

$$\Delta_i - \sqrt{\frac{2}{u_i}\log\frac{1}{\delta}} \ge c\Delta_i, \forall i = 2, \ldots, k \tag{1}$$

for some $c \in (0,1)$. At this moment we have

$$
\begin{aligned}
G^c &= \{\mu_1 \geq \min_{t \in \{1,\ldots,n\}} \mathrm{UCB}_1(t,\delta)\} \cup \bigcup_{i=2}^{k} \{\hat{\mu}_{iu_i} + \sqrt{\frac{2}{u_i} \log \frac{1}{\delta}} \geq \mu_1\} \\
&\subset \{\mu_1 \geq \min_{s \in \{1,\ldots,n\}} \hat{\mu}_1 + \sqrt{\frac{2}{s} \log \frac{1}{\delta}}\} \cup \bigcup_{i=2}^{k} \{\hat{\mu}_{iu_i} - \mu_i \geq \Delta_i - \sqrt{\frac{2}{u_i} \log \frac{1}{\delta}}\} \\
&\subset \bigcup_{s=1}^{n} \{\mu_1 \geq \hat{\mu}_{1s} + \sqrt{\frac{2}{s} \log \frac{1}{\delta}}\} \cup \bigcup_{i=2}^{k} \{\hat{\mu}_{iu_i} - \mu_i \geq c\Delta_i\}
\end{aligned}
$$

Hence we can obtain

$$
\mathbb{P}(G^c) \leq n\delta + \sum_{i=2}^{k} \exp(-\frac{u_i c^2 \Delta_i^2}{2})
$$

which holds under the restriction of (1). Assign $c = 1/2$ and $u_i, i = 2, \ldots, k$ to be the minimal feasible value, i.e.,

$$
u_i = \lceil \frac{8 \log \frac{1}{\delta}}{\Delta_i^2} \rceil
$$

we can obtain

$$
\begin{aligned}
&\mathbb{P}(\exists i \in \{2,\ldots,k\}, T_i \geq u_i + 1) \\
&\leq \mathbb{P}(G^c) \leq n\delta + (k-1)\exp(-\log \frac{1}{\delta}) = (n+k-1)\delta
\end{aligned} \tag{2}
$$

Meanwhile, for any real number $\Delta > 0$

$$
\begin{aligned}
&\mathbb{P}(\exists i \in \{2,\ldots,k\}, T_i \geq u_i + 1) \\
&\geq \mathbb{P}(\overline{R}_n \geq \sum_{i:\Delta_i < \Delta} n\Delta_i + \sum_{i:\Delta_i \geq \Delta} (u_i + 1)\Delta_i) \\
&\geq \mathbb{P}(\overline{R}_n \geq (k-1)n\Delta + \sum_{i:\Delta_i \geq \Delta} [(\frac{8 \log \frac{1}{\delta}}{\Delta_i^2} + 2)\Delta_i]) \\
&\geq \mathbb{P}(\overline{R}_n \geq (k-1)(n\Delta + \frac{8 \log \frac{1}{\delta}}{\Delta}) + 2\sum_{i=1}^{k} \Delta_i)
\end{aligned}
$$

By letting $\Delta = \sqrt{8\log(1/\delta)/n}$, we have

$$
\begin{aligned}
&\mathbb{P}(\exists i \in \{2,\ldots,k\}, T_i \geq u_i + 1) \\
&\geq \mathbb{P}(\overline{R}_n \geq (4k-4)\sqrt{2n \log \frac{1}{\delta}} + 2\sum_{i=1}^{k} \Delta_i)
\end{aligned} \tag{3}
$$

By combing equation (2) and (3) we obtain the desired conclusion. □

Theorem 4. *Suppose that RV X satisfies* $\mathrm{supp}(X) \subset [a,b]$ *and X is bounded by B with at least probability $1-\beta$, i.e.,*

$$\mathbb{P}(X \geq B) \leq \beta$$

then for any $\alpha \in [\beta, 1)$, the conditional value at risk at level α is bounded by $\frac{\beta}{\alpha}b + (1-\frac{\beta}{\alpha})B$.

Theorem 5. *Suppose that the bandit $\nu \in \mathcal{E}^k_{\mathrm{SG}}(1)$ and the suboptimality gap $\Delta_i, i = 1, \ldots, k$ is bounded by U, then for any $\alpha \in [(n+k-1)\delta, 1)$, the UCB depends on confidence parameter $\delta \in (0,1]$ satisfies that the conditional value at risk for the pseudo-regret $\overline{R}_n = \sum_{t=1}^n \Delta_{A_t}$ at level α is bounded by*

$$\frac{(n+k-1)\delta}{\alpha} nU + (1 - \frac{(n+k-1)\delta}{\alpha})[(4k-4)\sqrt{2n \log \frac{1}{\delta}} + 2kU]$$

References