Concentration Inequalities and Multi-Armed Bandits

Nan Jiang

September 6, 2018

1 Hoeffding's Inequality

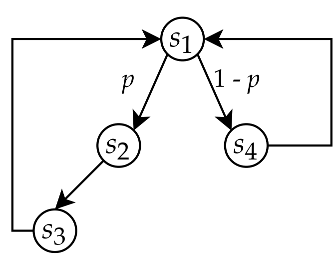
Theorem 1. *Let $X_1, \ldots, X_n$ be independent random variables on $\mathbb{R}$ such that X_i is bounded in the interval $[a_i, b_i]$. Let $S_n = \sum_{i=1}^n X_i$. Then for all $t > 0$,*

$$\Pr[S_n - \mathbb{E}[S_n] \geq t] \leq e^{-2t^2 / \sum_{i=1}^n (b_i - a_i)^2}, \tag{1}$$

$$\Pr[S_n - \mathbb{E}[S_n] \leq -t] \leq e^{-2t^2 / \sum_{i=1}^n (b_i - a_i)^2}. \tag{2}$$

Remarks:

- By union bound, we have $\Pr[|S_n - \mathbb{E}[S_n]| \geq t] \leq 2e^{-2t^2 / \sum_{i=1}^n (b_i - a_i)^2}$.
- We often care about the convergence of the empirical mean to the true average, so we can devide S_n by n: $\Pr\left[\left|\frac{S_n}{n} - \frac{\mathbb{E}[S_n]}{n}\right| \geq t\right] \leq 2e^{-2n^2t^2 / \sum_{i=1}^n (b_i - a_i)^2}$.
- A useful rephrase of the result when all variables share the same support $[a, b]$: with probability at least $1 - \delta$, $\left|\frac{S_n}{n} - \frac{\mathbb{E}[S_n]}{n}\right| \leq (b - a)\sqrt{\frac{1}{2n} \ln \frac{2}{\delta}}$.
- $X_1, \ldots, X_n$ are not necessarily identically distributed; they just have to be independent.
- The number of variables, n, is a constant in the theorem statement. When n is a random variable itself, for Hoeffding's inequality to apply, n cannot depend on the realization of $X_1, \ldots, X_n$. *Example:* Consider the following Markov chain:



Say we start at s_1 and sample a path of length T (T is a constant). Let n be the number of times we visit s_1, and we can use the transitions from s_1 to estimate p.

1. Can we directly apply Hoeffding's inequality here with n as the number of coin tosses? If you want to derive a concentration bound for this problem, look up Azuma's inequality.

2. What if we sample a path until we visit s_1 N times for some constant N? Can we apply Hoeffding's inequality with N as the number of random variables?

2 Multi-Armed Bandits (MAB)

2.1 Formulation

A MAB problem is specified by K distributions over $\mathbb{R}$, $\{R_i\}_{i=1}^K$. Each R_i has bounded supported $[0,1]$ and mean μ_i. Let $\mu^\star = \max_{i\in[K]} \mu_i$. For round $t = 1, 2, \ldots, T$, the learner

1. Chooses arm $i_t \in [K]$.
2. Receives reward $r_t \sim R_{i_t}$.

A popular objective for MAB is the pseudo-regret, which poses the *exploration-exploitation* challenge:

$$\text{Regret}_T = \sum_{t=1}^{T} (\mu^\star - \mu_{i_t}).$$

Another important objective is the simple regret:

$$\mu^\star - \mu_{\hat{i}},$$

where $\hat{i}$ is the arm that the learner picks after T rounds of interactions. This poses the "pure exploration" challenge, since all it matters is to make a good final guess and the regret incurred within the T rounds does not matter. A related objective is called Best-Arm Identification, which asks whether $\hat{i} \in \arg\max_{i\in[K]} \mu_i$; Best-Arm Identification results often require additional gap conditions.

2.2 Uniform sampling

We consider the simplest algorithm that chooses each arm the same number of times, and after T rounds selects the arm with the highest empirical mean. For simplicity let's assume that T/K is an integer. We will prove a high-probability bound on the simple regret. The analysis gives an example of the application of Hoeffiding's inequlaity to a learning problem; the algorithm itself is likely to be suboptimal.

For simplicity let's assume that T/K is an integer. After T rounds, each arm is chosen T/K times, and let $\hat{\mu}_i$ be the empirical average reward associated with arm i. By Hoeffding's inequality, we have:

$$\Pr[|\hat{\mu}_i - \mu_i| \geq \epsilon] \leq 2e^{-2T\epsilon^2/K}.$$

Now we want accurate estimation for *all* arms simultaneously. That is, we want to bound the probability of the event that *any* $\hat{\mu}_i$ deviating from μ_i too much. This is where union bound is useful:

$$\Pr\left[\bigcup_{i=1}^{K} \{|\hat{\mu}_i - \mu_i| \geq \epsilon\}\right] \qquad \text{(the event that estimation is } \epsilon\text{-inaccurate for at least 1 arm)}$$

$$\leq \sum_{i=1}^{K} \Pr\left[|\hat{\mu}_i - \mu_i| \geq \epsilon\right] \leq 2Ke^{-2T\epsilon^2/K}. \qquad \text{(union bound, then Hoeffding's inequality)}$$

To rephrase this result: with probability at least $1-\delta$, $|\hat{\mu}_i - \mu_i| \le \sqrt{\frac{K}{2T}\ln\frac{2K}{\delta}}$ holds for all i simultaneously.

Finally, we use the estimation error to bound the decision loss: recall that $\hat{i} = \arg\max_{i\in[K]} \hat{\mu}_i$, and let $i^\star = \arg\max_{i\in[K]} \mu_i$.

$$\begin{aligned}\mu^\star - \mu_{\hat{i}} &= \mu_{i^\star} - \hat{\mu}_{i^\star} + \hat{\mu}_{i^\star} - \mu_{\hat{i}} \\ &\le \mu_{i^\star} - \hat{\mu}_{i^\star} + \hat{\mu}_{\hat{i}} - \mu_{\hat{i}} \le 2\sqrt{\frac{K}{2T}\ln\frac{2K}{\delta}}.\end{aligned}$$

We can rephrase this result as a sample complexity statement: in order to guarantee that $\mu^\star - \mu_{\hat{i}} \le \epsilon$ with probablity at least $1-\delta$, we need $T = O\left(\frac{K}{\epsilon^2}\ln\frac{K}{\delta}\right)$.

2.3 Lower bound

The linear dependence of the sample complexity on K makes a lot of sense, as to choose a arm with high reward we have to try each arm at least once. Below we will see how to mathematically formalize this idea and prove a lower bound on the sample complexity of MAB.

Theorem 2. *For any $K \ge 2$, $\epsilon \le \sqrt{1/8}$, and any MAB algorithm, there exists an MAB instance where $\mu^\star$ is ϵ better than other arms, yet the algorithm identifies the best arm with no more than $2/3$ probability unless $T \ge \frac{K}{72\epsilon^2}$.*

The theorem itself is stated as a best-arm identification lower bound, but it is also a lower bound for simple regret minimization. This is because all arms except the best one is ϵ worse than $\mu^\star$, so missing the optimal arm means a simple regret of at least ϵ.

See the proof in [1] (Theorem 2); the technique is due to [2] and can be also used to prove the lower bound on the regret of MAB.

References

[1] Akshay Krishnamurthy, Alekh Agarwal, and John Langford. PAC reinforcement learning with rich observations. In *Advances in Neural Information Processing Systems*, pages 1840–1848, 2016.

[2] Peter Auer, Nicolò Cesa-Bianchi, and Paul Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine learning*, 47(2-3):235–256, 2002.