The Explore-Then-Commit Strategy for Multi-arm Bandits

COMP 767

Pascale Gourdeau

Material from:
http://banditalgs.com/2016/09/14/first-steps-explore-then-commit/

January 13th, 2017

The Strategy

The Explore-Then-Commit (ETC) strategy is very simple:

- First suppose we stop after n steps.
- Explore for a fixed time (each action is selected m times).
- Exploit for the rest of the time (choose one action and stick with it).

Goal: Present how to carry out regret analysis.

Notation and Definitions

Notation

n	—	# of steps/rounds
m	—	# of steps for which any action is explored
R_t	—	reward at time t
$q_*(a)$	—	expected reward of action a
$Q_t(a)$	—	estimated value of action a at time t

Definitions

$\Delta_a = \max_b q_*(b) - q_*(a)$:	immediate regret of action a
$J_n = \sum_{t=1}^{n} \mathbb{E}[\Delta_{A_t}]$:	expected regret
$R_t - \mathbb{E}[R_t]$:	noise for reward

The Theorem

Theorem

Assume that the noise of the reward of each arm in a k-armed stochastic bandit problem is 1-subgaussian. Then, after $n \geq mk$ rounds, the expected regret J_n of ETC which explores each arm exactly m times before committing is bounded as follows:

$$J_n \leq m \sum_{a=1}^{k} \Delta_a + (n - mk) \sum_{a=1}^{k} \Delta_a \exp\left(-\frac{m\Delta_a^2}{4}\right) \tag{1}$$

Subgaussian Random Variables

Definition

A random variable X is σ^2-subgaussian if for all $\lambda \in \mathbb{R}$

$$\mathbb{E}\left[e^{\lambda X}\right] \leq \exp\left(\frac{\lambda^2\sigma^2}{2}\right) \quad .$$

Lemma

Suppose that X is σ^2-subgaussian and X_1 and X_2 are independent and σ_1^2- and σ_2^2-subgaussian respectively. Then

1. $\mathbb{E}[X] = 0$ *and* $Var(X) \leq \sigma^2$.
2. *For all* $c \in \mathbb{R}$, cX *is* $c^2\sigma^2$*-subgaussian.*
3. $X_1 + X_2$ *is* $(\sigma_1^2 + \sigma_2^2)$*-subgaussian.*

Proof of Lemma

1. $\mathbb{E}[X] = 0$ *and* $Var(X) \leq \sigma^2$*:*

First note that by using Taylor series:

$$\mathbb{E}\left[e^{\lambda X}\right] = \sum_{n=0}^{\infty} \frac{\lambda^n}{n!}\mathbb{E}[X^n] \tag{2}$$

And since X is σ^2-subgaussian:

$$\mathbb{E}\left[e^{\lambda X}\right] \leq \exp\left(\frac{\lambda^2\sigma^2}{2}\right) = \sum_{n=0}^{\infty} \frac{(\lambda\sigma)^{2n}}{2^n n!} \tag{3}$$

Putting (2) and (3) together, and noting that the first term on each side of the sum is 1:

$$\sum_{n=1}^{\infty} \frac{\lambda^n}{n!}\mathbb{E}[X^n] \leq \sum_{n=1}^{\infty} \frac{(\lambda\sigma)^{2n}}{2^n n!} \tag{4}$$

Proof of Lemma Cont'd

$$\sum_{n=1}^{\infty} \frac{\lambda^n}{n!} \mathbb{E}[X^n] \leq \sum_{n=1}^{\infty} \frac{(\lambda\sigma)^{2n}}{2^n n!}$$

Now, since the above inequality is true for all $\lambda \in \mathbb{R}$, we only keep the terms for $n = 1$ and $n = 2$ on the LHS and let $\lambda > 0$, $\lambda \to 0$:

$$\lambda \mathbb{E}[X] + \frac{\lambda^2}{2} \mathbb{E}\left[X^2\right] \leq \frac{\sigma^2 \lambda^2}{2} + o(\lambda^2) \implies \mathbb{E}[X] \leq 0 \ .$$

Similarly, letting $\lambda < 0$ and $\lambda \to 0$, we get $\mathbb{E}[X] \geq 0$, so $\mathbb{E}[X] = 0$. Then, again letting $\lambda \to 0$:

$$\frac{\lambda^2}{2} \mathbb{E}\left[X^2\right] \leq \frac{\sigma^2 \lambda^2}{2} + o(\lambda^2) \implies \mathrm{Var}(X) \leq \mathbb{E}\left[X^2\right] \leq \sigma^2 \ .$$

Proof of Lemma Cont'd

2. *For all $c \in \mathbb{R}$, cX is $c^2\sigma^2$-subgaussian*: $c\lambda \in \mathbb{R}$, so it follows by definition that

$$\mathbb{E}\left[e^{(\lambda c)X}\right] \leq \exp\left(\frac{\lambda^2 c^2 \sigma^2}{2}\right) \quad .$$

3. *$X_1 + X_2$ is $(\sigma_1^2 + \sigma_2^2)$-subgaussian*: by independence,

$$\mathbb{E}\left[e^{\lambda(X_1+X_2)}\right] = \mathbb{E}\left[e^{\lambda X_1}\right]\mathbb{E}\left[e^{\lambda X_2}\right] \leq \exp\left(\frac{\lambda^2\sigma_1^2}{2}\right)\exp\left(\frac{\lambda^2\sigma_2^2}{2}\right) \quad .$$

Concentration Inequality

Theorem

If X is σ^2-subgaussian, then $\mathbb{P}(X \geq \epsilon) \leq \exp\left(-\frac{\epsilon^2}{2\sigma^2}\right)$.

Proof.

By Markov's inequality and subgaussianity:

$$\begin{aligned}
\mathbb{P}(X \geq \epsilon) &= \mathbb{P}\left(e^{\lambda X} \geq e^{\lambda \epsilon}\right) \\
&\leq \frac{\mathbb{E}\left[e^{\lambda X}\right]}{e^{\lambda \epsilon}} \\
&\leq \exp\left(\frac{\lambda^2 \sigma^2}{2} - \lambda \epsilon\right)
\end{aligned}$$

Since this relationship holds for all $\lambda \in \mathbb{R}$, we minimize the above w.r.t. λ and get $\lambda = \frac{\epsilon}{\sigma^2}$ and the result follows. $\square$

Concentration Inequality

Corollary (Hoeffding's Bound)

Let $X_1, \ldots, X_n$ be independent random variables with $\mathbb{E}[X_i] = \mu$. If $X_i - \mu$ are σ^2-subgaussian, then their sample mean $\hat{\mu}$ satisfies the following:

$$\mathbb{P}(\hat{\mu} - \mu \geq \epsilon) \leq \exp\left(-\frac{n\epsilon^2}{2\sigma^2}\right); \quad \mathbb{P}(\hat{\mu} - \mu \leq -\epsilon) \leq \exp\left(-\frac{n\epsilon^2}{2\sigma^2}\right) \quad .$$

Proof.

This is a direct consquence of the fact that $\hat{\mu} - \mu$ is $\frac{\sigma^2}{n}$-subgaussian. □

Why use this concentration inequality?

- We are interested in how far the sample mean is from the real mean (to estimate the average reward of each action), namely the tail probabilities.
- Chebyshev's inequality is too loose.
- CLT concerns the asymptotic behaviour of the estimate, and so it is not suitable for studying regret when we are limited to a finite amount of actions.

The Theorem

Theorem

Assume that the noise of the reward of each arm in a k-armed stochastic bandit problem is 1-subgaussian. Then, after $n \geq mk$ rounds, the expected regret J_n of ETC which explores each arm exactly m times before committing is bounded as follows:

$$J_n \leq m \sum_{a=1}^{k} \Delta_a + (n - mk) \sum_{a=1}^{k} \Delta_a \exp\left(-\frac{m\Delta_a^2}{4}\right) \tag{5}$$

Proof

First let us decompose the regret:

$$
\begin{aligned}
J_n &= \sum_{t=1}^{n} \mathbb{E}\left[\Delta_{A_t}\right] \\
&= m \sum_{a=1}^{k} \Delta_a + (n - mk) \sum_{a=1}^{k} \Delta_a \mathbb{P}\left(a = \arg\max_b Q_{mk}(b)\right) \quad .
\end{aligned}
$$

WLOG, let us assume that the optimal action is 1, namely $1 = \arg\max_a \max_b q_*(b) - q_*(a)$. Now,

$$
\begin{aligned}
\mathbb{P}\left(a = \arg\max_b Q_{mk}(b)\right) &\leq \mathbb{P}\left(Q_{mk}(a) - Q_{mk}(1) \geq 0\right) \\
&= \mathbb{P}\left(Q_{mk}(a) - q_*(a) - Q_{mk}(1) + q_*(1) \geq \Delta_a\right) \\
&\leq \exp\left(-\frac{m\Delta_a^2}{4}\right) \quad ,
\end{aligned}
$$

as $Q_{mk}(a) - q_*(a) - Q_{mk}(1) + q_*(1)$ is $\frac{2}{m}$-subgaussian and by the previous theorem.

Choosing m

$$J_n \leq m \sum_{a=1}^{k} \Delta_a + (n - mk) \sum_{a=1}^{k} \Delta_a \exp\left(-\frac{m\Delta_a^2}{4}\right)$$

- Large m: larger first term.
- Small m: larger second term, as the probability of committing to the wrong arm gets bigger.

Choosing m

Suppose $k = 2$, and 1 is the optimal arm. Then $\Delta_1 = 0$, and we let $\Delta_2 = \Delta$.

$$J_n \leq m\Delta + (n - 2m)\Delta \exp\left(-\frac{m\Delta^2}{4}\right) \leq m\Delta + n\Delta \exp\left(-\frac{m\Delta^2}{4}\right)$$

Minimizing the RHS wrt m, for n sufficiently large, we get

$$m = \left\lceil \frac{4}{\Delta^2} \ln\left(\frac{n\Delta^2}{4}\right) \right\rceil$$

and

$$J_n \leq \Delta + \frac{4}{\Delta}\left(1 + \ln\left(\frac{n\Delta^2}{4}\right)\right) \quad .$$

Choosing m

$$m = \left\lceil \frac{4}{\Delta^2} \ln\left(\frac{n\Delta^2}{4}\right) \right\rceil \; ; \qquad J_n \leq \Delta + \frac{4}{\Delta}\left(1 + \ln\left(\frac{n\Delta^2}{4}\right)\right)$$

- Issue: m depends on Δ (this is almost never known in practice) and n (reasonable only in certain occurences).
- Very small $\Delta \implies$ large J_n, so set

$$J_n \leq \min\left\{ n\Delta, \Delta + \frac{4}{\Delta}\left(1 + \ln\left(\frac{n\Delta^2}{4}\right)\right)\right\}$$

  to take into account the case for small n.