Stat 260/CS 294-102. Learning in Sequential Decision Problems.

Peter Bartlett

1. Multi-armed bandit algorithms.
   - Consistency: optimal per-round reward.
   - Robbins' consistent algorithm:
     vanishing exploration implies consistency.
   - Upper confidence bound (UCB) algorithms
     (and a foray into concentration inequalities).

Stochastic bandit problem.

- k arms.
- Arm j has unknown reward distribution P_{θ_j}, for $\theta_j \in \Theta$.
- Reward: $X_{j,t} \sim P_{\theta_j}$.
- Mean reward: $\mu_j = \mathbb{E} X_{j,1}$.
- Best: $\mu^* = \max_{j^*=1,\ldots,k} \mu_{j^*}$.
- Gap: $\Delta_j = \mu^* - \mu_j$.
- Number of plays: $T_j(s) = \sum_{t=1}^{s} 1[I_t = j]$.
- Pseudo-regret:
  $\overline{R}_n = n \max_{j^*=1,\ldots,k} \mu_{j^*} - \mathbb{E} \sum_{t=1}^{n} X_{I_t,t} = \sum_{j=1}^{k} \mathbb{E} T_j(n) \Delta_j$.

Consistency.

Call a strategy *consistent* if

$$\frac{\overline{R}_n}{n} \to 0.$$

How might we achieve consistency?

- Explore for a while, then exploit?

  But with positive probability, exploration will mislead us.

  $\Rightarrow$ Must explore forever.

Robbin's strategy.

Fix disjoint *exploration sequences*

$$
\begin{aligned}
1 &= e_1^1 < e_2^1 < \cdots < e_n^1 < \cdots \,, \\
2 &= e_1^2 < e_2^2 < \cdots < e_n^2 < \cdots \,, \\
&\vdots \\
k &= e_1^k < e_2^k < \cdots < e_n^k < \cdots \,.
\end{aligned}
$$

At time t, if some j, i has $t = e_i^j$, play $I_t = j$. Otherwise play

$$I_t = \hat{j}_t = \arg\max_j \frac{1}{T_j(t)} \sum_{s=1}^{t} X_{I_s,s} 1[I_s = j].$$

Robbin's strategy.

Since $e_n^j \to \infty$, $T_j(t) \to \infty$, so the strong law of large numbers shows that

$$\hat{\mu}_j(t) := \frac{1}{T_j(t)} \sum_{s=1}^{t} X_{I_s,s} 1[I_s = j] \stackrel{as}{\to} \mu_j,$$

hence $\hat{j}_t \to j^*$.

How often should we explore?

- Explore some fixed proportion of the time?
  But that proportion will always cost us.
  $\Rightarrow$ Must explore forever, but a vanishing fraction of the time.

Robbin's strategy.

Vanishing exploration implies consistency:

Theorem: If the *exploration set* up to time n,

$$E_n := \{t \le n : \text{ some } j, i \text{ has } t = e_i^j\},$$

satisfies $|E_n|/n \to 0$, then

$$\frac{\overline{R}_n}{n} = \sum_{j \neq j^*} \frac{\mathbb{E}T_j(n)}{n} \Delta_j \to 0.$$

Robbin's strategy.

Proof. With vanishing exploration, if $j \neq j^*$,

$$\frac{T_j(n)}{n} = \frac{1}{n}\sum_{t=1}^{n}\left(1[\exists i \text{ s.t. } t = x_i^j] + 1[t \notin E_t,\, \hat{j}_t = j]\right)$$

$$\leq \frac{|E_n|}{n} + \frac{1}{n}\sum_{t=1}^{n} 1[\hat{j}_t = j]$$

$$\stackrel{as}{\to} 0.$$

□

UCB strategy.

Upper Confidence Bounds:
Use data to define an upper bound on μ_j.
Choose the arm with the largest upper bound.

- Optimism in the face of uncertainty.
- Nicely balances exploration (few pulls $\Rightarrow$ loose upper bound $\Rightarrow$ more likely to try it) and exploitation (when confidence intervals are small, the best arm has the best upper bound).

UCB strategy.

- We want tight upper bounds (or we waste our time on a bad arm), but
- We don't want the bounds too tight (or we might miss a good arm).
- We shouldn't leave an arm untried for too long (since if we are misled to wrongfully neglect an arm with a very small probability, that becomes important again after a long period of neglect).

We'll consider estimates based on sample averages, $\hat{\mu}_j(t)$, and concentration inequalities in terms of *cumulant generating functions*. So we'll have a brief digression to look at concentration inequalities...

Concentration inequalities.

Definition: For a random variable X with mean μ, the moment-generating function is

$$M_{X-\mu}(\lambda) = \mathbb{E}\exp(\lambda(X - \mathbb{E}X)),$$

the cumulant-generating function is

$$\Gamma_{X-\mu}(\lambda) = \log M_{X-\mu}(\lambda).$$

Concentration inequalities.

Definition: For a random variable X, $\psi : \mathbb{R} \to \mathbb{R}$ is a *cumulant generating function upper bound* if, for $\lambda > 0$,

$$\psi(\lambda) \geq \max \left\{ \Gamma_X(\lambda), \Gamma_{-X}(\lambda) \right\},$$
$$\psi(-\lambda) = \psi(\lambda).$$

The *Legendre transform (convex conjugate)* of ψ is

$$\psi^*(\epsilon) = \sup_{\lambda \in \mathbb{R}} \left(\lambda\epsilon - \psi(\lambda) \right).$$

Concentration Inequalities.

Theorem:

$$\Gamma_{X+c}(\lambda) = \lambda c + \Gamma_X(\lambda),$$
$$\Gamma^*_{X+c}(\epsilon) = \Gamma^*_X(\epsilon - c).$$

(Easy to check.)

Concentration Inequalities.

Theorem: For $\epsilon \geq 0$, $\mathbb{P}\left(X - \mathbb{E}X \geq \epsilon\right) \leq \exp\left(-\psi^*_{X-\mathbb{E}X}(\epsilon)\right)$.

Concentration Inequality: Proof.

$$
\begin{aligned}
&\log \mathbb{P}\left(X - \mathbb{E}X \geq \epsilon\right) \\
&= \inf_{\lambda > 0} \log \mathbb{P}\left(\exp\left(\lambda\left(X - \mathbb{E}X - \epsilon\right)\right) \geq 1\right) && \text{(exp is monotonic)} \\
&\leq \inf_{\lambda > 0} \log \mathbb{E} \exp\left(\lambda\left(X - \mathbb{E}X - \epsilon\right)\right) && \text{(Markov's inequality)} \\
&\leq \inf_{\lambda > 0} \left(\psi_{X - \mathbb{E}X}(\lambda) - \lambda\epsilon\right) && \text{(cgf bound)} \\
&= \inf_{\lambda \in \mathbb{R}} \left(\psi_{X - \mathbb{E}X}(\lambda) - \lambda\epsilon\right) && \text{(from } \epsilon > 0\text{, definition of } \psi(-\lambda)\text{)} \\
&= -\psi^*_{X - \mathbb{E}X}(\epsilon).
\end{aligned}
$$

Concentration Inequalities.

Theorem: If $X_1, X_2, \ldots, X_n$ are mean zero, i.i.d. with cgf upper bound ψ, then $\bar{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$ has cgf bound

$$\psi_{\bar{X}_n}(\lambda) = n\psi\left(\frac{\lambda}{n}\right),$$

and
$$\psi^*_{\bar{X}_n}(\epsilon) = n\psi^*(\epsilon),$$

hence,
$$\mathbb{P}\left(\bar{X}_n \geq \epsilon\right) \leq \exp\left(-n\psi^*(\epsilon)\right),$$

(Easy to check.)

Example: Gaussian

For $X \sim N(\mu, \sigma^2)$,

$$\Gamma_{X-\mu}(\lambda) = \frac{\lambda^2\sigma^2}{2}, \qquad \Gamma^*_{X-\mu}(\epsilon) = \frac{\epsilon^2}{2\sigma^2}.$$

For $X_1, \ldots, X_n \sim N(\mu, \sigma^2)$, it's easy to check that the bound is tight:

$$\lim_{n\to\infty} \frac{1}{n} \ln P(\bar{X}_n - \mu \geq \epsilon) = -\frac{\epsilon^2}{2\sigma^2}.$$

Example: Bounded Support

Theorem: [Hoeffding's Inequality] For a random variable $X \in [a, b]$ with $\mathbb{E}X = \mu$ and $\lambda \in \mathbb{R}$,

$$\ln M_{X-\mu}(\lambda) \leq \frac{\lambda^2 (b-a)^2}{8}.$$

Note the resemblance to a Gaussian:

$$\frac{\lambda^2 \sigma^2}{2} \text{ vs } \frac{\lambda^2 (b-a)^2}{8}.$$

(And since P has support in $[a, b]$, $\mathrm{Var} X \leq (b-a)^2/4$.)

Example: Hoeffding's Inequality Proof

Define

$$A(\lambda) = \log\left(\mathbb{E}e^{\lambda X}\right) = \log\left(\int e^{\lambda x}\, dP(x)\right),$$

where $X \sim P$. Then A is the log normalization of the exponential family random variable X_λ with reference measure P and sufficient statistic x. Since P has bounded support, $A(\lambda) < \infty$ for all λ, and we know that

$$A'(\lambda) = \mathbb{E}(X_\lambda), \qquad\qquad A''(\lambda) = \mathrm{Var}(X_\lambda).$$

Since P has support in $[a, b]$, $\mathrm{Var}(X_\lambda) \leq (b-a)^2/4$. Then a Taylor expansion about $\lambda = 0$ (at this value of λ, X_λ has the same distribution as X, hence the same expectation) gives

$$A(\lambda) \leq \lambda \mathbb{E}X + \frac{\lambda^2}{2}\frac{(b-a)^2}{4}.$$

Sub-Gaussian Random Variables

Definition: X is **sub-Gaussian** with parameter σ^2 if, for all $\lambda \in \mathbb{R}$,

$$\ln M_{X-\mu}(\lambda) \leq \frac{\lambda^2 \sigma^2}{2}.$$

Note: Gaussian is sub-Gaussian. X sub-Gaussian iff $-X$ sub-Gaussian. X sub-Gaussian implies $P(X - \mu \geq t) \leq \exp(-t^2/(2\sigma^2))$.

Hoeffding Bound

Theorem: For $X_1, \ldots, X_n$ independent, $\mathbb{E}X_i = \mu$, X_i sub-Gaussian with parameter σ^2, then for all $t > 0$,

$$P\left(\frac{1}{n}\sum_{i=1}^{n} X_i - \mu \geq t\right) \leq \exp\left(-\frac{nt^2}{2\sigma^2}\right).$$