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

Peter Bartlett

1. Multi-armed bandit algorithms.

   - Concentration inequalities.
     - $\mathbb{P}(X \geq \epsilon) \leq \exp(-\psi^*(\epsilon)))$.
     - Cumulant generating function bounds.
     - Hoeffding's inequality for sub-Gaussian random variables.
   - Upper confidence bound (UCB) algorithms.
     - Compare upper bounds on means
       (based on sample averages and concentration inequalities)
     - Analysis: bound $\mathbb{E}T_j(n)$.
       Compare gap Δ_j to confidence interval width.

Recall: Concentration inequalities.

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

$$M_X(\lambda) = \mathbb{E}\exp(\lambda X),$$

the cumulant-generating function is

$$\Gamma_X(\lambda) = \log M_X(\lambda).$$

Recall: 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.)

Recall: Concentration Inequalities.

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

Recall: 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,

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

Recall: Concentration Inequalities.

And the exponent can't be improved.

Theorem: [Cramér-Chernoff] If $X_1, X_2, \ldots, X_n$ are iid and mean zero, and have cgf Γ, then for $\epsilon > 0$ and $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$,

$$\lim_{n\to\infty} \frac{1}{n} \log \mathbb{P}\left(\bar{X}_n \geq \epsilon\right) = -\Gamma^*(\epsilon).$$

(Lower bound is a change-of-measure argument plus central limit theorem.)

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:

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

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 $[a, b]$, $\text{Var}X \leq (b-a)^2/4$.)

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

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

Since P has support in $[a, b]$, $\text{Var}(X_\lambda) \leq (b-a)^2/4$. 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 $t > 0$,

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

Back to the 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:
  $\overline{R}_n = n \max_{j^*=1,\ldots,k} \mu_{j^*} - \mathbb{E} \sum_{t=1}^{n} X_{I_t,t} = \sum_{j=1}^{k} \mathbb{E}T_j(n)\Delta_j$.

UCB strategy

Define the sample averages

$$\hat{\mu}_j(t) = \frac{1}{T_j(t)} \sum_{s=1}^{t} X_{I_s,s} 1[I_s = j].$$

If $X_{j,s} - \mu_j$ has c.g.f. upper bound ψ,

$$\Pr\left(\frac{1}{n} \sum_{s=1}^{n} X_{j,s} \leq \mu_j - \epsilon \right) \leq e^{-n\psi^*(\epsilon)},$$

that is,

$$\Pr\left(\mu_j < \frac{1}{n} \sum_{s=1}^{n} X_{j,s} + (\psi^*)^{-1} \left(\frac{\log 1/\delta}{n} \right) \right) \geq 1 - \delta.$$

UCB strategy.

Suppose $X_{j,t} - \mu_j$ has c.g.f. upper bound ψ.

ψ-UCB Strategy:

$$I_t = \arg\max_{1 \le j \le k} \left(\hat{\mu}_{j,t-1} + (\psi^*)^{-1} \left(\frac{3 \log t}{T_j(t-1)} \right) \right).$$

e.g., $X_{j,t}$ sub-gaussian (with parameter σ^2),

$$\psi^*(\epsilon) = \frac{\epsilon^2}{2\sigma^2},$$

$$(\psi^*)^{-1} \left(\frac{3 \log t}{T_j(t-1)} \right) = \sqrt{\frac{6\sigma^2 \log t}{T_j(t-1)}}.$$

UCB strategy.

Theorem: If the reward distributions have cgf bound ψ, then the ψ-UCB Strategy satisfies

$$\overline{R}_n \leq \sum_{j:\Delta_j>0} \left(\frac{3\Delta_j \log n}{\psi^*(\Delta_j/2)} + o(1) \right).$$

Example: Rewards in $[0,1]$ are sub-Gaussian with $\sigma^2 = 1/4$, so

$$\overline{R}_n \leq \sum_{j:\Delta_j>0} \left(\frac{6 \log n}{\Delta_j} + o(1) \right).$$

This is within a constant factor of optimal ($\mu^*(1-\mu^*)$ versus 6).

UCB strategy: Proof

(Drop t indices.) Define $\epsilon_j = (\psi^*)^{-1}\left(\frac{3\log t}{T_j(t-1)}\right)$. So UCB chooses $\arg\max_j(\hat{\mu}_j + \epsilon_j)$. If

$$\begin{aligned}\hat{\mu}_{j^*} &> \mu_{j^*} - \epsilon_{j^*},\\ \hat{\mu}_j &< \mu_j + \epsilon_j,\\ \Delta_j &> 2\epsilon_j,\end{aligned}$$

then

$$\begin{aligned}\hat{\mu}_{j^*} + \epsilon_{j^*} &> \mu_{j^*}\\ &= \mu_j + \Delta_j\\ &> \hat{\mu}_j - \epsilon_j + \Delta_j\\ &> \hat{\mu}_j + \epsilon_j,\end{aligned}$$

so UCB will not choose $I_t = j$.

UCB strategy: Proof

So

$$\{I_t = j \text{ and } \Delta_j > 2\epsilon_j\} \subseteq \{\hat{\mu}_{j^*} \leq \mu_{j^*} - \epsilon_{j^*} \text{ or } \hat{\mu}_j \geq \mu_j + \epsilon_j\}\,.$$

Note that $\Delta_j > 2\epsilon_j$ for $T_j(t-1) \geq m := \frac{3\log n}{\psi^*(\Delta_j/2)} \geq \frac{3\log t}{\psi^*(\Delta_j/2)}$.

$$\begin{aligned}
\mathbb{E}T_j(n) &= \mathbb{E}\sum_{t=1}^{n} 1[I_t = j] \\
&\leq m + \mathbb{E}\sum_{t=m+1}^{n} 1[I_t = j \text{ and } T_j(t-1) \geq m] \\
&\leq m + \sum_{t=m+1}^{n} \left(\mathbb{P}(\hat{\mu}_{j^*} \leq \mu_{j^*} - \epsilon_{j^*}) + \mathbb{P}(\hat{\mu}_j \geq \mu_j + \epsilon_j)\right).
\end{aligned}$$

UCB strategy: Proof

$$\mathbb{P}(\hat{\mu}_{j^*} \leq \mu_{j^*} - \epsilon_{j^*})l \leq \mathbb{P}\left(\exists s \in \{1,\ldots,t\}\hat{\mu}_{j^*,s} + (\psi^*)^{-1}(3\ln t/s) \leq \mu_{j^*}\right)$$

$$\leq \sum_{s=1}^{t} \mathbb{P}\left(\hat{\mu}_{j^*,s} + (\psi^*)^{-1}(3\ln t/s) \leq \mu_{j^*}\right)$$

$$\leq \sum_{s=1}^{t} t^{-3} = t^{-2}.$$

Similarly for $\mathbb{P}(\hat{\mu}_j \geq \mu_j + \epsilon_j)$. So

$$\mathbb{E}T_j(n) \leq \frac{3\log n}{\psi^*(\Delta_j/2)} + O\left(\frac{1}{\log n}\right).$$

UCB strategy.

Theorem: If the reward distributions have cgf bound ψ, then the ψ-UCB Strategy satisfies

$$\overline{R}_n \leq \sum_{j:\Delta_j>0} \left(\frac{3\Delta_j \log n}{\psi^*(\Delta_j/2)} + o(1) \right).$$

Hence,

$$\overline{R}_n \leq \sum_{j:\Delta_j>0} \left(\frac{6 \log n}{\Delta_j} + o(1) \right),$$

for rewards in $[0, 1]$.

UCB strategy.

- This bound is on *expected* reward. Looking at the proof, the probability of bad behavior decays only polynomially with n (c.f. Hoeffding's inequality). This slow decay is not just an artefact of the analysis.
- For rewards in $[0, 1]$, the upper bound is

$$\overline{R}_n \leq \sum_{j:\Delta_j>0} \left(\frac{6 \log n}{\Delta_j} + o(1) \right).$$

  However, the lower bound is smaller: the leading term is of the form

$$\frac{\Delta_j \log n}{D_{KL}(P_{\theta_j}, P_{\theta^*})}.$$

  This is achievable, by considering a better c.g.f. bound ψ: the Bernoulli c.g.f.

UCB strategy: Some history.

- Index strategies (which involve a comparison of the value of an index for each arm, which depends only on observations at that arm) were introduced by Gittins and collaborators (in a Bayesian context).
- The idea of UCB strategies, where the index has the interpretation of an upper confidence bound, goes back to Lai and Robbins. They gave strategies and analysis where the leading term is asymptotically optimal (including the constant).
- Agrawal (1995) considered simpler strategies based only on the sample mean, and used c.g.f. bounds and large deviations theory to give asymptotic results.
- Auer, Cesa-Bianchi and Fischer (2002) used Hoeffding's inequality to give a finite time analysis.