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

Peter Bartlett

1. Contextual bandits.
   - Bandits with side information.
   - Model assumptions versus comparison class.
   - Woodroofe/Sarkar one-armed bandit with side information.
   - $|\mathcal{X}|$ distinct bandit problems.
   - Bandits with expert advice.

Contextual bandits

Bandits with side information.

Hope that the extra information will allow better decisions.

At each round:

- See $X_t \in \mathcal{X}$.
- Choose $I_t \in \{1, \ldots, k\}$.
- Receive reward $Y_{I_t,t} \in \mathbb{R}$.

Contextual bandits

1. Stochastic model:
   joint distribution $(X, Y) \sim P_\theta$, $X \in \mathcal{X}$, $Y \in \mathbb{R}^k$.
2. Game theoretic model:
   (X, Y) pairs chosen adversarially.
3. Mixture:
   e.g., X fixed design/adversarial, $Y \sim P_{\theta,X}$.

Regret and pseudo-regret:

$$R_n = \sup_f \sum_{t=1}^n Y_{f(X_t),t} - \sum_{t=1}^n Y_{I_t,t}.$$

$$\overline{R}_n = \sup_f \mathbb{E} \sum_{t=1}^n Y_{f(X_t),t} - \mathbb{E} \sum_{t=1}^n Y_{I_t,t}.$$

where $\sup$ is over *comparison class* F of functions $f : \mathcal{X} \to \{1, \ldots, k\}$.

Contextual bandits

1. If the comparison class F is all constant mappings

$$\{f_j : \mathcal{X} \to \{1, \ldots, k\} \text{ s.t. } j \in \{1, \ldots, k\}, \forall x \in \mathcal{X},\ f_j(x) = j\},$$

   then this is no harder than the multi-armed bandit problem (depending on $\{P_\theta\}$, it might be much easier, because of the extra information).

2. If F is all functions $f : \mathcal{X} \to \{1, \ldots, k\}$, then the aim is to predict, for each x, the maximizer j^* of $j \mapsto \mathbb{E}[Y_j | X = x]$.

   (a) If $\mathcal{X} = \{1, 2, \ldots, m\}$, then we can view it as m separate k-armed bandit problems. The X_t tells the strategy which bandit it is playing. And if $\{P_\theta\}$ is such that the distribution $Y|X$ gives no information about $Y|X'$, the k-armed bandit problems decouple in this way.

Contextual bandits

(b) If $\mathcal{X}$ is infinite (and F is all measurable functions $f : \mathcal{X} \to \{1, \ldots, k\}$), it may be more appropriate to view it as a pattern classification problem, but with limited information about the labels. (In the pattern classification setting, $f^*(x) = j^*$ is called the Bayes decision rule.)

3. If F is a family of prediction rules (such as linear threshold functions, or decision trees), then the aim is to accumulate almost as much reward as the best of these prediction rules.

4. We can also allow F to be a family of *randomized* functions, that is, functions that map from $\mathcal{X}$ to Δ_k, the set of probability distributions over the k arms. (In that case, we can interpret $Y_{f(X_t),t}$ as a random variable, and we're typically interested in maximizing the expectation of the sum of these.)

Contextual bandits

Two broad approaches:

1. Impose strong constraints on $\{P_\theta\}$, and aim for optimality (that is, use an unrestricted comparison class).
   (Woodroofe, 1979), (Sarkar, 1991), (Wang, Kulkarni, Poor, 2005), (Abe and Long, 1999), (Auer, 2002), (Li et al, 2010).

2. Impose few constraints on $\{P_\theta\}$, but strong constraints on the comparison policies.
   (Auer et al, 2002), (Dudik et al, 2011), (Agarwal et al, 2014).

Contextual bandits

(Woodroofe, 1979): Aim to maximize expected total discounted reward,

$$\mathbb{E}\sum_{t=1}^{\infty}\gamma^t Y_{f(X_t),t},$$

in a Bayesian setting.

Considered a one-armed bandit

- $\mathbb{E}[Y_0|X]$ is known,
- $(X, Y_1 - \mathbb{E}[Y_0|X]) \sim P_\theta$ with $\theta \sim \pi$ (prior π),

with a simple model P_θ.

Contextual bandits

(Sarkar, 1991) extended to the one-parameter exponential family model: $Y_1 - \mathbb{E}[Y_0|X]$ has density

$$f(y|x,\theta) = \exp\left(\theta' T(x,y) - A(x,\theta)\right).$$

Then (under suitable conditions on the distribution of X), the greedy policy,

$$I_t = \arg\max_{j\in\{1,\ldots,k\}} \mathbb{E}\left[Y_{j,t}\,|X_t,\ \text{history to } t-1\right],$$

is optimal asymptotically (as $\gamma \to 1$).

Resolves an ethical dilemma?

> Thus, the myopic procedure... fulfills the utilitarian goal as well as the individualistic one.

$|\mathcal{X}|$ distinct bandit problems

Suppose that $\mathcal{X} = \{1, 2, \ldots, m\}$, and we wish to compete with

$$F^{(m)} = \{f : \mathcal{X} \to \{1, \ldots, k\}\}.$$

We can think of each value of X as an index indicating which of the m bandit problems the strategy must play.

Weak constraints on data; comparison class is constrained because $\mathcal{X}$ is small.

$|\mathcal{X}|$ distinct bandit problems

There is a simple approach:

run m distinct multi-armed bandit strategies. For instance, for the EXP3 forecaster:

Theorem: Using EXP3 for each of the m bandits gives pseudo-regret

$$\overline{R}_n = \sup_{f \in F^{(m)}} \mathbb{E} \sum_{t=1}^{n} Y_{f(X_t),t} - \mathbb{E} \sum_{t=1}^{n} Y_{I_t,t} \leq 2\sqrt{nmk \log k}.$$

$|\mathcal{X}|$ distinct bandit problems

Proof. Recall:

Theorem: Exp3 with parameter $\eta = \sqrt{2 \log k/(nk)}$ incurs regret

$$\overline{R}_n \leq \sqrt{2nk \log k}.$$

Exp3 with parameter $\eta_t = \sqrt{\log k/(tk)}$ incurs regret

$$\overline{R}_n \leq 2\sqrt{nk \log k}.$$

Define the number of rounds of each of the m separate bandit problems,

$$n_i = \sum_{t=1}^{n} 1[X_t = i].$$

$|\mathcal{X}|$ distinct bandit problems

We have

$$\begin{aligned}
\overline{R}_n &= \sup_{f \in F^{(m)}} \mathbb{E} \sum_{t=1}^{n} \left(Y_{f(X_t),t} - Y_{I_t,t} \right) \\
&= \sup_{f \in F^{(m)}} \sum_{i=1}^{m} \mathbb{E} \sum_{t:X_t=i} \left(Y_{f(i),t} - Y_{I_t,t} \right) \\
&= \sum_{i=1}^{m} \max_{f(i) \in \{1,\ldots,k\}} \mathbb{E} \sum_{t:X_t=i} \left(Y_{f(i),t} - Y_{I_t,t} \right) \\
&\leq \sum_{i=1}^{m} 2\sqrt{n_i k \log k} \\
&\leq 2\sqrt{k \log k}\sqrt{m}\sqrt{n}. \qquad \text{(Cauchy-Schwarz)}
\end{aligned}$$

$|\mathcal{X}|$ distinct bandit problems

Theorem: For any strategy and any n, there is an oblivious adversary playing i.i.d. (product of uniform and Bernoullis)

$$(X_t, Y_{1,t}, \ldots, Y_{k,t}) \in \{1, \ldots, m\} \times \{0, 1\}^k$$

for which

$$\overline{R}_n = \Omega\left(\sqrt{nmk}\right).$$

$|\mathcal{X}|$ distinct bandit problems

Proof. Recall

Theorem: For any strategy and any n, there is an oblivious adversary playing Bernoulli rewards $y_{j,t} \in \{0, 1\}$ for which

$$\overline{R}_n \geq \frac{1}{18} \min\{\sqrt{nk}, n\}.$$

Under a uniform choice of X_t, $\mathbb{E}\,|\{i : n_i \geq n/(2m)\}|$ is $\Omega(m)$, provided $n = \Omega(m)$. For each of these $\Omega(m)$ (decoupled) games, we incur regret $\Omega(\sqrt{nk/m})$. In particular, for $n \geq 2m$, $\Pr\left(n_i \geq n/(2m)\right) \geq 1/2$, so:

$|\mathcal{X}|$ distinct bandit problems

$$
\begin{aligned}
\overline{R}_n &= \sup_{f \in F^{(m)}} \mathbb{E} \sum_{t=1}^{n} \left(Y_{f(X_t),t} - Y_{I_t,t} \right) \\
&= \sup_{f \in F^{(m)}} \sum_{i=1}^{m} \mathbb{E} \sum_{t:X_t=i} \left(Y_{f(i),t} - Y_{I_t,t} \right) \\
&= \sum_{i=1}^{m} \max_{f(i) \in \{1,\ldots,k\}} \mathbb{E} \sum_{t:X_t=i} \left(Y_{f(i),t} - Y_{I_t,t} \right) \\
&\geq \sum_{i=1}^{m} \mathbb{E} \left[1 \left[n_i \geq \frac{n}{2m} \right] \max_{f(i) \in \{1,\ldots,k\}} \mathbb{E} \sum_{t:X_t=i} \left(Y_{f(i),t} - Y_{I_t,t} \right) \right] \\
&\geq m \Pr \left(n_1 \geq \frac{n}{2m} \right) \Omega \left(\sqrt{kn/m} \right) \\
&= \Omega \left(\sqrt{mnk} \right) .
\end{aligned}
$$

Bandits with expert advice

Consider another setting where we impose only weak constraints on the process generating the data (we allow it to be adversarial), but constrain the comparison class by making it small, say cardinality N. We can even ignore the context X_t, and rely only on the 'expert advice' provided by functions in the comparison class. This is the setting of *bandits with expert advice*. We'll allow the comparison class to be distributions over the k arms. And we'll allow the choice of advice to be adversarial.

Bandits with expert advice

Repeated game:

1. Adversary chooses rewards $(y_{1,t}, \ldots, y_{k,t})$.
2. Adversary presents expert advice $\xi_t^1, \ldots, \xi_t^N \in \Delta_k$.
3. Strategy chooses the distribution of I_t.
4. Strategy receives reward $y_{I_t,t}$.

Aim to minimize psuedo-regret,

$$\overline{R}_n = \max_i \mathbb{E}\left[\sum_{t=1}^n \mathbb{E}_{J\sim\xi_t^i} y_{J,t} - \sum_{t=1}^n y_{I_t,t}\right].$$

Bandits with expert advice

We could treat each expert as an arm and use Exp3. This would give a distribution over experts in each round, and we could play the induced distribution over arms. By treating it as a random choice of expert, we can view the loss we observe as the loss of the chosen expert, and get a regret bound of $O\left(\sqrt{nN \log N}\right)$.

If k is large compared to N (many arms, few experts), this is a reasonable approach. But if not, a better regret is possible: $O\left(\sqrt{nk \log N}\right)$. Thus, it is possible to compete with a much larger family of experts.

Bandits with expert advice

The strategy corresponds to using Exp3 over experts, but computing the estimates of the experts losses from the (known) expert distributions. Computing expectations under the ξ_t^j, rather than using an unbiased estimate, saves some variance: The $\sqrt{k}$ term comes from the bound on the second moment of the estimated losses (it would be a $\sqrt{N}$ if we used the estimate), whereas the $\log N$ term comes from the initial value of the potential function.

Bandits with expert advice

Recall:

Strategy Exp3

set p_1 uniform on $\{1, \ldots, k\}$.

for $t = 1, 2, \ldots, n$, choose $I_t \sim p_t$, observe $\ell_{I_t,t}$.

$$\tilde{\ell}_{i,t} = \frac{\ell_{i,t}}{p_{i,t}} 1[I_t = i],$$

$$\tilde{L}_{i,t} = \sum_{s=1}^{t} \tilde{\ell}_{i,s},$$

$$p_{i,t+1} = \frac{\exp\left(-\eta \tilde{L}_{i,t}\right)}{\sum_{j=1}^{k} \exp\left(-\eta \tilde{L}_{j,t}\right)}.$$

Bandits with expert advice

Strategy Exp4

set q_1 uniform on $\{1, \ldots, N\}$.

for $t = 1, 2, \ldots, n$, observe $\xi_t^1, \ldots, \xi_t^N \in \Delta_k$;

choose $I_t \sim p_t$, where $p_{i,t} = \mathbb{E}_{J \sim q_t} \xi_{i,t}^J$; observe $\ell_{I_t,t}$.

$$\tilde{\ell}_{i,t} = \frac{\ell_{i,t}}{p_{i,t}} 1[I_t = i], \qquad \tilde{y}_{j,t} = \mathbb{E}_{I \sim \xi_t^j} \tilde{\ell}_{I,t},$$

$$\tilde{Y}_{j,t} = \sum_{s=1}^{t} \tilde{y}_{j,t}, \qquad q_{j,t+1} = \frac{\exp\left(-\eta \tilde{Y}_{j,t}\right)}{\sum_{i=1}^{N} \exp\left(-\eta \tilde{Y}_{i,t}\right)}.$$

Bandits with expert advice

Theorem: Exp4 with parameter η incurs regret

$$\overline{R}_n \leq \frac{n\eta k}{2} + \frac{\log N}{\eta}.$$

Choosing $\eta = \sqrt{2\log N/(nk)}$ gives $\overline{R}_n \leq \sqrt{2nk\log N}$.
(And choosing $\eta_t = \sqrt{\log N/(tk)}$ gives $\overline{R}_n \leq 2\sqrt{nk\log N}$.)

Bandits with expert advice

Proof. The regret is

$$\overline{R}_n = \min_j \mathbb{E} \sum_{t=1}^{n} \left(\ell_{I_t,t} - \mathbb{E}_{I \sim \xi_t^j} \ell_{I,t} \right).$$

We have

$$\ell_{I_t,t} = \mathbb{E}_{I \sim p_t} \tilde{\ell}_{I,t} = \mathbb{E}_{J \sim q_t} \mathbb{E}_{I \sim \xi_t^J} \tilde{\ell}_{I,t} = \mathbb{E}_{J \sim q_t} \tilde{y}_{J,t}.$$

$$y_{j,t} := \mathbb{E}_{I \sim \xi_t^j} \ell_{I,t} = \mathbb{E}_{I_t \sim p_t} \mathbb{E}_{I \sim \xi_t^j} \tilde{\ell}_{I,t} = \mathbb{E}_{I_t \sim p_t} \tilde{y}_{j,t}.$$

$$\mathbb{E} \sum_{t=1}^{n} y_{j,t} = \mathbb{E} \sum_{t=1}^{n} \mathbb{E}_{I_t \sim p_t} \left[\tilde{y}_{j,t} \middle| I_1, \ldots, I_{t-1} \right] = \mathbb{E} \tilde{Y}_{j,n}.$$

Bandits with expert advice

$$\mathbb{E}_{J\sim q_t}\tilde{y}_{J,t}^2 = \mathbb{E}_{J\sim q_t}\left(\mathbb{E}_{I\sim\xi_t^J}\tilde{\ell}_{I,t}\right)^2$$

$$\leq \mathbb{E}_{J\sim q_t}\mathbb{E}_{I\sim\xi_t^J}\tilde{\ell}_{I,t}^2 = \mathbb{E}_{I\sim p_t}\tilde{\ell}_{I,t}^2 = \frac{\ell_{I_t,t}^2}{p_{I_t,t}}.$$

Hence, as in the argument for Exp3, we exploit the one-sided sub-Gaussian behavior of $\tilde{y} \geq 0$ to show that, for any j,

$$\mathbb{E}\sum_{t=1}^n \ell_{I_t,t} = \mathbb{E}\sum_{t=1}^n \mathbb{E}_{J\sim q_t}\tilde{y}_{J,t}$$

$$\leq \mathbb{E}\sum_{t=1}^n \frac{\eta}{2}\mathbb{E}_{J\sim q_t}\tilde{y}_{J,t}^2 + \frac{\log N}{\eta} + \mathbb{E}\tilde{Y}_{j,n}$$

$$\leq \frac{\eta n k}{2} + \frac{\log N}{\eta} + \mathbb{E}\sum_{t=1}^n y_{j,t}.$$