Boltzmann Exploration Done Right

Nicolò Cesa-Bianchi
Università degli Studi di Milano
Milan, Italy

Claudio Gentile
INRIA Lille – Nord Europe
Villeneuve d'Ascq, France

Gábor Lugosi
ICREA & Universitat Pompeu Fabra
Barcelona, Spain

Gergely Neu
Universitat Pompeu Fabra
Barcelona, Spain

Stochastic multi-armed bandits

Protocol

Repeat for $t = 1, 2, \dots, T$:
- LEARNER plays action $I_t \in \{1, \dots, K\}$.
- ENVIRONMENT generates rewards $X_{t,i} \sim \nu_i$.
- LEARNER gains and observes reward X_{t,I_t}.

Notation:

- mean rewards: $\mu_i = \mathbb{E}[X_{t,i}]$
- best arm: $i^* = \arg\max_{i \in [K]} \mu_i$
- mean reward of best arm: $\mu^* = \max_{i \in [K]} \mu_i$
- suboptimality gaps: $\Delta_i = \mu^* - \mu_i$
- number of draws of arm i until round t: $N_{t,i}$

GOAL: minimize regret

$$R_T = \mu^* T - \sum_{t=1}^{T} \mathbb{E}[X_{t,I_t}] = \sum_{i=1}^{K} \Delta_i \mathbb{E}[N_{T,i}]$$

Assumption: σ^2-subgaussian rewards

$$\mathbb{E}\left[e^{y(X_{t,i} - \mathbb{E}[X_{t,i}])}\right] \le e^{\sigma^2 y^2/2}$$

Lower bound: For Gaussian rewards:

$$R_T \gtrsim \sum_{i \ne i^*} \frac{\sigma^2 \log T}{\Delta_i} \qquad R_T \gtrsim \sigma\sqrt{KT}$$

The elephant in the room

Boltzmann exploration

Initialize: $\hat{\mu}_{1,i} = 0$ for all $i \in [K]$.
Repeat for $t = 1, 2, \dots, T$:
- Compute distribution
$$p_{t,i} \propto e^{\eta_t \hat{\mu}_{t,i}}.$$
- Play action $I_t \sim p_t$ and observe r_{t,I_t}.
- Update empirical means
$$\hat{\mu}_{t,i} = \frac{\sum_{s=1}^{t} X_{s,i} \mathbb{I}\{I_s = i\}}{N_{t,i}}.$$

Broadly used exploration strategy in RL, but very little theory to support it!

Boltzmann exploration done wrong

Main result

For any monotone sequence of learning rates η_t, Boltzmann exploration will suffer suboptimal regret.

Two regimes with no middle ground:
- η_t grows too slowly $\Rightarrow$ too much time on exploration / too slow to zoom in on i^*
- η_t grows too quickly $\Rightarrow$ high probability of missing i^*

Proposition 1: over-exploration

Regret on any 2-armed bandit problem with known means:
- $\eta_t = \frac{\log(t\Delta^2)}{\Delta} \Rightarrow R_T \approx \frac{\log T}{\Delta}.$
- $\eta_t = \frac{\log(t\Delta^2)}{(1+\alpha)\Delta} \Rightarrow R_T \approx T^{\frac{\alpha}{(1+\alpha)}} \cdot \left(\frac{1}{\Delta}\right)^{\frac{1-\alpha}{1+\alpha}}.$

Proof idea:

$$R_T = \sum_{t=1}^{T} \mathbb{P}[I_t = 2] = \sum_{t=1}^{T} \frac{1}{1 + e^{\eta_t \Delta}} \ge \sum_{t=1}^{T} \frac{e^{-\eta_t \Delta}}{2}$$
$$= \frac{1}{2} \sum_{t=1}^{T} \left(t\Delta^2\right)^{1+\alpha}$$
$$\approx \begin{cases} \frac{\log T}{\Delta}, & \text{if } \alpha = 0, \\ T^{\frac{\alpha}{(1+\alpha)}} \cdot \left(\frac{1}{\Delta}\right)^{\frac{1-\alpha}{1+\alpha}}, & \text{if } \alpha > 0. \end{cases}$$

Proposition 2: under-exploration

There exists a 2-armed stochastic bandit problem where BE using any $\eta_t > 2\log t$ has regret $R_T = \Omega(T)$.

Proof idea:
- Two arms: $X_{t,2} = \frac{1}{2}$ and $X_{t,1} \sim \text{Bernoulli}\left(\frac{1}{2} + \Delta\right)$
- Bad event:
$$E_0 = \{\text{arm 1 gives 0 reward in first } t_0 \text{ rounds}\}$$
- Under E_0, BE will not draw arm 1 after round t_0 due to η_t growing too fast
- $\mathbb{P}[E_0] \ge \left(\frac{1}{2} - \Delta\right)^{t_0} = \text{const.}$

A quick fix

Theorem

Let $\tau = \frac{16eK\log T}{\Delta^2}$. Then the regret of BE with the learning rate sequence
$\eta_t = \mathbb{I}\{t < \tau\} + \frac{\log(t\Delta^2)}{\Delta}\mathbb{I}\{t \ge \tau\}$ satisfies

$$R_T \le \frac{16eK\log T}{\Delta^2} + \frac{9K}{\Delta^2}.$$

- near-optimal performance guarantees ☺
- requires prior knowledge of Δ and T ☹☹☹

Boltzmann exploration done right

What's wrong with Boltzmann exploration?

It doesn't reason about uncertainty of reward estimates!

Our solution: arm-dependent learning rates!

Key tool: "Gumbel–softmax trick"

$$I_t = \arg\max_{i \in [K]} \{\hat{\mu}_{t,i} + Z_{t,i}\},$$

follows Boltzmann distribution if $Z_{t,i}$ are i.i.d. standard Gumbel random variables.

Idea: account for uncertainty by scaling $Z_{t,i}$ differently for each arm!

Boltzmann–Gumbel exploration

Initialize: $\hat{\mu}_{1,i} = 0$ for all $i \in [K]$.
Repeat for $t = 1, 2, \dots, T$:
- Draw $Z_{t,i}$ i.i.d. from standard Gumbel distribution.
- Play action
$$I_t = \arg\max_{i \in [K]} \left\{\hat{\mu}_{t,i} + \sqrt{\frac{C^2}{N_{t,i}}} \cdot Z_{t,i}\right\}$$
- Observe r_{t,I_t} and update empirical means
$$\hat{\mu}_{t,i} = \frac{\sum_{s=1}^{t} X_{s,i}\mathbb{I}\{I_s = i\}}{N_{t,i}}$$

Analysis

Theorem

For σ^2-subgaussian rewards, the regret of BGE with $C = \sigma$ satisfies

$$R_T \lesssim \sum_{i \ne i^*} \frac{\sigma^2 \log^2\left(T\Delta_i^2/\sigma^2\right)}{\Delta_i} \qquad R_T \lesssim \sigma\sqrt{KT}\log K$$

Proof sketch:
- Let $\beta_{t,i} = \sqrt{C^2/N_{t,i}}$ and $\tilde{\mu}_{t,i} = \hat{\mu}_{t,i} + \beta_{t,i} Z_{t,i}$
- Set thresholds $x_i = \mu_i + \frac{\Delta}{3}$ and $y_i = \mu_1 - \frac{\Delta}{3}$
- Key events:
  - $E_{t,i}^{\hat{\mu}} = \{\hat{\mu}_{t,i} \le x_i\}$ ~ arm i well-estimated
  - $E_{t,i}^{\tilde{\mu}} = \{\tilde{\mu}_{t,i} \le y_i\}$ ~ small perturbation on arm i
- $\mathbb{E}[N_{t,i}]$ decomposed into 3 terms:
  - $\sum_{t=1}^{T} \mathbb{P}\left[I_t = i, E_{t,i}^{\tilde{\mu}}, E_{t,i}^{\hat{\mu}}\right]$ ~ interaction between perturbations $Z_{t,1}$ and fluctuations of $\hat{\mu}_{t,1}$
  - $\sum_{t=1}^{T} \mathbb{P}\left[I_t = i, \overline{E_{t,i}^{\tilde{\mu}}}, E_{t,i}^{\hat{\mu}}\right]$ ~ large perturbations
  - $\sum_{t=1}^{T} \mathbb{P}\left[I_t = i, \overline{E_{t,i}^{\hat{\mu}}}\right]$ ~ large deviations
- First and last terms bounded by
$$\sum_{k=0}^{T-1} \mathbb{E}\left[\exp\left(\frac{\mu_i - \hat{\mu}_{\tau_k,i}}{\beta_{\tau_k,i}}\right)\right] e^{-\frac{\Delta_i\sqrt{k}}{3C}} \le e^{\sigma^2/2C^2} \cdot \sum_{k=1}^{T-1} e^{-\frac{\Delta_i\sqrt{k}}{3C}} \le \frac{18C^2 e^{\sigma^2/2C^2}}{\Delta_i^2}$$
- Middle term bounded by
$$\frac{9C^2 \log_+^2\left(T\Delta_i^2/c^2\right) + c^2 e^{\gamma}}{\Delta_i^2}$$
for any $c > 0$
- Technique extends to other subgaussian mean estimators for heavy-tailed rewards

Empirical illustration

Sensitivity of BGE and BE with learning rates $\eta_t = C$, $\eta_t = C/\log t$, $\eta_t = C/\sqrt{t}$ to various settings of C: