Gergely Neu

INRIA Lille, SequeL

→ Universitat Pompeu Fabra, Barcelona

Explore no more

Improved high-probability regret bounds for non-stochastic bandits

Non-stochastic bandits

For each round $t = 1,2,\dots,T$
- Learner chooses action/arm $I_t \in \{1,2,\dots,K\}$
- Environment chooses losses $\ell_{t,i} \in [0,1] \quad (\forall i)$
- Learner suffers and observes loss ℓ_{t,I_t}

No assumptions about the environment → we need randomized algorithms

Goal: minimize regret in some probabilistic sense

Pseudo-regret:

$$\hat{R}_T = \mathbf{E}\left[\sum_{t=1}^{T} \ell_{t,I_t}\right] - \min_{i\in[K]} \mathbf{E}\left[\sum_{t=1}^{T} \ell_{t,i}\right]$$

Expected regret:

$$\mathbf{E}[R_T] = \mathbf{E}\left[\sum_{t=1}^{T} \ell_{t,I_t}\right] - \mathbf{E}\left[\min_{i\in[K]} \sum_{t=1}^{T} \ell_{t,i}\right]$$

Regret:

$$R_T = \sum_{t=1}^{T} \ell_{t,I_t} - \min_{i\in[K]} \sum_{t=1}^{T} \ell_{t,i}$$

Classical algorithms

EXP3 (Auer, Cesa-Bianchi, Freund and Schapire, 1995, 2002)

Parameters: $\eta > 0$.

Initialization: For all i, set $w_{1,i} = 1$.

For each round $t = 1,2,\dots,T$
- For all i, let
$$p_{t,i} = \frac{w_{t,i}}{\sum_j w_{t,j}}.$$
- Draw $I_t \sim \boldsymbol{p}_t$.
- For all i, let
$$\hat{\ell}_{t,i} = \frac{\ell_{t,i}}{p_{t,i}} \mathbf{1}_{\{I_t=i\}}.$$
- For all i, update weight as
$$w_{t+1,i} = w_{t,i} e^{-\eta \hat{\ell}_{t,i}}$$

Theorem: when tuned properly, EXP3 guarantees
$$\hat{R}_T \le \sqrt{2KT \log K}.$$

EXP3.P (Auer, Cesa-Bianchi, Freund and Schapire, 2002)

Parameters: $\eta > 0, \gamma \in [0,1], \beta > 0$.

Initialization: For all i, set $w_{1,i} = 1$.

For each round $t = 1,2,\dots,T$
- For all i, let
$$p_{t,i} = (1-\gamma)\frac{w_{t,i}}{\sum_j w_{t,j}} + \frac{\gamma}{K}.$$
- Draw $I_t \sim \boldsymbol{p}_t$.
- For all i, let
$$\hat{r}_{t,i} = \frac{r_{t,i}}{p_{t,i}} \mathbf{1}_{\{I_t=i\}} + \frac{\beta}{p_{t,i}}.$$
- For all i, update weight as
$$w_{t+1,i} = w_{t,i} e^{\eta \hat{r}_{t,i}}$$

Theorem: when tuned properly, EXP3.P guarantees w.p. at least $1-\delta$
$$R_T \le 5.25\sqrt{KT \log(K/\delta)}.$$

Can we...

- remove explicit exploration ($\gamma > 0$)? ✓
- work with losses? ✓
- improve the constants? ✓
- make it actually work well? ✓

The trick: Implicit eXploration (IX)

- Replace the standard loss estimate $\hat{\ell}_{t,i} = \frac{\ell_{t,i}}{p_{t,i}} \mathbf{1}_{\{I_t=i\}}$ in EXP3 by

$$\hat{\ell}_{t,i} = \frac{\ell_{t,i}}{p_{t,i} + \gamma} \mathbf{1}_{\{I_t=i\}}$$

- Our algorithm:

EXP3-IX

Parameters: $\eta > 0, \gamma > 0$.

Initialization: For all i, set $w_{1,i} = 1$.

For each round $t = 1,2,\dots,T$
- For all i, let
$$p_{t,i} = \frac{w_{t,i}}{\sum_j w_{t,j}}.$$
- Draw $I_t \sim \boldsymbol{p}_t$.
- For all i, let
$$\hat{\ell}_{t,i} = \frac{\ell_{t,i}}{p_{t,i} + \gamma} \mathbf{1}_{\{I_t=i\}}.$$
- For all i, update weight as
$$w_{t+1,i} = w_{t,i} e^{-\eta \hat{\ell}_{t,i}}$$

Main results

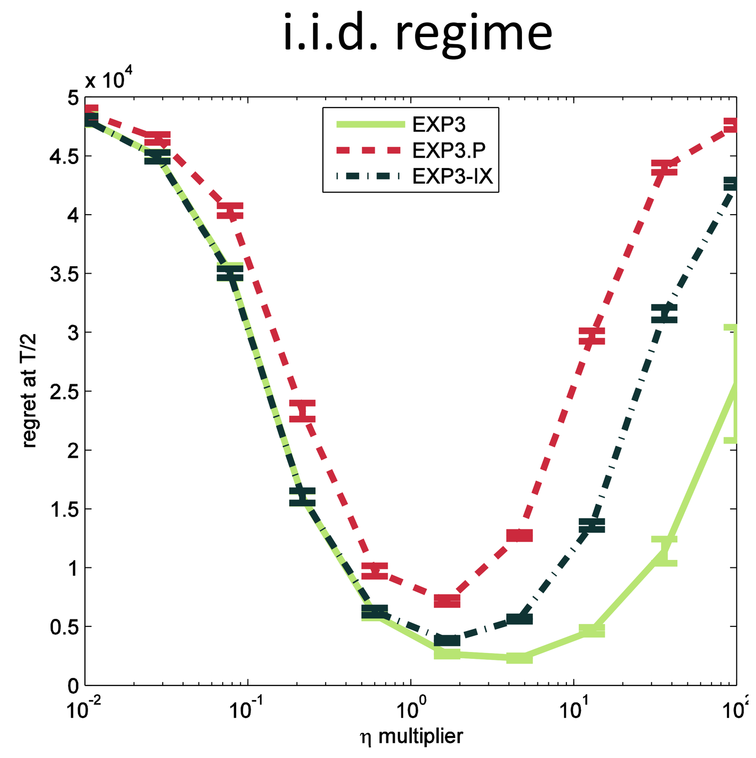
Setting	Best known bound	Our bound
Multi-armed bandits	$5.25\sqrt{KT\log(K/\delta)}$ (Bubeck and Cesa-Bianchi, 2012)	$2\sqrt{2KT\log(K/\delta)}$
Bandits with expert advice (N experts)	$6\sqrt{KT\log(N/\delta)}$ (Beygelzimer et al., 2011)	$2\sqrt{2KT\log(N/\delta)}$
Tracking the best arm (S switches)	$7\sqrt{KTS\log(KT/\delta S)}$ (Audibert and Bubeck, 2010)	$2\sqrt{2KTS\log(KT/\delta S)}$
Bandits with side observations	$\tilde{O}(\sqrt{mT})$ (Alon et al., 2014)	$\tilde{O}(\sqrt{\alpha T})$ $(\alpha \ll m)$

Experiments

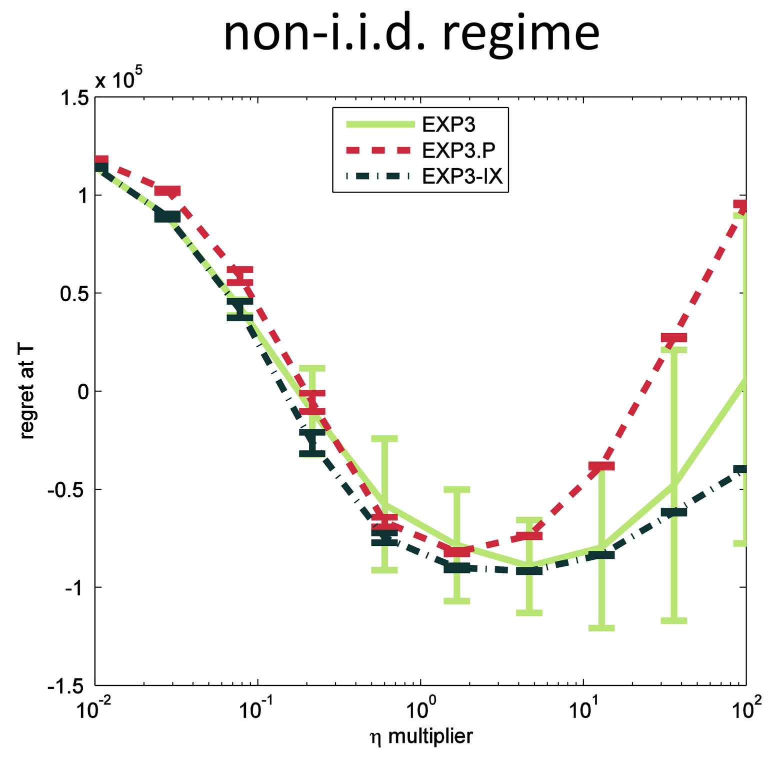
10-arm bandit, Bernoulli losses:
- arms 1-8 have mean 0.5
- arm 9 has mean 0.4
- arm 10 has mean 0.6 until $T/2$, then 0.1

Regret shown as function of learning rate

i.i.d. regime



non-i.i.d. regime



How does it work?

Lemma: With probability at least $1-\delta$,
$$\sum_{t=1}^{T}\left(\hat{\ell}_{t,i} - \ell_{t,i}\right) \le \frac{\log(K/\delta)}{2\gamma}$$
holds simultaneously for all $i \in \{1,2,\dots,K\}$.

Intuitive proof:
$$\hat{\ell}_{t,i} = \frac{\ell_{t,i}}{p_{t,i}+\gamma}\left(\mathbf{1}_{\{I_t=i\}} + \gamma\right) - \gamma\frac{\ell_{t,i}}{p_{t,i}+\gamma}$$
$$\approx \frac{\ell_{t,i}}{p_{t,i}}\mathbf{1}_{\{I_t=i\}} - \frac{\beta}{p_{t,i}}$$
... and then use Freedman's ineq.

A better proof:
- Let $\tilde{\ell}_{t,i} = \frac{\ell_{t,i}}{p_{t,i}}\mathbf{1}_{\{I_t=i\}}$ and show that
$$\hat{\ell}_{t,i} \le \frac{\ell_{t,i}}{p_{t,i} + \gamma\ell_{t,i}}\mathbf{1}_{\{I_t=i\}} \le \frac{1}{2\gamma}\log\left(1 + 2\gamma\tilde{\ell}_{t,i}\right)$$
- Show that
$$\mathbf{E}_t\left[e^{2\gamma\hat{\ell}_{t,i}}\right] = \mathbf{E}_t\left[1 + 2\gamma\tilde{\ell}_{t,i}\right] \le 1 + 2\gamma\ell_{t,i} \le e^{2\gamma\ell_{t,i}}$$
- This implies that $\mathbf{E}\left[\exp\left(2\gamma\sum_{t=1}^{T}\left(\hat{\ell}_{t,i} - \ell_{t,i}\right)\right)\right] \le 1$
- Thus, by Markov's inequality,
$$\mathbf{P}\left[\sum_{t=1}^{T}\left(\hat{\ell}_{t,i} - \ell_{t,i}\right) \ge \varepsilon\right] \le \exp(-2\gamma\varepsilon)$$

Extra panel

- "Optimal" parameters:
$$\eta = 2\gamma = \sqrt{2\log K / KT}$$
- Anytime version:
$$\eta_t = 2\gamma_t = \sqrt{\log K / Kt}$$
- Alternatives for $\hat{\ell}_{t,i}$:
$$\frac{\ell_{t,i}}{p_{t,i} + \gamma\ell_{t,i}}\mathbf{1}_{\{I_t=i\}}$$
$$\frac{\mathbf{1}_{\{I_t=i\}}}{2\gamma}\log\left(1 + 2\gamma\frac{\ell_{t,i}}{p_{t,i}}\right)$$