Follow-the-Perturbed-Leader Achieves Best-of-Both-Worlds for Bandit Problems

Junya Honda (Kyoto Univ. / RIKEN)
Shinji Ito (NEC)
Taira Tsuchiya (Kyoto Univ. / RIKEN)

ALT2023, Singapore
February 22th, 2023

K-armed Bandit Problems

- Model of a gambler pulling arms of K slot machines sequentially.
- At each round $t = 1, 2, \ldots, T$:
  - The environment sets loss $\ell_t = (\ell_{t,1}, \ldots, \ell_{t,K}) \in [0,1]^K$ of arms.
  - The player pulls an arm $I_t \in [K] = \{1, 2, \ldots, K\}$.
  - Loss ℓ_{t,I_t} only for the pulled arm is observed.
- (Pseudo-)regret: Gap of the cumulative loss compared with i^*:

$$\mathrm{Regret}(T) = \mathrm{E}\left[\sum_{t=1}^{T} \ell_{t,I_t} - \sum_{t=1}^{T} \ell_{t,i^*}\right], \quad i^* \in \operatorname*{argmin}_{i \in [K]} \mathrm{E}\left[\sum_{t=1}^{T} \ell_{t,i^*}\right].$$

- Goal: minimize the regret.

Settings of Environments

- Adversarial setting:
  - Loss is determined by an (adaptive) adversary.
  - Lower bound: $\mathrm{Regret}(T) \geq O(\sqrt{KT})$
- Stochastic setting:
  - Loss $\ell_{t,i} \in [0,1]$ from arm i is i.i.d. from unknown P_i.
  - Lower bound: $\mathrm{Regret}(T) \geq O(\sum_{i \neq i^*} \log T / \Delta_i)$ where $\mu_i = \mathrm{E}_{X \sim P_i}[X]$, $\Delta_i = \mu_i - \min_{j \in [K]} \mu_j$.
- A Best-of-Both-Worlds (BOBW) policy:
  - A policy simultaneously achieving both bounds

Follow-The-Regularized-Leader (FTRL)

- Pulls an arm randomly from $w_t = (w_{t,1}, w_{t,2}, \ldots, w_{t,K})$ for

$$w_t = \operatorname*{argmin}_{w \in \mathcal{P}_K} \left\{ w^\top \hat{L}_t + \phi(w)/\eta_t \right\}.$$

  - $\hat{L}_t = \sum_{s=1}^{t-1} \hat{\ell}_t \in \mathbb{R}^K$: unbiased cumulative loss estimator
  - $\phi(\cdot) \geq 0$: regularization function
  - η_t: learning rate
- Tsallis-INF [Zimmert+2021]: Tsallis-entropy $\phi(w) = \frac{1-\sum_i w_i^\alpha}{1-\alpha}$ with learning rate $\eta_t = O(t^{-1/2})$.
  - Achieves BOBW property.
  - Optimization with roughly $O(K)$ complexity
    - Becomes problematic for, e.g., combinatorial bandits.

Follow-The-Perturbed-Leader (FTPL)

- Generates a random perturbation vector $r_t \in \mathbb{R}^K$.
- Pulls the arm minimizing the perturbed estimated loss:

$$I_t = \operatorname*{argmin}_{i \in [K]} \{ \hat{L}_{t,i} - r_{t,i}/\eta_t \}.$$

- Gumbel distribution for r_t reproduces EXP3 [Auer+2002].
- No distribution reproduces Tsallis-INF [Kim+2019].
  - Conjecture: if FTPL achieves $O(\sqrt{KT})$ regret then the perturbation distribution would have Fréchet-type tail (that is, not Gumbel/Weibull-type tail).
- We derive $O(\sqrt{KT})$ bound for Fréchet distribution ($\subset$ Fréchet-type tail).

FTPL with Geometric Resampling (GR)

- Typical loss estimator: Importance-Weighted (IW) estimator

$$\hat{\ell}_{t,i} = \begin{cases} \ell_{t,i} w_{t,i}^{-1} & i \text{ is pulled,} \\ 0 & \text{otherwise.} \end{cases}$$

  - $w_{t,i}$ is not computed in FTPL.
- Geometric resampling [Neu+2016]: unbiased estimation of $w_{t,i}^{-1}$.
  - Used technique: if the success prob. of a trial is p then the expected # of trials until success is $1/p$.

☹ $O(K^2)$ complexity: seemingly worse than FTRL

☺ Optimization-free, no need to store w_t (potential applicability to combinatorial settings?)

Main Result

Theorem 1 (BOBW Property of Fréchet FTPL)

Under FTPL with Fréchet distribution (shape $\alpha = 2$), learning rate $\eta_t = c/\sqrt{t}$ and GR,

$$\mathrm{Regret}(T) \leq \begin{cases} \left(23c + \frac{4}{c}\right)\sqrt{KT} + o(\sqrt{KT}) & \text{(adversarial)}, \\ (4c + \frac{1}{c})^2 \sum_{i \neq i^*} \frac{\log T}{\Delta_i} + o(\log T) & \text{(stochastic, unique } i^*). \end{cases}$$

- First result to show $O(\sqrt{KT})$ regret of FTPL.
- Constant factors are large (particularly for the stochastic case).
  - $4c$ is improved to $2c$ if we exactly compute $w_{t,i}^{-1}$ without GR.

Analysis of FTRL and FTPL

- Regret is usually decomposed into two terms:
  - Penalty term: greediness for the estimated loss
  - Stability term: how stably the policy (or w_t) behaves
- Hard to analyze the stability of FTPL due to complicated w_t.
- Arm-selection probability of FTPL: $w_{t,i} = \phi_i(\eta_t \hat{L}_t)$ for

$$\phi_i(L) = \int_0^\infty \frac{2}{(z + L_i)^3} \exp\left(- \sum_j \frac{1}{(z + L_j)^2} \right) \mathrm{d}z, \quad L \in [0, \infty)^K.$$

  - Comes from the joint density of Fréchet RVs.
    (For simplicity $\min_j L_j = 0$ is assumed w.l.o.g.)

Stability Analysis of FTPL

- The stability term is roughly bounded by

$$\text{Stability} = \mathrm{E}\left[\sum_{t=1}^{T}\left\langle \hat{\ell}_t, w_{t,i} - w_{t+1,i}\right\rangle\right]$$
$$\lesssim \sum_{t=1}^{T}\eta_t \mathrm{E}\left[\sum_{i\in[K]}\frac{\phi_i'(\eta_t\hat{L}_t)}{\phi_i(\eta_t\hat{L}_t)}\right], \quad \text{where } \phi_i'(L) = \frac{\partial\phi(L)}{\partial L_i}.$$

- Ratio between $\phi_i(L)$ and $\phi_i'(L)$ is important ("hazard rate"):
  - [Abernethy+2015] uniformly bounds $\phi_i'(L)/\phi_i(L)$: Leads to a loose regret bound of $O(\sqrt{KT\log K})$.
  - [Bubeck2019] discussed that $\phi_i'(L)/\phi_i(L)^{3/2} = O(1)$ is sufficient for $O(\sqrt{KT})$ bound: difficult to show unlike FTRL.

Key Finding

Lemma 2 (informal)

Under Fréchet FTPL, for any $L \in \mathbb{R}^K$, if L_i is the σ_i-th smallest then

$$\frac{\phi_i'(L)}{\phi_i(L)} \leq O(\sqrt{1/\sigma_i}).$$

- Technical tool: monotonicity analysis by the technique of symmetric polynomials.
  - Heavily depends on the form of the Fréchet distribution.
  - Unclear on the extension to other perturbations.
- $O(\sqrt{KT})$ adversarial bound is immediate from this lemma.

Transformation into a Stochastic Bound for FTRL

- Self-bounding technique [Zimmert+2021]: Typical tool for transforming the adversarial bound into the stochastic bound.
- Key argument: w_t results in regret increases of
  - at most $O\left(\sum_{i \neq i^*} \sqrt{w_{t,i}/t}\right)$,
  - at least $\sum_{i \neq i^*} w_{t,i} \Delta_i$.
- Worst case of $\{w_t\}_{t=1}^T$ without contradiction: $w_{t,i} = O(\frac{1}{t\Delta_i^2})$

  $\rightarrow \mathrm{Regret}(T) \leq O\left(\sum_{i \neq i^*} \sum_{t=1}^T \frac{1}{t\Delta_i}\right) = O\left(\sum_{i \neq i^*} \frac{\log T}{\Delta_i}\right)$.

Transformation into a Stochastic Bound for FTPL

- Hard to apply the same discussion.
- Can still derive some bounds depending on event $A_t = \{\text{estimated losses } \hat{L}_{t,i} \text{ for } i \neq i^* \text{ are large enough}\}$:
  - If A_t holds (well converged) then the regret increase is
    - at most $O\left(\sum_{i \neq i^*} \sqrt{w_{t,i}/t}\right)$,
    - at least $\sum_{i \neq i^*} w_{t,i} \Delta_i$.
  - Otherwise (not converged):
    - at most $O(\sqrt{K/t})$,
    - at least $O(\min_{i \neq i^*} \Delta_i)$.
- Taking the worst case on $\{(w_{t,i},\ \mathbb{1}[A_t])\}_{t=1}^{T}$:

$$\text{Regret}(T) \leq O\left(\sum_{i \neq i^*} \frac{\log T}{\Delta_i}\right).$$

Comparison with Other Policies (1/2)

Consider Fréchet FTPL with IW estimator and

- GR: Geometric resampling,
- GR10: GR is run 10 times (x10 smaller variance of $\widehat{w_{t,i}^{-1}}$), almost equivalent to exact computation of $1/w_{t,i}$.

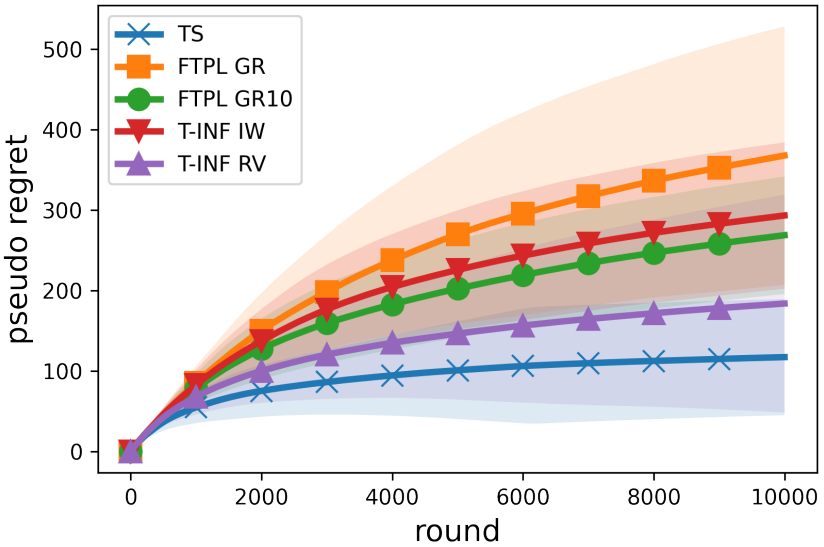
Compared with:

- Thompson Sampling (TS): designed only for the stochastic case.
- Tsallis-INF (T-INF) with two loss estimators:
  - Importance Weighted (IW) estimator (same as FTPL),
  - Reduced Variance (RV) estimator (not applicable to FTPL).

Environments: 8-armed Bernoulli bandits under

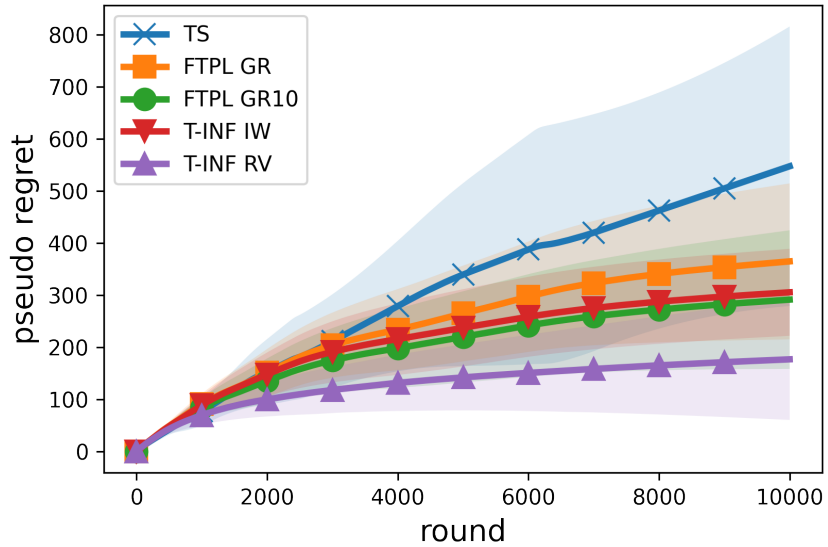
- Stochastic setting,
- Stochastically constrained adversarial setting.

Comparison with Other Policies (2/2)



Stochastic

better



Adversarial

- FTPL GR10 performs similarly to Tsallis-INF IW (as expected).
- FTPL GR performs a little worse due to the variance of $\widehat{w_{t,i}^{-1}}$.
- Tsallis-INF RV performs much better.
- Thompson sampling only works for the stochastic case.

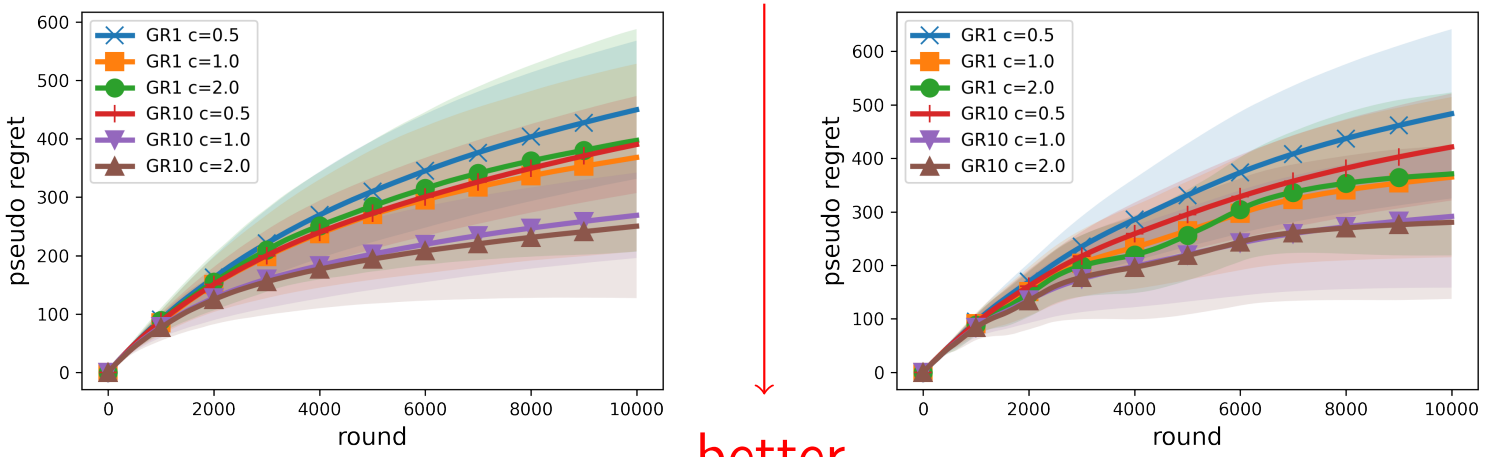
Conclusion

- Showed that FTPL with Fréchet perturbation achieves $O(\sqrt{KT})$ adversarial regret and $O(\sum_{i \neq i^*} \frac{\log T}{\Delta_i})$ stochastic regret.
- The constant factors in the bounds are large.
- Smaller variance of $\hat{\ell}_t$ leads to better performance.

Future directions:

- General sufficient conditions on the perturbation distribution to achieve $O(\sqrt{KT})$ regret.
- Extension to other (especially combinatorial) settings.
  - Adversarial: maybe possible but still nontrivial.
  - BOBW: difficult since known FTRLs use hybrid regularizers.
- Loss estimators with small variances.

Effect of Parameters and Variance Reduction



better

Stochastic Adversarial

- Parameter c controls the learning rate:
  - Small c: stably explores arms,
  - Large c: aggressively exploits seemingly good arms.
- Recommended c from the regret bound: around 0.2–0.5.
- Performs much better for 1.0–2.0.
  - Theoretically suggested c makes the policy too stable.
  - Suggests looseness of the analysis on the stability.