Adversarially Robust Multi-Armed Bandit Algorithm with Variance-Dependent Regret Bounds

Shinji Ito[1,3], Taira Tsuchiya[2,3], Junya Honda[2,3]

1. NEC Corporation, 2. Kyoto University, 3. RIKEN AIP

COLT2022@London, 2022.07.04

Summary

- We consider the multi-armed bandit problem with K arms and T rounds
- We propose a best-of-both-worlds algorithm for **three regimes**:

Summary

- We consider the multi-armed bandit problem with K arms and T rounds
- We propose a best-of-both-worlds algorithm for **three regimes**:
  - Stochastic regime: $R(T) = O\left(\sum_{i \neq i^*} \left(\frac{\sigma_i^2}{\Delta_i} + 1\right) \log T\right)$
    - Δ_i: suboptimality gap of arm i
    - σ_i^2: variance of arm i

Summary

- We consider the multi-armed bandit problem with K arms and T rounds
- We propose a best-of-both-worlds algorithm for **three regimes**:
  - Stochastic regime: $R(T) = O\left(\sum_{i \neq i^*} \left(\frac{\sigma_i^2}{\Delta_i} + 1\right) \log T\right)$
    - Δ_i: suboptimality gap of arm i
    - σ_i^2: variance of arm i
  - Adversarial regime: $R(T) = O\left(\sqrt{K \log T \cdot \min\{T, L^*, Q_\infty\}}\right)$
    - $L^* = \min_{i^* \in [K]} \mathbf{E}\left[\sum_{t=1}^{T} \ell_{i^*}(t)\right]$: cumulative loss for the optimal arm
    - $Q_\infty = \min_{\bar{\ell}} \mathbf{E}\left[\sum_{t=1}^{T} \left\| \ell(t) - \bar{\ell} \right\|_\infty^2\right]$: variation of loss (w.r.t. L^∞-norm)

Summary

- We consider the multi-armed bandit problem with K arms and T rounds
- We propose a best-of-both-worlds algorithm for **three regimes**:
  - Stochastic regime: $R(T) = O\left(\sum_{i \neq i^*} \left(\frac{\sigma_i^2}{\Delta_i} + 1\right) \log T\right)$
    - Δ_i: suboptimality gap of arm i
    - σ_i^2: variance of arm i
  - Adversarial regime: $R(T) = O\left(\sqrt{K \log T \cdot \min\{T, L^*, Q_\infty\}}\right)$
    - $L^* = \min_{i^* \in [K]} \mathbf{E}[\sum_{t=1}^T \ell_{i^*}(t)]$: cumulative loss for the optimal arm
    - $Q_\infty = \min_{\bar{\ell}} \mathbf{E}\left[\sum_{t=1}^T \|\ell(t) - \bar{\ell}\|_\infty^2\right]$: variation of loss (w.r.t. L^∞-norm)
  - Stochastic regime w/ adversarial corruption:
    $R(T) = O\left(\sum_{i \neq i^*} \left(\frac{\sigma_i^2}{\Delta_i} + 1\right) \log T + \sqrt{C \sum_{i \neq i^*} \left(\frac{\sigma_i^2}{\Delta_i} + 1\right) \log T}\right)$
    - C: corruption level

Outline

- Introduction
- Problem setting
- Regret bounds
- Proposed algorithm
- Regret Analysis
- Numerical examples

Multi-armed bandits

- K: the number arms; $[K] = \{1,2,\dots,K\}$: set of arms
- T: the number of rounds
- For $t = 1,2,\dots,T$:
  - The environment chooses a *loss vector* $\ell(t) = \left(\ell_1(t), \ell_2(t), \dots, \ell_K(t)\right)^\top \in [0,1]^K$
  - The player chooses an arm $I(t) \in [K]$ and observes the incurred loss $\ell_{I(t)}(t)$

- Performance metric: the *regret* $R(T)$ defined as

$$R_{i^*}(T) = \mathbf{E}\left[\sum_{t=1}^{T} \ell_{I(t)}(t) - \sum_{t=1}^{T} \ell_{i^*}(t)\right], \qquad R(T) = \max_{i^* \in [K]} R_{i^*}(T)$$

Three regimes for environment

- **Stochastic regime**:
  - Assume $\ell(t)$ is i.i.d. for $t = 1,2,\dots,T$
  - $\mu_i = \mathbf{E}[\ell_i(t)], \quad i^* \in \arg\min_{i\in[K]} \mu_i\,, \quad \Delta_i = \mu_i - \mu_{i^*}, \quad \sigma_i^2 = E[(\ell_i(t) - \mu_i)^2]$
  - We assume the best arm i^* is unique, i.e., $\Delta_i > 0$ for all $i \neq i^*$

Three regimes for environment

- **Stochastic regime**:
  - Assume $\ell(t)$ is i.i.d. for $t = 1,2,\dots,T$
  - $\mu_i = \mathbf{E}[\ell_i(t)], \quad i^* \in \arg\min_{i\in[K]} \mu_i, \quad \Delta_i = \mu_i - \mu_{i^*}, \quad \sigma_i^2 = E[(\ell_i(t) - \mu_i)^2]$
  - We assume the best arm i^* is unique, i.e., $\Delta_i > 0$ for all $i \neq i^*$
- **Adversarial regime**:
  - The environment chooses $\ell(t) \in [0,1]^K$ depending on $\{(\ell(s), I(s))\}_{s=1}^{t-1}$

Three regimes for environment

- **Stochastic regime**:
  - Assume $\ell(t)$ is i.i.d. for $t = 1,2,\ldots,T$
  - $\mu_i = \mathbf{E}[\ell_i(t)], \quad i^* \in \arg\min_{i\in[K]} \mu_i\,, \quad \Delta_i = \mu_i - \mu_{i^*}, \quad \sigma_i^2 = E[(\ell_i(t) - \mu_i)^2]$
  - We assume the best arm i^* is unique, i.e., $\Delta_i > 0$ for all $i \neq i^*$
- **Adversarial regime**:
  - The environment chooses $\ell(t) \in [0,1]^K$ depending on $\{(\ell(s), I(s))\}_{s=1}^{t-1}$
- **Stochastic regime** with **adversarial corruption**:
  - The loss is expressed as $\ell(t) = \ell'(t) + c(t) \in [0,1]^K$
    - $\ell'(t) \in [0,1]^K$: stochastic (i.i.d.)
    - $c(t) \in [-1,1]^K$: adversarial noise
  - *Corruption level* $C := \sum_{t=1}^{T} \mathbf{E}[\|c(t)\|_\infty]$
    - $C = 0 \Rightarrow$ stochastic regimes,
      C: unbounded $\Rightarrow$ adversarial regimes

stochastic regime $(C = 0)$

$\cap$

stochastic regime w/ adversarial constraints

$(C = \sum_{t=1}^{T} \|\bar{\ell}_t - \ell_t\|_\infty)$

$\cap$

adversarial regime

Regret bounds: existing studies

Variance-dependent regret bound

	Stochastic	Adversarial	Stochastic with adversarial corruption
UCB-V [Audibert+, 2009]	$O\left(\sum_{i:\Delta_i>0}\left(\frac{\sigma_i^2}{\Delta_i}+1\right)\log T\right)$	NA	NA

Regret bounds: existing studies

Variance-dependent regret bound

Best of both worlds

	Stochastic	Adversarial	Stochastic with adversarial corruption
UCB-V [Audibert+, 2009]	$O\left(\sum_{i:\Delta_i>0}\left(\frac{\sigma_i^2}{\Delta_i}+1\right)\log T\right)$	NA	NA
Tsallis-INF [Zimmert&Seldin, 2021]	$O\left(\sum_{i:\Delta_i>0}\frac{1}{\Delta_i}\log T\right)$	$O(\sqrt{KT})$	$O\left(\sum_{i:\Delta_i>0}\frac{1}{\Delta_i}\log T+\sqrt{C\sum_{i:\Delta_i>0}\frac{1}{\Delta_i}\log T}\right)$

Regret bounds: existing studies

Variance-dependent regret bound

Best of both worlds

Robust against corruption

	Stochastic	Adversarial	Stochastic with adversarial corruption
UCB-V [Audibert+, 2009]	$O\left(\sum_{i:\Delta_i>0}\left(\frac{\sigma_i^2}{\Delta_i}+1\right)\log T\right)$	NA	NA
Tsallis-INF [Zimmert&Seldin, 2021]	$O\left(\sum_{i:\Delta_i>0}\frac{1}{\Delta_i}\log T\right)$	$O(\sqrt{KT})$	$O\left(\sum_{i:\Delta_i>0}\frac{1}{\Delta_i}\log T+\sqrt{C\sum_{i:\Delta_i>0}\frac{1}{\Delta_i}\log T}\right)$
LB-INF [Ito, 2021]	$O\left(\sum_{i\neq i^*}\frac{1}{\Delta_i}\log T\right)$	$O(\sqrt{K\log T\cdot\min\{T,\boldsymbol{L^*},\boldsymbol{Q_\infty},\boldsymbol{V_1}\}})$	$O\left(\sum_{i\neq i^*}\frac{1}{\Delta_i}\log T+\sqrt{C\sum_{i\neq i^*}\frac{1}{\Delta_i}\log T}\right)$

Regret bounds: existing studies

Variance-dependent regret bound

	Stochastic	Adversarial	Stochastic with adversarial corruption
UCB-V [Audibert+, 2009]	$O\left(\sum_{i:\Delta_i>0}\left(\frac{\sigma_i^2}{\Delta_i}+1\right)\log T\right)$	NA	NA
Tsallis-INF [Zimmert&Seldin, 2021]	$O\left(\sum_{i:\Delta_i>0}\frac{1}{\Delta_i}\log T\right)$	$O(\sqrt{KT})$	$O\left(\sum_{i:\Delta_i>0}\frac{1}{\Delta_i}\log T+\sqrt{C\sum_{i:\Delta_i>0}\frac{1}{\Delta_i}\log T}\right)$
LB-INF [Ito, 2021]	$O\left(\sum_{i\neq i^*}\frac{1}{\Delta_i}\log T\right)$	$O\left(\sqrt{K\log T\cdot\min\{T,\boldsymbol{L^*},\boldsymbol{Q_\infty},\boldsymbol{V_1}\}}\right)$	$O\left(\sum_{i\neq i^*}\frac{1}{\Delta_i}\log T+\sqrt{C\sum_{i\neq i^*}\frac{1}{\Delta_i}\log T}\right)$

Best of both worlds

Robust against corruption

Data-dependent regret bound

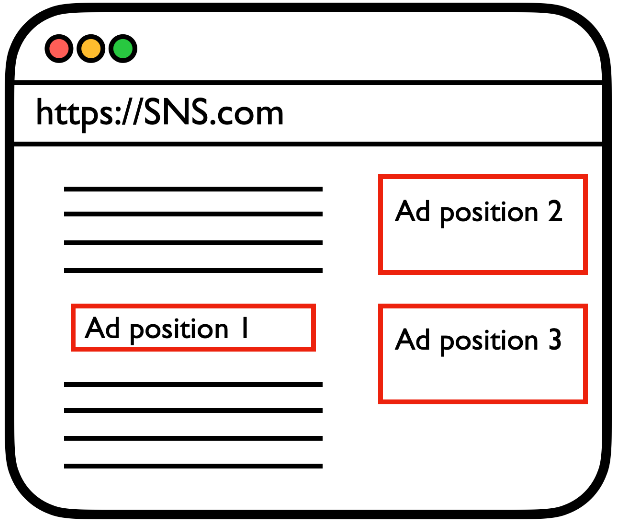
- $L^* = \min_{i^*\in[K]} \mathbf{E}[\sum_{t=1}^{T}\ell_i(t)]$: cumulative loss for the optimal arm
- $Q_\infty = \min_{\bar{\ell}} \mathbf{E}\left[\sum_{t=1}^{T}\left\|\ell_{i^*}(t)-\bar{\ell}\right\|_\infty^2\right]$: variation of loss (w.r.t. L^∞-norm)
- $V_1 = \mathbf{E}[\sum_{t=1}^{T-1}\|\ell(t)-\ell(t-1)\|_1]$: path-length (w.r.t. L^1-norm)

Can we go further?

In many applications such as recommender systems…

- Positive label feedback (e.g., to purchase or click) is rare
  ⇒ small variance σ_i^2
  ⇒ algorithm w/ variance-dependent regret bound
  can perform very well

- Losses / rewards are not always i.i.d.
  ⇒ BOBW and corruption-robustness are important



<u>Research question</u>: any BOBW algorithm with a variance-dependent regret bound?

Regret bounds: this study

	Stochastic	Adversarial	Stochastic with adversarial corruption
UCB-V [Audibert+, 2009]	$O\left(\sum_{i:\Delta_i>0}\left(\frac{\sigma_i^2}{\Delta_i}+1\right)\log T\right)$	NA	NA
Tsallis-INF [Zimmert&Seldin, 2021]	$O\left(\sum_{i:\Delta_i>0}\frac{1}{\Delta_i}\log T\right)$	$O(\sqrt{KT})$	$O\left(\sum_{i:\Delta_i>0}\frac{1}{\Delta_i}\log T+\sqrt{C\sum_{i:\Delta_i>0}\frac{1}{\Delta_i}\log T}\right)$
LB-INF [Ito, 2021]	$O\left(\sum_{i\neq i^*}\frac{1}{\Delta_i}\log T\right)$	$O(\sqrt{K\log T\cdot\min\{T,L^*,Q_\infty,V_1\}})$	$O\left(\sum_{i\neq i^*}\frac{1}{\Delta_i}\log T+\sqrt{C\sum_{i\neq i^*}\frac{1}{\Delta_i}\log T}\right)$
LB-INF-V (This work)	$O\left(\sum_{i\neq i^*}\left(\frac{\sigma_i^2}{\Delta_i}+1\right)\log T\right)$	$O(\sqrt{K\log T\cdot\min\{T,L^*,Q_\infty\}})$	$O\left(\sum_{i\neq i^*}\left(\frac{\sigma_i^2}{\Delta_i}+1\right)\log T+\sqrt{C\sum_{i\neq i^*}\left(\frac{\sigma_i^2}{\Delta_i}+1\right)\log T}\right)$

- This study proposes the first BOBW algorithm with variance-dependent regret bounds
- The proposed algorithm (LB-INF-V) is corruption-robust and has data-dependent regret bounds

Regret bounds: this study

	Stochastic	Gap from lower bound	Adversarial
UCB-V [Audibert+, 2009]	$\sum_{i:\Delta_i>0}\left(10\frac{\sigma_i^2}{\Delta_i}+20\right)\log T$	≈ 5	NA
Tsallis-INF [Zimmert&Seldin, 2021]	$\approx \sum_{i:\Delta_i>0}\frac{1}{\Delta_i}\log T$	≈ 2	$O\left(\sqrt{KT}\right)$
LB-INF [Ito, 2021]	$\approx 36\sum_{i\neq i^*}\frac{1}{\Delta_i}\log T$	≈ 72	$O\left(\sqrt{K\log T\cdot\min\{T,L^*,Q_\infty,V_1\}}\right)$
LB-INF-V (This work)	$\approx \sum_{i\neq i^*}\max\left\{4\frac{\sigma_i^2}{\Delta_i},2\right\}\log T$	≈ 2	$O\left(\sqrt{K\log T\cdot\min\{T,L^*,Q_\infty\}}\right)$

The leading constant of the regret upper bound is close to the lower bound (gap ≈ 2)

Regret bounds: this study

	Stochastic	Gap from lower bound	Adversarial
UCB-V [Audibert+, 2009]	$\sum_{i:\Delta_i>0}\left(10\frac{\sigma_i^2}{\Delta_i}+20\right)\log T$	≈ 5	NA
Tsallis-INF [Zimmert&Seldin, 2021]	$\approx \sum_{i:\Delta_i>0}\frac{1}{\Delta_i}\log T$	≈ 2	$O\left(\sqrt{KT}\right)$
LB-INF [Ito, 2021]	$\approx 36\sum_{i\neq i^*}\frac{1}{\Delta_i}\log T$	≈ 72	$O\left(\sqrt{K\log T\cdot\min\{T,L^*,Q_\infty,V_1\}}\right)$
LB-INF-V [This work]	$\approx \sum_{i\neq i^*}\max\left\{4\frac{\sigma_i^2}{\Delta_i},2\right\}\log T$	≈ 2	$O\left(\sqrt{K\log T\cdot\min\{T,L^*,Q_\infty\}}\right)$
LB-INF-V-mod **(This work)**	$\approx \sum_{i\neq i^*}\max\left\{8\frac{\sigma_i^2}{\Delta_i},4\right\}\log T$	≈ 4	$O\left(\sqrt{K\log T\cdot\min\{T,L^*,Q_\infty,V_1\}}\right)$

- The leading constant of the regret upper bound is close to the lower bound (gap ≈ 2)
- Modifications to the algorithm yield a **path-length regret bound** in exchange for a larger constant

Outline

- Introduction
- Problem setting
- Regret bounds
- Proposed algorithm
- Regret Analysis
- Numerical examples

Proposed algorithm

- **Optimistic follow the regularized leader** (cf. [Rakhlin & Sridharan, 2013], [Wei & Luo, 2018])
  - For each t, choose $I(t) \in [K]$ according to the distribution $p(t)$ such that

$$p(t) \in \arg\min_{p \in \mathcal{P}_K} \left\{ \left\langle m(t) + \sum_{s=1}^{t-1} \hat{\ell}_s , p \right\rangle + \psi_t(p) \right\}$$

    - $m(t) \in [0,1]^K$: optimistic prediction for $\ell(t)$
    - $\hat{\ell}_t \in \mathbb{R}^K$: unbiased estimator of ℓ_t
    - ψ_t: convex regularization function

Proposed algorithm

- **Optimistic follow the regularized leader** (cf. [Rakhlin & Sridharan, 2013], [Wei & Luo, 2018])
  - For each t, choose $I(t) \in [K]$ according to the distribution $p(t)$ such that

$$p(t) \in \arg\min_{p \in \mathcal{P}_K} \left\{ \left\langle m(t) + \sum_{s=1}^{t-1} \hat{\ell}_s , p \right\rangle + \psi_t(p) \right\}$$

    - $m(t) \in [0,1]^K$: optimistic prediction for $\ell(t)$
    - $\hat{\ell}_t \in \mathbb{R}^K$: unbiased estimator of ℓ_t
    - ψ_t: convex regularization function

Converges to μ_i

- **Optimistic prediction**: empirical mean of observed data of losse $m_i(t) = \frac{\frac{1}{2} + \sum_{s=1}^{t-1} \mathbf{1}[I(s)=i] \ell_i(s)}{1 + \sum_{s=1}^{t-1} \mathbf{1}[I(s)=i]}$

Proposed algorithm

- **Optimistic follow the regularized leader** (cf. [Rakhlin & Sridharan, 2013], [Wei & Luo, 2018])
  - For each t, choose $I(t) \in [K]$ according to the distribution $p(t)$ such that

$$p(t) \in \arg\min_{p \in \mathcal{P}_K} \left\{ \left\langle m(t) + \sum_{s=1}^{t-1} \hat{\ell}_s, p \right\rangle + \psi_t(p) \right\}$$

    - $m(t) \in [0,1]^K$: optimistic prediction for $\ell(t)$
    - $\hat{\ell}_t \in \mathbb{R}^K$: unbiased estimator of ℓ_t
    - ψ_t: convex regularization function

Converges to μ_i

- **Optimistic prediction**: empirical mean of observed data of losse $m_i(t) = \frac{\frac{1}{2} + \sum_{s=1}^{t-1} \mathbf{1}[I(s)=i]\ell_i(s)}{1 + \sum_{s=1}^{t-1} \mathbf{1}[I(s)=i]}$

- **Unbiased estimator**: $\hat{\ell}_i(t) = m_i(t) + \frac{\mathbf{1}[I(t)=i]}{p_i(t)}\left(\ell_i(t) - m_i(t)\right)$ **Reduce variances using $\mathbf{m_i(t)}$**

Proposed algorithm

- **Optimistic follow the regularized leader**
  - For each t, choose $I(t) \in [K]$ according to the distribution $p(t)$ such that

$$p(t) \in \arg\min_{p \in \mathcal{P}_K} \left\{ \left\langle m(t) + \sum_{s=1}^{t-1} \hat{\ell}_s , p \right\rangle + \psi_t(p) \right\}$$

    - ψ_t: convex regularization function

- **Regularization function**: $\psi_t(p) = \sum_{i=1}^{K} \beta_i(t) \phi(p_i)$, where
  - $\phi(x) = x - 1 \underline{- \log x} + \log T \cdot (x + \underline{(1-x)\log(1-x)})$

Log-barrier regularization
cf. BROAD [Wei&Luo, 2018], LB-INF [Ito, 2021]

Entropy regularization for $(1 - \mathrm{x})$:
used to handle the impact of the variance of the optimal arm

Proposed algorithm

- **Optimistic follow the regularized leader**
  - For each t, choose $I(t) \in [K]$ according to the distribution $p(t)$ such that

$$p(t) \in \arg\min_{p \in \mathcal{P}_K} \left\{ \left\langle m(t) + \sum_{s=1}^{t-1} \hat{\ell}_s , p \right\rangle + \psi_t(p) \right\}$$

    - ψ_t: convex regularization function

- **Regularization function**: $\psi_t(p) = \sum_{i=1}^{K} \beta_i(t) \phi(p_i)$, where
  - $\phi(x) = x - 1 \underline{-\log x} + \log T \cdot (x + \underline{(1-x)\log(1-x)})$

    Log-barrier regularization
    cf. BROAD [Wei&Luo, 2018], LB-INF [Ito, 2021]

    Entropy regularization for $(1 - \mathrm{x})$:
    used to handle the impact of the variance of the optimal arm

  - $\beta_i(t)$: adaptively chosen based on $\underline{\text{squared prediction error } \left(\ell_{I(s)} - m_{I(s)}(s)\right)^2}$ of $m(s)$

    $\xrightarrow[s \to \infty]{} \sigma_{I(s)}^2$

Outline

- Introduction
- Problem setting
- Regret bounds
- Proposed algorithm
- Regret Analysis
- Numerical examples

Regret analysis: stochastic regime

- Definition of ψ_t and a standard analysis technique for OFTRL yield:

Lem. 1 For sufficiently large T, $R(T) \simeq O\left(\sum_{i \neq i^*} \sqrt{\sum_{t=1}^{T} 1[I(t)=i]\boxed{(\ell_i(t)-m_i(t))^2}\log(T)}\right)$

- Definition of $m(t)$ yields:

Lem. 2 $\mathbb{E}\left[\sum_{t=1}^{T} 1[I(t)=i]\boxed{(\ell_i(t)-m_i(t))^2}\right] = O\left(\boxed{\sigma_i^2}\mathbb{E}[\sum_{t=1}^{T} p_i(t)] + \log(T)\right)$

- Combining the above two lemmas and Jensen's inequality, we obtain:

Prop. 1 For sufficiently large T, $R(T) = O\left(\sum_{i \neq i^*} \sqrt{\boxed{\sigma_i^2}\mathbb{E}[\sum_{t=1}^{T} p_i(t)]\log(T)} + K\log(T)\right)$

Regret analysis: stochastic regime

Prop. 1 For sufficiently large T, $R(T) = O\left(\sum_{i \neq i^*} \sqrt{\sigma_i^2 \mathbb{E}[\sum_{t=1}^T p_i(t)] \log(T)} + K \log(T) \right)$

$+$

Self-bounding constraint. $R(T) = \sum_{i \neq i^*} \Delta_i \mathbb{E}\left[\sum_{t=1}^T p_i(t) \right]$

Self-bounding technique
cf. [Zimmert & Seldin, 2021],
[Wei & Luo, 2018], [Gaillard+, 2014]

$\Downarrow$

Thm. 1 $R(T) = O\left(\sum_{i \neq i^*} \left(\frac{\sigma_i^2}{\Delta_i} + 1 \right) \log(T) \right)$

Regret analysis: adversarial regime

- Definition of ψ_t and a standard analysis technique for OFTRL yield:

Lem. 1 For sufficiently large T, $R(T) \simeq O\left(\sum_{i \neq i^*} \sqrt{\sum_{t=1}^{T} 1[I(t)=i](\ell_i(t)-m_i(t))^2 \log(T)}\right)$

- Definition of $m(t)$ yields:

Lem. 3 It holds for any $\ell^* \in [0,1]^K$ that

$$\mathbb{E}\left[\sum_{t=1}^{T} 1[I(t)=i](\ell_i(t)-m_i(t))^2\right] = \mathbb{E}\left[\sum_{t=1}^{T} 1[I(t)=i](\ell_i(t)-\ell_i^*)^2\right] + O(K\log(T))$$

Consequently,

$$\mathbb{E}\left[\sum_{t=1}^{T} 1[I(t)=i](\ell_i(t)-m_i(t))^2\right] = \min\{Q_\infty, L^* + R(T), T - L^* - R(T)\} + O(K\log(T))$$

- Combining the above two lemmas 1 and 3, we obtain $R(T) = O\left(\sqrt{K \min\{Q_\infty, L^*, T - L^*\}\log(T)} + K\log(T)\right)$

Numerical Comparison with Thompson Sampling & Tsallis-INF w/ RV-estimator

Setting: Bernoulli bandits with K = 5

<u>Experiment 1.</u>

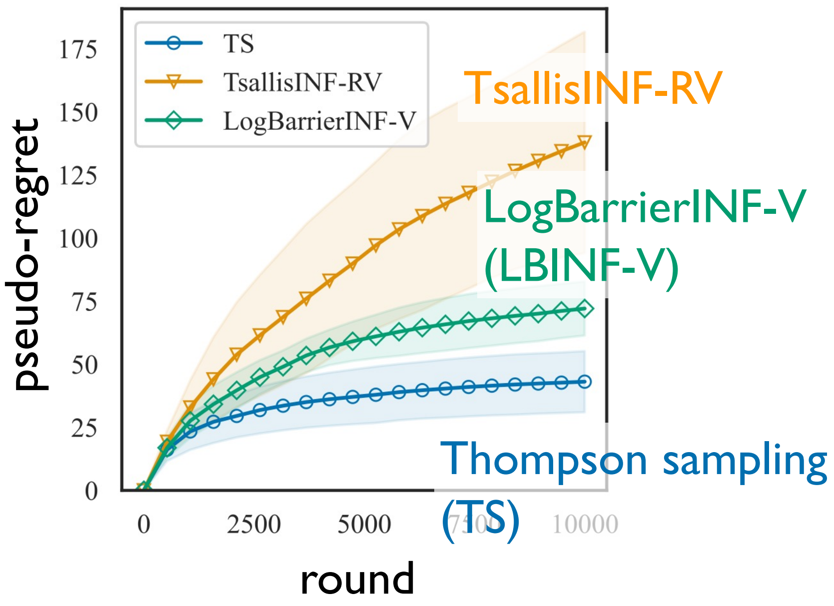
- Stochastic regime
- $\mu = (0.5, 0.9, \ldots, 0.9)$ $\rightarrow$ small σ_i^2

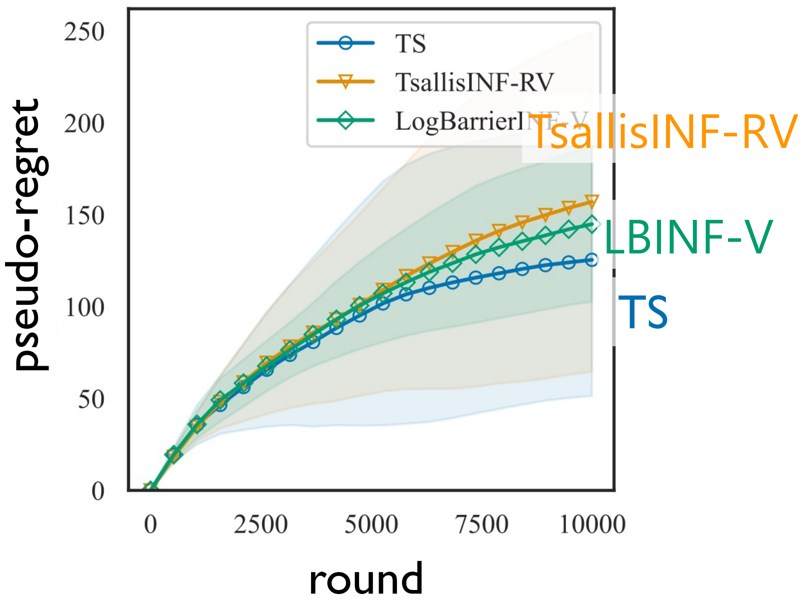
<u>Experiment 2.</u>

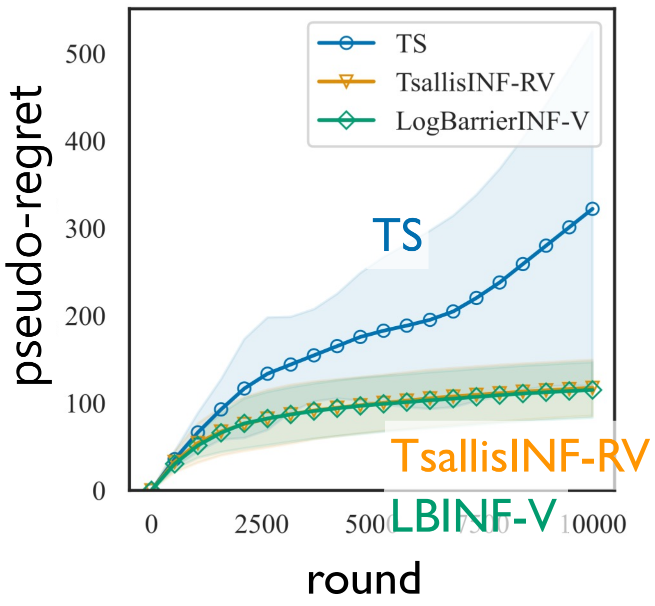
- Stochastic regime
- $\mu = (0.5, 0.55, \ldots, 0.55)$ $\rightarrow$ large σ_i^2

<u>Experiment 3.</u>

- Stochastically constrained adversarial regime
- $\Delta = 0.1$(same as Figure 3 in [Zimmert & Seldin 2021])







Conclusion

- OFTRL with adaptive learning rate achieves

stochastic regime $(C = 0)$

$\cap$

stochastic regime w/ adversarial constraints $(C = \sum_{t=1}^{T} \|\bar{\ell}_t - \ell_t\|_\infty)$

$\cap$

adversarial regime

$O\left(\sqrt{K \min\{T, L^*, Q_\infty\} \log T}\right)$

$\subset$

adversarial regime w/ self-bounding constraints

$$O\left(\sum_{i \neq i^*} \left(\frac{\sigma_i^2}{\Delta_i} + 1\right)\log T + \sqrt{C \sum_{i \neq i^*} \left(\frac{\sigma_i^2}{\Delta_i} + 1\right)\log T}\right)$$

$\boxed{\sigma_i^2 : \text{ variance of arm } i}$

The leading constant of the regret upper bound is close to the lower bound (**gap** ≈ 2)

Conclusion

- OFTRL with adaptive learning rate achieves

stochastic regime $(C = 0)$

$\cap$

stochastic regime w/ adversarial constraints

$(C = \sum_{t=1}^{T} \|\bar{\ell}_t - \ell_t\|_\infty)$

$\cap$

adversarial regime

$O\left(\sqrt{K \min\{T, L^*, Q_\infty\} \log T}\right)$

$\subset$

adversarial regime w/
self-bounding constraints

$$O\left(\sum_{i \neq i^*} \left(\frac{\sigma_i^2}{\Delta_i} + 1\right) \log T + \sqrt{C \sum_{i \neq i^*} \left(\frac{\sigma_i^2}{\Delta_i} + 1\right) \log T}\right)$$

σ_i^2 : variance of arm i

The leading constant of the regret upper bound is close to the lower bound (gap ≈ 2)

- Open questions and future directions:
  - Can we achieve a gap < 2 while preserving BOBW and/or corruption-robustness?
  - Can we remove the assumption that the optimal arm is unique?