First Order Bayesian Regret Analysis of Thompson Sampling

Mark Sellke

Joint with Sébastien Bubeck (MSR)

ALT 2020

Overview

Multi-armed Bandits

- Finite set $[n] = \{1, 2, \ldots, n\}$ of arms.

Multi-armed Bandits

- Finite set $[n] = \{1, 2, \ldots, n\}$ of arms.
- Oblivious adversary chooses T loss functions $\ell_t : [n] \to [0, 1]$.

Multi-armed Bandits

- Finite set $[n] = \{1, 2, \ldots, n\}$ of arms.
- Oblivious adversary chooses T loss functions $\ell_t : [n] \to [0, 1]$.
- Each round, player picks $i_t \in [n]$, pays loss $\ell_t(i_t)$.

Multi-armed Bandits

- Finite set $[n] = \{1, 2, \ldots, n\}$ of arms.
- Oblivious adversary chooses T loss functions $\ell_t : [n] \to [0, 1]$.
- Each round, player picks $i_t \in [n]$, pays loss $\ell_t(i_t)$.
- Full-feedback: observe the entire vector ℓ_t.
  Bandit: observe only $\ell_t(i_t)$.
  Semi-bandit: choose m arms at once and observe all m losses.

Multi-armed Bandits

- Finite set $[n] = \{1, 2, \ldots, n\}$ of arms.
- Oblivious adversary chooses T loss functions $\ell_t : [n] \to [0, 1]$.
- Each round, player picks $i_t \in [n]$, pays loss $\ell_t(i_t)$.
- Full-feedback: observe the entire vector ℓ_t.
  Bandit: observe only $\ell_t(i_t)$.
  Semi-bandit: choose m arms at once and observe all m losses.
- Arm i has a total loss $L_{i,T} = \sum_{t=1}^{T} \ell_t(i)$.
  Player has total loss $L_T = \sum_{t=1}^{T} \ell_t(i_t)$.

Multi-armed Bandits

- Finite set $[n] = \{1, 2, \dots, n\}$ of arms.
- Oblivious adversary chooses T loss functions $\ell_t : [n] \to [0, 1]$.
- Each round, player picks $i_t \in [n]$, pays loss $\ell_t(i_t)$.
- Full-feedback: observe the entire vector ℓ_t.
  Bandit: observe only $\ell_t(i_t)$.
  Semi-bandit: choose m arms at once and observe all m losses.
- Arm i has a total loss $L_{i,T} = \sum_{t=1}^{T} \ell_t(i)$.
  Player has total loss $L_T = \sum_{t=1}^{T} \ell_t(i_t)$.
- Let i^* be the arm with smallest loss L^*.

Multi-armed Bandits

- Finite set $[n] = \{1, 2, \dots, n\}$ of arms.
- Oblivious adversary chooses T loss functions $\ell_t : [n] \to [0, 1]$.
- Each round, player picks $i_t \in [n]$, pays loss $\ell_t(i_t)$.
- Full-feedback: observe the entire vector ℓ_t.
  Bandit: observe only $\ell_t(i_t)$.
  Semi-bandit: choose m arms at once and observe all m losses.
- Arm i has a total loss $L_{i,T} = \sum_{t=1}^{T} \ell_t(i)$.
  Player has total loss $L_T = \sum_{t=1}^{T} \ell_t(i_t)$.
- Let i^* be the arm with smallest loss L^*.
- Goal: small expected regret $\mathbb{E}[R_T] = \mathbb{E}[L_T - L^*]$.

Bayesian Regret and Thompson Sampling

Bayesian Regret and Thompson Sampling

- Adversarial Bayesian setting: known prior over oblivious adversaries. Goal is low Bayesian regret $\mathbb{E}[R_T]$ against this prior.

Bayesian Regret and Thompson Sampling

- Adversarial Bayesian setting: known prior over oblivious adversaries. Goal is low Bayesian regret $\mathbb{E}[R_T]$ against this prior.
- Minimax theorem says adversarial Bayesian guarantees must hold for some frequentist algorithm.

Bayesian Regret and Thompson Sampling

- Adversarial Bayesian setting: known prior over oblivious adversaries. Goal is low Bayesian regret $\mathbb{E}[R_T]$ against this prior.
- Minimax theorem says adversarial Bayesian guarantees must hold for some frequentist algorithm.
- Thompson sampling (TS): play next action from current distribution $p_t(\cdot)$ of the (eventual) best arm i^*.

Bayesian Regret and Thompson Sampling

- Adversarial Bayesian setting: known prior over oblivious adversaries. Goal is low Bayesian regret $\mathbb{E}[R_T]$ against this prior.
- Minimax theorem says adversarial Bayesian guarantees must hold for some frequentist algorithm.
- Thompson sampling (TS): play next action from current distribution $p_t(\cdot)$ of the (eventual) best arm i^*.
- Easy to simulate, used lots in practice. TS known to attain optimal $O(\sqrt{T})$ and gap-dependent guarantees in various settings, and more. [RVR16, AG12, KMN12].

Bayesian Regret and Thompson Sampling

- Adversarial Bayesian setting: known prior over oblivious adversaries. Goal is low Bayesian regret $\mathbb{E}[R_T]$ against this prior.
- Minimax theorem says adversarial Bayesian guarantees must hold for some frequentist algorithm.
- Thompson sampling (TS): play next action from current distribution $p_t(\cdot)$ of the (eventual) best arm i^*.
- Easy to simulate, used lots in practice. TS known to attain optimal $O(\sqrt{T})$ and gap-dependent guarantees in various settings, and more. [RVR16, AG12, KMN12].
- Some frequentist algorithms (EXP3, MD with log barrier) achieve more refined *first order* regret $\mathbb{E}[R_T] = O(\sqrt{L^*})$. What about TS?

Main Result

Main Theorem

Thompson sampling has optimal first order regret guarantees in full-feedback, bandit, and semi-bandit setting.

Main Result

Main Theorem

Thompson sampling has optimal first order regret guarantees in full-feedback, bandit, and semi-bandit setting.

Proofs build on $O(\sqrt{T})$ entropic analysis of [RVR16].

Main Result

Main Theorem

Thompson sampling has optimal first order regret guarantees in full-feedback, bandit, and semi-bandit setting.

Proofs build on $O(\sqrt{T})$ entropic analysis of [RVR16].
Three Key Techniques:

Main Result

Main Theorem

Thompson sampling has optimal first order regret guarantees in full-feedback, bandit, and semi-bandit setting.

Proofs build on $O(\sqrt{T})$ entropic analysis of [RVR16].

Three Key Techniques:

- Refinement of Pinsker's inequality turns $O(\sqrt{T})$ arguments into $O(\sqrt{L^*})$ arguments.

Main Result

Main Theorem

Thompson sampling has optimal first order regret guarantees in full-feedback, bandit, and semi-bandit setting.

Proofs build on $O(\sqrt{T})$ entropic analysis of [RVR16].

Three Key Techniques:

- Refinement of Pinsker's inequality turns $O(\sqrt{T})$ arguments into $O(\sqrt{L^*})$ arguments.
- "Self-awareness" of TS: shows that arm i is discarded when its total (observed or unobserved) loss is much larger than L^*.

Main Result

Main Theorem

Thompson sampling has optimal first order regret guarantees in full-feedback, bandit, and semi-bandit setting.

Proofs build on $O(\sqrt{T})$ entropic analysis of [RVR16].

Three Key Techniques:

- Refinement of Pinsker's inequality turns $O(\sqrt{T})$ arguments into $O(\sqrt{L^*})$ arguments.
- "Self-awareness" of TS: shows that arm i is discarded when its total (observed or unobserved) loss is much larger than L^*.
- Extensions of Shannon entropy for the semi-bandit setting.

More Precise Results

	Full Feedback	Bandit	Semi-bandit
Minimax	$\sqrt{T \log n}$	$\sqrt{Tn}$	$\sqrt{Tnm}$
First Order	$\sqrt{L^* \log n}$	$\sqrt{L^* n}$	$\tilde{O}(\sqrt{L^* n}))$
TS	$\sqrt{TH(p_0)}$	$\sqrt{Tn}$	$\tilde{O}(\sqrt{Tnm})$
First Order TS	$\sqrt{L^* H(p_0)}$	$\sqrt{L^* n}$	$\tilde{O}(\sqrt{L^* n})$

More Precise Results

	Full Feedback	Bandit	Semi-bandit
Minimax	$\sqrt{T \log n}$	$\sqrt{Tn}$	$\sqrt{Tnm}$
First Order	$\sqrt{L^* \log n}$	$\sqrt{L^* n}$	$\tilde{O}(\sqrt{L^* n}))$
TS	$\sqrt{TH(p_0)}$	$\sqrt{Tn}$	$\tilde{O}(\sqrt{Tnm})$
First Order TS	$\sqrt{L^* H(p_0)}$	$\sqrt{L^* n}$	$\tilde{O}(\sqrt{L^* n})$

- Bottom line is new for TS.

More Precise Results

	Full Feedback	Bandit	Semi-bandit
Minimax	$\sqrt{T \log n}$	$\sqrt{Tn}$	$\sqrt{Tnm}$
First Order	$\sqrt{L^* \log n}$	$\sqrt{L^* n}$	$\tilde{O}(\sqrt{L^* n}))$
TS	$\sqrt{TH(p_0)}$	$\sqrt{Tn}$	$\tilde{O}(\sqrt{Tnm})$
First Order TS	$\sqrt{L^* H(p_0)}$	$\sqrt{L^* n}$	$\tilde{O}(\sqrt{L^* n})$

- Bottom line is new for TS.
- $\tilde{O}(\sqrt{Tnm})$ matches previous work in more generality: previously prior had to be independent over arms. Same generalization for contextual bandit, but first order bound does not hold for TS.

More Precise Results

	Full Feedback	Bandit	Semi-bandit
Minimax	$\sqrt{T \log n}$	$\sqrt{Tn}$	$\sqrt{Tnm}$
First Order	$\sqrt{L^* \log n}$	$\sqrt{L^* n}$	$\tilde{O}(\sqrt{L^* n}))$
TS	$\sqrt{TH(p_0)}$	$\sqrt{Tn}$	$\tilde{O}(\sqrt{Tnm})$
First Order TS	$\sqrt{L^* H(p_0)}$	$\sqrt{L^* n}$	$\tilde{O}(\sqrt{L^* n})$

- Bottom line is new for TS.
- $\tilde{O}(\sqrt{Tnm})$ matches previous work in more generality: previously prior had to be independent over arms. Same generalization for contextual bandit, but first order bound does not hold for TS.
- First column highlights that TS adapts to informative priors if $H(p_0) \ll \log(n)$. Similar results hold for partial feedback.

Two Meanings of “First Order Regret Bound”

Two Meanings of "First Order Regret Bound"

- We show two types of first order regret bound.

Two Meanings of “First Order Regret Bound”

- We show two types of first order regret bound.
- Easier formulation: suppose that (an upper bound on) L^* is known, show $O(\sqrt{L^*})$ regret.

Two Meanings of "First Order Regret Bound"

- We show two types of first order regret bound.
- Easier formulation: suppose that (an upper bound on) L^* is known, show $O(\sqrt{L^*})$ regret.
- All technical points today will be for this version.

Two Meanings of “First Order Regret Bound”

- We show two types of first order regret bound.
- Easier formulation: suppose that (an upper bound on) L^* is known, show $O(\sqrt{L^*})$ regret.
- All technical points today will be for this version.
- If L^* is known, can achieve fully T-independent regret using *thresholded TS*. Never play arm i if $p_t(i) \leq \gamma$ for γ a small constant. Ordinary TS has mild $\log(T)$ dependence. Same story already existed for (thresholded) EXP3.

Two Meanings of "First Order Regret Bound"

- We show two types of first order regret bound.
- Easier formulation: suppose that (an upper bound on) L^* is known, show $O(\sqrt{L^*})$ regret.
- All technical points today will be for this version.
- If L^* is known, can achieve fully T-independent regret using *thresholded TS*. Never play arm i if $p_t(i) \leq \gamma$ for γ a small constant. Ordinary TS has mild $\log(T)$ dependence. Same story already existed for (thresholded) EXP3.
- Harder formulation: unknown L^*, prove $O(\sqrt{\mathbb{E}[L^*]})$ regret. We show this for TS also, seems to require log-barrier instead of entropy.

Two Meanings of "First Order Regret Bound"

- We show two types of first order regret bound.
- Easier formulation: suppose that (an upper bound on) L^* is known, show $O(\sqrt{L^*})$ regret.
- All technical points today will be for this version.
- If L^* is known, can achieve fully T-independent regret using *thresholded TS*. Never play arm i if $p_t(i) \leq \gamma$ for γ a small constant. Ordinary TS has mild $\log(T)$ dependence. Same story already existed for (thresholded) EXP3.
- Harder formulation: unknown L^*, prove $O(\sqrt{\mathbb{E}[L^*]})$ regret. We show this for TS also, seems to require log-barrier instead of entropy.
- $O(\sqrt{\mathbb{E}[L^*]})$ analysis is based on recent connection between TS and mirror descent from [LZ19]. For known L^*, can similarly remove logs with Tsallis entropy instead of Shannon.

Russo and van Roy Entropic Analysis for Full Feedback

Russo and van Roy Entropic Analysis for Full Feedback

- Denote the regret incurred in timestep t by

$$r_t = \ell_t(i_t) - \ell_t(i^*).$$

Russo and van Roy Entropic Analysis for Full Feedback

- Denote the regret incurred in timestep t by

$$r_t = \ell_t(i_t) - \ell_t(i^*).$$

- By linearity, $\mathbb{E}[\sum_{t \leq T} r_t] = \mathbb{E}[R_T]$ is the expected total regret.

Russo and van Roy Entropic Analysis for Full Feedback

- Denote the regret incurred in timestep t by

$$r_t = \ell_t(i_t) - \ell_t(i^*).$$

- By linearity, $\mathbb{E}[\sum_{t \leq T} r_t] = \mathbb{E}[R_T]$ is the expected total regret.
- For full-feedback, $\mathbb{E}[r_t]$ is the inner product of loss and probability movement:

$$\mathbb{E}[r_t] = \mathbb{E}[\langle \ell_t, p_t - p_{t+1} \rangle] = \mathbb{E} \sum_i \Big(\ell_t(i) \cdot (p_t(i) - p_{t+1}(i)) \Big).$$

Russo and van Roy Entropic Analysis for Full Feedback

- Denote the regret incurred in timestep t by

$$r_t = \ell_t(i_t) - \ell_t(i^*).$$

- By linearity, $\mathbb{E}[\sum_{t \leq T} r_t] = \mathbb{E}[R_T]$ is the expected total regret.
- For full-feedback, $\mathbb{E}[r_t]$ is the inner product of loss and probability movement:

$$\mathbb{E}[r_t] = \mathbb{E}[\langle \ell_t, p_t - p_{t+1} \rangle] = \mathbb{E} \sum_i \Big(\ell_t(i) \cdot (p_t(i) - p_{t+1}(i)) \Big).$$

- Reason: player's average loss is $\mathbb{E}^t \langle \ell_t, p_t \rangle$. After $\ell_t(\cdot)$ is known, average loss of i^* is $\mathbb{E}^{t+1}[\ell_t(i^*)] = \langle \ell_t, p_{t+1} \rangle$.

Russo and van Roy Entropic Analysis for Full Feedback

- Denote the regret incurred in timestep t by

$$r_t = \ell_t(i_t) - \ell_t(i^*).$$

- By linearity, $\mathbb{E}[\sum_{t \le T} r_t] = \mathbb{E}[R_T]$ is the expected total regret.
- For full-feedback, $\mathbb{E}[r_t]$ is the inner product of loss and probability movement:

$$\mathbb{E}[r_t] = \mathbb{E}[\langle \ell_t, p_t - p_{t+1} \rangle] = \mathbb{E} \sum_i \Big(\ell_t(i) \cdot (p_t(i) - p_{t+1}(i)) \Big).$$

- Reason: player's average loss is $\mathbb{E}^t \langle \ell_t, p_t \rangle$. After $\ell_t(\cdot)$ is known, average loss of i^* is $\mathbb{E}^{t+1}[\ell_t(i^*)] = \langle \ell_t, p_{t+1} \rangle$.
- Because $\ell_t(i) \in [0, 1]$, we have

$$\mathbb{E}[r_t] \le \mathbb{E}[||p_t - p_{t+1}||_{\ell^1}].$$

Russo and van Roy Entropic Analysis for Full Feedback

- Denote the regret incurred in timestep t by
$$r_t = \ell_t(i_t) - \ell_t(i^*).$$
- By linearity, $\mathbb{E}[\sum_{t\le T} r_t] = \mathbb{E}[R_T]$ is the expected total regret.
- For full-feedback, $\mathbb{E}[r_t]$ is the inner product of loss and probability movement:
$$\mathbb{E}[r_t] = \mathbb{E}[\langle \ell_t, p_t - p_{t+1}\rangle] = \mathbb{E}\sum_i \Big(\ell_t(i)\cdot(p_t(i) - p_{t+1}(i))\Big).$$
- Reason: player's average loss is $\mathbb{E}^t\langle \ell_t, p_t\rangle$. After $\ell_t(\cdot)$ is known, average loss of i^* is $\mathbb{E}^{t+1}[\ell_t(i^*)] = \langle \ell_t, p_{t+1}\rangle$.
- Because $\ell_t(i)\in[0,1]$, we have
$$\mathbb{E}[r_t] \le \mathbb{E}[||p_t - p_{t+1}||_{\ell^1}].$$
- To bound regret, estimate ℓ^1 (= TV) movement of p_t.

Russo and van Roy Entropic Analysis for Full Feedback

- Now entropy appears via Pinsker's Inequality:

$$||p_t - p_{t+1}||_{\ell^1}^2 \leq KL[p_{t+1}; p_t].$$

Russo and van Roy Entropic Analysis for Full Feedback

- Now entropy appears via Pinsker's Inequality:

$$||p_t - p_{t+1}||_{\ell^1}^2 \leq KL[p_{t+1}; p_t].$$

- Cauchy-Schwarz:

$$\mathbb{E}\left[\sum_{t=0}^{T-1} ||p_t - p_{t+1}||_{\ell^1}\right] \leq \sqrt{T \cdot \mathbb{E}\left[\sum_{t=0}^{T-1} ||p_t - p_{t+1}||_{\ell^1}^2\right]}$$

Russo and van Roy Entropic Analysis for Full Feedback

- Now entropy appears via Pinsker's Inequality:

$$||p_t - p_{t+1}||_{\ell^1}^2 \le KL[p_{t+1}; p_t].$$

- Cauchy-Schwarz:

$$\mathbb{E}\left[\sum_{t=0}^{T-1} ||p_t - p_{t+1}||_{\ell^1}\right] \le \sqrt{T \cdot \mathbb{E}\left[\sum_{t=0}^{T-1} ||p_t - p_{t+1}||_{\ell^1}^2\right]}$$

- Pinsker:

$$\mathbb{E}\left[\sum_{t=0}^{T-1} ||p_t - p_{t+1}||_{\ell^1}^2\right] \le \mathbb{E}\left[\sum_{t\ge 0} KL[p_{t+1}; p_t]\right] \le H(p_0) \le \log(n).$$

Russo and van Roy Entropic Analysis for Full Feedback

Theorem [RVR16]

In the full-feedback setting, Thompson Sampling has expected regret $\mathbb{E}[R_T] \le \sqrt{T \cdot H(p_0)} \le \sqrt{T \log(n)}$.

Russo and van Roy Entropic Analysis for Full Feedback

Theorem [RVR16]

In the full-feedback setting, Thompson Sampling has expected regret $\mathbb{E}[R_T] \leq \sqrt{T \cdot H(p_0)} \leq \sqrt{T \log(n)}$.

- Upper bound relied on Pinsker's inequality:

$$||p_t - p_{t+1}||_{\ell^1}^2 \leq KL[p_{t+1}; p_t].$$

Russo and van Roy Entropic Analysis for Full Feedback

Theorem [RVR16]

In the full-feedback setting, Thompson Sampling has expected regret $\mathbb{E}[R_T] \leq \sqrt{T \cdot H(p_0)} \leq \sqrt{T \log(n)}$.

- Upper bound relied on Pinsker's inequality:

$$||p_t - p_{t+1}||_{\ell^1}^2 \leq KL[p_{t+1}; p_t].$$

- 2nd order Taylor expansion of KL suggests refined Pinsker:

$$\sum_i \frac{(p_t(i) - p_{t+1}(i))^2}{p_t(i)} \leq \mathbb{E}[KL[p_{t+1}; p_t]].$$

Russo and van Roy Entropic Analysis for Full Feedback

Theorem [RVR16]

In the full-feedback setting, Thompson Sampling has expected regret $\mathbb{E}[R_T] \leq \sqrt{T \cdot H(p_0)} \leq \sqrt{T \log(n)}$.

- Upper bound relied on Pinsker's inequality:

$$||p_t - p_{t+1}||_{\ell^1}^2 \leq KL[p_{t+1}; p_t].$$

- 2nd order Taylor expansion of KL suggests refined Pinsker:

$$\sum_i \frac{(p_t(i) - p_{t+1}(i))^2}{p_t(i)} \leq \mathbb{E}[KL[p_{t+1}; p_t]].$$

- This inequality is false. Fix by using $(p_t(i) - p_{t+1}(i))_+$ throughout (gives positive part of regret). Let's pretend it is true.

First Order Regret for Full-Feedback

- The new calculation, with Cauchy-Schwarz and refined Pinsker:

First Order Regret for Full-Feedback

- The new calculation, with Cauchy-Schwarz and refined Pinsker:

$$(\mathbb{E}[R_t])^2 = \left(\sum_{t,i} \mathbb{E}\left[\ell_t(i) \cdot (p_t(i) - p_{t+1}(i))\right] \right)^2$$

First Order Regret for Full-Feedback

- The new calculation, with Cauchy-Schwarz and refined Pinsker:

$$(\mathbb{E}[R_t])^2 = \left(\sum_{t,i} \mathbb{E}\left[\ell_t(i) \cdot (p_t(i) - p_{t+1}(i))\right] \right)^2$$

$$\leq \left(\mathbb{E}\left[\sum_{t,i} \ell_t(i)^2 \cdot p_t(i) \right] \right) \left(\mathbb{E}\left[\sum_{t,i} \frac{(p_t(i) - p_{t+1}(i))^2}{p_t(i)} \right] \right)$$

First Order Regret for Full-Feedback

- The new calculation, with Cauchy-Schwarz and refined Pinsker:

$$(\mathbb{E}[R_t])^2 = \left(\sum_{t,i} \mathbb{E}\left[\ell_t(i) \cdot (p_t(i) - p_{t+1}(i))\right] \right)^2$$

$$\leq \left(\mathbb{E}\left[\sum_{t,i} \ell_t(i)^2 \cdot p_t(i) \right] \right) \left(\mathbb{E}\left[\sum_{t,i} \frac{(p_t(i) - p_{t+1}(i))^2}{p_t(i)} \right] \right)$$

$$\leq \mathbb{E}[L_T] \cdot \mathbb{E} \sum_t KL[p_{t+1}; p_t] \leq \mathbb{E}[L_T] \cdot H(p_0).$$

First Order Regret for Full-Feedback

- The new calculation, with Cauchy-Schwarz and refined Pinsker:

$$(\mathbb{E}[R_t])^2 = \left(\sum_{t,i} \mathbb{E}\left[\ell_t(i) \cdot (p_t(i) - p_{t+1}(i))\right]\right)^2$$

$$\leq \left(\mathbb{E}\left[\sum_{t,i} \ell_t(i)^2 \cdot p_t(i)\right]\right)\left(\mathbb{E}\left[\sum_{t,i} \frac{(p_t(i) - p_{t+1}(i))^2}{p_t(i)}\right]\right)$$

$$\leq \mathbb{E}[L_T] \cdot \mathbb{E}\sum_t KL[p_{t+1}; p_t] \leq \mathbb{E}[L_T] \cdot H(p_0).$$

- Hence $\mathbb{E}[R_T] \leq \sqrt{\mathbb{E}[L_T]H(p_0)}$. Almost what we want.

First Order Regret for Full-Feedback

- The new calculation, with Cauchy-Schwarz and refined Pinsker:

$$(\mathbb{E}[R_t])^2 = \left(\sum_{t,i} \mathbb{E}\left[\ell_t(i) \cdot (p_t(i) - p_{t+1}(i))\right] \right)^2$$

$$\leq \left(\mathbb{E}\left[\sum_{t,i} \ell_t(i)^2 \cdot p_t(i) \right] \right) \left(\mathbb{E}\left[\sum_{t,i} \frac{(p_t(i) - p_{t+1}(i))^2}{p_t(i)} \right] \right)$$

$$\leq \mathbb{E}[L_T] \cdot \mathbb{E} \sum_t KL[p_{t+1}; p_t] \leq \mathbb{E}[L_T] \cdot H(p_0).$$

- Hence $\mathbb{E}[R_T] \leq \sqrt{\mathbb{E}[L_T] H(p_0)}$. Almost what we want.
- Recalling $R_T = L_T - L^*$, easy algebra shows what we want: $\mathbb{E}[R_T] = O(\sqrt{\mathbb{E}[L^*] H(p_0)})$

The Bandit Case

The Bandit Case

- Bandit: use refined Pinsker again. Analysis is a bit different.

The Bandit Case

- Bandit: use refined Pinsker again. Analysis is a bit different.

Key Lemma for Bandit (roughly)

In the bandit setting with known L^*, once arm i has total (observed plus unobserved) loss $\sum_{s\leq t} \ell_s(i) \gg L^*$, we will have $p_t(i) \approx 0$.

The Bandit Case

- Bandit: use refined Pinsker again. Analysis is a bit different.

Key Lemma for Bandit (roughly)

In the bandit setting with known L^*, once arm i has total (observed plus unobserved) loss $\sum_{s \le t} \ell_s(i) \gg L^*$, we will have $p_t(i) \approx 0$.

- With full-feedback, $p_t(i) = 0$ as soon as loss crosses L^*. Need to show TS can estimate unseen losses accurately.

The Bandit Case

- Bandit: use refined Pinsker again. Analysis is a bit different.

Key Lemma for Bandit (roughly)

In the bandit setting with known L^*, once arm i has total (observed plus unobserved) loss $\sum_{s\le t} \ell_s(i) \gg L^*$, we will have $p_t(i) \approx 0$.

- With full-feedback, $p_t(i) = 0$ as soon as loss crosses L^*. Need to show TS can estimate unseen losses accurately.
- A frequentist can use unbiased estimator $\sum_{t:a_t=i} \frac{\ell_t(i)}{p_t(i)}$ for total loss, and tight frequentist confidence intervals (CIs) around it.

The Bandit Case

- Bandit: use refined Pinsker again. Analysis is a bit different.

Key Lemma for Bandit (roughly)

In the bandit setting with known L^*, once arm i has total (observed plus unobserved) loss $\sum_{s\le t} \ell_s(i) \gg L^*$, we will have $p_t(i) \approx 0$.

- With full-feedback, $p_t(i) = 0$ as soon as loss crosses L^*. Need to show TS can estimate unseen losses accurately.
- A frequentist can use unbiased estimator $\sum_{t:a_t=i} \frac{\ell_t(i)}{p_t(i)}$ for total loss, and tight frequentist confidence intervals (CIs) around it.
- Turn out TS implicitly knows the CIs.

The Bandit Case

- Bandit: use refined Pinsker again. Analysis is a bit different.

Key Lemma for Bandit (roughly)

In the bandit setting with known L^*, once arm i has total (observed plus unobserved) loss $\sum_{s\le t} \ell_s(i) \gg L^*$, we will have $p_t(i) \approx 0$.

- With full-feedback, $p_t(i) = 0$ as soon as loss crosses L^*. Need to show TS can estimate unseen losses accurately.
- A frequentist can use unbiased estimator $\sum_{t:a_t=i} \frac{\ell_t(i)}{p_t(i)}$ for total loss, and tight frequentist confidence intervals (CIs) around it.
- Turn out TS implicitly knows the CIs.
- In general, $\mathbb{E}[\mathbb{P}^t[X]] = \mathbb{P}[X]$ for any event X.
  Letting $X =$ "the CIs are accurate" $\implies$ TS believes CIs w.h.p.

The Bandit Case

- Bandit: use refined Pinsker again. Analysis is a bit different.

Key Lemma for Bandit (roughly)

In the bandit setting with known L^*, once arm i has total (observed plus unobserved) loss $\sum_{s \leq t} \ell_s(i) \gg L^*$, we will have $p_t(i) \approx 0$.

- With full-feedback, $p_t(i) = 0$ as soon as loss crosses L^*. Need to show TS can estimate unseen losses accurately.
- A frequentist can use unbiased estimator $\sum_{t:a_t=i} \frac{\ell_t(i)}{p_t(i)}$ for total loss, and tight frequentist confidence intervals (CIs) around it.
- Turn out TS implicitly knows the CIs.
- In general, $\mathbb{E}[\mathbb{P}^t[X]] = \mathbb{P}[X]$ for any event X.
  Letting $X =$ "the CIs are accurate" $\implies$ TS believes CIs w.h.p.
- When loss $\gg L^*$, lower confidence bound for loss is $> L^*$ w.h.p.

The Bandit Case

- Bandit: use refined Pinsker again. Analysis is a bit different.

Key Lemma for Bandit (roughly)

In the bandit setting with known L^*, once arm i has total (observed plus unobserved) loss $\sum_{s\leq t} \ell_s(i) \gg L^*$, we will have $p_t(i) \approx 0$.

- With full-feedback, $p_t(i) = 0$ as soon as loss crosses L^*. Need to show TS can estimate unseen losses accurately.
- A frequentist can use unbiased estimator $\sum_{t:a_t=i} \frac{\ell_t(i)}{p_t(i)}$ for total loss, and tight frequentist confidence intervals (CIs) around it.
- Turn out TS implicitly knows the CIs.
- In general, $\mathbb{E}[\mathbb{P}^t[X]] = \mathbb{P}[X]$ for any event X. Letting $X =$ "the CIs are accurate" $\implies$ TS believes CIs w.h.p.
- When loss $\gg L^*$, lower confidence bound for loss is $> L^*$ w.h.p.
- Now $p_t(i) \leq \mathbb{P}[L_{i,T} \leq L^*] \leq 1 - \mathbb{P}^t[X] \approx 0$ proving the lemma.

Semi-bandit

- Semi-bandit: choose, pay, and observe an m-set A_t of m arms.

Semi-bandit

- Semi-bandit: choose, pay, and observe an m-set A_t of m arms.
- (Maybe restricted to subset $\mathcal{A} \subseteq \binom{[n]}{m}$ of m-sets.) E.g. online shortest path.

Semi-bandit

- Semi-bandit: choose, pay, and observe an m-set A_t of m arms.
- (Maybe restricted to subset $\mathcal{A} \subseteq \binom{[n]}{m}$ of m-sets.) E.g. online shortest path.
- Natural to use $H(A^*)$, viewing each m-set separately. If arms are independent and we aim for $O(\sqrt{T})$ this is fine. Need changes for correlated priors and $O(\sqrt{L^*})$ regret.

Semi-bandit

- Semi-bandit: choose, pay, and observe an m-set A_t of m arms.
- (Maybe restricted to subset $\mathcal{A} \subseteq \binom{[n]}{m}$ of m-sets.) E.g. online shortest path.
- Natural to use $H(A^*)$, viewing each m-set separately. If arms are independent and we aim for $O(\sqrt{T})$ this is fine. Need changes for correlated priors and $O(\sqrt{L^*})$ regret.
- Correlations: decouple arms via *coordinate entropy* $H^c(A^*) = -\sum_i p_t(i \in A^*) \log(p_t(i \in A^*))$. These probabilities add to m, not 1.

Semi-bandit

- Semi-bandit: choose, pay, and observe an m-set A_t of m arms.
- (Maybe restricted to subset $\mathcal{A} \subseteq \binom{[n]}{m}$ of m-sets.) E.g. online shortest path.
- Natural to use $H(A^*)$, viewing each m-set separately. If arms are independent and we aim for $O(\sqrt{T})$ this is fine. Need changes for correlated priors and $O(\sqrt{L^*})$ regret.
- Correlations: decouple arms via *coordinate entropy* $H^c(A^*) = -\sum_i p_t(i \in A^*) \log(p_t(i \in A^*))$. These probabilities add to m, not 1.
- $O(\sqrt{L^*})$: rank the m arms in A^*: $L_T(a_1) \geq L_T(a_2) \geq L_T(a_m)$.

Semi-bandit

- Semi-bandit: choose, pay, and observe an m-set A_t of m arms.
- (Maybe restricted to subset $\mathcal{A} \subseteq \binom{[n]}{m}$ of m-sets.) E.g. online shortest path.
- Natural to use $H(A^*)$, viewing each m-set separately. If arms are independent and we aim for $O(\sqrt{T})$ this is fine. Need changes for correlated priors and $O(\sqrt{L^*})$ regret.
- Correlations: decouple arms via *coordinate entropy* $H^c(A^*) = -\sum_i p_t(i \in A^*) \log(p_t(i \in A^*))$. These probabilities add to m, not 1.
- $O(\sqrt{L^*})$: rank the m arms in A^*: $L_T(a_1) \geq L_T(a_2) \geq L_T(a_m)$.
- Use coordinate entropy on dyadic subsets $S_0 = \{a_1\}, S_1 = \{a_2, a_3\}, \dots, S_k = \{a_{2^k}, a_{2^{k+1}-1}\}$.

Semi-bandit

- Semi-bandit: choose, pay, and observe an m-set A_t of m arms.
- (Maybe restricted to subset $\mathcal{A} \subseteq \binom{[n]}{m}$ of m-sets.) E.g. online shortest path.
- Natural to use $H(A^*)$, viewing each m-set separately. If arms are independent and we aim for $O(\sqrt{T})$ this is fine. Need changes for correlated priors and $O(\sqrt{L^*})$ regret.
- Correlations: decouple arms via *coordinate entropy* $H^c(A^*) = -\sum_i p_t(i \in A^*) \log(p_t(i \in A^*))$. These probabilities add to m, not 1.
- $O(\sqrt{L^*})$: rank the m arms in A^*: $L_T(a_1) \geq L_T(a_2) \geq L_T(a_m)$.
- Use coordinate entropy on dyadic subsets $S_0 = \{a_1\}, S_1 = \{a_2, a_3\}, \dots, S_k = \{a_{2^k}, a_{2^{k+1}-1}\}$.
- If $i \in S_k$ then $L_T(i) \leq \frac{L^*}{2^k}$. So $p_t(i \in S_k) \approx 0$ quickly for larger k as in the previous slide. This means entropy is depleted fast for most of the ranks.

Open Problems

Open Problems

- For adversarial priors, TS can have $\Omega(T)$ regret with constant probability (and also $-\Omega(T)$). Also true for EXP3, but EXP3.P has low regret with high probability. Any corresponding TS variant?

Open Problems

- For adversarial priors, TS can have $\Omega(T)$ regret with constant probability (and also $-\Omega(T)$). Also true for EXP3, but EXP3.P has low regret with high probability. Any corresponding TS variant?
- Same story for contextual bandit. TS achieves $O(\sqrt{T})$ but not $O(\sqrt{L^*})$. But there is an algorithm with first order regret [ABL18]. Is there an analog of TS achieving this?

References

Lattimore and Zimmert
Connections Between Mirror Descent, Thompson sampling and the Information Ratio.
NeurIPS 2019.

Allen-Zhu, Bubeck, and Li
Make the Minority Great Again: first Order Regret Bound for Contextual Bandits.
ICML 2018.

Russo and Van Roy
An Information-Theoretic Analysis of Thompson Sampling
The Journal of Machine Learning Research The Journal of Machine Learning Research 17.1 (2016): 2442-2471.

Kaufmann, Kordan, and Munos
Thompson Sampling: An Asymptotically Optimal Finite-Time Analysis.
ALT 2012.

Agrawal and Goyal
Analysis of Thompson Sampling for the Multi-Armed Bandit Problem.
COLT 2012.