Small Loss Regret Bounds for Thompson Sampling

Mark Sellke

with Sébastien Bubeck

August 29, 2018

Overview

1. Learning from Experts
2. Analysis of Thompson Sampling for Full Feedback
3. Small Loss Bound for Full Feedback
4. Bandit and Semibandit Cases
5. Open Problems

Learning from Experts

- Finite (static) set $[n] = \{1, 2, \ldots, n\}$ of possible actions each round.

Learning from Experts

- Finite (static) set $[n] = \{1, 2, \ldots, n\}$ of possible actions each round.
- Beforehand, an (oblivious) adversary chooses a sequence $(\ell_t(i))_{t=1}^T$ of loss functions $\ell_t : [n] \to [0, 1]$.

Learning from Experts

- Finite (static) set $[n] = \{1, 2, \ldots, n\}$ of possible actions each round.
- Beforehand, an (oblivious) adversary chooses a sequence $(\ell_t(i))_{t=1}^T$ of loss functions $\ell_t : [n] \to [0, 1]$.
- Each round, the player chooses an action i_t and pays loss $\ell_t(i_t)$.

Learning from Experts

- Finite (static) set $[n] = \{1, 2, \ldots, n\}$ of possible actions each round.
- Beforehand, an (oblivious) adversary chooses a sequence $(\ell_t(i))_{t=1}^T$ of loss functions $\ell_t : [n] \to [0, 1]$.
- Each round, the player chooses an action i_t and pays loss $\ell_t(i_t)$.
- In the *full feedback* setting we observe the entire vector $\vec{\ell}_t$ after timestep t, while in the *bandit* setting we observe only $\ell_t(i_t)$.

Learning from Experts

- Finite (static) set $[n] = \{1, 2, \ldots, n\}$ of possible actions each round.
- Beforehand, an (oblivious) adversary chooses a sequence $(\ell_t(i))_{t=1}^T$ of loss functions $\ell_t : [n] \to [0, 1]$.
- Each round, the player chooses an action i_t and pays loss $\ell_t(i_t)$.
- In the *full feedback* setting we observe the entire vector $\vec{\ell}_t$ after timestep t, while in the *bandit* setting we observe only $\ell_t(i_t)$.
- At the end, each action has a total loss $L_{i,T} = \sum_{t=1}^T \ell_t(i)$. Player has total loss $L_T = \sum_{t=1}^T \ell_t(i_t)$. Denote by $i^\star = \arg\min_i L_{i,T}$ the coordinate with smallest loss $L^\star = \sum_{t=1}^T \ell_t(i^\star)$.

Learning from Experts

- Finite (static) set $[n] = \{1, 2, \ldots, n\}$ of possible actions each round.
- Beforehand, an (oblivious) adversary chooses a sequence $(\ell_t(i))_{t=1}^T$ of loss functions $\ell_t : [n] \to [0, 1]$.
- Each round, the player chooses an action i_t and pays loss $\ell_t(i_t)$.
- In the *full feedback* setting we observe the entire vector $\vec{\ell}_t$ after timestep t, while in the *bandit* setting we observe only $\ell_t(i_t)$.
- At the end, each action has a total loss $L_{i,T} = \sum_{t=1}^T \ell_t(i)$. Player has total loss $L_T = \sum_{t=1}^T \ell_t(i_t)$. Denote by $i^\star = \arg\min_i L_{i,T}$ the coordinate with smallest loss $L^\star = \sum_{t=1}^T \ell_t(i^\star)$.
- The *regret* is $R_T = L_T - L^\star$.

Learning from Experts

- Finite (static) set $[n] = \{1, 2, \dots, n\}$ of possible actions each round.
- Beforehand, an (oblivious) adversary chooses a sequence $(\ell_t(i))_{t=1}^{T}$ of loss functions $\ell_t : [n] \to [0, 1]$.
- Each round, the player chooses an action i_t and pays loss $\ell_t(i_t)$.
- In the *full feedback* setting we observe the entire vector $\vec{\ell}_t$ after timestep t, while in the *bandit* setting we observe only $\ell_t(i_t)$.
- At the end, each action has a total loss $L_{i,T} = \sum_{t=1}^{T} \ell_t(i)$. Player has total loss $L_T = \sum_{t=1}^{T} \ell_t(i_t)$. Denote by $i^\star = \arg\min_i L_{i,T}$ the coordinate with smallest loss $L^\star = \sum_{t=1}^{T} \ell_t(i^\star)$.
- The *regret* is $R_T = L_T - L^\star$.
- Player's objective: minimize expected regret $\mathbb{E}[R_T]$.

Learning from Experts

- Finite (static) set $[n] = \{1, 2, \dots, n\}$ of possible actions each round.
- Beforehand, an (oblivious) adversary chooses a sequence $(\ell_t(i))_{t=1}^T$ of loss functions $\ell_t : [n] \to [0, 1]$.
- Each round, the player chooses an action i_t and pays loss $\ell_t(i_t)$.
- In the *full feedback* setting we observe the entire vector $\vec{\ell}_t$ after timestep t, while in the *bandit* setting we observe only $\ell_t(i_t)$.
- At the end, each action has a total loss $L_{i,T} = \sum_{t=1}^T \ell_t(i)$. Player has total loss $L_T = \sum_{t=1}^T \ell_t(i_t)$. Denote by $i^\star = \arg\min_i L_{i,T}$ the coordinate with smallest loss $L^\star = \sum_{t=1}^T \ell_t(i^\star)$.
- The *regret* is $R_T = L_T - L^\star$.
- Player's objective: minimize expected regret $\mathbb{E}[R_T]$.
- Later we also consider the *semibandit* case: player picks m of n actions each round.

The Bayesian Approach

- As stated, we have no probability distribution over adversaries; we have to do well on average against all adversaries at once.

The Bayesian Approach

- As stated, we have no probability distribution over adversaries; we have to do well on average against all adversaries at once.
- If we only care about the value of this player-adversary game, by the mini-max theorem, suffices to do well against any *fixed* probability distribution over adversaries.

The Bayesian Approach

- As stated, we have no probability distribution over adversaries; we have to do well on average against all adversaries at once.
- If we only care about the value of this player-adversary game, by the mini-max theorem, suffices to do well against any *fixed* probability distribution over adversaries.
- Hence we may assume we are given a prior distribution over loss sequences. If we can always keep $\mathbb{E}[R_T]$ low, great!

The Bayesian Approach

- As stated, we have no probability distribution over adversaries; we have to do well on average against all adversaries at once.
- If we only care about the value of this player-adversary game, by the mini-max theorem, suffices to do well against any *fixed* probability distribution over adversaries.
- Hence we may assume we are given a prior distribution over loss sequences. If we can always keep $\mathbb{E}[R_T]$ low, great!
- As we observe losses, we update our distribution to obtain a new posterior distribution each round.

The Bayesian Approach

- As stated, we have no probability distribution over adversaries; we have to do well on average against all adversaries at once.
- If we only care about the value of this player-adversary game, by the mini-max theorem, suffices to do well against any *fixed* probability distribution over adversaries.
- Hence we may assume we are given a prior distribution over loss sequences. If we can always keep $\mathbb{E}[R_T]$ low, great!
- As we observe losses, we update our distribution to obtain a new posterior distribution each round.
- Given initial prior distribution, finding optimal play is a complicated, deterministic computation.

The Bayesian Approach

- As stated, we have no probability distribution over adversaries; we have to do well on average against all adversaries at once.
- If we only care about the value of this player-adversary game, by the mini-max theorem, suffices to do well against any *fixed* probability distribution over adversaries.
- Hence we may assume we are given a prior distribution over loss sequences. If we can always keep $\mathbb{E}[R_T]$ low, great!
- As we observe losses, we update our distribution to obtain a new posterior distribution each round.
- Given initial prior distribution, finding optimal play is a complicated, deterministic computation.
- Thompson Sampling is a simple strategy for any probability distribution. Not exactly optimal, but it does very well and is feasible in practice.

Thompson Sampling

Thompson Sampling Procedure

At each time t, compute the posterior distribution p_t for the best coordinate $i^\star = \arg\min_{i\in[n]} L_{i,T}$. Then pick the next action i_t according to the distribution p_t.

Thompson Sampling

Thompson Sampling Procedure

At each time t, compute the posterior distribution p_t for the best coordinate $i^\star = \arg\min_{i \in [n]} L_{i,T}$. Then pick the next action i_t according to the distribution p_t.

- In the full-information case, this strategy intuitively hedges by softly following the leader. In the bandit case, this strategy intuitively balances explore/exploit similarly to multiplicative weights or upper confidence bound algorithms.

Thompson Sampling

Thompson Sampling Procedure

At each time t, compute the posterior distribution p_t for the best coordinate $i^\star = \arg\min_{i\in[n]} L_{i,T}$. Then pick the next action i_t according to the distribution p_t.

- In the full-information case, this strategy intuitively hedges by softly following the leader. In the bandit case, this strategy intuitively balances explore/exploit similarly to multiplicative weights or upper confidence bound algorithms.
- Unlike the exact optimal strategy, Thompson Sampling is often efficient to simulate, and is amenable to analysis.

Small Loss Regret Bounds

- Standard regret bounds show $\mathbb{E}[R_T] = O(\sqrt{T})$.

Small Loss Regret Bounds

- Standard regret bounds show $\mathbb{E}[R_T] = O(\sqrt{T})$.
- However, if $L^\star$ is very small, not so impressive. Why should we do worse when the same loss comes more slowly?

Small Loss Regret Bounds

- Standard regret bounds show $\mathbb{E}[R_T] = O(\sqrt{T})$.
- However, if $L^\star$ is very small, not so impressive. Why should we do worse when the same loss comes more slowly?
- Slow accumulation of small losses? But we can always assume losses are binary, either 0 or 1.

Small Loss Regret Bounds

- Standard regret bounds show $\mathbb{E}[R_T] = O(\sqrt{T})$.
- However, if $L^\star$ is very small, not so impressive. Why should we do worse when the same loss comes more slowly?
- Slow accumulation of small losses? But we can always assume losses are binary, either 0 or 1.
- Hence, interest in showing more refined $O(\sqrt{L^\star})$ regret bounds.

What Was Known?

- Optimal $O(\sqrt{T})$ regret bounds are well established. E.g Multiplicative weights, online mirror descent.

What Was Known?

- Optimal $O(\sqrt{T})$ regret bounds are well established. E.g Multiplicative weights, online mirror descent.
- Multiplicative weights gives $O(\sqrt{L^\star})$ small loss bound in the full feedback setting. In the bandit/semibandit cases, modify to never play bad actions ([Lykouris, Sridharan, and Tardos '18]).

What Was Known?

- Optimal $O(\sqrt{T})$ regret bounds are well established. E.g Multiplicative weights, online mirror descent.
- Multiplicative weights gives $O(\sqrt{L^\star})$ small loss bound in the full feedback setting. In the bandit/semibandit cases, modify to never play bad actions ([Lykouris, Sridharan, and Tardos '18]).
- Russo and Van Roy show that Thompson Sampling achieves regret bounds of the form $O(\sqrt{T})$ in a variety of situations including those we consider here [Russo and Van Roy '16].

What Was Known?

	Full Feedback	Bandit	Semibandit
Regret	$\sqrt{T \log n}$	$\sqrt{nT}$	$\sqrt{nmT}$
Small Loss	$\sqrt{L^\star \log n}$	$\sqrt{nL^\star \log n}$	$\tilde{O}(\sqrt{L^\star(m^3+n)\log(T)})$
Thompson	$\sqrt{TH(\vec{p}_0)}$	$\tilde{O}(\sqrt{nT})$	$\tilde{O}(\sqrt{nmT})$ $(*)$

- Optimal $O(\sqrt{T})$ regret bounds are well established. E.g Multiplicative weights, online mirror descent.
- Multiplicative weights gives $O(\sqrt{L^\star})$ small loss bound in the full feedback setting. In the bandit/semibandit cases, modify to never play bad actions ([Lykouris, Sridharan, and Tardos '18]).
- Russo and Van Roy show that Thompson Sampling achieves regret bounds of the form $O(\sqrt{T})$ in a variety of situations including those we consider here [Russo and Van Roy '16].

What Was Known?

	Full Feedback	Bandit	Semibandit
Regret	$\sqrt{T \log n}$	$\sqrt{nT}$	$\sqrt{nmT}$
Small Loss	$\sqrt{L^\star \log n}$	$\sqrt{nL^\star \log n}$	$\tilde{O}(\sqrt{L^\star(m^3+n)\log(T)})$
Thompson	$\sqrt{TH(\vec{p}_0)}$	$\tilde{O}(\sqrt{nT})$	$\tilde{O}(\sqrt{nmT}) \quad (*)$

- Optimal $O(\sqrt{T})$ regret bounds are well established. E.g Multiplicative weights, online mirror descent.
- Multiplicative weights gives $O(\sqrt{L^\star})$ small loss bound in the full feedback setting. In the bandit/semibandit cases, modify to never play bad actions ([Lykouris, Sridharan, and Tardos '18]).
- Russo and Van Roy show that Thompson Sampling achieves regret bounds of the form $O(\sqrt{T})$ in a variety of situations including those we consider here [Russo and Van Roy '16].
- (*) Thompson Sampling for semibandits was only analyzed when different coordinate losses are independent of each other.

What's New?

	Full Feedback	Bandit	Semibandit
Regret	$\sqrt{T \log n}$	$\sqrt{nT}$	$\sqrt{nmT}$
Small Loss	$\sqrt{L^\star \log n}$	$\sqrt{nL^\star \log n}$	$\tilde{O}(\sqrt{L^\star(m^3+n)\log(T)})$
Thompson	$\sqrt{TH(\vec{p}_0)}$	$\tilde{O}(\sqrt{nT})$	$\tilde{O}(\sqrt{nmT})$
Thompson	$\sqrt{L^\star H(\vec{p}_0)}$	$\tilde{O}(\sqrt{nL^\star})$	$\tilde{O}(\sqrt{nL^\star})$

What's New?

	Full Feedback	Bandit	Semibandit
Regret	$\sqrt{T \log n}$	$\sqrt{nT}$	$\sqrt{nmT}$
Small Loss	$\sqrt{L^\star \log n}$	$\sqrt{nL^\star \log n}$	$\tilde{O}(\sqrt{L^\star(m^3+n)\log(T)})$
Thompson	$\sqrt{TH(\vec{p}_0)}$	$\tilde{O}(\sqrt{nT})$	$\tilde{O}(\sqrt{nmT})$
Thompson	$\sqrt{L^\star H(\vec{p}_0)}$	$\tilde{O}(\sqrt{nL^\star})$	$\tilde{O}(\sqrt{nL^\star})$

- We prove analogous small loss regret bounds for Thompson Sampling. The first two cases match existing tight upper bounds while the semi-bandit upper bound is the best known.

What's New?

	Full Feedback	Bandit	Semibandit
Regret	$\sqrt{T \log n}$	$\sqrt{nT}$	$\sqrt{nmT}$
Small Loss	$\sqrt{L^\star \log n}$	$\sqrt{nL^\star \log n}$	$\tilde{O}(\sqrt{L^\star(m^3+n)\log(T)})$
Thompson	$\sqrt{TH(\vec{p}_0)}$	$\tilde{O}(\sqrt{nT})$	$\tilde{O}(\sqrt{nmT})$
Thompson	$\sqrt{L^\star H(\vec{p}_0)}$	$\tilde{O}(\sqrt{nL^\star})$	$\tilde{O}(\sqrt{nL^\star})$

- We prove analogous small loss regret bounds for Thompson Sampling. The first two cases match existing tight upper bounds while the semi-bandit upper bound is the best known.
- We also address the semibandit case with arbitrary priors.

What's New?

	Full Feedback	Bandit	Semibandit
Regret	$\sqrt{T \log n}$	$\sqrt{nT}$	$\sqrt{nmT}$
Small Loss	$\sqrt{L^\star \log n}$	$\sqrt{nL^\star \log n}$	$\tilde{O}(\sqrt{L^\star(m^3+n)}\log(T))$
Thompson	$\sqrt{TH(\vec{p}_0)}$	$\tilde{O}(\sqrt{nT})$	$\tilde{O}(\sqrt{nmT})$
Thompson	$\sqrt{L^\star H(\vec{p}_0)}$	$\tilde{O}(\sqrt{nL^\star})$	$\tilde{O}(\sqrt{nL^\star})$

- We prove analogous small loss regret bounds for Thompson Sampling. The first two cases match existing tight upper bounds while the semi-bandit upper bound is the best known.
- We also address the semibandit case with arbitrary priors.
- To achieve full T-independence in the bandit/semibandit settings, we have to modify Thompson Sampling by never playing low probability actions. Otherwise upper bounds have $\log(T)$ terms.

Information Theoretic Analysis for Full Feedback

- Denote the regret incurred in timestep t by

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

Information Theoretic Analysis for Full Feedback

- Denote the regret incurred in timestep t by
$$r_t = \ell_t(i_t) - \ell_t(i^\star).$$
- By linearity, $\mathbb{E}[\sum_{t\leq T} r_t] = \mathbb{E}[R_T]$ is the expected total regret.

Information Theoretic Analysis for Full Feedback

- Denote the regret incurred in timestep t by

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

- By linearity, $\mathbb{E}[\sum_{t \le T} r_t] = \mathbb{E}[R_T]$ is the expected total regret.
- Claim: $\mathbb{E}^{p_t}[r_t]$ is given by

$$\mathbb{E}^{p_t}[r_t] = \mathbb{E}^{p_t}\left[\sum_i \ell_t(i) \cdot (p_t(i) - p_{t+1}(i))\right].$$

Information Theoretic Analysis for Full Feedback

- Denote the regret incurred in timestep t by

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

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

$$\mathbb{E}^{p_t}[r_t] = \mathbb{E}^{p_t}\left[\sum_i \ell_t(i) \cdot (p_t(i) - p_{t+1}(i))\right].$$

- Proof: the expected loss from round t is $\sum_i p_t(i)\mathbb{E}[\ell_t(i)]$ while the in-hindsight expected loss of the best action from that round is $\mathbb{E}^{p_t}[\sum_i p_{t+1}(i)\ell_t(i)]$.

Information Theoretic Analysis for Full Feedback

- Denote the regret incurred in timestep t by

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

- By linearity, $\mathbb{E}[\sum_{t \le T} r_t] = \mathbb{E}[R_T]$ is the expected total regret.
- Claim: $\mathbb{E}^{p_t}[r_t]$ is given by

$$\mathbb{E}^{p_t}[r_t] = \mathbb{E}^{p_t}\left[\sum_i \ell_t(i) \cdot (p_t(i) - p_{t+1}(i))\right].$$

- Proof: the expected loss from round t is $\sum_i p_t(i)\mathbb{E}[\ell_t(i)]$ while the in-hindsight expected loss of the best action from that round is $\mathbb{E}^{p_t}[\sum_i p_{t+1}(i)\ell_t(i)]$.
- As $\ell_t(i) \in [0,1]$, we obtain

$$\mathbb{E}[r_t] \le \frac{1}{2}\mathbb{E}[||\vec{p}_t - \vec{p}_{t+1}||_{\ell^1}].$$

Information Theoretic Analysis for Full Feedback

- Denote the regret incurred in timestep t by

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

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

$$\mathbb{E}^{p_t}[r_t] = \mathbb{E}^{p_t}\left[\sum_i \ell_t(i) \cdot (p_t(i) - p_{t+1}(i))\right].$$

- Proof: the expected loss from round t is $\sum_i p_t(i)\mathbb{E}[\ell_t(i)]$ while the in-hindsight expected loss of the best action from that round is $\mathbb{E}^{p_t}[\sum_i p_{t+1}(i)\ell_t(i)]$.
- As $\ell_t(i) \in [0,1]$, we obtain

$$\mathbb{E}[r_t] \leq \frac{1}{2}\mathbb{E}[||\vec{p}_t - \vec{p}_{t+1}||_{\ell^1}].$$

- New goal: estimate ℓ^1 movement of $\vec{p}_t$.

Information Theoretic Analysis for Full-Feedback

- Now the information theory appears. Review:

Information Theoretic Analysis for Full-Feedback

- Now the information theory appears. Review:
- Entropy is $H(\vec{p}) = -\sum_i p(i) \log p(i)$.

Information Theoretic Analysis for Full-Feedback

- Now the information theory appears. Review:
- Entropy is $H(\vec{p}) = -\sum_i p(i) \log p(i)$.
- The KL divergence or *relative entropy* between distributions $\vec{q}, \vec{p}$ is $Ent[\vec{q}; \vec{p}] = \sum_i q(i) \log\left(\frac{q(i)}{p(i)}\right)$.

Information Theoretic Analysis for Full-Feedback

- Now the information theory appears. Review:
- Entropy is $H(\vec{p}) = -\sum_i p(i) \log p(i)$.
- The KL divergence or *relative entropy* between distributions $\vec{q}, \vec{p}$ is $Ent[\vec{q}; \vec{p}] = \sum_i q(i) \log \left(\frac{q(i)}{p(i)} \right)$.
- Since $\mathbb{E}^{p_t}[\vec{p}_{t+1}] = \vec{p}_t$, we have

$$\mathbb{E}^{p_t}[H(\vec{p}_t) - H(\vec{p}_{t+1})] = \mathbb{E}^{p_t}[Ent[\vec{p}_{t+1}; \vec{p}_t]].$$

Information Theoretic Analysis for Full-Feedback

- Now the information theory appears. Review:
- Entropy is $H(\vec{p}) = -\sum_i p(i) \log p(i)$.
- The KL divergence or *relative entropy* between distributions $\vec{q}, \vec{p}$ is $Ent[\vec{q}; \vec{p}] = \sum_i q(i) \log \left(\frac{q(i)}{p(i)} \right)$.
- Since $\mathbb{E}^{p_t}[\vec{p}_{t+1}] = \vec{p}_t$, we have

$$\mathbb{E}^{p_t}[H(\vec{p}_t) - H(\vec{p}_{t+1})] = \mathbb{E}^{p_t}[Ent[\vec{p}_{t+1}; \vec{p}_t]].$$

- Expected information on $i^\star$ is also the mutual information

$$I_t := I_t[i^\star, (i_t, \ell_t(i_t))] = \mathbb{E}^{p_t}[H(\vec{p}_{t+1}) - H(\vec{p}_t)] = \mathbb{E}^{p_t}[Ent[\vec{p}_{t+1}; \vec{p}_t]].$$

Information Theoretic Analysis for Full-Feedback

- Now the information theory appears. Review:
- Entropy is $H(\vec{p}) = -\sum_i p(i) \log p(i)$.
- The KL divergence or *relative entropy* between distributions $\vec{q}, \vec{p}$ is $Ent[\vec{q}; \vec{p}] = \sum_i q(i) \log\left(\frac{q(i)}{p(i)}\right)$.
- Since $\mathbb{E}^{p_t}[\vec{p}_{t+1}] = \vec{p}_t$, we have

$$\mathbb{E}^{p_t}[H(\vec{p}_t) - H(\vec{p}_{t+1})] = \mathbb{E}^{p_t}[Ent[\vec{p}_{t+1}; \vec{p}_t]].$$

- Expected information on $i^\star$ is also the mutual information

$$I_t := I_t[i^\star, (i_t, \ell_t(i_t))] = \mathbb{E}^{p_t}[H(\vec{p}_{t+1}) - H(\vec{p}_t)] = \mathbb{E}^{p_t}[Ent[\vec{p}_{t+1}; \vec{p}_t]].$$

- Pinsker's Inequality controls the movement:

$$||\vec{p}_t - \vec{p}_{t+1}||_{\ell^1}^2 \leq 2 \cdot Ent[\vec{p}_{t+1}; \vec{p}_t]$$

Information Theoretic Analysis for Full-Feedback

- Goal: estimate

$$\mathbb{E}\left[\sum_{t=0}^{T-1} ||\vec{p}_t - \vec{p}_{t+1}||_{\ell^1}\right].$$

We will just use that $(\vec{p}_t)_{t\geq 0}$ is a martingale in the simplex.

Information Theoretic Analysis for Full-Feedback

- Goal: estimate
$$\mathbb{E}\left[\sum_{t=0}^{T-1} ||\vec{p}_t - \vec{p}_{t+1}||_{\ell^1}\right].$$
We will just use that $(\vec{p}_t)_{t\geq 0}$ is a martingale in the simplex.
- Cauchy-Schwarz:

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

Information Theoretic Analysis for Full-Feedback

- Goal: estimate

$$\mathbb{E}\left[\sum_{t=0}^{T-1} ||\vec{p}_t - \vec{p}_{t+1}||_{\ell^1}\right].$$

We will just use that $(\vec{p}_t)_{t\geq 0}$ is a martingale in the simplex.

- Cauchy-Schwarz:

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

- Pinsker:

$$\mathbb{E}\left[\sum_{t=0}^{T-1} ||\vec{p}_t - \vec{p}_{t+1}||_{\ell^1}^2\right] \leq 2 \cdot \mathbb{E}\left[\sum_{t\geq 0} Ent[\vec{p}_{t+1}; \vec{p}_t]\right] \leq 2H(\vec{p}_0) \leq 2\log(n).$$

Information Theoretic Analysis for Full-Feedback

Theorem [Russo and Van Roy '16]

Thompson Sampling gives expected regret $\mathbb{E}[R_T] \leq \sqrt{\frac{T \cdot H(\vec{p}_0)}{2}} \leq \sqrt{\frac{T \log(n)}{2}}$.

Information Theoretic Analysis for Full-Feedback

Theorem [Russo and Van Roy '16]

Thompson Sampling gives expected regret
$\mathbb{E}[R_T] \le \sqrt{\frac{T \cdot H(\vec{p}_0)}{2}} \le \sqrt{\frac{T \log(n)}{2}}$.

- Another way to understand this:

Information Ratio

For time-step t, define the *information ratio* between squared regret and information obtained to be

$$\Gamma_t := \frac{\mathbb{E}^{p_t}[r_t]^2}{I_t(i^\star, (i_t, \ell_t(i_t)))}.$$

Here the mutual information I_t is the average amount of new information we obtain about the best coordinate $i^\star$ on round t.

Information Theoretic Analysis for Full-Feedback

Information Ratio

For time-step t, define the *information ratio* between squared regret and information obtained to be

$$\Gamma_t := \frac{\mathbb{E}^{p_t}[r_t]^2}{I_t(i^\star, (i_t, \ell_t(i_t)))}.$$

Here the mutual information I_t is the average amount of new information we obtain about the best coordinate $i^\star$ on round t.

Information Theoretic Analysis for Full-Feedback

Information Ratio

For time-step t, define the *information ratio* between squared regret and information obtained to be

$$\Gamma_t := \frac{\mathbb{E}^{p_t}[r_t]^2}{I_t(i^\star, (i_t, \ell_t(i_t)))}.$$

Here the mutual information I_t is the average amount of new information we obtain about the best coordinate $i^\star$ on round t.

- The goal is to upper-bound Γ_t; then regret implies learning.
- Pinsker tells us that $\Gamma_t \leq \frac{1}{2}$ for Thompson Sampling.

Information Theoretic Analysis for Full-Feedback

Information Ratio

For time-step t, define the *information ratio* between squared regret and information obtained to be

$$\Gamma_t := \frac{\mathbb{E}^{p_t}[r_t]^2}{I_t(i^\star, (i_t, \ell_t(i_t)))}.$$

Here the mutual information I_t is the average amount of new information we obtain about the best coordinate $i^\star$ on round t.

- The goal is to upper-bound Γ_t; then regret implies learning.
- Pinsker tells us that $\Gamma_t \leq \frac{1}{2}$ for Thompson Sampling.
- In general, if $\mathbb{E}[\Gamma_t] \leq a_t$, we obtain:

$$\mathbb{E}[R_T]^2 = \left(\mathbb{E}\left[\sum_t \mathbb{E}^{p_t}[r_t]\right]\right)^2 \leq \mathbb{E}\left[\sum_t I_t\right]\mathbb{E}\left[\sum_t \Gamma_t\right] \leq H(\vec{p}_0)\sum_t a_t$$

Obtaining a Small Loss Bound

Theorem

Thompson Sampling satisfies

$$\mathbb{E}[R_T] = O(\sqrt{\mathbb{E}[L^\star]H(\vec{p}_0)}).$$

- To prove this, we could try to prove $\Gamma_t = O(\bar{\ell}_t)$ where $\bar{\ell}_t = \mathbb{E}^{p_t}[\ell_t(i_t)] = \sum_i p_t(i)\bar{\ell}_t(i)$. Then $\mathbb{E}[\Gamma_t] = O(\mathbb{E}^{p_0}[\ell_t])$ and we'd obtain:

$$\mathbb{E}[L_T - L^\star] = \mathbb{E}[R_T] \leq O\left(\sqrt{H(\vec{p}_0)\sum_t \mathbb{E}^{p_0}[\ell_t]}\right) = O\left(\sqrt{H(\vec{p}_0)\mathbb{E}[L_T]}\right).$$

- Simple algebra would now yield the result.

Obtaining a Small-Loss Bound

- We try to prove $\Gamma_t = O(\overline{\ell}_t)$:

Obtaining a Small-Loss Bound

- We try to prove $\Gamma_t = O(\overline{\ell}_t)$:

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

Obtaining a Small-Loss Bound

- We try to prove $\Gamma_t = O(\bar{\ell}_t)$:

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

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

Obtaining a Small-Loss Bound

- We try to prove $\Gamma_t = O(\overline{\ell}_t)$:

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

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

$$= 2\overline{\ell}_t \cdot \mathbb{E}^{p_t}[\chi^2[\vec{p}_{t+1}; \vec{p}_t]].$$

Obtaining a Small-Loss Bound

- We try to prove $\Gamma_t = O(\overline{\ell}_t)$:

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

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

$$= 2\overline{\ell}_t \cdot \mathbb{E}^{p_t}[\chi^2[\vec{p}_{t+1}; \vec{p}_t]].$$

- Does $\chi^2[p_{t+1}; p_t] \leq Ent[p_{t+1}; p_t]$ hold? Then we'd be done.

Obtaining a Small-Loss Bound

- We try to prove $\Gamma_t = O(\bar{\ell}_t)$:

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

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

$$= 2\bar{\ell}_t \cdot \mathbb{E}^{p_t}[\chi^2[\vec{p}_{t+1}; \vec{p}_t]].$$

- Does $\chi^2[p_{t+1}; p_t] \leq Ent[p_{t+1}; p_t]$ hold? Then we'd be done.
- No! In fact $\chi^2[p_{t+1}; p_t] \geq 2Ent[p_{t+1}; p_t]$ always holds.

Obtaining a Small-Loss Bound

Lemma

For any probability distributions p_{t+1}, p_t we have

$$\sum_{i: p_t(i) \geq p_{t+1}(i)} \frac{(p_t(i) - p_{t+1}(i))^2}{2p_t(i)} \leq Ent[p_{t+1}; p_t].$$

Obtaining a Small-Loss Bound

Lemma

For any probability distributions p_{t+1}, p_t we have

$$\sum_{i:p_t(i)\geq p_{t+1}(i)} \frac{(p_t(i)-p_{t+1}(i))^2}{2p_t(i)} \leq Ent[p_{t+1};p_t].$$

- Proof: second order Taylor expansion in $\vec{q}$ for KL divergence

$$Ent[\vec{q};\vec{p}] = \sum_i q(i)\log\left(\frac{q(i)}{p(i)}\right).$$

Obtaining a Small-Loss Bound

Lemma

For any probability distributions p_{t+1}, p_t we have

$$\sum_{i:p_t(i)\geq p_{t+1}(i)} \frac{(p_t(i)-p_{t+1}(i))^2}{2p_t(i)} \leq Ent[p_{t+1}; p_t].$$

- Proof: second order Taylor expansion in $\vec{q}$ for KL divergence

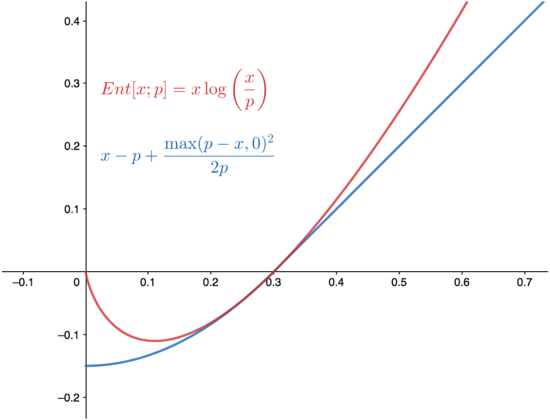
$$Ent[\vec{q};\vec{p}] = \sum_i q(i)\log\left(\frac{q(i)}{p(i)}\right).$$

- The first-order terms cancel because p_t, p_{t+1} are both probability distributions. The second-order derivatives are given by

$$\partial^2_{q(i)}\left[q(i)\log\left(\frac{q(i)}{p(i)}\right)\right] = \frac{1}{q(i)}.$$

Obtaining a Small-Loss Bound

- Since the 2nd derivative is decreasing and positive, lower bound $x \log\left(\frac{x}{p(i)}\right)$ quadratically for $x < p(i)$ and linearly for $x > p(i)$.



Obtaining a Small-Loss Bound

- Terms with $p_{t+1}(i) \geq p_t(i)$ only reduce the regret, so we ignore them.

Obtaining a Small-Loss Bound

- Terms with $p_{t+1}(i) \geq p_t(i)$ only reduce the regret, so we ignore them.
- Putting it together:

$$\mathbb{E}^{p_t}[r_t] = \left(\mathbb{E}^{p_t} \sum_i \bar{\ell}_t(i)(p_t(i) - p_{t+1}(i)) \right)$$

Obtaining a Small-Loss Bound

- Terms with $p_{t+1}(i) \geq p_t(i)$ only reduce the regret, so we ignore them.
- Putting it together:

$$\mathbb{E}^{p_t}[r_t] = \left(\mathbb{E}^{p_t} \sum_i \bar{\ell}_t(i)(p_t(i) - p_{t+1}(i)) \right)$$

$$\leq \left(\mathbb{E}^{p_t} \sum_{i: p_t(i) \geq p_{t+1}(i)} \ell_t(i)(p_t(i) - p_{t+1}(i)) \right)$$

Obtaining a Small-Loss Bound

- Terms with $p_{t+1}(i) \geq p_t(i)$ only reduce the regret, so we ignore them.
- Putting it together:

$$\mathbb{E}^{p_t}[r_t] = \left(\mathbb{E}^{p_t} \sum_i \bar{\ell}_t(i)(p_t(i) - p_{t+1}(i)) \right)$$

$$\leq \left(\mathbb{E}^{p_t} \sum_{i: p_t(i) \geq p_{t+1}(i)} \ell_t(i)(p_t(i) - p_{t+1}(i)) \right)$$

$$\leq \left(\mathbb{E}^{p_t} \sum_i p_t(i) \ell_t(i)^2 \right)^{1/2} \left(\mathbb{E}^{p_t} \sum_{i: p_t(i) \geq p_{t+1}(i)} \frac{(p_t(i) - p_{t+1}(i))^2}{p_t(i)} \right)^{1/2}$$

Obtaining a Small-Loss Bound

- Terms with $p_{t+1}(i) \geq p_t(i)$ only reduce the regret, so we ignore them.
- Putting it together:

$$\mathbb{E}^{p_t}[r_t] = \left(\mathbb{E}^{p_t} \sum_i \bar{\ell}_t(i)(p_t(i) - p_{t+1}(i))\right)$$

$$\leq \left(\mathbb{E}^{p_t} \sum_{i: p_t(i) \geq p_{t+1}(i)} \ell_t(i)(p_t(i) - p_{t+1}(i))\right)$$

$$\leq \left(\mathbb{E}^{p_t} \sum_i p_t(i)\ell_t(i)^2\right)^{1/2} \left(\mathbb{E}^{p_t} \sum_{i: p_t(i) \geq p_{t+1}(i)} \frac{(p_t(i) - p_{t+1}(i))^2}{p_t(i)}\right)^{1/2}$$

$$\leq \sqrt{\bar{\ell}_t \cdot \mathbb{E}[Ent[\vec{p}_{t+1}; \vec{p}_t]]} \leq \sqrt{\bar{\ell}_t \cdot I_t}.$$

Obtaining a Small-Loss Bound

Theorem

Thompson Sampling satisfies

$$\frac{\mathbb{E}[(r_t)_+]^2}{I_t(i^\star)} \leq \bar{\ell}_t.$$

Obtaining a Small-Loss Bound

Theorem

Thompson Sampling satisfies

$$\frac{\mathbb{E}[(r_t)_+]^2}{I_t(i^\star)} \leq \bar{\ell}_t.$$

Theorem

Thompson Sampling satisfies

$$\mathbb{E}[R_T] = O(\sqrt{\mathbb{E}[L^\star] H(\vec{p}_0)}).$$

Obtaining a Small-Loss Bound

Theorem

Thompson Sampling satisfies

$$\frac{\mathbb{E}[(r_t)_+]^2}{I_t(i^\star)} \leq \bar{\ell}_t.$$

Theorem

Thompson Sampling satisfies

$$\mathbb{E}[R_T] = O(\sqrt{\mathbb{E}[L^\star]H(\vec{p}_0)}).$$

Remark

We can also show that $\Gamma_t = O(\bar{\ell}_t + \bar{\ell}_t^\star)$ by using a cruder entropy inequality.

The Bandit Case

- In the bandit case, the expected regret for time t is

$$\mathbb{E}[r_t] = \sum_i p_t(i) \cdot (\overline{\ell}_t(i) - \overline{\ell}_t(i,i))$$

where $\overline{\ell}_t(i,i) = \mathbb{E}[\ell_t(i) | i^\star = i]$.

The Bandit Case

- In the bandit case, the expected regret for time t is

$$\mathbb{E}[r_t] = \sum_i p_t(i) \cdot (\bar{\ell}_t(i) - \bar{\ell}_t(i,i))$$

where $\bar{\ell}_t(i,i) = \mathbb{E}[\ell_t(i) | i^\star = i]$.

- The information gain $I_t(i^\star)$ is lower-bounded by

$$I_t(i^\star) \geq \sum_i p_t(i)^2 Ent[(\ell_t(i) | i^\star = i); \ell_t(i)].$$

This is the information gain when $i^\star = i_t = i$.

The Bandit Case

- In the bandit case, the expected regret for time t is

$$\mathbb{E}[r_t] = \sum_i p_t(i) \cdot (\bar{\ell}_t(i) - \bar{\ell}_t(i,i))$$

  where $\bar{\ell}_t(i,i) = \mathbb{E}[\ell_t(i) | i^\star = i]$.

- The information gain $I_t(i^\star)$ is lower-bounded by

$$I_t(i^\star) \geq \sum_i p_t(i)^2 Ent[(\ell_t(i) | i^\star = i); \ell_t(i)].$$

  This is the information gain when $i^\star = i_t = i$.

- A similar Pinsker's inequality argument shows that the expected regret is $O(\sqrt{nT \cdot H(\vec{p}_0)})$. Optimal is $O(\sqrt{nT})$ and most methods (e.g. multiplicative weights) give $O(\sqrt{nT \log(n)})$.

The Bandit Case

- Ordinary regret estimate:

$$\mathbb{E}^{p_t}[r_t]^2 = \left(\mathbb{E}^{p_t}\left[\sum_i p_t(i)\cdot(\bar{\ell}_t(i) - \bar{\ell}_t(i,i))\right]\right)^2$$

The Bandit Case

- Ordinary regret estimate:

$$\mathbb{E}^{p_t}[r_t]^2 = \left(\mathbb{E}^{p_t}\left[\sum_i p_t(i) \cdot (\bar{\ell}_t(i) - \bar{\ell}_t(i,i)\right]\right)^2$$

$$\leq \left(\mathbb{E}^{p_t}\sum_i 1\right)\left(\mathbb{E}^{p_t}\sum_i p_t(i)^2 \cdot (\bar{\ell}_t(i) - \bar{\ell}_t(i,i))^2\right)$$

The Bandit Case

- Ordinary regret estimate:

$$\mathbb{E}^{p_t}[r_t]^2 = \left(\mathbb{E}^{p_t}\left[\sum_i p_t(i)\cdot(\bar{\ell}_t(i) - \bar{\ell}_t(i,i)\right]\right)^2$$

$$\leq \left(\mathbb{E}^{p_t}\sum_i 1\right)\left(\mathbb{E}^{p_t}\sum_i p_t(i)^2\cdot(\bar{\ell}_t(i) - \bar{\ell}_t(i,i))^2\right)$$

$$\leq n\cdot\mathbb{E}^{p_t}\left[\sum_i p_t(i)^2 Ent[\ell_t(i,i);\ell(i)]\right] \leq nI_t(i^\star).$$

The Bandit Case

- Ordinary regret estimate:

$$\mathbb{E}^{p_t}[r_t]^2 = \left(\mathbb{E}^{p_t}\left[\sum_i p_t(i)\cdot(\bar{\ell}_t(i) - \bar{\ell}_t(i,i)\right]\right)^2$$

$$\leq \left(\mathbb{E}^{p_t}\sum_i 1\right)\left(\mathbb{E}^{p_t}\sum_i p_t(i)^2\cdot(\bar{\ell}_t(i) - \bar{\ell}_t(i,i))^2\right)$$

$$\leq n\cdot\mathbb{E}^{p_t}\left[\sum_i p_t(i)^2 Ent[\ell_t(i,i);\ell(i)]\right] \leq nI_t(i^\star).$$

- Hence $\Gamma_t \leq n$.

The Bandit Case

- Ordinary regret estimate:

$$\mathbb{E}^{p_t}[r_t]^2 = \left(\mathbb{E}^{p_t}\left[\sum_i p_t(i)\cdot(\bar{\ell}_t(i) - \bar{\ell}_t(i,i)\right]\right)^2$$

$$\leq \left(\mathbb{E}^{p_t}\sum_i 1\right)\left(\mathbb{E}^{p_t}\sum_i p_t(i)^2\cdot(\bar{\ell}_t(i) - \bar{\ell}_t(i,i))^2\right)$$

$$\leq n\cdot\mathbb{E}^{p_t}\left[\sum_i p_t(i)^2 Ent[\ell_t(i,i);\ell(i)]\right] \leq nI_t(i^\star).$$

- Hence $\Gamma_t \leq n$.

Theorem [Russo and Van Roy '16]

Thompson Sampling for bandits satisfies $\mathbb{E}[R_T] = O(\sqrt{nT\cdot H(\vec{p}_0)})$.

The Bandit Case

- For small-loss bound, we assume $L^\star$ is given (or bounded) and aim for $\mathbb{E}[R_T] = \tilde{O}(\sqrt{nL^\star})$. A first attempt:

The Bandit Case

- For small-loss bound, we assume $L^\star$ is given (or bounded) and aim for $\mathbb{E}[R_T] = \tilde{O}(\sqrt{nL^\star})$. A first attempt:

$$\mathbb{E}[R_T]^2 = \left(\mathbb{E}\left[\sum_{i,t} p_t(i) \cdot (\bar{\ell}_t(i) - \bar{\ell}_t(i,i))\right]\right)^2$$

The Bandit Case

- For small-loss bound, we assume $L^\star$ is given (or bounded) and aim for $\mathbb{E}[R_T] = \tilde{O}(\sqrt{nL^\star})$. A first attempt:

$$\mathbb{E}[R_T]^2 = \left(\mathbb{E}\left[\sum_{i,t} p_t(i) \cdot (\bar{\ell}_t(i) - \bar{\ell}_t(i,i))\right]\right)^2$$

$$\leq \left(\mathbb{E}\sum_{i,t} \ell_t(i)\right)\left(\mathbb{E}\sum_{i,t} p_t(i)^2 \cdot \frac{(\bar{\ell}_t(i) - \bar{\ell}_t(i,i))^2}{\ell_t(i)}\right)$$

The Bandit Case

- For small-loss bound, we assume $L^\star$ is given (or bounded) and aim for $\mathbb{E}[R_T] = \tilde{O}(\sqrt{nL^\star})$. A first attempt:

$$\mathbb{E}[R_T]^2 = \left(\mathbb{E}\left[\sum_{i,t} p_t(i) \cdot (\bar{\ell}_t(i) - \bar{\ell}_t(i,i))\right]\right)^2$$

$$\leq \left(\mathbb{E}\sum_{i,t} \ell_t(i)\right)\left(\mathbb{E}\sum_{i,t} p_t(i)^2 \cdot \frac{(\bar{\ell}_t(i) - \bar{\ell}_t(i,i))^2}{\ell_t(i)}\right)$$

$$\leq \left(\mathbb{E}\sum_{i,t} \bar{\ell}_t(i)\right) H(\vec{p}_0).$$

The Bandit Case

- For small-loss bound, we assume $L^\star$ is given (or bounded) and aim for $\mathbb{E}[R_T] = \tilde{O}(\sqrt{nL^\star})$. A first attempt:

$$\mathbb{E}[R_T]^2 = \left(\mathbb{E}\left[\sum_{i,t} p_t(i) \cdot (\bar{\ell}_t(i) - \bar{\ell}_t(i,i))\right]\right)^2$$

$$\leq \left(\mathbb{E}\sum_{i,t} \ell_t(i)\right)\left(\mathbb{E}\sum_{i,t} p_t(i)^2 \cdot \frac{(\bar{\ell}_t(i) - \bar{\ell}_t(i,i))^2}{\ell_t(i)}\right)$$

$$\leq \left(\mathbb{E}\sum_{i,t} \bar{\ell}_t(i)\right) H(\vec{p}_0).$$

- If the player could track this sum, then $p_t(i) = 0$ once $\sum_{s \leq t} \ell_s(i) \geq L^\star$. After that, ignore coordinate i. This would result in $\mathbb{E}[R_T] \leq \sqrt{nL^\star \cdot H(\vec{p}_0)}$.

The Bandit Case

- Actually Thompson Sampling does NOT give T-independent small-loss regret for bandits.

The Bandit Case

- Actually Thompson Sampling does NOT give T-independent small-loss regret for bandits.
- If T is VERY large (think Tower($L^\star$)), turns out you can force Thompson Sampling to keep playing every action until it is 100% sure that this action had loss at least $L^\star$.

The Bandit Case

- Actually Thompson Sampling does NOT give T-independent small-loss regret for bandits.
- If T is VERY large (think Tower($L^\star$)), turns out you can force Thompson Sampling to keep playing every action until it is 100% sure that this action had loss at least $L^\star$.
- Hence you will eventually pay $nL^\star$ total cost. Not good!

The Bandit Case

- Actually Thompson Sampling does NOT give T-independent small-loss regret for bandits.
- If T is VERY large (think Tower($L^\star$)), turns out you can force Thompson Sampling to keep playing every action until it is 100% sure that this action had loss at least $L^\star$.
- Hence you will eventually pay $nL^\star$ total cost. Not good!

Thresholded Thompson Sampling

In Thresholded Thompson Sampling, we pick small $\gamma > 0$ and play from p_t but restrict to actions i with probability $p_t(i) \geq \gamma$ to be optimal.

The Bandit Case

- Actually Thompson Sampling does NOT give T-independent small-loss regret for bandits.
- If T is VERY large (think Tower($L^\star$)), turns out you can force Thompson Sampling to keep playing every action until it is 100% sure that this action had loss at least $L^\star$.
- Hence you will eventually pay $nL^\star$ total cost. Not good!

Thresholded Thompson Sampling

In Thresholded Thompson Sampling, we pick small $\gamma > 0$ and play from p_t but restrict to actions i with probability $p_t(i) \geq \gamma$ to be optimal.

- Thresholded Thompson Sampling avoids the bad case above. It also parallels the thresholded EXP3 algorithm which circumvents the same issue for multiplicative weights.

The Bandit Case

- Threshold with $\gamma = \frac{1}{L^\star}$.

The Bandit Case

- Threshold with $\gamma = \frac{1}{L^\star}$.
- Redo the algebra, but no real change. Main thing is to estimate:

$$\mathbb{E}\left[\sum_{i,t:p_t(i)\geq\gamma} \ell_t(i)\right].$$

The Bandit Case

- Threshold with $\gamma = \frac{1}{L^\star}$.
- Redo the algebra, but no real change. Main thing is to estimate:

$$\mathbb{E}\left[\sum_{i,t:p_t(i)\geq\gamma} \ell_t(i)\right].$$

- Claim: for each i, we have $\mathbb{E}\left[\sum_{t:p_t(i)\geq\gamma} \ell_t(i)\right] \leq L^\star + \tilde{O}(\sqrt{L^\star/\gamma})$.

The Bandit Case

- Threshold with $\gamma = \frac{1}{L^\star}$.
- Redo the algebra, but no real change. Main thing is to estimate:

$$\mathbb{E}\left[\sum_{i,t:p_t(i)\geq\gamma} \ell_t(i)\right].$$

- Claim: for each i, we have $\mathbb{E}\left[\sum_{t:p_t(i)\geq\gamma} \ell_t(i)\right] \leq L^\star + \tilde{O}(\sqrt{L^\star/\gamma})$.
- From observing $\ell_t(i_t)$, we can compute an unbiased estimator for this sum via importance sampling; when you play an i_t which had probability $p_t(i_t)$ you should count the loss as $\tilde{\ell}_t(i_t) = \frac{\ell_t(i_t)}{p_t(i_t)}$.

The Bandit Case

- Threshold with $\gamma = \frac{1}{L^\star}$.
- Redo the algebra, but no real change. Main thing is to estimate:

$$\mathbb{E}\left[\sum_{i,t:p_t(i)\geq\gamma} \ell_t(i)\right].$$

- Claim: for each i, we have $\mathbb{E}\left[\sum_{t:p_t(i)\geq\gamma} \ell_t(i)\right] \leq L^\star + \tilde{O}(\sqrt{L^\star/\gamma})$.
- From observing $\ell_t(i_t)$, we can compute an unbiased estimator for this sum via importance sampling; when you play an i_t which had probability $p_t(i_t)$ you should count the loss as $\tilde{\ell}_t(i_t) = \frac{\ell_t(i_t)}{p_t(i_t)}$.
- Since $p_t(i) \geq \gamma$ the unbiased estimator is a sum of bounded random variables in $[0, \frac{1}{\gamma}]$, so it is concentrated near the true value.

The Bandit Case

- Threshold with $\gamma = \frac{1}{L^\star}$.
- Redo the algebra, but no real change. Main thing is to estimate:

$$\mathbb{E}\left[\sum_{i,t:p_t(i)\geq\gamma} \ell_t(i)\right].$$

- Claim: for each i, we have $\mathbb{E}\left[\sum_{t:p_t(i)\geq\gamma} \ell_t(i)\right] \leq L^\star + \tilde{O}(\sqrt{L^\star/\gamma})$.
- From observing $\ell_t(i_t)$, we can compute an unbiased estimator for this sum via importance sampling; when you play an i_t which had probability $p_t(i_t)$ you should count the loss as $\tilde{\ell}_t(i_t) = \frac{\ell_t(i_t)}{p_t(i_t)}$.
- Since $p_t(i) \geq \gamma$ the unbiased estimator is a sum of bounded random variables in $[0, \frac{1}{\gamma}]$, so it is concentrated near the true value.
- Hence for each i, the player has a good estimate for $\sum_{s\leq t:p_s(i)\geq\gamma} \ell_s(i)$. When this sum gets significantly above $L^\star$, $p_t(i)$ will usually be very small.

The Bandit Case

Theorem

Thresholded Thompson Sampling achieves T-independent regret

$$\tilde{O}(\sqrt{nL^\star} + n).$$

The Bandit Case

Theorem

Thresholded Thompson Sampling achieves T-independent regret

$$\tilde{O}(\sqrt{nL^\star} + n).$$

We can also analyze ordinary Thompson Sampling the same way. Separately estimate the expected loss $p_t(i)\ell_t(i)$ for $p_t(i) < \gamma$. When the observed loss exceeds $\tilde{O}(\log T)$, we expect $p_t(i) \leq \frac{1}{T}$. So the total small-probability contribution should be $\tilde{O}(\log T)$ per action.

The Bandit Case

Theorem

Thresholded Thompson Sampling achieves T-independent regret

$$\tilde{O}(\sqrt{nL^\star} + n).$$

We can also analyze ordinary Thompson Sampling the same way. Separately estimate the expected loss $p_t(i)\ell_t(i)$ for $p_t(i) < \gamma$. When the observed loss exceeds $\tilde{O}(\log T)$, we expect $p_t(i) \leq \frac{1}{T}$. So the total small-probability contribution should be $\tilde{O}(\log T)$ per action.

Theorem

Ordinary Thompson Sampling achieves regret

$$\tilde{O}(\sqrt{nL^\star} + n\log(T)).$$

The Semibandit Case

- Now we play a subset $A = (i_1, \ldots, i_m)$ of size m from a given collection $\mathcal{A} \subseteq \binom{[n]}{m}$. We observe and pay all m losses $\ell_t(i_k)$ each turn.

The Semibandit Case

- Now we play a subset $A = (i_1, \ldots, i_m)$ of size m from a given collection $\mathcal{A} \subseteq \binom{[n]}{m}$. We observe and pay all m losses $\ell_t(i_k)$ each turn.
- Example: $\mathcal{A}$ consists of shortest paths in a graph from $u \to v$, the player chooses such a path, each round there is a loss from each edge.

The Semibandit Case

- Now we play a subset $A = (i_1, ..., i_m)$ of size m from a given collection $\mathcal{A} \subseteq \binom{[n]}{m}$. We observe and pay all m losses $\ell_t(i_k)$ each turn.
- Example: $\mathcal{A}$ consists of shortest paths in a graph from $u \to v$, the player chooses such a path, each round there is a loss from each edge.
- Again need to fix $L^\star$.

The Semibandit Case

- Now we play a subset $A = (i_1, ..., i_m)$ of size m from a given collection $\mathcal{A} \subseteq \binom{[n]}{m}$. We observe and pay all m losses $\ell_t(i_k)$ each turn.
- Example: $\mathcal{A}$ consists of shortest paths in a graph from $u \to v$, the player chooses such a path, each round there is a loss from each edge.
- Again need to fix $L^\star$.
- Not obvious what entropy to use. Entropy of $A^\star$ or probabilities $p_t(i \in A^\star)$?

The Semibandit Case

- Now we play a subset $A = (i_1, ..., i_m)$ of size m from a given collection $\mathcal{A} \subseteq \binom{[n]}{m}$. We observe and pay all m losses $\ell_t(i_k)$ each turn.
- Example: $\mathcal{A}$ consists of shortest paths in a graph from $u \to v$, the player chooses such a path, each round there is a loss from each edge.
- Again need to fix $L^\star$.
- Not obvious what entropy to use. Entropy of $A^\star$ or probabilities $p_t(i \in A^\star)$?
- We simply add the entropies for probabilities that the individual coordinates are in $A^\star$:

$$H_t = \sum_i H(p_t(i \in A^\star)).$$

The Semibandit Case

- Now we play a subset $A = (i_1, ..., i_m)$ of size m from a given collection $\mathcal{A} \subseteq \binom{[n]}{m}$. We observe and pay all m losses $\ell_t(i_k)$ each turn.
- Example: $\mathcal{A}$ consists of shortest paths in a graph from $u \to v$, the player chooses such a path, each round there is a loss from each edge.
- Again need to fix $L^\star$.
- Not obvious what entropy to use. Entropy of $A^\star$ or probabilities $p_t(i \in A^\star)$?
- We simply add the entropies for probabilities that the individual coordinates are in $A^\star$:

$$H_t = \sum_i H(p_t(i \in A^\star)).$$

- By doing separate information theory for each coordinate, we can handle arbitrary priors unlike previous analysis.

The Semibandit Case

- Now we play a subset $A = (i_1, \ldots, i_m)$ of size m from a given collection $\mathcal{A} \subseteq \binom{[n]}{m}$. We observe and pay all m losses $\ell_t(i_k)$ each turn.
- Example: $\mathcal{A}$ consists of shortest paths in a graph from $u \to v$, the player chooses such a path, each round there is a loss from each edge.
- Again need to fix $L^\star$.
- Not obvious what entropy to use. Entropy of $A^\star$ or probabilities $p_t(i \in A^\star)$?
- We simply add the entropies for probabilities that the individual coordinates are in $A^\star$:

$$H_t = \sum_i H(p_t(i \in A^\star)).$$

- By doing separate information theory for each coordinate, we can handle arbitrary priors unlike previous analysis.
- Total entropy at most $m \log(n)$. Similarly obtain $\tilde{O}(\sqrt{nmT})$ regret.

The Semibandit Case

- Now we play a subset $A = (i_1, ..., i_m)$ of size m from a given collection $\mathcal{A} \subseteq \binom{[n]}{m}$. We observe and pay all m losses $\ell_t(i_k)$ each turn.
- Example: $\mathcal{A}$ consists of shortest paths in a graph from $u \to v$, the player chooses such a path, each round there is a loss from each edge.
- Again need to fix $L^\star$.
- Not obvious what entropy to use. Entropy of $A^\star$ or probabilities $p_t(i \in A^\star)$?
- We simply add the entropies for probabilities that the individual coordinates are in $A^\star$:

$$H_t = \sum_i H(p_t(i \in A^\star)).$$

- By doing separate information theory for each coordinate, we can handle arbitrary priors unlike previous analysis.
- Total entropy at most $m \log(n)$. Similarly obtain $\tilde{O}(\sqrt{nmT})$ regret.
- For small loss, new $\tilde{O}(\sqrt{nL^\star})$ bound! No contradiction, $L^\star \leq mT$.

The Semibandit Case

- If we copy bandit proof, again for each i: $\mathbb{E}[\sum_{t:p_t(i)\geq\gamma} \ell_t(i)] \lesssim L^\star$.
- Initial entropy $m\log(n)$. Hence $\mathbb{E}[R_T] = \tilde{O}(\sqrt{nmL^\star})$.

The Semibandit Case

- If we copy bandit proof, again for each i: $\mathbb{E}[\sum_{t:p_t(i)\geq\gamma} \ell_t(i)] \lesssim L^\star$.
- Initial entropy $m\log(n)$. Hence $\mathbb{E}[R_T] = \tilde{O}(\sqrt{nmL^\star})$.
- Idea: sort $A^\star = (i_1^\star, i_2^\star, \dots, i_m^\star)$ so that $L_{i_1^\star,T} \geq L_{i_2^\star,T} \geq \cdots \geq L_{i_m^\star,T}$.

The Semibandit Case

- If we copy bandit proof, again for each i: $\mathbb{E}[\sum_{t:p_t(i)\geq\gamma} \ell_t(i)] \lesssim L^\star$.
- Initial entropy $m\log(n)$. Hence $\mathbb{E}[R_T] = \tilde{O}(\sqrt{nmL^\star})$.
- Idea: sort $A^\star = (i_1^\star, i_2^\star, \dots, i_m^\star)$ so that $L_{i_1^\star,T} \geq L_{i_2^\star,T} \geq \cdots \geq L_{i_m^\star,T}$.
- We know that $L_{i_k^\star,T} \leq \frac{L^\star}{k}$.

The Semibandit Case

- If we copy bandit proof, again for each i: $\mathbb{E}[\sum_{t:p_t(i)\geq\gamma} \ell_t(i)] \lesssim L^\star$.
- Initial entropy $m\log(n)$. Hence $\mathbb{E}[R_T] = \tilde{O}(\sqrt{nmL^\star})$.
- Idea: sort $A^\star = (i_1^\star, i_2^\star, \dots, i_m^\star)$ so that $L_{i_1^\star,T} \geq L_{i_2^\star,T} \geq \cdots \geq L_{i_m^\star,T}$.
- We know that $L_{i_k^\star,T} \leq \frac{L^\star}{k}$.
- Now we can do information theory separately for each of the unknown coordinates $i_k^\star$. Setting $\bar{\ell}_t(i,k) = \mathbb{E}[\ell_t(i) | i_k^\star = i]$:

The Semibandit Case

- If we copy bandit proof, again for each i: $\mathbb{E}[\sum_{t:p_t(i)\geq\gamma} \ell_t(i)] \lesssim L^\star$.
- Initial entropy $m\log(n)$. Hence $\mathbb{E}[R_T] = \tilde{O}(\sqrt{nmL^\star})$.
- Idea: sort $A^\star = (i_1^\star, i_2^\star, \ldots, i_m^\star)$ so that $L_{i_1^\star,T} \geq L_{i_2^\star,T} \geq \cdots \geq L_{i_m^\star,T}$.
- We know that $L_{i_k^\star,T} \leq \frac{L^\star}{k}$.
- Now we can do information theory separately for each of the unknown coordinates $i_k^\star$. Setting $\bar{\ell}_t(i,k) = \mathbb{E}[\ell_t(i)|i_k^\star = i]$:

$$\mathbb{E}[R_T] = \sum_{i,k,t} \mathbb{E}^{p_t}[p_t(i = i_k^\star)(\bar{\ell}_t(i) - \bar{\ell}_t(i,k))].$$

The Semibandit Case

- If we copy bandit proof, again for each i: $\mathbb{E}[\sum_{t:p_t(i)\geq\gamma} \ell_t(i)] \lesssim L^\star$.
- Initial entropy $m\log(n)$. Hence $\mathbb{E}[R_T] = \tilde{O}(\sqrt{nmL^\star})$.
- Idea: sort $A^\star = (i_1^\star, i_2^\star, \ldots, i_m^\star)$ so that $L_{i_1^\star,T} \geq L_{i_2^\star,T} \geq \cdots \geq L_{i_m^\star,T}$.
- We know that $L_{i_k^\star,T} \leq \frac{L^\star}{k}$.
- Now we can do information theory separately for each of the unknown coordinates $i_k^\star$. Setting $\bar{\ell}_t(i,k) = \mathbb{E}[\ell_t(i)|i_k^\star = i]$:

$$\mathbb{E}[R_T] = \sum_{i,k,t} \mathbb{E}^{p_t}[p_t(i = i_k^\star)(\bar{\ell}_t(i) - \bar{\ell}_t(i,k))].$$

- Now do the same argument for each k, gaining information on $i_k^\star$ in exchange for regret. Then sum over k.

The Semibandit Case

- If we copy bandit proof, again for each i: $\mathbb{E}[\sum_{t:p_t(i)\geq\gamma} \ell_t(i)] \lesssim L^\star$.
- Initial entropy $m\log(n)$. Hence $\mathbb{E}[R_T] = \tilde{O}(\sqrt{nmL^\star})$.
- Idea: sort $A^\star = (i_1^\star, i_2^\star, \ldots, i_m^\star)$ so that $L_{i_1^\star,T} \geq L_{i_2^\star,T} \geq \cdots \geq L_{i_m^\star,T}$.
- We know that $L_{i_k^\star,T} \leq \frac{L^\star}{k}$.
- Now we can do information theory separately for each of the unknown coordinates $i_k^\star$. Setting $\bar{\ell}_t(i,k) = \mathbb{E}[\ell_t(i)|i_k^\star = i]$:

$$\mathbb{E}[R_T] = \sum_{i,k,t} \mathbb{E}^{p_t}[p_t(i = i_k^\star)(\bar{\ell}_t(i) - \bar{\ell}_t(i,k))].$$

- Now do the same argument for each k, gaining information on $i_k^\star$ in exchange for regret. Then sum over k.
- This also doesn't work: we pay $\tilde{O}\left(\sqrt{\frac{nL^\star}{k}}\right)$ for $i_k^\star$, these sum to $\tilde{O}(\sqrt{nmL^\star})$. No change!

The Semibandit Case

- If we copy bandit proof, again for each i: $\mathbb{E}[\sum_{t:p_t(i)\geq\gamma} \ell_t(i)] \lesssim L^\star$.
- Initial entropy $m\log(n)$. Hence $\mathbb{E}[R_T] = \tilde{O}(\sqrt{nmL^\star})$.
- Idea: sort $A^\star = (i_1^\star, i_2^\star, \ldots, i_m^\star)$ so that $L_{i_1^\star,T} \geq L_{i_2^\star,T} \geq \cdots \geq L_{i_m^\star,T}$.
- We know that $L_{i_k^\star,T} \leq \frac{L^\star}{k}$.
- Now we can do information theory separately for each of the unknown coordinates $i_k^\star$. Setting $\bar{\ell}_t(i,k) = \mathbb{E}[\ell_t(i)|i_k^\star = i]$:

$$\mathbb{E}[R_T] = \sum_{i,k,t} \mathbb{E}^{p_t}[p_t(i = i_k^\star)(\bar{\ell}_t(i) - \bar{\ell}_t(i,k))].$$

- Now do the same argument for each k, gaining information on $i_k^\star$ in exchange for regret. Then sum over k.
- This also doesn't work: we pay $\tilde{O}\left(\sqrt{\frac{nL^\star}{k}}\right)$ for $i_k^\star$, these sum to $\tilde{O}(\sqrt{nmL^\star})$. No change!
- Detail: we need to threshold separately for each $i_k^\star$.

The Semibandit Case

- Splitting up $A^\star$ gained efficiency in thresholding.

The Semibandit Case

- Splitting up $A^\star$ gained efficiency in thresholding.
- But the information usage is worse. Now when we pick coordinate i, we guess that $i = i_k^\star$ for some fixed k, gain info when this holds.

The Semibandit Case

- Splitting up $A^\star$ gained efficiency in thresholding.
- But the information usage is worse. Now when we pick coordinate i, we guess that $i = i_k^\star$ for some fixed k, gain info when this holds.
- Chance to learn decreases: $\sum_i p_t(i \in A^\star)^2 \to \sum_{i,k} p_t(i = i_k^\star)^2$.

The Semibandit Case

- Splitting up $A^\star$ gained efficiency in thresholding.
- But the information usage is worse. Now when we pick coordinate i, we guess that $i = i_k^\star$ for some fixed k, gain info when this holds.
- Chance to learn decreases: $\sum_i p_t(i \in A^\star)^2 \to \sum_{i,k} p_t(i = i_k^\star)^2$.
- We can do the same argument for any partition of $[m]$. Find a partition with good thresholding and information properties.

The Semibandit Case

- Splitting up $A^\star$ gained efficiency in thresholding.
- But the information usage is worse. Now when we pick coordinate i, we guess that $i = i_k^\star$ for some fixed k, gain info when this holds.
- Chance to learn decreases: $\sum_i p_t(i \in A^\star)^2 \to \sum_{i,k} p_t(i = i_k^\star)^2$.
- We can do the same argument for any partition of $[m]$. Find a partition with good thresholding and information properties.
- For a index subset $S \subseteq [m]$, consider the set

$$A_S^\star = \{i_s^\star : s \in S\}.$$

The Semibandit Case

- Splitting up $A^\star$ gained efficiency in thresholding.
- But the information usage is worse. Now when we pick coordinate i, we guess that $i = i_k^\star$ for some fixed k, gain info when this holds.
- Chance to learn decreases: $\sum_i p_t(i \in A^\star)^2 \to \sum_{i,k} p_t(i = i_k^\star)^2$.
- We can do the same argument for any partition of $[m]$. Find a partition with good thresholding and information properties.
- For a index subset $S \subseteq [m]$, consider the set

$$A_S^\star = \{i_s^\star : s \in S\}.$$

- Essentially, $A_S^\star$ has entropy $|S| \log n$. For each i the sum

$$\sum_{t: p_t(i \in A_S^\star) \geq \gamma} \ell_t(i)$$

reaches $\frac{L^\star}{\min_{s \in S} s}$ before $p_t(i)$ becomes small and the sum freezes.

The Semibandit Case

- As a result, if we partition $[m]$ into subsets $S_1, \ldots, S_k$ we get a regret bound like

$$\tilde{O}\left(\sqrt{nL^\star}\sum_{j=1}^{k}\sqrt{\frac{|S_j|}{\min_{s\in S_j} s}}\right).$$

The Semibandit Case

- As a result, if we partition $[m]$ into subsets $S_1, \ldots, S_k$ we get a regret bound like

$$\tilde{O}\left(\sqrt{nL^\star}\sum_{j=1}^{k}\sqrt{\frac{|S_j|}{\min_{s\in S_j} s}}\right).$$

- We saw $S_1 = [m]$ gives $\tilde{O}(\sqrt{nmL^\star})$, as does $S_j = \{j\}$.

The Semibandit Case

- As a result, if we partition $[m]$ into subsets $S_1, \ldots, S_k$ we get a regret bound like

$$\tilde{O}\left(\sqrt{nL^\star}\sum_{j=1}^{k}\sqrt{\frac{|S_j|}{\min_{s\in S_j} s}}\right).$$

- We saw $S_1 = [m]$ gives $\tilde{O}(\sqrt{nmL^\star})$, as does $S_j = \{j\}$.
- However, taking a dyadic decomposition $S_j = [2^{j-1}, 2^j - 1]$ gives $\tilde{O}(\sqrt{nL^\star \log(m)}) = \tilde{O}(\sqrt{nL^\star})$.

The Semibandit Case

- As a result, if we partition $[m]$ into subsets $S_1, \ldots, S_k$ we get a regret bound like

$$\tilde{O}\left(\sqrt{nL^\star}\sum_{j=1}^{k}\sqrt{\frac{|S_j|}{\min_{s\in S_j} s}}\right).$$

- We saw $S_1 = [m]$ gives $\tilde{O}(\sqrt{nmL^\star})$, as does $S_j = \{j\}$.
- However, taking a dyadic decomposition $S_j = [2^{j-1}, 2^j - 1]$ gives $\tilde{O}(\sqrt{nL^\star \log(m)}) = \tilde{O}(\sqrt{nL^\star})$.

Theorem

A variant of thresholded Thompson Sampling achieves T-independent $\tilde{O}(\sqrt{nL^\star})$ regret in the semibandit setting. Without thresholding the regret is $\tilde{O}(\sqrt{nL^\star})$ with T-dependence.

Open Problems

Open Problems

- What about Contextual Bandits? A small-loss algorithm was found here last year: [Allen-Zhu, Bubeck, and Li '18]

Open Problems

- What about Contextual Bandits? A small-loss algorithm was found here last year: [Allen-Zhu, Bubeck, and Li '18]
- General graph feedback? An $O(\sqrt{T})$ analysis for Thompson Sampling has been done. [Liu, Zheng, and Shroff '18]

Open Problems

- What about Contextual Bandits? A small-loss algorithm was found here last year: [Allen-Zhu, Bubeck, and Li '18]
- General graph feedback? An $O(\sqrt{T})$ analysis for Thompson Sampling has been done. [Liu, Zheng, and Shroff '18]
- Can we turn our semibandit analysis into a non-Bayesian algorithm achieving the optimal small loss bound?

Open Problems

- What about Contextual Bandits? A small-loss algorithm was found here last year: [Allen-Zhu, Bubeck, and Li '18]
- General graph feedback? An $O(\sqrt{T})$ analysis for Thompson Sampling has been done. [Liu, Zheng, and Shroff '18]
- Can we turn our semibandit analysis into a non-Bayesian algorithm achieving the optimal small loss bound?
- Any formal connections between Thompson Sampling and other algorithms?

References

Russo and Van Roy (2016)
An information-theoretic analysis of Thompson sampling
The Journal of Machine Learning Research The Journal of Machine Learning Research 17.1 (2016): 2442-2471.

Lykouris, Sridharan, and Tardos (2018)
Small-loss bounds for online learning with partial information
arXiv:1711.03639.

Liu, Zheng, and Shroff (2018)
Analysis of Thompson Sampling for Graphical Bandits Without the Graphs.
arXiv:1805.08930.]

Allen-Zhu, Bubeck, and Li (2018)
Make the Minority Great Again: First-Order Regret Bound for Contextual Bandits.
arXiv:1802.03386.

Thank You