Improved Path-length Regret Bounds for Bandits

Sébastien Bubeck (Microsoft Research)
Yuanzhi Li (Stanford University)
Haipeng Luo (University of Southern California)
Chen-Yu Wei (University of Southern California)

Multi-armed Bandits

$$\text{regret} = \mathbb{E}\left[\sum_{t=1}^{T} \ell_{t,a_t}\right] - \min_i \mathbb{E}\left[\sum_{t=1}^{T} \ell_{t,i}\right]$$

- K: number of arms
  a_t: the arm the learner chooses at time t
- Minimax regret bound: $\Theta\left(\sqrt{KT}\right)$

Main Theme: Path-length Regret Bound

- **Path length bound**: regret only depends on $\sum_{t=1}^{T} \|\ell_t - \ell_{t-1}\|$

Main Theme: Path-length Regret Bound

- **Path length bound**: regret only depends on $\sum_{t=1}^{T} \|\ell_t - \ell_{t-1}\|$
- First path-length regret bound for bandits [Wei&Luo18]:

$$\widetilde{O}\left(\sqrt{K \sum_{t=1}^{T} \|\ell_t - \ell_{t-1}\|_1}\right)$$

Main Theme: Path-length Regret Bound

- **Path length bound**: regret only depends on $\sum_{t=1}^{T} \|\ell_t - \ell_{t-1}\|$
- First path-length regret bound for bandits [Wei&Luo18]:

$$\widetilde{O}\left(\sqrt{K\sum_{t=1}^{T}\|\ell_t - \ell_{t-1}\|_1}\right)$$

- Main results (notation: $V_p := \sum_{t=1}^{T} \|\ell_t - \ell_{t-1}\|_p$)

Main Theme: Path-length Regret Bound

- **Path length bound**: regret only depends on $\sum_{t=1}^{T} \|\ell_t - \ell_{t-1}\|$
- First path-length regret bound for bandits [Wei&Luo18]:

$$\tilde{O}\left(\sqrt{K\sum_{t=1}^{T}\|\ell_t - \ell_{t-1}\|_1}\right)$$

- Main results (notation: $V_p := \sum_{t=1}^{T} \|\ell_t - \ell_{t-1}\|_p$)
  - [WL18]'s $O(\sqrt{KV_1})$ is **tight** when the adversary is adaptive, but the dependency on K is **improvable** when the adversary is oblivious and when V_1 is large.

Main Theme: Path-length Regret Bound

- **Path length bound**: regret only depends on $\sum_{t=1}^{T} \|\ell_t - \ell_{t-1}\|$
- First path-length regret bound for bandits [Wei&Luo18]:

$$\widetilde{O}\left(\sqrt{K \sum_{t=1}^{T} \|\ell_t - \ell_{t-1}\|_1}\right)$$

- Main results (notation: $V_p := \sum_{t=1}^{T} \|\ell_t - \ell_{t-1}\|_p$)
  - [WL18]'s $O(\sqrt{KV_1})$ is **tight** when the adversary is adaptive, but the dependency on K is **improvable** when the adversary is oblivious and when V_1 is large.
  - $O(\sqrt{KV_1})$ can be improved to $O(\sqrt{KV_\infty})$

Main Theme: Path-length Regret Bound

- **Path length bound**: regret only depends on $\sum_{t=1}^{T} \|\ell_t - \ell_{t-1}\|$
- First path-length regret bound for bandits [Wei&Luo18]:

$$\widetilde{O}\left(\sqrt{K \sum_{t=1}^{T} \|\ell_t - \ell_{t-1}\|_1}\right)$$

- Main results (notation: $V_p := \sum_{t=1}^{T} \|\ell_t - \ell_{t-1}\|_p$)
  - [WL18]'s $O(\sqrt{KV_1})$ is **tight** when the adversary is adaptive, but the dependency on K is **improvable** when the adversary is oblivious and when V_1 is large.
  - $O(\sqrt{KV_1})$ can be improved to $O(\sqrt{KV_\infty})$
  - First path-length bound for **linear bandits**

1. $\sqrt{KV_1} \to \sqrt{KV_\infty}$ for Multi-armed Bandits

Wei&Luo'18: Optimistic FTRL

Regularizer: $\psi(x) = \sum_i \log \frac{1}{x_i}$
For $t = 1, 2, \dots, T$**:**

1. Solve

$$x_t = \operatorname*{argmin}_{x \in \Delta^K} \left\{ \sum_{\tau=1}^{t-1} \langle x, \hat{\ell}_\tau \rangle + \langle x, m_t \rangle + \psi(x) \right\}$$

   where $m_{t,i}$ = last observed loss of arm i

2. $a_t \sim x_t$

3. Construct variance-reduced loss estimator $\hat{\ell}_t$ centered at m_t

Wei&Luo'18: Optimistic FTRL

This Paper: FTRL with biased distribution

Regularizer: $\psi(x) = \sum_i \log \frac{1}{x_i}$

For $t = 1, 2, \ldots, T$**:**

1. Solve

$$x_t = \operatorname*{argmin}_{x \in \Delta^K} \left\{ \sum_{\tau=1}^{t-1} \langle x, \hat{\ell}_\tau \rangle + \langle x, m_t \rangle + \psi(x) \right\}$$

   where $m_{t,i}$ = last observed loss of arm i

2. $a_t \sim x_t$

3. Construct variance-reduced loss estimator $\hat{\ell}_t$ centered at m_t

Wei&Luo'18: Optimistic FTRL

This Paper: FTRL with biased distribution

Regularizer: $\psi(x) = \sum_i \log \frac{1}{x_i}$

For $t = 1, 2, \ldots, T$**:**

1. Solve

$$x_t = \operatorname*{argmin}_{x \in \Delta^K} \left\{ \sum_{\tau=1}^{t-1} \langle x, \hat{\ell}_\tau \rangle \qquad\qquad + \psi(x) \right\}$$

2. $a_t \sim x_t$

3. Construct variance-reduced loss estimator $\hat{\ell}_t$ centered at m_t

Wei&Luo'18: Optimistic FTRL

This Paper: FTRL with biased distribution

Regularizer: $\psi(x) = \sum_i \log \frac{1}{x_i}$
For $t = 1, 2, \ldots, T$**:**

1. Solve
$$x_t = \operatorname*{argmin}_{x \in \Delta^K} \left\{ \sum_{\tau=1}^{t-1} \langle x, \hat{\ell}_\tau \rangle \qquad\qquad + \psi(x) \right\}$$

2. $\cancel{a_t \sim x_t}$
$$\begin{cases} a_t \sim x_t & w.p.\ 1 - \alpha_t \\ a_t = a_{t-1} & w.p.\ \alpha_t \end{cases}$$
where $\alpha_t \approx \alpha(1 - \ell_{t-1,a_{t-1}})$

3. Construct variance-reduced loss estimator $\hat{\ell}_t$ centered at m_t

Wei&Luo'18: Optimistic FTRL

This Paper: FTRL with biased distribution

Regularizer: $\psi(x) = \sum_i \log \frac{1}{x_i}$

For $t = 1, 2, \ldots, T$**:**

1. Solve

$$x_t = \operatorname*{argmin}_{x \in \Delta^K} \left\{ \sum_{\tau=1}^{t-1} \langle x, \hat{\ell}_\tau \rangle \qquad + \psi(x) \right\}$$

2. $\cancel{a_t \sim x_t}$

$$\begin{cases} a_t \sim x_t & w.p.\ 1 - \alpha_t \\ a_t = a_{t-1} & w.p.\ \alpha_t \end{cases} \qquad \text{where } \alpha_t \approx \alpha(1 - \ell_{t-1,a_{t-1}})$$

3. Construct variance-reduced loss estimator $\hat{\ell}_t$ centered at $\ell_{t-1,a_{t-1}}$

2. A Gap Between Oblivious and Adaptive Adversaries In Path-length Bound

Oblivious and Adaptive Adversaries

- Oblivious: $\ell_1, \ldots, \ell_T$ are all selected before the game starts.
- Adaptive: ℓ_t may depend on learner's previous actions.

Theorems

Lower bound. For any $V_1 \leq T$, the regret is $\Omega(\sqrt{KV_1})$ when the adversary is adaptive.

- when $V_1 = T \Rightarrow \Omega(\sqrt{KT})$

Upper bound. For any V_1, the regret is $O(K^{\frac{1}{3}} V_1^{\frac{1}{3}} T^{\frac{1}{6}})$ when the adversary is oblivious.

- when $V_1 = T \Rightarrow O(K^{\frac{1}{3}} \sqrt{T})$

The algorithm with $O(K^{\frac{1}{3}} V_1^{\frac{1}{3}} T^{\frac{1}{6}})$ upper bound against oblivious adversary

- FTRL with regularizer: $\psi(x) = \sum_{i=1}^{K} x_i \log(x_i) + \frac{1}{K} \sum_{i=1}^{K} \log \frac{1}{x_i}$

The algorithm with $O(K^{\frac{1}{3}} V_1^{\frac{1}{3}} T^{\frac{1}{6}})$ upper bound against oblivious adversary

- FTRL with regularizer: $\psi(x) = \sum_{i=1}^{K} x_i \log(x_i) + \frac{1}{K} \sum_{i=1}^{K} \log \frac{1}{x_i}$
- For **heavy arms** (i.e., arms with larger $x_{t,i}$),
  - use the old mechanism (optimistic prediction with last observed loss) as in [WL18]
  - the $x_i \log(x_i)$ regularizer part is taking effect

The algorithm with $O(K^{\frac{1}{3}} V_1^{\frac{1}{3}} T^{\frac{1}{6}})$ upper bound against oblivious adversary

- FTRL with regularizer: $\psi(x) = \sum_{i=1}^{K} x_i \log(x_i) + \frac{1}{K} \sum_{i=1}^{K} \log \frac{1}{x_i}$
- For **heavy arms** (i.e., arms with larger $x_{t,i}$),
  - use the old mechanism (optimistic prediction with last observed loss) as in [WL18]
  - the $x_i \log(x_i)$ regularizer part is taking effect
- For **light arms** (i.e., arms with smaller $x_{t,i}$)
  - use the new mechanism (biased distribution) as in the previously introduced algorithm
  - the $\log \frac{1}{x_i}$ regularizer part is taking effect

The algorithm with $O(K^{\frac{1}{3}} V_1^{\frac{1}{3}} T^{\frac{1}{6}})$ upper bound against oblivious adversary

- FTRL with regularizer: $\psi(x) = \sum_{i=1}^{K} x_i \log(x_i) + \frac{1}{K} \sum_{i=1}^{K} \log \frac{1}{x_i}$
- For **heavy arms** (i.e., arms with larger $x_{t,i}$),
  - use the old mechanism (optimistic prediction with last observed loss) as in [WL18]
  - the $x_i \log(x_i)$ regularizer part is taking effect
- For **light arms** (i.e., arms with smaller $x_{t,i}$)
  - use the new mechanism (biased distribution) as in the previously introduced algorithm
  - the $\log \frac{1}{x_i}$ regularizer part is taking effect
- **Open question**: Is $O\left(\sqrt{V_1}\right)$ possible for oblivious adversary? (recall: $\Omega(\sqrt{KV_1})$ for adaptive adversary)

3. Path-length Bounds for Linear Bandits

Linear Bandits

For $t = 1, 2, \ldots, T$**:**

- Adversary decides loss vector ℓ_t
- Learner picks an **action** $a_t \in \mathcal{A} \subseteq \{x : \|x\| \leq 1\}$.
- Learner observes **loss** $a_t^\top \ell_t$

Optimistic SCRiBLe [Rakhlin&Sridharan'13]

Regularizer: a self-concordant barrier ψ for conv$(\mathcal{A})$
For $t = 1, 2, \ldots, T$**:**

1. Solve

$$x_t = \operatorname*{argmin}_{x \in \text{conv}(\mathcal{A})} \left\{ \sum_{\tau=1}^{t-1} \langle x, \hat{\ell}_\tau \rangle + \underbrace{\langle x, m_t \rangle}_{\text{optimistic prediction}} + \psi(x) \right\}$$

2. Sample a_t from the *Dikin ellipsoid* centered at x_t.
3. Construct loss unbiased estimator $\hat{\ell}_t$.

Optimistic SCRiBLe [Rakhlin&Sridharan'13]

Regularizer: a self-concordant barrier ψ for $\mathrm{conv}(\mathcal{A})$
For $t = 1, 2, \ldots, T$**:**

1. Solve

$$x_t = \operatorname*{argmin}_{x \in \mathrm{conv}(\mathcal{A})} \left\{ \sum_{\tau=1}^{t-1} \langle x, \hat{\ell}_\tau \rangle + \underbrace{\langle x, m_t \rangle}_{\text{optimistic prediction}} + \psi(x) \right\}$$

2. Sample a_t from the *Dikin ellipsoid* centered at x_t.
3. Construct loss unbiased estimator $\hat{\ell}_t$.

$$\text{regret} = \mathbb{E}\left[\sum_{t=1}^{T} \langle a_t, \ell_t \rangle\right] - \min_{a \in \mathcal{A}} \mathbb{E}\left[\sum_{t=1}^{T} \langle a, \ell_t \rangle\right] \le \widetilde{O}\left(d^{\frac{3}{2}} \sqrt{\sum_{t=1}^{T} \langle a_t, \ell_t - m_t \rangle^2}\right)$$

$m_t = ?$

$$\widetilde{O}\left(\sqrt{\sum_{t=1}^{T} \langle a_t, \ell_t - m_t \rangle^2}\right) \quad \xrightarrow{?} \quad \widetilde{O}\left(\sqrt{\sum_{t=1}^{T} \|\ell_t - \ell_{t-1}\|}\right)$$

How to set m_t?

- $m_t = \ell_{t-1}$ is not feasible – the learner does not know ℓ_{t-1}.

Our Solutions

$$\widetilde{O}\left(\sqrt{\sum_{t=1}^{T}\langle a_t, \ell_t - m_t\rangle^2}\right) \quad \xrightarrow{?} \quad \widetilde{O}\left(\sqrt{\sum_{t=1}^{T}\|\ell_t - \ell_{t-1}\|}\right)$$

Our Solutions

$$\widetilde{O}\left(\sqrt{\sum_{t=1}^{T}\langle a_t, \ell_t - m_t\rangle^2}\right) \quad \xrightarrow{?} \quad \widetilde{O}\left(\sqrt{\sum_{t=1}^{T}\|\ell_t - \ell_{t-1}\|}\right)$$

- **For** $\|\cdot\| = \|\cdot\|_2$: Greedy projection

$$m_{t+1} = \Pi_{\mathcal{C}_t}\left(m_t\right), \qquad \text{where } \mathcal{C}_t = \Big\{m : \langle a_t, \ell_t - m\rangle = 0\Big\}.$$

Our Solutions

$$\widetilde{O}\left(\sqrt{\sum_{t=1}^{T}\langle a_t, \ell_t - m_t\rangle^2}\right) \quad \xrightarrow{?} \quad \widetilde{O}\left(\sqrt{\sum_{t=1}^{T}\|\ell_t - \ell_{t-1}\|}\right)$$

- **For** $\|\cdot\| = \|\cdot\|_2$: Greedy projection

$$m_{t+1} = \Pi_{\mathcal{C}_t}(m_t), \qquad \text{where } \mathcal{C}_t = \Big\{m : \langle a_t, \ell_t - m\rangle = 0\Big\}.$$

$$\Rightarrow \sum_{t=1}^{T}\langle a_t, \ell_t - m_t\rangle^2 = O\left(\sum_{t=1}^{T}\|\ell_t - \ell_{t-1}\|_2\right).$$

Our Solutions

$$\widetilde{O}\left(\sqrt{\sum_{t=1}^{T}\langle a_t, \ell_t - m_t\rangle^2}\right) \quad \xrightarrow{?} \quad \widetilde{O}\left(\sqrt{\sum_{t=1}^{T}\|\ell_t - \ell_{t-1}\|}\right)$$

- **For** $\|\cdot\| = \|\cdot\|_2$: Greedy projection

$$m_{t+1} = \Pi_{\mathcal{C}_t}(m_t), \qquad \text{where } \mathcal{C}_t = \Big\{m : \langle a_t, \ell_t - m\rangle = 0\Big\}.$$

$$\Rightarrow \sum_{t=1}^{T}\langle a_t, \ell_t - m_t\rangle^2 = O\left(\sum_{t=1}^{T}\|\ell_t - \ell_{t-1}\|_2\right).$$

- **For general** $\|\cdot\|$: reduction to the **Convex Body Chasing** problem [Friedman&Linial'93] (using [Sellke'19]'s algorithm)

Summary

- A gap between adaptive adversary and oblivious adversary settings in path-length bound
- Improving $O(\sqrt{KV_1})$ to $O(\sqrt{KV_\infty})$
- First path-length bound for **linear bandits**