Lecture 10. Adversarial Bandits

Advanced Optimization (Fall 2024)

Peng Zhao

zhaop@lamda.nju.edu.cn

Nanjing University

Outline

- Bandit Problems and Adversarial Bandits
- Multi-Armed Bandits
- Bandit Convex Optimization
- BCO with Smooth Functions

Part 1. Bandits

- Bandit Problems

- Adversarial Bandits

Online Learning

At each round $t = 1, 2, \cdots$

(1) the player first picks a model $\mathbf{x}_t$ from a feasible set $\mathcal{X} \subseteq \mathbb{R}^d$;

(2) and environments pick an online function $f_t : \mathcal{X} \to \mathbb{R}$;

(3) the player suffers loss $f_t(\mathbf{x}_t)$, observes some information about f_t and updates the model.

on the feedback information:

- ***full information***: observe entire f_t (or at least gradient $\nabla f_t(\mathbf{x}_t)$)

- ***partial information (bandits)***: observe function value $f_t(\mathbf{x}_t)$ only

less information

full information

horse racing

partial information

multi-armed bandits

Bandits

- Bandit problems
  - named after a ***one-armed bandit***
  - ***arm***: a colloquial term for a slot machine that is pulled to try to win
  - ***bandit***: comes from the idea that the machine is a "thief" that takes your money without offering a guaranteed return

- Multi-armed bandits
  - Context: there are multiple slot machines, each with its own probability of payout
  - Goal: the player (gambler) places her bets on a slot machine to maximize the total reward
  - **Exploration-Exploitation tradeoff**

Bandits: history bit

- **Bandit problems** were introduced for the clinical trial design by **William R. Thompson** in an article published in 1933 [Thompson, 1933].

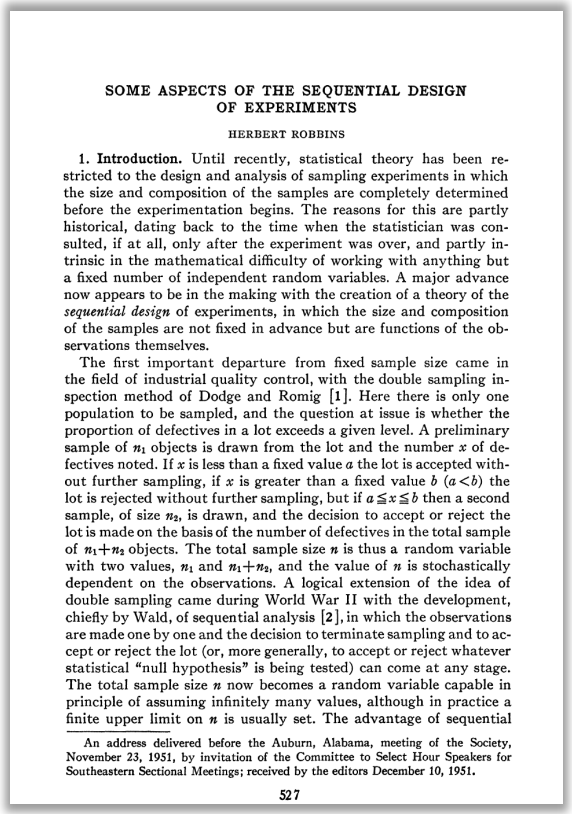
ON THE LIKELIHOOD THAT ONE UNKNOWN PROBABILITY EXCEEDS ANOTHER IN VIEW OF THE EVIDENCE OF TWO SAMPLES.

By WILLIAM R. THOMPSON. From the Department of Pathology, Yale University.

- Thompson Sampling (TS) was originally described in this paper but has been largely ignored by the artificial intelligence community.
- TS was subsequently rediscovered numerous times independently in the context of reinforcement learning.

Bandits: history bit

- **Bandit problems** were later formally restated in a short but influential paper [Robbins, 1952] and further developed in [Lai and Robbins, 1985].



SOME ASPECTS OF THE SEQUENTIAL DESIGN OF EXPERIMENTS

HERBERT ROBBINS

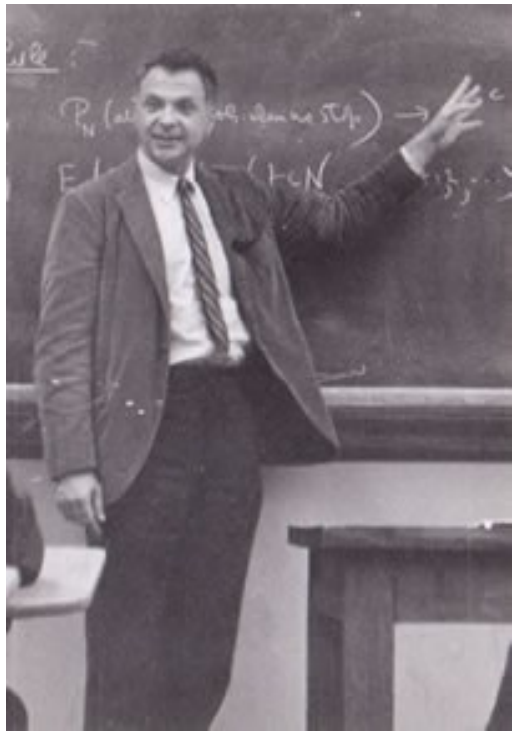
1. **Introduction.** Until recently, statistical theory has been restricted to the design and analysis of sampling experiments in which the size and composition of the samples are completely determined before the experimentation begins. The reasons for this are partly historical, dating back to the time when the statistician was consulted, if at all, only after the experiment was over, and partly intrinsic in the mathematical difficulty of working with anything but a fixed number of independent random variables. A major advance now appears to be in the making with the creation of a theory of the *sequential design* of experiments, in which the size and composition of the samples are not fixed in advance but are functions of the observations themselves.

The first important departure from fixed sample size came in the field of industrial quality control, with the double sampling inspection method of Dodge and Romig [1]. Here there is only one population to be sampled, and the question at issue is whether the proportion of defectives in a lot exceeds a given level. A preliminary sample of n_1 objects is drawn from the lot and the number x of defectives noted. If x is less than a fixed value a the lot is accepted without further sampling, if x is greater than a fixed value b $(a<b)$ the lot is rejected without further sampling, but if $a \leqq x \leqq b$ then a second sample, of size n_2, is drawn, and the decision to accept or reject the lot is made on the basis of the number of defectives in the total sample of n_1+n_2 objects. The total sample size n is thus a random variable with two values, n_1 and n_1+n_2, and the value of n is stochastically dependent on the observations. A logical extension of the idea of double sampling came during World War II with the development, chiefly by Wald, of sequential analysis [2], in which the observations are made one by one and the decision to terminate sampling and to accept or reject the lot (or, more generally, to accept or reject whatever statistical "null hypothesis" is being tested) can come at any stage. The total sample size n now becomes a random variable capable in principle of assuming infinitely many values, although in practice a finite upper limit on n is usually set. The advantage of sequential

An address delivered before the Auburn, Alabama, meeting of the Society, November 23, 1951, by invitation of the Committee to Select Hour Speakers for Southeastern Sectional Meetings; received by the editors December 10, 1951.

527

H. Robbins. Some aspects of the sequential design of experiments. Bulletin of the American Mathematical Society, 58(5):527–535, 1952.

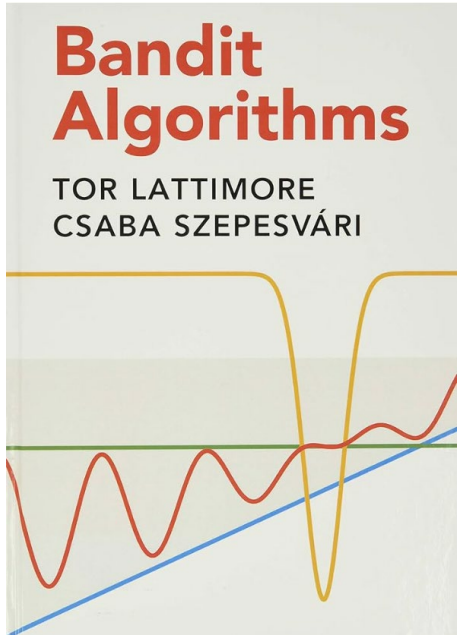


Herbert Ellis Robbins (1915 - 2001)

Bandit Problems

- Also called ***partial-information*** online learning.

- There are a variety of bandit problems:
  - ❑ multi-armed bandits (MAB)
  - ❑ linear bandits/convex bandits
  - ❑ generalized linear bandits/graph bandits
  - ❑ contextual bandits
  - ❑ partial monitoring
  - ❑ … (even MDP for RL)



Bandit Algorithms
Tor Lattimore, Csaba Szepesvári
Cambridge University Press, 2021

Online Learning

At each round $t = 1, 2, \cdots$

(1) the player first picks a model $\mathbf{x}_t$ from a feasible set $\mathcal{X} \subseteq \mathbb{R}^d$;

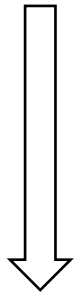
(2) and environments pick an online function $f_t : \mathcal{X} \to \mathbb{R}$;

(3) the player suffers loss $f_t(\mathbf{x}_t)$, observes some information about f_t and updates the model.

on the difficulty of environments:

- stochastic setting

- adversarial setting { oblivious; adaptive (non-oblivious)

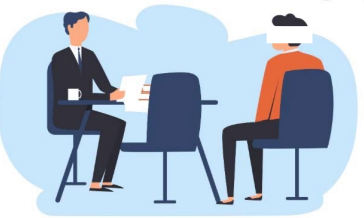


less restricted but harder

oblivious adversary

examination

adaptive adversary



interview

Bandit Problems

- Also called ***partial-information*** online learning.

- According to the environments, it can be roughly classified as
  - ❑ ***Stochastic bandits***: environment is generated by a stochastic model
  - ❑ ***Adversarial bandits***: environment can be chosen against the learner
    - oblivious adversary: thinking of the final exam
    - non-oblivious adversary: thinking of the online games

Adversarial Bandits

- Continuing the OCO protocol:

At each round $t = 1, 2, \cdots$

(1) the player first picks a model $\mathbf{x}_t$ from a convex set $\mathcal{X} \subseteq \mathbb{R}^d$;

(2) and environments pick an online convex function $f_t : \mathcal{X} \to \mathbb{R}$;

(3) the player suffers loss $f_t(\mathbf{x}_t)$, observes some information about f_t and updates the model.

⇨ at round t, the learner can only observe the function value $f_t(\mathbf{x}_t)$ information

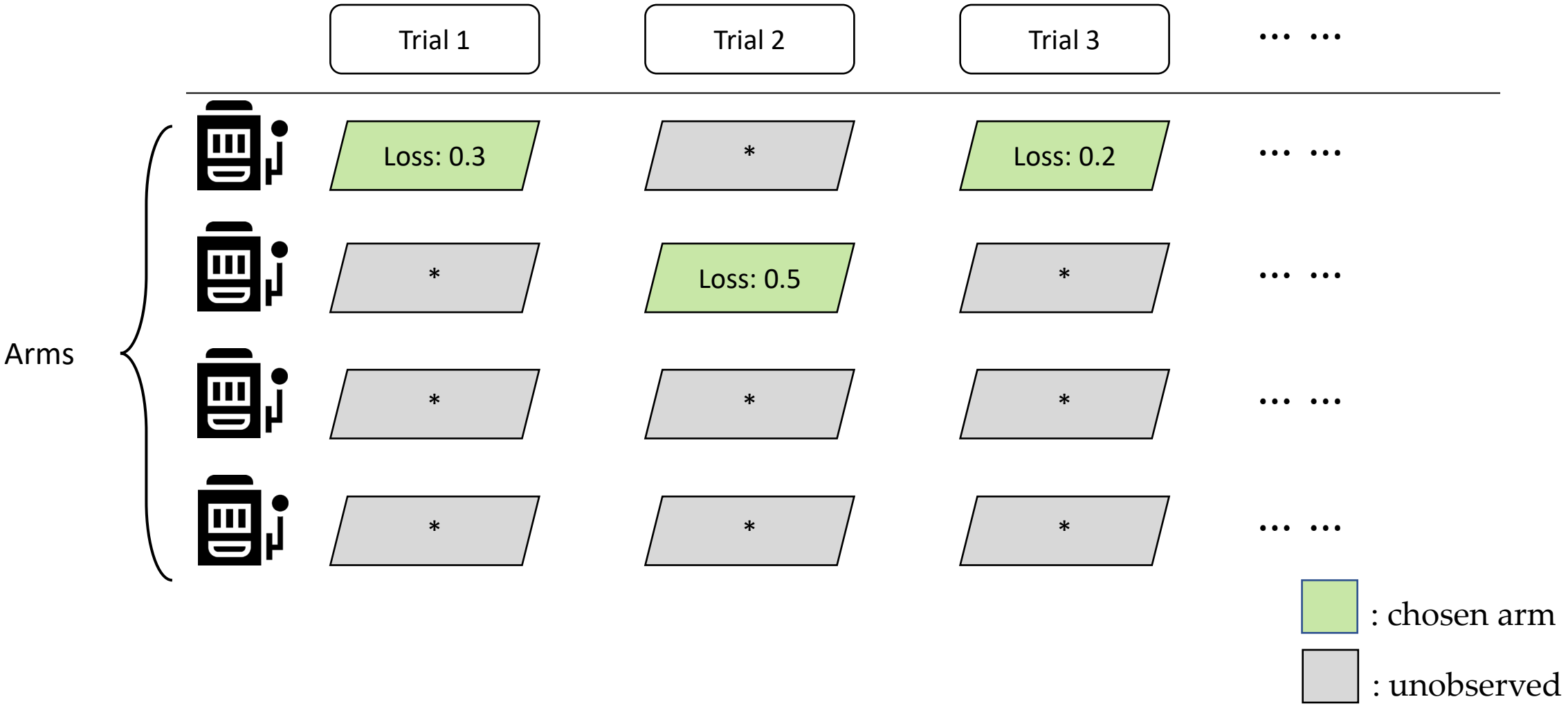
We focus on the ***oblivious*** setting (non-oblivious bandits are usually challenging)

i.e., environments decide online functions of all the rounds before the online game starts.

Part 2. (Adversarial) Multi-Armed Bandits

- Formulation
- Loss Estimator
- Exp3 and Regret Analysis

Multi-Armed Bandit



Formulation

At each round $t = 1, 2, \cdots$

(1) the player first picks an arm $a_t \in [K]$ from K candidate arms;

(2) and simultaneously environments pick a loss vector $\boldsymbol{\ell}_t \in [0, 1]^K$;

(3) the player suffers and only observes loss ℓ_{t,a_t}, then updates the model.

on the difficulty of environments:

- ***adversarial*** setting
  - ***oblivious***: $\{\boldsymbol{\ell}_t\}_{t=1}^T$ are chosen before the game starts.
  - ***non-oblivious***: $\boldsymbol{\ell}_t(a_1, \ell_{1,a_1}, \ldots, a_{t-1}, \ell_{t-1,a_{t-1}})$ can depend on the past history.
- ***stochastic*** setting: $\boldsymbol{\ell}_t \overset{\text{i.i.d.}}{\sim} \mathcal{D}$, where $\mathcal{D}$ is a fixed unknown distribution.

Formulation

At each round $t = 1, 2, \cdots$

(1) the player first picks an arm $a_t \in [K]$ from K candidate arms;

(2) and simultaneously environments pick a loss vector $\boldsymbol{\ell}_t \in [0, 1]^K$;

(3) the player suffers and only observes loss ℓ_{t,a_t}, then updates the model.

Goal: to minimize ***expected regret***

$$\mathbb{E}[\text{Regret}_T] = \mathbb{E}\left[\sum_{t=1}^{T} \ell_{t,a_t}\right] - \min_{a\in[K]} \sum_{t=1}^{T} \ell_{t,a},$$

where the expectation is taken over the ***randomness of algorithms***.

deterministic algorithms will suffer an $\Omega(T)$ regret in the worst case under the bandit setting!

Comparison

Full-Information Problem	Domain	Loss Functions	Feedback
Prediction with Experts' Advice	Δ_d	$f_t(\boldsymbol{p}_t) = \langle \boldsymbol{\ell}_t, \boldsymbol{p}_t \rangle$	$f_t(\boldsymbol{p}_t),\ \boldsymbol{\ell}_t$
Online Convex Optimization	$\mathcal{X}$	$f_t(\cdot)$	$f_t(\mathbf{x}_t),\ \nabla f_t(\mathbf{x}_t), \dots$

Bandit Problem	Domain	Loss Functions	Feedback
Multi-Armed Bandits	$\{\mathbf{e}_1, \dots, \mathbf{e}_K\}$	$f_t(\mathbf{e}_{a_t}) = \langle \boldsymbol{\ell}_t, \mathbf{e}_{a_t} \rangle$	$f_t(\mathbf{e}_{a_t}) = \ell_{t,a_t}$
Bandit Convex Optimization	$\mathcal{X}$	$f_t(\cdot)$	$f_t(\mathbf{x}_t)$

Notation: $\mathbf{e}_i \in \mathbb{R}^K$ is the one-hot vector, with i-th entry being 1.

Caveat: the feasible domain of MAB is actually ***not*** convex.

(simplex is the convex hull of $\{\mathbf{e}_1, \dots, \mathbf{e}_K\}$)

Comparison

Full-Information Problem	Domain	Loss Functions	Feedback
Prediction with Experts' Advice	Δ_d	$f_t(\boldsymbol{p}_t) = \langle \boldsymbol{\ell}_t, \boldsymbol{p}_t \rangle$	$f_t(\boldsymbol{p}_t),\ \boldsymbol{\ell}_t$
Online Convex Optimization	$\mathcal{X}$	$f_t(\cdot)$	$f_t(\mathbf{x}_t),\ \nabla f_t(\mathbf{x}_t), \dots$

Bandit Problem	Domain	Loss Functions	Feedback
Multi-Armed Bandits	$\{\mathbf{e}_1, \dots, \mathbf{e}_K\}$	$f_t(\mathbf{e}_{a_t}) = \langle \boldsymbol{\ell}_t, \mathbf{e}_{a_t} \rangle$	$f_t(\mathbf{e}_{a_t}) = \ell_{t,a_t}$
Bandit Convex Optimization	$\mathcal{X}$	$f_t(\cdot)$	$f_t(\mathbf{x}_t)$

Notation: $\mathbf{e}_i \in \mathbb{R}^K$ is the one-hot vector, with i-th entry being 1.

Caveat: the feasible domain of MAB is actually ***not*** convex.

(simplex is the convex hull of $\{\mathbf{e}_1, \dots, \mathbf{e}_K\}$)

A Natural Solution for MAB

- MAB bares much similarity with the PEA problem (except for the amount feedback information).

⇨ Deploying Hedge to MAB problem.

Hedge for PEA

At each round $t = 1, 2, \cdots$

(1) compute $\boldsymbol{p}_t \in \Delta_K$ such that $p_{t,i} \propto \exp\left(-\eta L_{t-1,i}\right)$ for $i \in [K]$

(2) the player submits $\boldsymbol{p}_t$, suffers loss $\langle \boldsymbol{p}_t, \boldsymbol{\ell}_t \rangle$, and observes loss $\boldsymbol{\ell}_t \in \mathbb{R}^K$

(3) update $\boldsymbol{L}_t = \boldsymbol{L}_{t-1} + \boldsymbol{\ell}_t$

A Natural Solution for MAB

- However, Hedge does not fit for MAB setting due to ***limited feedback***.

> Hedge requires $\ell_{t,i}$ for all $i \in [K]$, but only ℓ_{t,a_t} is available in MAB.

Hedge for PEA

At each round $t = 1, 2, \cdots$

(1) compute $\boldsymbol{p}_t \in \Delta_K$ such that $p_{t,i} \propto \exp\left(-\eta L_{t-1,i}\right)$ for $i \in [K]$

(2) the player submits $\boldsymbol{p}_t$, suffers loss $\langle \boldsymbol{p}_t, \boldsymbol{\ell}_t \rangle$, and observes loss $\boldsymbol{\ell}_t \in \mathbb{R}^K$

(3) update $\boldsymbol{L}_t = \boldsymbol{L}_{t-1} + \boldsymbol{\ell}_t$

Reduction of MAB to PEA

- Given the similarity of MAB and PEA, can we realize the reduction?

- $\boldsymbol{p}_t \in \Delta_K$ denotes the distribution over arms, and sample an arm $a_t \sim \boldsymbol{p}_t$
- $\widehat{\boldsymbol{\ell}}_t \in \mathbb{R}_+^K$ is the estimated loss fed to Hedge

$$\Rightarrow \quad \boxed{\mathrm{Regret}_T^{\mathrm{MAB}} \overset{\text{by reduction}}{\sim} \mathrm{Regret}_T^{\mathrm{PEA}} = \sum_{t=1}^{T} \langle \boldsymbol{p}_t, \widehat{\boldsymbol{\ell}}_t \rangle - \sum_{t=1}^{T} \widehat{\ell}_{t,i} \le \mathcal{O}(\sqrt{T})}$$

Reduction of MAB to PEA

- **Importance-Weighted (IW)** Loss Estimator

Define $\widehat{\ell}_t \in \mathbb{R}^K$, for all $a \in [K]$,

$$\widehat{\ell}_{t,a} = \frac{\ell_{t,a_t}}{p_{t,a}} \mathbb{1}\{a = a_t\} = \begin{cases} \frac{\ell_{t,a_t}}{p_{t,a_t}} & \text{if } a = a_t; \\ 0 & \text{else.} \end{cases}$$

Loss Estimator

IW Loss Estimator $$\widehat{\ell}_{t,a} = \frac{\ell_{t,a_t}}{p_{t,a}} \mathbb{1}\{a = a_t\} = \begin{cases} \frac{\ell_{t,a_t}}{p_{t,a_t}} & \text{if } a = a_t; \\ 0 & \text{else.} \end{cases}$$

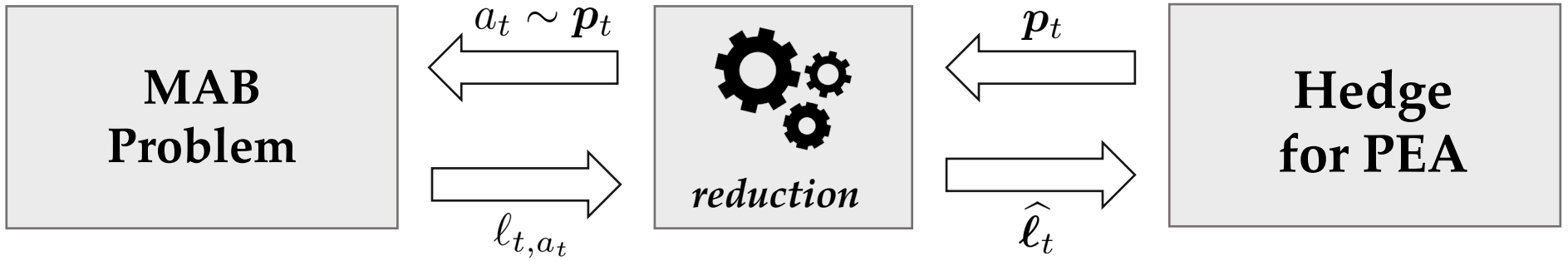
- Property 1. $\ell_{t,a_t} = \langle \boldsymbol{p}_t, \widehat{\boldsymbol{\ell}}_t \rangle$

- Property 2. $\mathbb{E}_{a_t \sim \boldsymbol{p}_t}[\widehat{\ell}_{t,a}] = \ell_{t,a}, \ \forall a \in [K]$ ***unbiasedness***

Proof. $$\mathbb{E}_{a_t \sim \boldsymbol{p}_t}[\widehat{\ell}_{t,a}] = \mathbb{E}_{a_t \sim \boldsymbol{p}_t}\left[\frac{\ell_{t,a_t}}{p_{t,a}} \mathbb{1}\{a = a_t\}\right] = \mathbb{E}_{a_t \sim \boldsymbol{p}_t}\left[\frac{\ell_{t,a}}{p_{t,a}} \mathbb{1}\{a = a_t\}\right]$$

$$= \frac{\ell_{t,a}}{p_{t,a}} \mathbb{E}_{a_t \sim \boldsymbol{p}_t}[\mathbb{1}\{a = a_t\}] = \frac{\ell_{t,a}}{p_{t,a}} p_{t,a} = \ell_{t,a}. \qquad \square$$

Other Choice



- Other estimators coming to mind,

$$\widehat{\boldsymbol{\ell}}_t = [0, \ldots, 0, \underbrace{\ell_{t,a_t}}_{a_t\text{-th entry}}, 0, \ldots, 0]^\top$$

$$\Longrightarrow \quad \underbrace{\langle \boldsymbol{p}_t, \widehat{\boldsymbol{\ell}}_t \rangle}_{\text{loss for Hedge}} = p_{t,a_t} \ell_{t,a_t} \neq \underbrace{\ell_{t,a_t}}_{\text{loss for MAB}}$$

cannot apply Hedge

Importance-Weighted Loss Estimator

- Importance weighting estimator,

$$\widehat{\ell}_t = [0, \ldots, 0, \underbrace{\frac{\ell_{t,a_t}}{p_{t,a_t}}}_{a_t\text{-th entry}}, 0, \ldots, 0]^\top$$

$\Longrightarrow$ balancing exploitation ℓ_{t,a_t} and exploration p_{t,a_t}

Exp3: Algorithm

Exp3 (Exponential-weight for Exploration and Exploitation)

At each round $t = 1, 2, \cdots$

(1) compute $\boldsymbol{p}_t \in \Delta_K$ such that $p_{t,i} \propto \exp\left(-\eta \widehat{L}_{t-1,a}\right)$ for $a \in [K]$

(2) chooses $a_t \sim \boldsymbol{p}_t$, suffers and observe loss ℓ_{t,a_t}, and construct loss estimator $\widehat{\boldsymbol{\ell}}_t \in \mathbb{R}^K$ as

$$\widehat{\ell}_{t,a} = \frac{\ell_{t,a_t}}{p_{t,a}} \mathbb{1}\{a = a_t\} = \begin{cases} \frac{\ell_{t,a_t}}{p_{t,a_t}} & \text{if } a = a_t; \\ 0 & \text{else.} \end{cases}$$

(3) update $\widehat{\boldsymbol{L}}_t = \widehat{\boldsymbol{L}}_{t-1} + \widehat{\boldsymbol{\ell}}_t$

Exp3: Regret Bound

Theorem 1. *Suppose that* $\forall t \in [T]$ *and* $a \in [K], 0 \leq \ell_{t,a} \leq 1$, *then Exp3 with learning rate* $\eta = \sqrt{(\ln K)/(TK)}$ *guarantees*

$$\mathbb{E}[\text{Regret}_T] = \mathbb{E}\left[\sum_{t=1}^{T} \ell_{t,a_t}\right] - \min_{a\in[K]} \sum_{t=1}^{T} \ell_{t,a} \leq \mathcal{O}\left(\sqrt{TK \log K}\right),$$

where the expectation is taken over the randomness of the algorithm.

Comparision:

Hedge for PEA	***Exp3 for MAB***
full-information feedback	bandit feedback
$\text{Regret}_T \leq \mathcal{O}(\sqrt{T \log K})$	$\mathbb{E}[\text{Regret}_T] \leq \mathcal{O}(\sqrt{TK \log K})$

suffer a larger arm dependence

Proof of Exp3 Regret Bound

Proof.

Recall that (Lecture 6, OMD), Hedge under PEA setting guarantees,

$$\sum_{t=1}^{T}\langle \boldsymbol{p}_t, \widehat{\boldsymbol{\ell}}_t\rangle - \sum_{t=1}^{T}\widehat{\ell}_{t,a} \le \frac{\ln K}{\eta} + \eta\sum_{t=1}^{T}\sum_{a=1}^{K} p_{t,a}\left(\widehat{\ell}_{t,a}\right)^2,\ \forall a\in[K]$$

(potential-based analysis allows $\widehat{\ell}_t \in \mathbb{R}_+^K$.)

Note that our previous reduction ensures $\ell_{t,a_t} = \langle \boldsymbol{p}_t, \widehat{\boldsymbol{\ell}}_t\rangle$,

$$\Longrightarrow \quad \sum_{t=1}^{T}\ell_{t,a_t} - \sum_{t=1}^{T}\widehat{\ell}_{t,a} \le \frac{\ln K}{\eta} + \eta\sum_{t=1}^{T}\sum_{a=1}^{K} p_{t,a}\left(\widehat{\ell}_{t,a}\right)^2$$

Proof of Exp3 Regret Bound

Proof. $$\sum_{t=1}^{T} \ell_{t,a_t} - \sum_{t=1}^{T} \widehat{\ell}_{t,a} \leq \frac{\ln K}{\eta} + \eta \sum_{t=1}^{T} \sum_{a=1}^{K} p_{t,a} \left(\widehat{\ell}_{t,a}\right)^2$$

We have the following upper bound for the variance,

$$\mathbb{E}_{a_t \sim \boldsymbol{p}_t} \left[\left(\widehat{\ell}_{t,a}\right)^2 \right] = \mathbb{E}_{a_t \sim \boldsymbol{p}_t} \left[\left(\frac{\ell_{t,a_t}}{p_{t,a}}\right)^2 \mathbb{1}\left\{a = a_t\right\} \right] = \mathbb{E}_{a_t \sim \boldsymbol{p}_t} \left[\left(\frac{\ell_{t,a}}{p_{t,a}}\right)^2 \mathbb{1}\left\{a = a_t\right\} \right]$$

$$= \left(\frac{\ell_{t,a}}{p_{t,a}}\right)^2 \mathbb{E}_{a_t \sim \boldsymbol{p}_t} \left[\mathbb{1}\left\{a = a_t\right\}\right] = \frac{\left(\ell_{t,a}\right)^2}{p_{t,a}}.$$

Proof of Exp3 Regret Bound

Proof. $$\sum_{t=1}^{T} \ell_{t,a_t} - \sum_{t=1}^{T} \widehat{\ell}_{t,a} \leq \frac{\ln K}{\eta} + \eta \sum_{t=1}^{T} \sum_{a=1}^{K} p_{t,a} \left(\widehat{\ell}_{t,a}\right)^2$$

By the Law of total expectation and the above inequality,

$$\mathbb{E}\left[\sum_{t=1}^{T} \ell_{t,a_t} - \sum_{t=1}^{T} \widehat{\ell}_{t,a}\right] = \sum_{t=1}^{T} \mathbb{E}\left[\mathbb{E}_t\left[\ell_{t,a_t} - \widehat{\ell}_{t,a}\right]\right] = \sum_{t=1}^{T} \mathbb{E}\left[\mathbb{E}_t\left[\ell_{t,a_t}\right] - \ell_{t,a}\right] \qquad (\mathbb{E}_t[\widehat{\ell}_{t,a}] = \ell_{t,a})$$

regret bound

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

$$\leq \frac{\ln K}{\eta} + \eta \sum_{t=1}^{T} \sum_{a=1}^{K} \mathbb{E}\left[\mathbb{E}_t\left[p_{t,a} \cdot \left(\widehat{\ell}_{t,a}\right)^2\right]\right]$$

Proof of Exp3 Regret Bound

Proof.

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

$$\leq \frac{\ln K}{\eta}+\eta\sum_{t=1}^{T}\sum_{a=1}^{K}\mathbb{E}\left[\mathbb{E}_t\left[p_{t,a}\cdot\left(\widehat{\ell}_{t,a}\right)^2\right]\right]$$

$$= \frac{\ln K}{\eta}+\eta\sum_{t=1}^{T}\sum_{a=1}^{K}p_{t,a}\cdot\frac{\ell_{t,a}^2}{p_{t,a}} \qquad \left(\mathbb{E}_t\left[p_{t,a}\cdot\widehat{\ell}_{t,a}^2\right]=p_{t,a}\cdot\mathbb{E}_t\left[\widehat{\ell}_{t,a}^2\right]=p_{t,a}\cdot\frac{\ell_{t,a}^2}{p_{t,a}}\right)$$

$$= \frac{\ln K}{\eta}+\eta\sum_{t=1}^{T}\sum_{a=1}^{K}\ell_{t,a}^2$$

$$\leq \frac{\ln K}{\eta}+\eta TK\leq\mathcal{O}(\sqrt{TK\log K}) \quad \left(\ell_{t,a}^2\leq 1,\eta=\sqrt{(\ln K)/TK}\right)$$

□

Lower Bound for MAB

- As above, we have proved the regret upper bound for Exp3:

$$\mathbb{E}\left[\text{Regret}_T\right] \leq \mathcal{O}\left(\sqrt{TK\log K}\right)$$

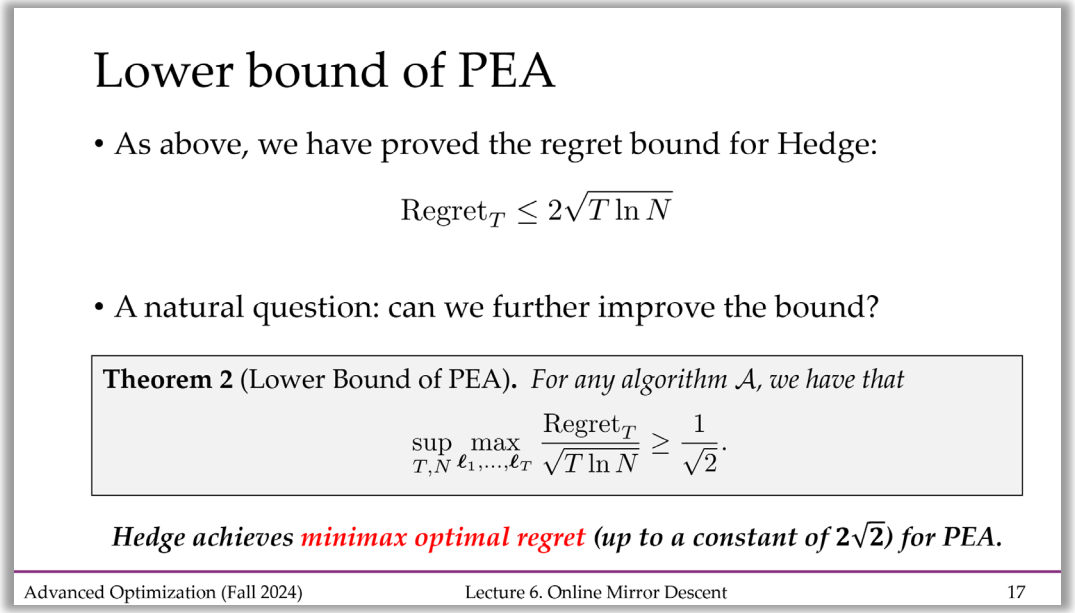
- Can we further improve the regret bound?

It turns out that Exp3 ***doesn't*** achieve minimax optimal regret for MAB.

Lower Bound for MAB

Theorem 2 (Lower Bound for MAB). *For any algorithm $\mathcal{A}$, there exists a sequence of loss vectors $\boldsymbol{\ell}_1, \boldsymbol{\ell}_2, \cdots, \boldsymbol{\ell}_T$ constituing an MAB problem such that*

$$\inf_{\mathcal{A}} \sup_{\boldsymbol{\ell}_1,\ldots,\boldsymbol{\ell}_T} \mathbb{E}\left[\mathrm{Regret}_T\right] = \Omega(\sqrt{TK})$$



Lower bound of PEA

- As above, we have proved the regret bound for Hedge:

$$\mathrm{Regret}_T \le 2\sqrt{T \ln N}$$

- A natural question: can we further improve the bound?

Theorem 2 (Lower Bound of PEA). *For any algorithm $\mathcal{A}$, we have that*

$$\sup_{T,N} \max_{\boldsymbol{\ell}_1,\ldots,\boldsymbol{\ell}_T} \frac{\mathrm{Regret}_T}{\sqrt{T \ln N}} \ge \frac{1}{\sqrt{2}}.$$

Hedge achieves minimax optimal regret (up to a constant of $2\sqrt{2}$) for PEA.

Advanced Optimization (Fall 2024) Lecture 6. Online Mirror Descent 17

MAB Problem $\Omega(\sqrt{TK})$

PEA Problem $\Omega(\sqrt{T \log K})$

Proof Sketch

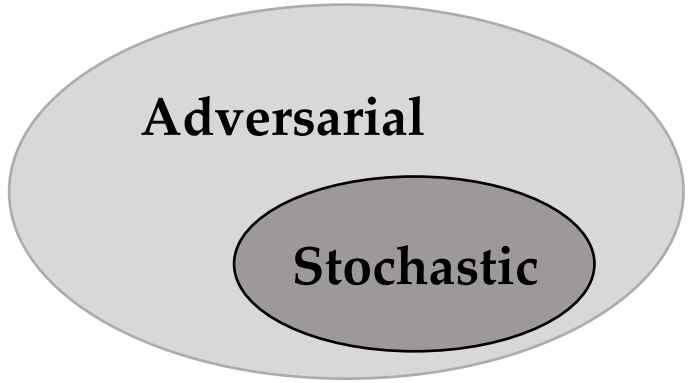
Proof (Sketch).

We prove the theorem under stochastic MAB setting, since the stochastic setting is strictly easier than the adversarial one.

We construct two hard distributions over arms $\mathcal{D}_1, \mathcal{D}_2$ and show that,

$\forall A \in \mathcal{A}$, the following holds

$$\max\left\{\mathbb{E}[\mathrm{Regret_T(A)}; \mathcal{D}_1], \mathbb{E}[\mathrm{Regret_T(A)}; \mathcal{D}_2]\right\} = \Omega(\sqrt{TK})$$



Upper and Lower Bounds for MAB

Theorem 1 (Upper Bound for Exp3). *Suppose that* $\forall t \in [T]$ *and* $a \in [K]$, $0 \le \ell_{t,a} \le 1$, *then Exp3 with learning rate* $\eta = \sqrt{(\ln K)/(TK)}$ *guarantees*

$$\mathbb{E}[\text{Regret}_T] = \mathbb{E}\left[\sum_{t=1}^{T} \ell_{t,a_t}\right] - \min_{a\in[K]} \sum_{t=1}^{T} \ell_{t,a} \le \mathcal{O}\left(\sqrt{TK \log K}\right),$$

where the expectation is taken over the randomness of the algorithm.

Theorem 2 (Lower Bound for MAB). *For any algorithm* $\mathcal{A}$, *there exists a sequence of loss vectors* $\boldsymbol{\ell}_1, \boldsymbol{\ell}_2, \cdots, \boldsymbol{\ell}_T$ *constituing an MAB problem such that*

$$\inf_{\mathcal{A}} \sup_{\boldsymbol{\ell}_1,\ldots,\boldsymbol{\ell}_T} \mathbb{E}\left[\text{Regret}_T\right] = \Omega(\sqrt{TK}).$$

Advanced Topics

- How to shave off the extra $\sqrt{\log K}$ factor?

⇨ Using OMD with ***Tsallis entropy*** regularizer, also using the IW estimator

$$\psi(\boldsymbol{p}) = \frac{1 - \sum_{a=1}^{K} p_a^{\beta}}{1 - \beta}$$

which is actually a *generalization* of negative-entropy used in Hedge, as we have the following fact due to the L'Hôpital'srule

$$\lim_{\beta \to 1} \frac{1 - \sum_a p_a^{\beta}}{1 - \beta} = \sum_a p_a \ln(p_a).$$

Reference: Jean-Yves Audibert and Sebastien Bubeck. Regret bounds and minimax policies under partial monitoring. Journal of Machine Learning Research, 11(Oct):2785–2836, 2010.

Advanced Topics

- How to boost from expected guarantee to a ***high-probability*** one?

⇨ Using an improved estimator: **Implicit eXploration (IX) Loss Estimator**

IW Loss Estimator

$$\widehat{\ell}_{t,a} = \frac{\ell_{t,a_t}}{p_{t,a}} \mathbb{1}\{a = a_t\} = \begin{cases} \frac{\ell_{t,a_t}}{p_{t,a_t}} & \text{if } a = a_t; \\ 0 & \text{else.} \end{cases}$$

IX Loss Estimator

$$\widehat{\ell}_{t,a} = \frac{\ell_{t,a_t}}{p_{t,a} + \gamma} \mathbb{1}\{a = a_t\} = \begin{cases} \frac{\ell_{t,a_t}}{p_{t,a_t} + \gamma} & \text{if } a = a_t; \\ 0 & \text{else.} \end{cases}$$

Reference: Gergely Neu. Explore no more: Improved high-probability regret bounds for non-stochastic bandits. NIPS 2015.

SIAM J. COMPUT.
Vol. 32, No. 1, pp. 48–77

THE NONSTOCHASTIC MULTIARMED BANDIT PROBLEM*

PETER AUER†, NICOLÒ CESA-BIANCHI‡, YOAV FREUND§, AND ROBERT E. SCHAPIRE¶

Abstract. In the multiarmed bandit problem, a gambler must decide which arm of K non-identical slot machines to play in a sequence of trials so as to maximize his reward. This classical problem has received much attention because of the simple model it provides of the trade-off between exploration (trying out each arm to find the best one) and exploitation (playing the arm believed to give the best payoff). Past solutions for the bandit problem have almost always relied on assumptions about the statistics of the slot machines.

In this work, we make no statistical assumptions whatsoever about the nature of the process generating the payoffs of the slot machines. We give a solution to the bandit problem in which an adversary, rather than a well-behaved stochastic process, has complete control over the payoffs. In a sequence of T plays, we prove that the per-round payoff of our algorithm approaches that of the best arm at the rate $O(T^{-1/2})$. We show by a matching lower bound that this is the best possible.

We also prove that our algorithm approaches the per-round payoff of *any* set of strategies at a similar rate: if the best strategy is chosen from a pool of N strategies, then our algorithm approaches the per-round payoff of the strategy at the rate $O((\log N)^{1/2}T^{-1/2})$. Finally, we apply our results to the problem of playing an unknown repeated matrix game. We show that our algorithm approaches the minimax payoff of the unknown game at the rate $O(T^{-1/2})$.

Key words. adversarial bandit problem, unknown matrix games

AMS subject classifications. 68Q32, 68T05, 91A20

PII. S0097539701398375

1. Introduction. In the multiarmed bandit problem, originally proposed by Robbins [17], a gambler must choose which of K slot machines to play. At each time step, he pulls the arm of one of the machines and receives a reward or payoff (possibly zero or negative). The gambler's purpose is to maximize his return, i.e., the sum of the rewards he receives over a sequence of pulls. In this model, each arm is assumed to deliver rewards that are independently drawn from a fixed and unknown distribution. As reward distributions differ from arm to arm, the goal is to find the arm with the highest expected payoff as early as possible and then to keep gambling using that best arm.

The problem is a paradigmatic example of the trade-off between exploration and exploitation. On the one hand, if the gambler plays exclusively on the machine that he thinks is best ("exploitation"), he may fail to discover that one of the other arms actually has a higher expected payoff. On the other hand, if he spends too much time

*Received by the editors November 18, 2001; accepted for publication (in revised form) July 7, 2002; published electronically November 19, 2002. An early extended abstract of this paper appeared in *Proceedings of the 36th Annual Symposium on Foundations of Computer Science*, 1995, IEEE Computer Society, pp. 322–331.
http://www.siam.org/journals/sicomp/32-1/39837.html
†Institute for Theoretical Computer Science, Graz University of Technology, A-8010 Graz, Austria (pauer@igi.tu-graz.ac.at). This author gratefully acknowledges the support of ESPRIT Working Group EP 27150, Neural and Computational Learning II (NeuroCOLT II).
‡Department of Information Technology, University of Milan, I-26013 Crema, Italy (cesa-bianchi@dti.unimi.it). This author gratefully acknowledges the support of ESPRIT Working Group EP 27150, Neural and Computational Learning II (NeuroCOLT II).
§Banter Inc. and Hebrew University, Jerusalem, Israel (yoavf@cs.huji.ac.il).
¶AT&T Labs – Research, Shannon Laboratory, Florham Park, NJ 07932-0971 (schapire@research.att.com).

The non-stochastic multi-armed bandit problem*

Peter Auer
Institute for Theoretical Computer Science
Graz University of Technology
A-8010 Graz (Austria)
pauer@igi.tu-graz.ac.at

Nicolò Cesa-Bianchi
Department of Computer Science
Università di Milano
I-20135 Milano (Italy)
cesabian@dsi.unimi.it

Yoav Freund **Robert E. Schapire**
AT&T Labs
180 Park Avenue
Florham Park, NJ 07932-0971
{yoav, schapire}@research.att.com

November 18, 2001

Finite-time analysis of the multiarmed bandit problem P Auer, N Cesa-Bianchi, P Fischer Machine learning 47 (2), 235-256	8565	2002
The nonstochastic multiarmed bandit problem P Auer, N Cesa-Bianchi, Y Freund, RE Schapire SIAM Journal on Computing 32 (1), 48-77	3179	2003
Using confidence bounds for exploitation-exploration trade-offs P Auer Journal of Machine Learning Research 3 (Nov), 397-422	2420	2002
Near-optimal regret bounds for reinforcement learning T Jaksch, R Ortner, P Auer The Journal of Machine Learning Research 11, 1563-1600	1606 *	2010
Gambling in a rigged casino: The adversarial multi-armed bandit problem P Auer, N Cesa-Bianchi, Y Freund, RE Schapire Foundations of Computer Science, 1995. Proceedings., 36th Annual Symposium …	1209	1995

The Nonstochastic Multiarmed Bandit Problem.
SIAM Journal on Computing (SICOMP). 2002.

Part 3. Bandit Convex Optimization

- Problem Formulation
- Gradient Estimator
- Bandit Gradient Descent
- Regret Analysis

Bandit Convex Optimization

- One of the most general forms for bandits, hence very fundamental.

Full-Information Problem	Domain	Loss Functions	Feedback
Prediction with Experts' Advice	Δ_d	$f_t(\boldsymbol{p}_t) = \langle \boldsymbol{\ell}_t, \boldsymbol{p}_t \rangle$	$f_t(\boldsymbol{p}_t),\ \boldsymbol{\ell}_t$
Online Convex Optimization	$\mathcal{X}$	$f_t(\cdot)$	$f_t(\mathbf{x}_t),\ \nabla f_t(\mathbf{x}_t), \ldots$

Bandit Problem	Domain	Loss Functions	Feedback
Multi-Armed Bandits	$\{\mathbf{e}_1, \ldots, \mathbf{e}_K\}$	$f_t(\mathbf{e}_{a_t}) = \langle \boldsymbol{\ell}_t, \mathbf{e}_{a_t} \rangle$	$f_t(\mathbf{e}_{a_t}) = \ell_{t,a_t}$
Bandit Convex Optimization	$\mathcal{X}$	$f_t(\cdot)$	$f_t(\mathbf{x}_t)$

Formulation

At each round $t = 1, 2, \cdots$

(1) the player first picks a model $\mathbf{x}_t$ from a convex set $\mathcal{X} \subseteq \mathbb{R}^d$;

(2) and environments pick an online convex function $f_t : \mathcal{X} \to \mathbb{R}$;

(3) the player suffers loss $f_t(\mathbf{x}_t)$, observes loss value only to update the model.

Goal: to optimize ***expected regret***,

$$\mathbb{E}[\text{Regret}_T] = \mathbb{E}\left[\sum_{t=1}^{T} f_t(\mathbf{x}_t)\right] - \min_{\mathbf{u}\in\mathcal{X}} \sum_{t=1}^{T} f_t(\mathbf{u}),$$

where the expectation is taken over the ***randomness of algorithms***.

A Natural Solution for BCO

- BCO bares much similarity with the OCO problem.

$\Longrightarrow$ Deploying OGD to BCO problem.

Online Gradient Descent

At each round $t = 1, 2, \cdots$

$$\mathbf{x}_{t+1} = \Pi_{\mathcal{X}} \left[\mathbf{x}_t - \eta \nabla f_t(\mathbf{x}_t) \right],$$

where $\Pi_{\mathcal{X}}[\cdot]$ denotes the projection onto the feasible domain $\mathcal{X}$.

However, we don't have the gradient information due to the *limited feedback*.

Gradient Estimator

- Construct the final decision via the ***perturbation technique***.

$$\mathbf{x}_t \triangleq \mathbf{y}_t + \delta \mathbf{s}_t$$

where $\mathbf{s}_t$ is sampled from unit sphere $\mathbb{S} = \{\mathbf{x} \in \mathbb{R}^d \mid \|\mathbf{x}\|_2 = 1\}$.

Definition 1 (Gradient Estimator). The gradient estimator is defined as

$$\widetilde{\mathbf{g}}_t = \frac{d}{\delta} f_t(\mathbf{y}_t + \delta \mathbf{s}_t) \cdot \mathbf{s}_t$$

with $\mathbf{s}_t$ sampled from the unit sphere $\mathbb{S} = \{\mathbf{x} \in \mathbb{R}^d \mid \|\mathbf{x}\|_2 = 1\}$.

Gradient Estimator

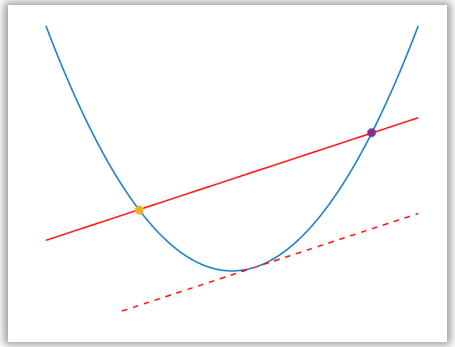
Lemma 1 *For any convex (but not necessarily differentiable) function $f : \mathcal{X} \mapsto \mathbb{R}$, define its smoothed version $\widehat{f}^{\delta}(\mathbf{x}) = \mathbb{E}_{\mathbf{v}\in\mathbb{B}}[f(\mathbf{x}+\delta\mathbf{v})]$. Then for any $\delta > 0$, we have*

$$\mathbb{E}_{\mathbf{s}\in\mathbb{S}}\left[\frac{d}{\delta} f(\mathbf{x}+\delta\mathbf{s})\cdot\mathbf{s}\right] = \nabla\widehat{f}^{\delta}(\mathbf{x}),$$

where $\mathbb{B} = \{\mathbf{x}\in\mathbb{R}^d \mid \|\mathbf{x}\|_2 \le 1\}$ is the unit ball and $\mathbb{S} = \{\mathbf{x}\in\mathbb{R}^d \mid \|\mathbf{x}\|_2 = 1\}$ is the unit sphere.

- Consider the 1-dim case ($d = 1$).

$$\mathbb{E}_{\mathbf{s}\in\mathbb{S}}\left[\frac{d}{\delta} f(\mathbf{x}+\delta\mathbf{s})\cdot\mathbf{s}\right] = \frac{1}{2\delta} f(x+\delta) - \frac{1}{2\delta} f(x-\delta)$$



Gradient Estimator

Lemma 1 *For any convex (but not necessarily differentiable) function $f : \mathcal{X} \mapsto \mathbb{R}$, define its smoothed version $\widehat{f}^{\delta}(\mathbf{x}) = \mathbb{E}_{\mathbf{v}\in\mathbb{B}}[f(\mathbf{x}+\delta\mathbf{v})]$. Then for any $\delta > 0$, we have*

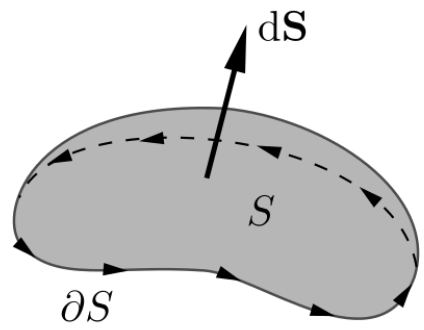
$$\mathbb{E}_{\mathbf{s}\in\mathbb{S}}\left[\frac{d}{\delta} f(\mathbf{x}+\delta\mathbf{s})\cdot\mathbf{s}\right] = \nabla\widehat{f}^{\delta}(\mathbf{x}),$$

where $\mathbb{B} = \{\mathbf{x}\in\mathbb{R}^d \mid \|\mathbf{x}\|_2 \le 1\}$ is the unit ball and $\mathbb{S} = \{\mathbf{x}\in\mathbb{R}^d \mid \|\mathbf{x}\|_2 = 1\}$ is the unit sphere.

- General d-dim case.

  Non-trivial, can be proved by Stokes equation.

  See [Flaxman et al., SODA'05; Proof of Lemma 2.1].



$$\begin{aligned}
&\iint_S \left(\frac{\partial R}{\partial y} - \frac{\partial Q}{\partial z}\right) \mathrm{d}y\,\mathrm{d}z \\
&+ \left(\frac{\partial P}{\partial z} - \frac{\partial R}{\partial x}\right) \mathrm{d}z\,\mathrm{d}x \\
&+ \left(\frac{\partial Q}{\partial x} - \frac{\partial P}{\partial y}\right) \mathrm{d}x\,\mathrm{d}y \\
&= \oint_\Gamma P\,\mathrm{d}x + Q\,\mathrm{d}y + R\,\mathrm{d}z
\end{aligned}$$

Bandit Gradient Descent

- **BGD**: deploying **OGD** to BCO problem using the gradient estimator.

At each round $t = 1, 2, \cdots$

(1) sample a unit vector $\mathbf{s}_t \in \mathbb{S}$;

(2) submit $\mathbf{x}_t = \mathbf{y}_t + \delta \mathbf{s}_t$;

(3) receive feedback $f_t(\mathbf{x}_t)$;

(4) construct gradient estimator $\widetilde{\mathbf{g}}_t = \frac{\delta}{d} f_t(\mathbf{y}_t + \delta \mathbf{s}_t) \cdot \mathbf{s}_t$;

(5) $\mathbf{y}_{t+1} = \Pi_{(1-\alpha)\mathcal{X}}[\mathbf{y}_t - \eta \widetilde{\mathbf{g}}_t]$.

where $(1-\alpha)\mathcal{X} \triangleq \{\mathbf{x} \in \mathbb{R}^d \mid \frac{1}{1-\alpha}\mathbf{x} \in \mathcal{X}\}$.

Note that the gradient estimator satisfies $\mathbb{E}[\widetilde{\mathbf{g}}_t] = \nabla \widehat{f}_t^{\delta}(\mathbf{y}_t)$, with $\widehat{f}_t^{\delta}(\mathbf{x}) = \mathbb{E}_{\mathbf{v} \in \mathbb{B}}[f_t(\mathbf{x} + \delta \mathbf{v})]$.

Bandit Gradient Descent

Theorem 3 *Suppose the online function* $f_t : \mathcal{X} \mapsto \mathbb{R}$ *is G-Lipschitz,* $\max_{\mathbf{x}\in\mathcal{X}} |f_t(\mathbf{x})| \leq C$, *and* $r\cdot\mathbb{B} \subseteq \mathcal{X} \subseteq R\cdot\mathbb{B}$. *For the* ***oblivious*** *adversary setting, the BGD algorithm satisfies*

$$\mathbb{E}\left[\sum_{t=1}^{T} f_t(\mathbf{x}_t) - \min_{\mathbf{x}\in\mathcal{X}} \sum_{t=1}^{T} f_t(\mathbf{x})\right] \leq \frac{R^2}{2\eta} + \frac{\eta d^2 C^2 T}{2\delta^2} + 3G\delta T + \alpha GRT = \mathcal{O}\left(T^{\frac{3}{4}}\right),$$

the last step holds by setting the step size $\eta = \mathcal{O}((R^2/T)^{3/4})$, *the pertubation parameter* $\delta = \eta^{1/3}$ *and the shrink parameter* $\alpha = \delta/r$.

Proof of BGD

$$\mathbf{y}_{t+1} = \Pi_{(1-\alpha)\mathcal{X}}[\mathbf{y}_t - \eta\widetilde{\mathbf{g}}_t], \quad \mathbb{E}[\widetilde{\mathbf{g}}_t] = \nabla\widehat{f}_t^{\delta}(\mathbf{y}_t)$$

smoothed function $\widehat{f}_t^{\delta}(\mathbf{x}) = \mathbb{E}_{\mathbf{v}\in\mathbb{B}}[f_t(\mathbf{x}+\delta\mathbf{v})]$

Proof. For a *fixed* comparator $\mathbf{u} \in \mathcal{X}$ (i.e., it depends on $f_1, \cdots, f_T$ but won't depend on the algorithmic outputs), we can decompose the regret as follows:

$$\mathbb{E}\left[\sum_{t=1}^{T} f_t(\mathbf{x}_t)\right] - \sum_{t=1}^{T} f_t(\mathbf{u}) = \underbrace{\mathbb{E}\left[\sum_{t=1}^{T} \widehat{f}_t^{\delta}(\mathbf{y}_t) - \widehat{f}_t^{\delta}((1-\alpha)\mathbf{u})\right]}_{\text{TERM (A)}} + \underbrace{\mathbb{E}\left[\sum_{t=1}^{T} f_t(\mathbf{x}_t) - f_t(\mathbf{y}_t)\right]}_{\text{TERM (B)}}$$

$$+ \underbrace{\mathbb{E}\left[\sum_{t=1}^{T} f_t(\mathbf{y}_t) - \widehat{f}_t^{\delta}(\mathbf{y}_t)\right]}_{\text{TERM (C)}} + \underbrace{\mathbb{E}\left[\sum_{t=1}^{T} \widehat{f}_t^{\delta}((1-\alpha)\mathbf{u}) - f_t((1-\alpha)\mathbf{u})\right]}_{\text{TERM (D)}} + \underbrace{\mathbb{E}\left[\sum_{t=1}^{T} f_t((1-\alpha)\mathbf{u}) - f_t(\mathbf{u})\right]}_{\text{TERM (E)}}$$

Proof of BGD

Lemma 2 (smoothing error) *Suppose $f : \mathbb{R}^d \mapsto \mathbb{R}$ is G-Lipschitz, and its smoothed version $\widehat{f}^{\delta}(\mathbf{x}) = \mathbb{E}_{\mathbf{v}\in\mathbb{B}}[f(\mathbf{x}+\delta\mathbf{v})]$ satisfies that for any $\mathbf{x} \in \mathbb{R}^d$,*

$$\left|\widehat{f}^{\delta}(\mathbf{x}) - f(\mathbf{x})\right| \leq G \cdot \delta.$$

Proof. $|\widehat{f}^{\delta}(\mathbf{x}) - f(\mathbf{x})| = \mathbb{E}_{\mathbf{v}\in\mathbb{B}}\left[f(\mathbf{x}+\delta\mathbf{v}) - f(\mathbf{x})\right] \leq \mathbb{E}_{\mathbf{v}\in\mathbb{B}}\left[L\|\delta\mathbf{v}\|_2\right] \leq L\delta.$ □

$$\textsc{Term(C)}: \ \mathbb{E}\left[\sum_{t=1}^{T} f_t(\mathbf{y}_t) - \widehat{f}_t^{\delta}(\mathbf{y}_t)\right] \leq TG\delta$$

$$\textsc{Term(D)}: \ \mathbb{E}\left[\sum_{t=1}^{T} \widehat{f}_t^{\delta}\big((1-\alpha)\mathbf{u}\big) - f_t((1-\alpha)\mathbf{u})\right] \leq TG\delta$$

Note: the feasile domain of $\widehat{f}_t^{\delta}$ needs to be within $(1-\alpha)\mathcal{X}$ for a proper α.

Proof of BGD

Proof. Further exploiting the G-Lipschitzness of f_t, we can bound term (A) and term (E) as follows.

$$\textsc{Term}(\text{B}): \ \mathbb{E}\left[\sum_{t=1}^{T} f_t(\mathbf{x}_t) - f_t(\mathbf{y}_t)\right] \leq \mathbb{E}\left[\sum_{t=1}^{T} G\|\mathbf{x}_t - \mathbf{y}_t\|_2\right] \leq \mathbb{E}\left[\sum_{t=1}^{T} G\delta\|\mathbf{s}_t\|_2\right] \leq G\delta T$$

(recall that $\mathbf{x}_t \triangleq \mathbf{y}_t + \delta\mathbf{s}_t$)

$$\textsc{Term}(\text{E}): \ \mathbb{E}\left[\sum_{t=1}^{T} f_t\big((1-\alpha)\mathbf{u}\big) - f_t(\mathbf{u})\right] \leq \mathbb{E}\left[\sum_{t=1}^{T} G\|(1-\alpha)\mathbf{u} - \mathbf{u}\|_2\right] \leq \mathbb{E}\left[\sum_{t=1}^{T} G\alpha\|\mathbf{u}\|_2\right] \leq GR\alpha T$$

(recall that $\mathcal{X} \subseteq R \cdot \mathbb{B}$)

Proof of BGD

Proof. Therefore, it suffices to further bound $\textsc{Term}(\mathrm{A})$.

$$\mathbb{E}\left[\sum_{t=1}^{T} f_t(\mathbf{x}_t)\right] - \sum_{t=1}^{T} f_t(\mathbf{u}) \leq \underbrace{\mathbb{E}\left[\sum_{t=1}^{T} \widehat{f}_t^{\delta}(\mathbf{y}_t) - \widehat{f}_t^{\delta}((1-\alpha)\mathbf{u})\right]}_{\textsc{Term}(\mathrm{A})} + 3G\delta T + GR\alpha T$$

Define $h_t(\mathbf{x}) \triangleq \widehat{f}_t^{\delta}(\mathbf{x}) + \boldsymbol{\xi}_t^\top \mathbf{x}, \quad \boldsymbol{\xi}_t = \widetilde{\mathbf{g}}_t - \nabla \widehat{f}_t^{\delta}(\mathbf{y}_t)$

- It is obvious that $\nabla h_t(\mathbf{y}_t) = \widetilde{\mathbf{g}}_t$
- for any *fixed* $\mathbf{x} \in \mathcal{X}$, we have $\mathbb{E}[h_t(\mathbf{x})] = \mathbb{E}[\widehat{f}_t^{\delta}(\mathbf{x})]$

$$\begin{aligned}\mathbb{E}\left[h_t(\mathbf{x})\right] &= \mathbb{E}\left[\widehat{f}_t^{\delta}(\mathbf{x})\right] + \mathbb{E}\left[\boldsymbol{\xi}_t^\top \mathbf{x}\right] = \mathbb{E}\left[\widehat{f}_t^{\delta}(\mathbf{x})\right] + \mathbb{E}\left[\mathbb{E}[\boldsymbol{\xi}_t^\top \mathbf{x} \mid \mathbf{x}_1, f_1, \ldots, \mathbf{x}_t, f_t]\right] \\ &= \mathbb{E}\left[\widehat{f}_t^{\delta}(\mathbf{x})\right] + \mathbb{E}\left[\mathbb{E}[\boldsymbol{\xi}_t \mid \mathbf{x}_1, f_1, \ldots, \mathbf{x}_t, f_t]^\top \mathbf{x}\right] = \mathbb{E}\left[\widehat{f}_t^{\delta}(\mathbf{x})\right]\end{aligned}$$

Proof of BGD

(for any *fixed* $\mathbf{x} \in \mathcal{X}$, we have $\mathbb{E}[h_t(\mathbf{x})] = \mathbb{E}[\widehat{f}_t^{\delta}(\mathbf{x})]$)

Proof. $\textsc{Term(A)}: \mathbb{E}\left[\sum_{t=1}^{T} \widehat{f}_t^{\delta}(\mathbf{y}_t) - \widehat{f}_t^{\delta}((1-\alpha)\mathbf{u})\right] = \mathbb{E}\left[\underbrace{\sum_{t=1}^{T} h_t(\mathbf{y}_t) - h_t((1-\alpha)\mathbf{u})}\right]$

(recall that $\nabla h_t(\mathbf{y}_t) = \widetilde{\mathbf{g}}_t$)

$\{\mathbf{y}_t\}_{t=1}^T$ is performing OGD on h_t over $(1-\alpha)\mathcal{X}$.

Theorem 4 (Regret of OGD) *Suppose $\{h_t\}_{t=1}^T$ are convex functions, the feasible set $\mathcal{X}$ is closed and convex, $\forall \mathbf{x}, \mathbf{y} \in \mathcal{X}, \|\mathbf{x}-\mathbf{y}\|_2 \le R$, and $\forall \mathbf{x} \in \mathcal{X}, \|\nabla h_t(\mathbf{x})\| \le G$. The OGD algorithm with fixed step size η satisfies that for any $\mathbf{u} \in \mathcal{X}$, $t \in [T]$:*

$$\sum_{t=1}^{T} h_t(\mathbf{y}_t) - \sum_{t=1}^{T} h_t((1-\alpha)\mathbf{u}) \le \frac{R^2}{2\eta} + \frac{G^2}{2}\eta T$$

$$\|\nabla h_t(\mathbf{x}_t)\|_2 = \|\widetilde{\mathbf{g}}_t\|_2 = \left\|\frac{d}{\delta} f_t(\mathbf{y}_t + \delta \mathbf{s}_t) \cdot \mathbf{s}_t\right\|_2 = \frac{dC}{\delta} \quad \Rightarrow \quad \textsc{Term(A)} \le \frac{R^2}{2\eta} + \frac{\eta d^2 C^2 T}{2\delta^2}$$

Proof of BGD

Proof. $\textsc{Term}(\mathrm{A}): \mathbb{E}\left[\sum_{t=1}^{T} \widehat{f}_t^{\delta}(\mathbf{y}_t) - \widehat{f}_t^{\delta}((1-\alpha)\mathbf{u})\right] = \mathbb{E}\left[\sum_{t=1}^{T} h_t(\mathbf{y}_t) - h_t((1-\alpha)\mathbf{u})\right]$

(for any *fixed* $\mathbf{x} \in \mathcal{X}$, we have $\mathbb{E}[h_t(\mathbf{x})] = \mathbb{E}[\widehat{f}_t^{\delta}(\mathbf{x})]$)

Since $\nabla h_t(\mathbf{y}_t) = \widetilde{\mathbf{g}}_t$, $\{\mathbf{y}_t\}_{t=1}^T$ can be regarded as performing OGD on $\{h_t\}_{t=1}^T$ over $(1-\alpha)\mathcal{X}$.

Theorem 4 (Regret of OGD) *Suppose $\{h_t\}_{t=1}^T$ are convex functions, the feasible set $\mathcal{X}$ is closed and convex, $\forall \mathbf{x}, \mathbf{y} \in \mathcal{X}, \|\mathbf{x} - \mathbf{y}\|_2 \le R$, and $\forall \mathbf{x} \in \mathcal{X}, \|\nabla h_t(\mathbf{x})\| \le G$. The OGD algorithm with fixed step size η satisfies that for any $\mathbf{u} \in \mathcal{X}$, $t \in [T]$:*

$$\sum_{t=1}^{T} h_t(\mathbf{y}_t) - \sum_{t=1}^{T} h_t((1-\alpha)\mathbf{u}) \le \frac{R^2}{2\eta} + \frac{G^2}{2}\eta T$$

$$\Rightarrow \quad \|\nabla h_t(\mathbf{x}_t)\|_2 = \|\widetilde{\mathbf{g}}_t\|_2 = \left\|\frac{d}{\delta} f_t(\mathbf{y}_t + \delta \mathbf{s}_t) \cdot \mathbf{s}_t\right\|_2 = \frac{dC}{\delta}$$

$$\Rightarrow \quad \textsc{Term}(\mathrm{A}) \le \frac{R^2}{2\eta} + \frac{\eta d^2 C^2 T}{2\delta^2}$$

Proof of BGD

Proof. To summarize, we have the following expected regret upper bound:

$$\mathbb{E}\left[\sum_{t=1}^{T} f_t(\mathbf{x}_t)\right] - \sum_{t=1}^{T} f_t(\mathbf{u})$$

$$= \textsc{Term}(\mathrm{A}) + \textsc{Term}(\mathrm{B}) + \textsc{Term}(\mathrm{C}) + \textsc{Term}(\mathrm{D}) + \textsc{Term}(\mathrm{E})$$

$$\leq \frac{R^2}{2\eta} + \frac{\eta d^2 C^2 T}{2\delta^2} + 3G\delta T + \alpha G R T$$

$$\leq \frac{R^2}{2\eta} + \frac{\eta d^2 C^2 T}{2\delta^2} + \left(3G + \frac{GR}{r}\right)\delta T$$

$$\leq \mathcal{O}\left(\sqrt{RdC\widetilde{G}}T^{\frac{3}{4}}\right) \qquad \text{with } \widetilde{G} \triangleq 3G + \tfrac{GR}{r}. \qquad \square$$

$$\alpha = \frac{\delta}{r}$$

$$\eta = \left(dC\widetilde{G}\right)^{-\frac{1}{2}} \left(\frac{R^2}{T}\right)^{\frac{3}{4}}$$

$$\delta = \left(\frac{dC}{\widetilde{G}}\right)^{\frac{1}{2}} \left(\frac{R^2}{T}\right)^{\frac{1}{4}}$$

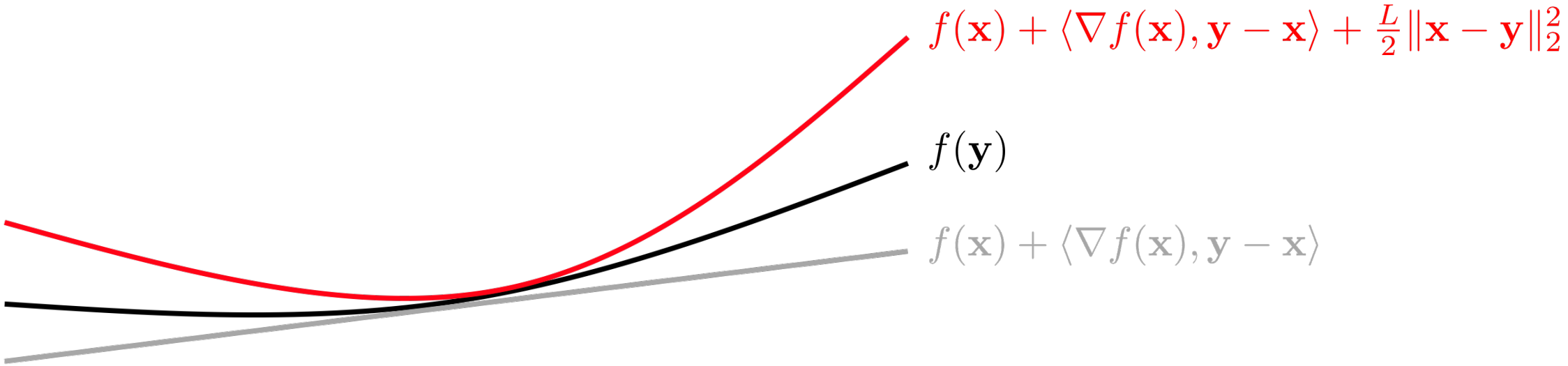
Part 4. BCO with Smooth Functions

- Problem Formulation
- Gradient Estimator
- Self-concordant Barrier
- Regret Analysis

BCO with Smooth Functions

Definition 2 (Smooth). Let f be an L-smooth function over a given convex set $\mathcal{X}$. Then for any $\mathbf{x}, \mathbf{y} \in \mathcal{X}$

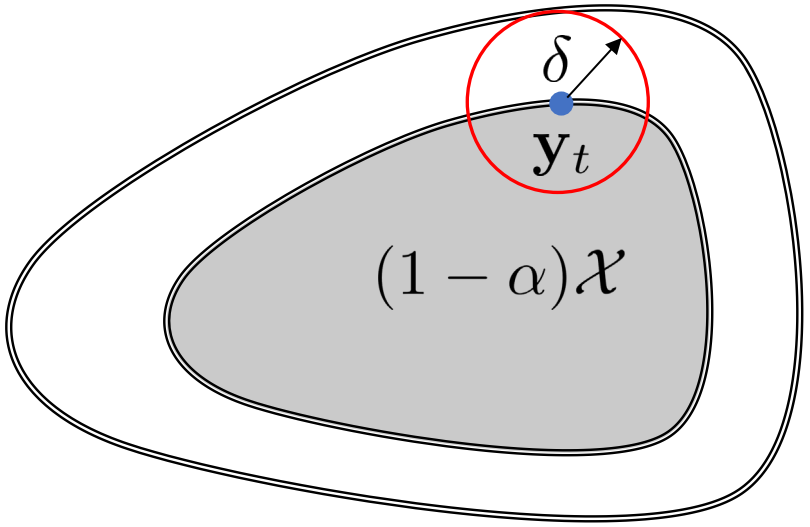
$$f(\mathbf{y}) \leq f(\mathbf{x}) + \nabla f(\mathbf{x})^\top (\mathbf{y} - \mathbf{x}) + \frac{L}{2}\|\mathbf{y} - \mathbf{x}\|^2.$$



Exploration

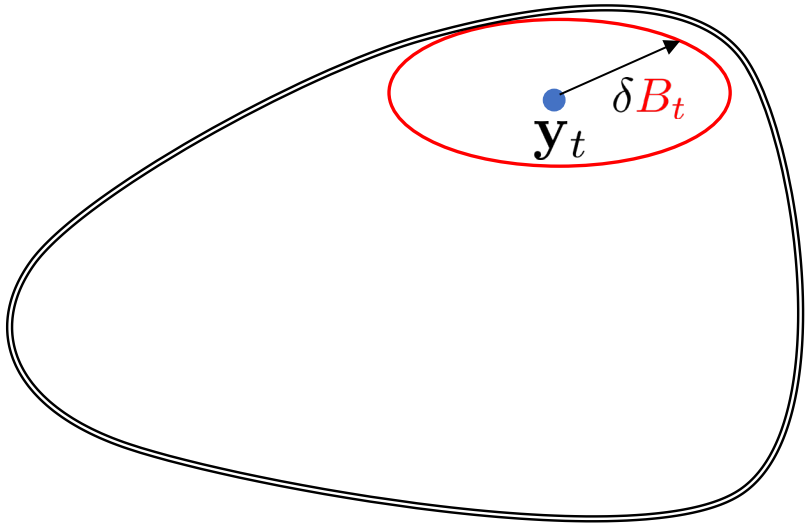
isotropy exploration strategy

$$\mathbf{x}_t = \mathbf{y}_t + \delta \mathbf{s}_t$$



anisotropy exploration strategy

$$\mathbf{x}_t = \mathbf{y}_t + \delta B_t \mathbf{s}_t$$



An Anisotropy Exploration

- We use several key tools in convex geometry and analysis, including the ***self-concordant barrier*** (formally defined later), ***Dikin ellipsoid***

Fact 1. *Let $\mathcal{R}$ be a self-concordant barrier on the closed convex set $\mathcal{X} \subseteq \mathbb{R}^d$. For every $\mathbf{x} \in \operatorname{int}(\mathcal{X})$, the unit Dikin ellipsoid defined as*

$$\mathcal{E}_1(\mathbf{x}) \triangleq \left\{ \boldsymbol{\xi} \in \mathbb{R}^d \mid \|\boldsymbol{\xi} - \mathbf{x}\|_{\nabla^2 \mathcal{R}(\mathbf{x})} \leq 1 \right\},$$

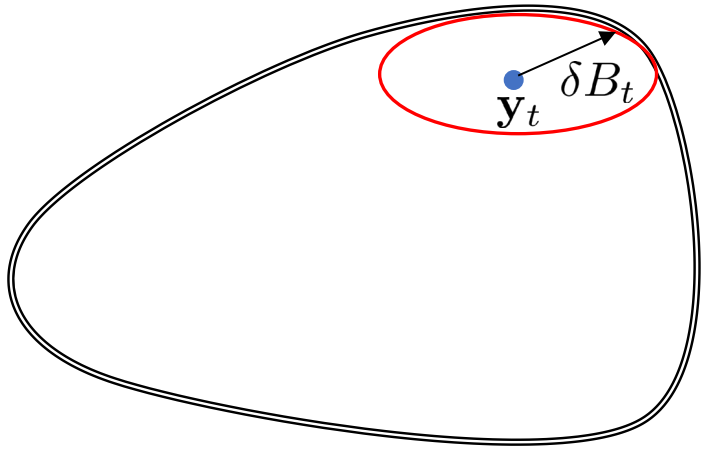
is completely contained in $\mathcal{X}$.

An Anisotropy Exploration

- Exploration using ***local norm*** with Dikin ellipsoid.

$$\mathbf{x}_t = \mathbf{y}_t + \delta B_t \mathbf{s}_t$$

$$B_t^2 = [\nabla^2 \mathcal{R}(\mathbf{y}_t)]^{-1}$$



Anisotropy exploration strategy
δ: magnitude, B_t: direction

Analysis:

$$\|\mathbf{x}_t - \mathbf{y}_t\|_{\nabla^2 \mathcal{R}(\mathbf{y}_t)} = \delta \sqrt{\mathbf{s}_t^\top B_t \nabla^2 \mathcal{R}(\mathbf{y}_t) B_t \mathbf{s}_t} \leq 1,$$

and recall in Lemma 2, we know $\mathcal{E}_1(\mathbf{x}) \triangleq \left\{ \boldsymbol{\xi} \in \mathbb{R}^d \mid \|\boldsymbol{\xi} - \mathbf{x}\|_{\nabla^2 \mathcal{R}(\mathbf{x})} \leq 1 \right\} \subseteq \mathcal{X}$,

so it has guaranteed the feasibility of $\mathbf{x}_t$ within $\mathcal{X}$.

FTRL with Self-Concordant Barrier

Input: a ν-self-concordant barrier $\mathcal{R}$ associated with the feasible domain $\mathcal{X}$

At each round $t = 1, 2, \cdots$

(1) define $B_t^2 = \nabla^{-2}\mathcal{R}(\mathbf{y}_t)$

(2) sample a unit vector $\mathbf{s}_t \in \mathbb{S}$;

(3) submit $\mathbf{x}_t = \mathbf{y}_t + \delta B_t \mathbf{s}_t$;

(4) receive feedback $f_t(\mathbf{x}_t)$;

(5) construct gradient estimator $\widetilde{\mathbf{g}}_t = \frac{d}{\delta} f_t(\mathbf{y}_t + \delta B_t \mathbf{s}_t) \cdot B_t^{-1} \mathbf{s}_t$;

(6) FTRL update

$$\mathbf{y}_{t+1} = \arg\min_{\mathbf{x} \in \mathcal{X}} \left\{ \sum_{s=1}^{t} \langle \widetilde{\mathbf{g}}_t, \mathbf{x} \rangle + \frac{1}{\eta} \mathcal{R}(\mathbf{x}) \right\}.$$

Invent the algorithm from the analysis

- We focus on the FTRL framework (OMD can be similarly obtained).

Theorem 4 (FTRL Regret). *Consider the FTRL update* $\mathbf{x}_t = \arg\min_{\mathbf{x}\in\mathcal{X}} F_t(\mathbf{x})$ *with* $F_t(\mathbf{x}) \triangleq \psi(\mathbf{x}) + \eta \sum_{s=1}^{t-1} \langle \mathbf{g}_s, \mathbf{x} \rangle$. *Then, for any* $\mathbf{u} \in \mathcal{X}$*:*

$$\sum_{t=1}^{T} \langle \mathbf{g}_t, \mathbf{x}_t \rangle - \sum_{t=1}^{T} \langle \mathbf{g}_t, \mathbf{u} \rangle \leq \underbrace{\sum_{t=1}^{T} \langle \mathbf{g}_t, \mathbf{x}_t - \mathbf{x}_{t+1} \rangle}_{\text{stability term}} + \underbrace{\frac{\psi(\mathbf{u}) - \psi(\mathbf{x}_1)}{\eta}}_{\text{range term}}.$$

We will choose the regularizer $\psi(\mathbf{x})$ as a ***self-concordant barrier*** $\mathcal{R} : \mathcal{X} \mapsto \mathbb{R}$.

Invent the algorithm from the analysis

FTRL update: $\mathbf{y}_t = \arg\min_{\mathbf{x}\in\mathcal{X}} \left\{ \mathcal{R}(\mathbf{x}) + \eta \sum_{s=1}^{t-1} \langle \mathbf{g}_s, \mathbf{x} \rangle \right\}$

$$\sum_{t=1}^{T} \langle \mathbf{g}_t, \mathbf{y}_t \rangle - \sum_{t=1}^{T} \langle \mathbf{g}_t, \mathbf{u} \rangle \leq \underbrace{\sum_{t=1}^{T} \langle \mathbf{g}_t, \mathbf{y}_t - \mathbf{y}_{t+1} \rangle}_{\text{stability term}} + \underbrace{\frac{\psi(\mathbf{u}) - \psi(\mathbf{y}_1)}{\eta}}_{\text{range term}}$$

When using barrier functions as the regularizer

- Stability term: need suitable local norm

  ⇨ *Newton decrement, suitable gradient estimator*

- Range term: $\psi(\mathbf{u})$ could be infinite

  ⇨ *shifting comparators, controlling the shifting cost*

Stability Term and Newton Decrement

FTRL update: $\mathbf{y}_t = \arg\min_{\mathbf{x}\in\mathcal{X}} \left\{\mathcal{R}(\mathbf{x}) + \eta \sum_{s=1}^{t-1} \langle \mathbf{g}_s, \mathbf{x}\rangle\right\}$

$$\sum_{t=1}^{T}\langle \mathbf{g}_t, \mathbf{y}_t\rangle - \sum_{t=1}^{T}\langle \mathbf{g}_t, \mathbf{u}\rangle \le \underbrace{\sum_{t=1}^{T}\langle \mathbf{g}_t, \mathbf{y}_t - \mathbf{y}_{t+1}\rangle}_{\text{stability term}} + \underbrace{\frac{\psi(\mathbf{u}) - \psi(\mathbf{y}_1)}{\eta}}_{\text{range term}}$$

Fact 2. *Let $\mathcal{R}$ be a self-concordant function on $\mathcal{X}$ and $\mathbf{x}^*$ be its minimizer. For every $\mathbf{x} \in \mathcal{X}$, if its Newton decrement $\lambda_{\mathcal{R}}(\mathbf{x}) \triangleq \|\nabla\mathcal{R}(\mathbf{x})\|_{\nabla^{-2}\mathcal{R}(\mathbf{x})} \le 1/2$, then we have $\|\mathbf{x} - \mathbf{x}^*\|_{\nabla^2\mathcal{R}(\mathbf{x})} \le 2\lambda_{\mathcal{R}}(\mathbf{x})$.*

Netwon decrement vanishes exactly at the (unique, if any) minimizer $\mathbf{x}^*$ of f on $\mathrm{int}(\mathcal{X})$, and this function can be considered as an "observable" measure of proximity of $\mathbf{x}$ to $\mathbf{x}^*$.

Stability Term and Newton Decrement

Fact 2. *Let $\mathcal{R}$ be a self-concordant function on $\mathcal{X}$ and $\mathbf{x}^*$ be its minimizer. For every $\mathbf{x} \in \mathcal{X}$, if its Newton decrement $\lambda_{\mathcal{R}}(\mathbf{x}) \triangleq \|\nabla \mathcal{R}(\mathbf{x})\|_{\nabla^{-2}\mathcal{R}(\mathbf{x})} \leq 1/2$, then we have $\|\mathbf{x} - \mathbf{x}^*\|_{\nabla^2 \mathcal{R}(\mathbf{x})} \leq 2\lambda_{\mathcal{R}}(\mathbf{x})$.*

FTRL update: $\mathbf{y}_t = \arg\min_{\mathbf{x} \in \mathcal{X}} F_t(\mathbf{x}) \triangleq \left\{ \mathcal{R}(\mathbf{x}) + \eta \sum_{s=1}^{t-1} \langle \mathbf{g}_s, \mathbf{x} \rangle \right\}$

Analysis: $\lambda_{F_{t+1}}(\mathbf{y}_t) = \left\| \nabla \mathcal{R}(\mathbf{y}_t) + \eta \sum_{s=1}^{t} \mathbf{g}_s \right\|_{\nabla^{-2}\mathcal{R}(\mathbf{y}_t)} = \eta \|\mathbf{g}_t\|_{\nabla^{-2}\mathcal{R}(\mathbf{y}_t)} \quad (\nabla F_t(\mathbf{y}_t) = 0)$

If $\lambda_{F_{t+1}}(\mathbf{y}_t) \leq \frac{1}{2}$, we have $\|\mathbf{y}_t - \mathbf{y}_{t+1}\|_{\nabla^2 F_t(\mathbf{y}_t)} \leq 2\lambda_{F_{t+1}}(\mathbf{y}_t) = 2\eta \|\mathbf{g}_t\|_{\nabla^{-2}\mathcal{R}(\mathbf{y}2_t)}$.

$\Rightarrow \quad \langle \mathbf{g}_t, \mathbf{y}_t - \mathbf{y}_{t+1} \rangle \leq \|\mathbf{g}_t\|_{\nabla^{-2}\mathcal{R}(\mathbf{y}_t)} \|\mathbf{y}_t - \mathbf{y}_{t+1}\|_{\nabla^2 \mathcal{R}(\mathbf{y}_t)} \leq 2\eta \|\mathbf{g}_t\|^2_{\nabla^{-2}\mathcal{R}(\mathbf{y}_t)}$

Range Term and Shifting Cost

FTRL update: $\mathbf{y}_t = \arg\min_{\mathbf{x}\in\mathcal{X}} \left\{\mathcal{R}(\mathbf{x}) + \eta \sum_{s=1}^{t-1} \langle \mathbf{g}_s, \mathbf{x}\rangle\right\}$

$$\sum_{t=1}^{T} \langle \mathbf{g}_t, \mathbf{x}_t\rangle - \sum_{t=1}^{T} \langle \mathbf{g}_t, \mathbf{u}\rangle \le \underbrace{\sum_{t=1}^{T} \langle \mathbf{g}_t, \mathbf{x}_t - \mathbf{x}_{t+1}\rangle}_{\text{stability term}} + \underbrace{\frac{\psi(\mathbf{u}) - \psi(\mathbf{x}_1)}{\eta}}_{\text{range term}}$$

Fact 3. *If $\mathcal{R}$ is a ν-self-concordant barrier on $\mathcal{X}$, then for any $\mathbf{x}, \mathbf{y} \in \operatorname{int}(\mathcal{X})$, we have $\mathcal{R}(\mathbf{y}) - \mathcal{R}(\mathbf{x}) \le \nu \log \frac{1}{1-\pi_{\mathcal{X}}(\mathbf{y})}$, where $\pi_{\mathbf{x}}(\mathbf{y}) = \inf\{t \ge 0 \mid \mathbf{x} + t^{-1}(\mathbf{y}-\mathbf{x}) \in \mathcal{X}\}$ is called the Minkowski function of $\mathcal{X}$ with respect to $\mathbf{x}$, which is always in $[0, 1]$. Moreover, as $\mathbf{y}$ approaches to $\partial\mathcal{X}$, $\pi_{\mathcal{X}}(\mathbf{y}) \to 0$.*

Range Term and Shifting Cost

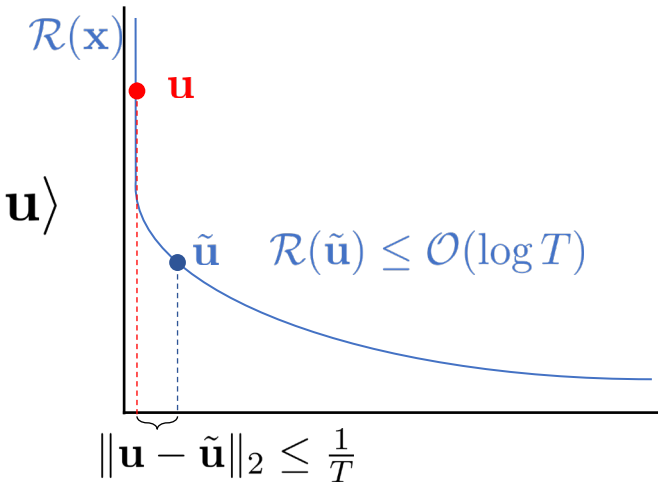
Analysis: For comparator $\mathbf{u} \in \mathcal{X}$, we shift it as $\tilde{\mathbf{u}} = \frac{1}{1+\epsilon}\mathbf{u} + \frac{\epsilon}{1+\epsilon}\mathbf{x}_1$ for some $\epsilon > 0$.

Then, Fact 3 ensures $\mathcal{R}(\tilde{\mathbf{u}}) - \mathcal{R}(\mathbf{x}_1) \leq \nu \ln\left(\frac{1}{\epsilon} + 1\right)$.

$$\mathcal{R}(\mathbf{y}) - \mathcal{R}(\mathbf{x}) \leq \nu \log \frac{1}{1-\pi_{\mathcal{X}}(\mathbf{y})}, \text{ with } \pi_{\mathbf{x}}(\mathbf{y}) = \inf\{t \geq 0 \mid \mathbf{x} + t^{-1}(\mathbf{y} - \mathbf{x}) \in \mathcal{X}\}$$

Let $\epsilon = \frac{1}{T-1}$, then $\mathcal{R}(\widetilde{\mathbf{u}}) - \mathcal{R}(\mathbf{x}_1) \leq \nu \ln T$ and $\|\mathbf{u} - \widetilde{\mathbf{u}}\|_2 \leq \frac{D}{T}$

$$\begin{aligned}\sum_{t=1}^{T}\langle \mathbf{g}_t, \mathbf{y}_t\rangle - \sum_{t=1}^{T}\langle \mathbf{g}_t, \mathbf{u}\rangle &= \sum_{t=1}^{T}\langle \mathbf{g}_t, \mathbf{y}_t\rangle - \sum_{t=1}^{T}\langle \mathbf{g}_t, \widetilde{\mathbf{u}}\rangle + \sum_{t=1}^{T}\langle \mathbf{g}_t, \widetilde{\mathbf{u}} - \mathbf{u}\rangle \\ &\leq 2\eta \sum_{t=1}^{T} \|\mathbf{g}_t\|^2_{\nabla^{-2}\mathcal{R}(\mathbf{x}_t)} + \frac{\nu \log T}{\eta} + \mathcal{O}(1)\end{aligned}$$



FTRL with Self-Concordant Barrier

- Putting above components together yields the following results.

Theorem 5 (Regret Bound for FTRL). *Assume $\mathcal{R}$ is a ν-self-concordant barrier on $\mathcal{X}$, if we run FTRL algorithm over the online functions $\{f_1, \ldots, f_T\}$ and obtains*

$$\mathbf{y}_t = \arg\min_{\mathbf{x}\in\mathcal{X}} \left\{ \mathcal{R}(\mathbf{x}) + \eta \sum_{s=1}^{t-1} \langle \mathbf{g}_t, \mathbf{x} \rangle \right\}$$

with gradient estimates $\mathbb{E}[\mathbf{g}_t] = \nabla h_t(\mathbf{y}_t)$, then it holds that for any $\mathbf{u} \in \mathcal{X}$,

$$\mathbb{E}\left[\sum_{t=1}^{T} h_t(\mathbf{y}_t) - \sum_{t=1}^{T} h_t(\mathbf{u}) \right] \leq \frac{\nu \log T}{\eta} + 2\eta \sum_{t=1}^{T} \mathbb{E}\left[\|\mathbf{g}_t\|^2_{\nabla^{-2}\mathcal{R}(\mathbf{y}_t)} \right] + \mathcal{O}(1).$$

Gradient Estimator

Definition 4 (Gradient Estimator). The gradient estimator is defined as

$$\widetilde{\mathbf{g}}_t = \frac{d}{\delta} f_t(\mathbf{y}_t + \delta B_t \mathbf{s}_t) \cdot B_t^{-1} \mathbf{s}_t, \qquad B_t^2 = \nabla^{-2} \mathcal{R}(\mathbf{y}_t),$$

where $\mathbf{s}_t$ is sampled from unit sphere $\mathbb{S} = \{\mathbf{x} \in \mathbb{R}^d \mid \|\mathbf{x}\|_2 = 1\}$.

Lemma 3 *For any convex function $f_t : \mathcal{X} \mapsto \mathbb{R}$, define its smoothed version $\widehat{f}_t(\mathbf{x}) = \mathbb{E}_{\mathbf{v} \in \mathbb{B}^d}[f_t(\mathbf{x} + \delta B_t \mathbf{v})]$. Then for any $\delta > 0$, we have*

$$\mathbb{E}[\widetilde{\mathbf{g}}_t] = \mathbb{E}_{\mathbf{s} \in \mathbb{S}} \left[\frac{d}{\delta} f_t(\mathbf{x}_t + \delta B_t \mathbf{s}_t) \cdot B_t^{-1} \mathbf{s}_t \right] = \nabla \widehat{f}_t(\mathbf{x}_t),$$

where $\mathbb{B} = \{\mathbf{x} \in \mathbb{R}^d \mid \|\mathbf{x}\|_2 \leq 1\}$ is the unit ball and $\mathbb{S}$ is the unit sphere.

Gradient Estimator

Definition 4 (Gradient Estimator). The gradient estimator is defined as

$$\widetilde{\mathbf{g}}_t = \frac{d}{\delta} f_t(\mathbf{y}_t + \delta B_t \mathbf{s}_t) \cdot B_t^{-1} \mathbf{s}_t, \qquad B_t^2 = \nabla^{-2} \mathcal{R}(\mathbf{y}_t),$$

where $\mathbf{s}_t$ is sampled from unit sphere $\mathbb{S} = \{\mathbf{x} \in \mathbb{R}^d \mid \|\mathbf{x}\|_2 = 1\}$.

Lemma 4 (Gradient Local Norm) *The elliptical graident estimator satisfies*

$$\|\widetilde{\mathbf{g}}_t\|^2_{\nabla^{-2}\mathcal{R}(\mathbf{y}_t)} = \left\| \frac{d}{\delta} f_t(\mathbf{y}_t + \delta B_t \mathbf{s}_t) \cdot B_t^{-1} \mathbf{s}_t \right\|^2_{\nabla^{-2}\mathcal{R}(\mathbf{y}_t)} \leq \frac{d^2 C^2}{\delta^2}.$$

FTRL with Self-Concordant Barrier

Input: a ν-self-concordant barrier $\mathcal{R}$ associated with the feasible domain $\mathcal{X}$

At each round $t = 1, 2, \cdots$

(1) define $B_t^2 = \nabla^{-2}\mathcal{R}(\mathbf{y}_t)$

(2) sample a unit vector $\mathbf{s}_t \in \mathbb{S}$;

(3) submit $\mathbf{x}_t = \mathbf{y}_t + \delta B_t \mathbf{s}_t$;

(4) receive feedback $f_t(\mathbf{x}_t)$;

(5) construct gradient estimator $\widetilde{\mathbf{g}}_t = \frac{d}{\delta} f_t(\mathbf{y}_t + \delta B_t \mathbf{s}_t) \cdot B_t^{-1}\mathbf{s}_t$;

(6) FTRL update

$$\mathbf{y}_{t+1} = \arg\min_{\mathbf{x}\in\mathcal{X}} \left\{ \sum_{s=1}^{t} \langle \widetilde{\mathbf{g}}_t, \mathbf{x} \rangle + \frac{1}{\eta}\mathcal{R}(\mathbf{x}) \right\}.$$

FTRL with Self-Concordant Barrier

Theorem 6 *Suppose the online function $f_t : \mathcal{X} \mapsto \mathbb{R}$ is L-smooth, $\max_{\mathbf{x}\in\mathcal{X}} |f_t(\mathbf{x})| \leq C$, and for any $\mathbf{x}_1, \mathbf{x}_2 \in \mathcal{X}$, $\|\mathbf{x}_1 - \mathbf{x}_2\|_2 \leq D$. Then, FTRL with a self-concordant barrier algorithm satisfies*

$$\begin{aligned}\mathbb{E}\left[\sum_{t=1}^{T} f_t(\mathbf{x}_t)\right] - \sum_{t=1}^{T} f_t(\mathbf{u}) &\leq \mathcal{O}\left(\frac{\nu \log T}{\eta} + \frac{\eta T}{\delta^2} + \delta^2 T\right) \\ &= \mathcal{O}\left(T^{\frac{2}{3}} (\log T)^{\frac{1}{3}}\right),\end{aligned}$$

where the last step holds by setting the step size $\eta = \mathcal{O}((\log T/T)^{2/3})$ and the perturbation parameter $\delta = \eta^{1/4}$.

Proof Sketch

$$\mathbb{E}\left[\sum_{t=1}^{T} f_t(\mathbf{x}_t)\right] - \sum_{t=1}^{T} f_t(\mathbf{u})$$

$$= \mathbb{E}\left[\sum_{t=1}^{T} f_t(\mathbf{x}_t) - f_t(\widetilde{\mathbf{u}})\right] + \mathbb{E}\left[\sum_{t=1}^{T} f_t(\widetilde{\mathbf{u}}) - f_t(\mathbf{u})\right]$$

$$\leq \underbrace{\mathbb{E}\left[\sum_{t=1}^{T} \widehat{f}_t(\mathbf{y}_t) - \widehat{f}_t(\widetilde{\mathbf{u}})\right]}_{\leq \frac{\nu \log T}{\eta} + \frac{\eta T}{\delta^2}} + \underbrace{\mathbb{E}\left[\sum_{t=1}^{T} \widehat{f}_t(\mathbf{x}_t) - \widehat{f}_t(\mathbf{y}_t)\right] + \mathbb{E}\left[\sum_{t=1}^{T} \widehat{f}_t(\widetilde{\mathbf{u}}) - f_t(\widetilde{\mathbf{u}})\right]}_{\leq LD\delta^2 T} + 0 + \underbrace{\mathbb{E}\left[\sum_{t=1}^{T} f_t(\widetilde{\mathbf{u}}) - f_t(\mathbf{u})\right]}_{\leq \widetilde{L}}$$

Regret of FTRL with Self-concordant barrier

Smoothness introduces an additional $\frac{L}{2}\|\delta B_t \mathbf{s}_t\|^2$, which improves δ to δ^2

Comparator shifting causes additional constant $\widetilde{L} \triangleq \left(\frac{C}{D} + \frac{LD}{2}\right)$

Self-Concordant Functions/Barriers

Fact 1. *Let $\mathcal{R}$ be a self-concordant barrier on the closed convex set $\mathcal{X} \subseteq \mathbb{R}^d$. For every $\mathbf{x} \in \operatorname{int}(\mathcal{X})$, the unit Dikin ellipsoid defined as $\mathcal{E}_1(\mathbf{x}) \triangleq \left\{\boldsymbol{\xi} \in \mathbb{R}^d \mid \|\boldsymbol{\xi} - \mathbf{x}\|_{\nabla^2 \mathcal{R}(\mathbf{x})} \leq 1\right\}$ is completely contained in $\mathcal{X}$.*

Fact 2. *Let $\mathcal{R}$ be a self-concordant function on $\mathcal{X}$ and $\mathbf{x}^*$ be its minimizer. For every $\mathbf{x} \in \mathcal{X}$, if its Newton decrement $\lambda_{\mathcal{R}}(\mathbf{x}) \triangleq \|\nabla \mathcal{R}(\mathbf{x})\|_{\nabla^{-2} \mathcal{R}(\mathbf{x})} \leq 1/2$, then we have $\|\mathbf{x} - \mathbf{x}^*\|_{\nabla^2 \mathcal{R}(\mathbf{x})} \leq 2\lambda_{\mathcal{R}}(\mathbf{x})$.*

Fact 3. *If $\mathcal{R}$ is a ν-self-concordant barrier on $\mathcal{X}$, then for any $\mathbf{x}, \mathbf{y} \in \operatorname{int}(\mathcal{X})$, we have $\mathcal{R}(\mathbf{y}) - \mathcal{R}(\mathbf{x}) \leq \nu \log \frac{1}{1 - \pi_{\mathcal{X}}(\mathbf{y})}$, where $\pi_{\mathbf{x}}(\mathbf{y}) = \inf\{t \geq 0 \mid \mathbf{x} + t^{-1}(\mathbf{y} - \mathbf{x}) \in \mathcal{X}\}$ is called the Minkowski function of $\mathcal{X}$ with respect to $\mathbf{x}$, which is always in $[0, 1]$. Moreover, as $\mathbf{y}$ approaches to $\partial \mathcal{X}$, $\pi_{\mathcal{X}}(\mathbf{y}) \to 0$.*

Self-Concordant Functions

Definition 3 (Self-Concordant Functions)**.** Let $\mathcal{X} \subseteq \mathbb{R}^d$ be a closed convex domain with a nonempty interior $\text{int}(\mathcal{X})$. A function $\mathcal{R} : \text{int}(\mathcal{X}) \mapsto \mathbb{R}$ is called *self-concordant* on $\mathcal{X}$ if

(i) $\mathcal{R}$ is a three times continuously differentiable convex function, and approaches infinity along any sequence of points approaching $\partial\mathcal{X}$; and

(ii) $\mathcal{R}$ satisfies the differential inequality: for every $\mathbf{h} \in \mathbb{R}^d$ and $\mathbf{x} \in \text{int}(\mathcal{X})$,

$$|D^3\mathcal{R}(\mathbf{x})[\mathbf{h},\mathbf{h},\mathbf{h}]| \leq 2(D^2\mathcal{R}(\mathbf{x})[\mathbf{h},\mathbf{h}])^{\frac{3}{2}},$$

where the third-order differential is defined as

$$D^3\mathcal{R}(\mathbf{x})[\mathbf{h},\mathbf{h},\mathbf{h}] \triangleq \frac{\partial^3}{\partial t_1 \partial t_2 \partial t_3} \mathcal{R}\left(\mathbf{x} + t_1\mathbf{h} + t_2\mathbf{h} + t_3\mathbf{h}\right)\Big|_{t_1=t_2=t_3=0}.$$

Self-Concordant Barrier

Definition 4 (ν-Self-Concordant Barrier)**.** Given a real $\nu \geq 1$, $\mathcal{R}$ is called a *ν-self-concordant barrier* (ν-SCB) for $\mathcal{X}$ if $\mathcal{R}$ is self-concordant on $\mathcal{X}$ and, in addition, for every $\mathbf{h} \in \mathbb{R}^d$ and $\mathbf{x} \in \operatorname{int}(\mathcal{X})$,

$$|D\mathcal{R}(\mathbf{x})[\mathbf{h}]| \leq \nu^{\frac{1}{2}} \left(D^2\mathcal{R}(\mathbf{x})[\mathbf{h}, \mathbf{h}]\right)^{\frac{1}{2}}.$$

The notion of self-concordant barrier is defined based on the notion of self-concordant function. Thus, a self-concordant function is ***not*** necessarily a self-concordant barrier.

The self-concordant barrier is associated with the (convex) feasible domain.

Example 1. The function $f(\mathbf{x}) = \text{constant}$ is a 0-self-concordant barrier for $\mathbb{R}^d$.

Example 2. The function $f(\mathbf{x}) = -\ln \mathbf{x}$ is a 1-self-concordant barrier for the non-negative half-axis.

Self-Concordant Barrier

Definition 4 (ν-Self-Concordant Barrier)**.** Given a real $\nu \geq 1$, $\mathcal{R}$ is called a *ν-self-concordant barrier* (ν-SCB) for $\mathcal{X}$ if $\mathcal{R}$ is self-concordant on $\mathcal{X}$ and, in addition, for every $\mathbf{h} \in \mathbb{R}^d$ and $\mathbf{x} \in \operatorname{int}(\mathcal{X})$,

$$|D\mathcal{R}(\mathbf{x})[\mathbf{h}]| \leq \nu^{\frac{1}{2}} \left(D^2\mathcal{R}(\mathbf{x})[\mathbf{h},\mathbf{h}]\right)^{\frac{1}{2}}.$$

2.5 Universal barrier

In this section, we demonstrate that an arbitrary n-dimensional closed convex domain admits an $O(n)$-self-concordant barrier. This barrier is given by certain universal construction and, for this reason, will be called *universal*. In fact, the universal barrier usually is too complicated to be used in interior-point algorithms, so that what follows should be regarded as nothing but an existence theorem. At the same time, this existence theorem is very important theoretically, since it means that the approach we are developing *in principle* can be applied to *any* convex problem.

History bits: Self-Concordant Barrier

Definition 3 (Self-Concordant Functions). Let $\mathcal{X} \subseteq \mathbb{R}^d$ be a closed convex domain with a nonempty interior $\mathrm{int}(\mathcal{X})$. A function $\mathcal{R} : \mathrm{int}(\mathcal{X}) \mapsto \mathbb{R}$ is called self-concordant on $\mathcal{X}$ if

(i) $\mathcal{R}$ is a three times continuously differentiable convex function, and approaches infinity along any sequence of points approaching $\partial\mathcal{X}$;

(ii) $\mathcal{R}$ satisfies the differential inequality: for every $\mathbf{h} \in \mathbb{R}^d$ and $\mathbf{x} \in \mathrm{int}(\mathcal{X})$,

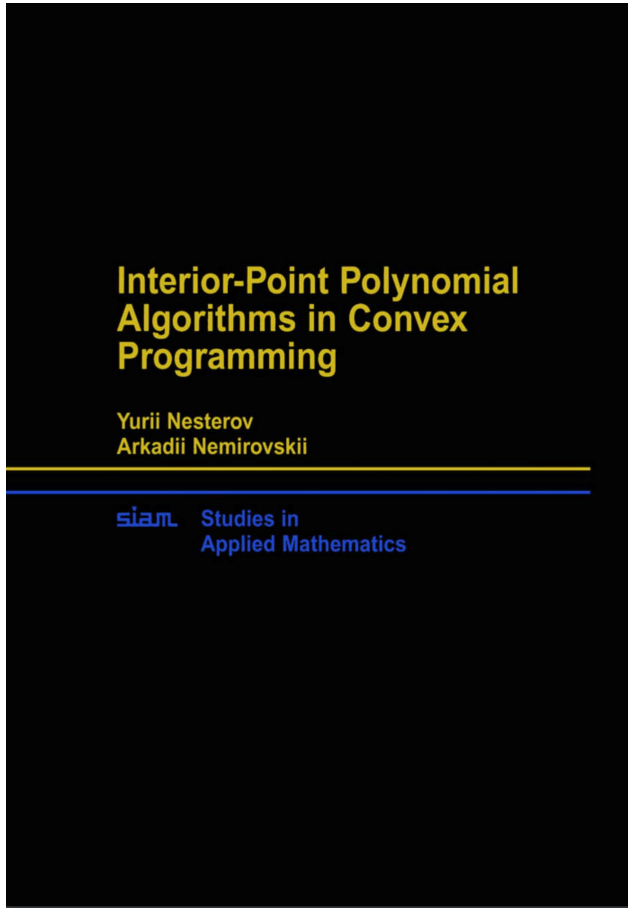
$$\left|D^3\mathcal{R}(\mathbf{x})[\mathbf{h},\mathbf{h},\mathbf{h}]\right| \leq 2\left(D^2\mathcal{R}(\mathbf{x})[\mathbf{h},\mathbf{h}]\right)^{\frac{3}{2}},$$

where the third-order differential is defined as

$$D^3\mathcal{R}(\mathbf{x})[\mathbf{h},\mathbf{h},\mathbf{h}] \triangleq \frac{\partial^3}{\partial t_1 \partial t_2 \partial t_3}\mathcal{R}\left(\mathbf{x}+t_1\mathbf{h}+t_2\mathbf{h}+t_3\mathbf{h}\right)\Big|_{t_1=t_2=t_3=0}.$$

At the same time, Nesterov and Nemirovski were investigating the new methods from a more fundamental viewpoint: what are the basic properties that lead to polynomial-time complexity? It turned out that the key property is that the barrier function should be *self-concordant*. This seemed to provide a clear, complexity-based criterion to delineate the class of optimization problems that could be solved in a provably efficient way using the new methods. The culmination of this work was the book (Nesterov and Nemirovski 1994), whose complexity emphasis contrasted with the classic text on barrier methods by Fiacco and McCormick (1968).

Arkadi S. Nemirovski and Michael J. Todd, Interior-point methods for optimization, Acta Numerica, 2008



Yurii Nesterov and Arkadi S. Nemirovski, Interior Point Polynomial Methods in Convex Programming, SIAM, 1994.

Beyond

- Can we further improve the dependence on T?

⇨ If loss function is ***linear***, then using FTRL with self-concordant barrier on $\mathcal{X}$

$$\mathbb{E}\left[\sum_{t=1}^{T} f_t(\mathbf{x}_t)\right] - \min_{\mathbf{x}\in\mathcal{X}} \sum_{t=1}^{T} f_t(\mathbf{x}) \leq \widetilde{\mathcal{O}}\left(T^{1/2}\right)$$

⇨ If loss function is ***strongly convex and smooth***, then using FTRL with self-concordant barrier on $\mathcal{X}$

$$\mathbb{E}\left[\sum_{t=1}^{T} f_t(\mathbf{x}_t)\right] - \min_{\mathbf{x}\in\mathcal{X}} \sum_{t=1}^{T} f_t(\mathbf{x}) \leq \widetilde{\mathcal{O}}\left(T^{1/2}\right)$$

Online convex optimization in the bandit setting: gradient descent without a gradient

Abraham D. Flaxman * Adam Tauman Kalai † H. Brendan McMahan ‡

November 30, 2004

Abstract

We study a general online convex optimization problem. We have a convex set S and an unknown sequence of cost functions $c_1, c_2, \ldots$, and in each period, we choose a feasible point x_t in S, and learn the cost $c_t(x_t)$. If the function c_t is also revealed after each period then, as Zinkevich shows in [25], gradient descent can be used on these functions to get regret bounds of $O(\sqrt{n})$. That is, after n rounds, the total cost incurred will be $O(\sqrt{n})$ more than the cost of the best single feasible decision chosen with the benefit of hindsight, $\min_x \sum c_t(x)$.

We extend this to the "bandit" setting, where, in each period, only the cost $c_t(x_t)$ is revealed, and bound the expected regret as $O(n^{3/4})$.

Our approach uses a simple approximation of the gradient that is computed from evaluating c_t at a single (random) point. We show that this biased estimate is sufficient to approximate gradient descent on the sequence of functions. In other words, it is possible to use gradient descent without seeing anything more than the value of the functions at a single point. The guarantees hold even in the most general case: online against an adaptive adversary.

For the online linear optimization problem [15], algorithms with low regrets in the bandit setting have recently been given against oblivious [1] and adaptive adversaries [19]. In contrast to these algorithms, which distinguish between explicit *explore* and *exploit* periods, our algorithm can be interpreted as doing a small amount of exploration in each period.

1 Introduction

Consider three optimization settings where one would like to minimize a convex function (equivalently maximize a concave function). In all three settings, gradient descent is one of the most popular methods.

1. Offline: Minimize a fixed convex cost function $c\colon \mathbb{R}^d \to \mathbb{R}$. In this case, gradient descent is $x_{t+1} = x_t - \eta \nabla c(x_t)$.
2. Stochastic: Minimize a fixed convex cost function c given only "noisy" access to c. For example, at time $T = t$, we may only have access to $c_t(x) = c(x) + \epsilon_t(x)$, where $\epsilon_t(x)$ is a random sampling error. Here, stochastic gradient descent is $x_{t+1} = x_t - \eta \nabla c_t(x_t)$. (The intuition is that the expected gradient is correct; $\mathbf{E}[\nabla c_t(x)] = \nabla \mathbf{E}[c_t(x)] = \nabla c(x)$.) In non-convex cases, the additional randomness may actually help avoid local minima [3], in a manner similar to Simulated Annealing [13].
3. Online: Minimize an adversarially generated sequence of convex functions, $c_1, c_2, \ldots$. This requires that we choose a sequence $x_1, x_2, \ldots$ where each x_t is selected based only on $x_1, x_2, \ldots, x_{t-1}$ and $c_1, c_2, \ldots, c_{t-1}$. The goals is to have low *regret* $\sum c_t(x_t) - \min_{x \in S} \sum c_t(x)$ for not using the best single point, chosen with the benefit of hindsight. In this setting, Zinkevich analyzes the regret of gradient descent given by $x_{t+1} = x_t - \eta \nabla c_t(x_t)$.

We will focus on gradient descent in a "bandit" version of the online setting. As a motivating example, consider a company that has to decide, every week, how much to spend advertising on each of d different channels, represented as a vector $x_t \in \mathbb{R}^d$. At the end of each week, they calculate their total profit $p_t(x_t)$. In the offline case, one might assume that each week the function $p_1, p_2, \ldots$ are identical. In the stochastic case, one might assume that in different weeks the profit functions $p_t(x)$ will be noisy realizations of some underlying "true" profit function, for example $p_t(x) = p(x) + \epsilon_t(x)$, where $\epsilon_t(x)$ has mean 0. In the online case, *no assumptions* are made about a distribution over convex profit functions and instead they are modeled as the malicious choices of an adversary. This allows, for example, for more complicated time-dependent random noise or the effects of a bad economy, or even an environment that responds to the choices we make (an adaptive adversary).

*http://www.math.cmu.edu/~adf, Department of Mathematical Sciences, Carnegie Mellon University.

†http://people.cs.uchicago.edu/~kalai, Toyota Technical Institute at Chicago.

‡http://www.cs.cmu.edu/~mcmahan, Department of Computer Science, Carnegie Mellon University.

Abraham D. Flaxman, Adam Tauman Kalai, and H. Brendan McMahan. Online convex optimization in the bandit setting: gradient descent without a gradient. SODA, 2004.

Competing in the Dark: An Efficient Algorithm for Bandit Linear Optimization

Jacob Abernethy
Computer Science Division
UC Berkeley
jake@cs.berkeley.edu
(eligible for best student paper award)

Elad Hazan
IBM Almaden
hazan@us.ibm.com

Alexander Rakhlin
Computer Science Division
UC Berkeley
rakhlin@cs.berkeley.edu

Abstract

We introduce an *efficient* algorithm for the problem of online linear optimization in the bandit setting which achieves the optimal $O^*(\sqrt{T})$ regret. The setting is a natural generalization of the non-stochastic multi-armed bandit problem, and the existence of an efficient optimal algorithm has been posed as an open problem in a number of recent papers. We show how the difficulties encountered by previous approaches are overcome by the use of a self-concordant potential function. Our approach presents a novel connection between online learning and interior point methods.

1 Introduction

One's ability to learn and make decisions rests heavily on the availability of feedback. Indeed, an agent may only improve itself when it can reflect on the outcomes of its own taken actions. In many environments feedback is readily available: a gambler, for example, can observe entirely the outcome of a horse race regardless of where he placed his bet. But such perspective is not always available in hindsight. When the same gambler chooses his route to travel to the race track, perhaps at a busy hour, he will likely never learn the outcome of possible alternatives. When betting on horses, the gambler has thus the benefit (or perhaps the detriment) to muse *"I should have done..."*, yet when betting on traffic he can only think *"the result was..."*.

This problem of sequential decision making was stated by Robbins [19] in 1952 and was later termed "the multi-armed bandit problem". The name inherits from the model whereby, on each of a sequence of rounds, a gambler must pull the arm on one of several slot machines ("one-armed bandits") that each returns a reward chosen stochastically from a fixed distribution. Of course, an ideal strategy would simply be to pull the arm of the machine with the greatest rewards. However, as the gambler does not know the best arm a priori, his goal is then to maximize the reward of his strategy relative to reward he would receive had he known the optimal arm. This problem has gained nuch interest over the past 20 years in a number of fields, as it presents a very natural model of an agent seeking to simultaneously explore the world while exploiting high-reward actions.

As early as 1990 [8, 13] the sequential decision problem was studied under *adversarial* assumptions, where we assume the environment may even try to hurt the learner. The multi-armed bandit problem was brought into the adversarial learning model in 2002 by Auer et al [1], who showed that one may obtain nontrivial guarantees on the gambler's performance relative to the best arm *even when the arm values are chosen by adversary*! In particular, Auer et al [1] showed that the gambler's *regret*, i.e. the difference between the gain of the best arm minus the gain of the gambler, can be bounded by $O(\sqrt{NT})$ where N is the number of bandit arms, and T is the length of the game. In comparison to the game where the gambler is given full information about alternative arms (such as the horse racing example mentioned above), it is possible to obtain $O(\sqrt{T \log N})$, which scales better in N but identically in T.

One natural and well studied problem which escapes the Auer et al result, is online shortest path. In this problem the decision set is exponentially large (i.e. set of all paths in a given graph), and the straightforward reduction of modeling each path as an arm for the multi-armed bandit problem suffers from both efficiency issues as well as exponential regret. To cope with these issues, several authors [2, 9, 14] have recently proposed a very natural generalization of the multi-armed bandit problem to field of Convex Optimization, and we will call this "bandit linear optimization". In this setting we imagine that, on each round t, an adversary chooses some linear function $f_t(\cdot)$ which is not revealed to the player. The player then chooses a point $\mathbf{x}_t$ within some given convex set[1] $\mathcal{K} \subset \mathbb{R}^n$. The player then suffers $f_t(\mathbf{x}_t)$ and this quantity is revealed to him. This process continues for T rounds, and at the end the learner's payoff is his *regret*:

$$R_T = \sum_{t=1}^{T} f_t(\mathbf{x}_t) - \min_{\mathbf{x}^* \in \mathcal{K}} \sum_{t=1}^{T} f_t(\mathbf{x}^*).$$

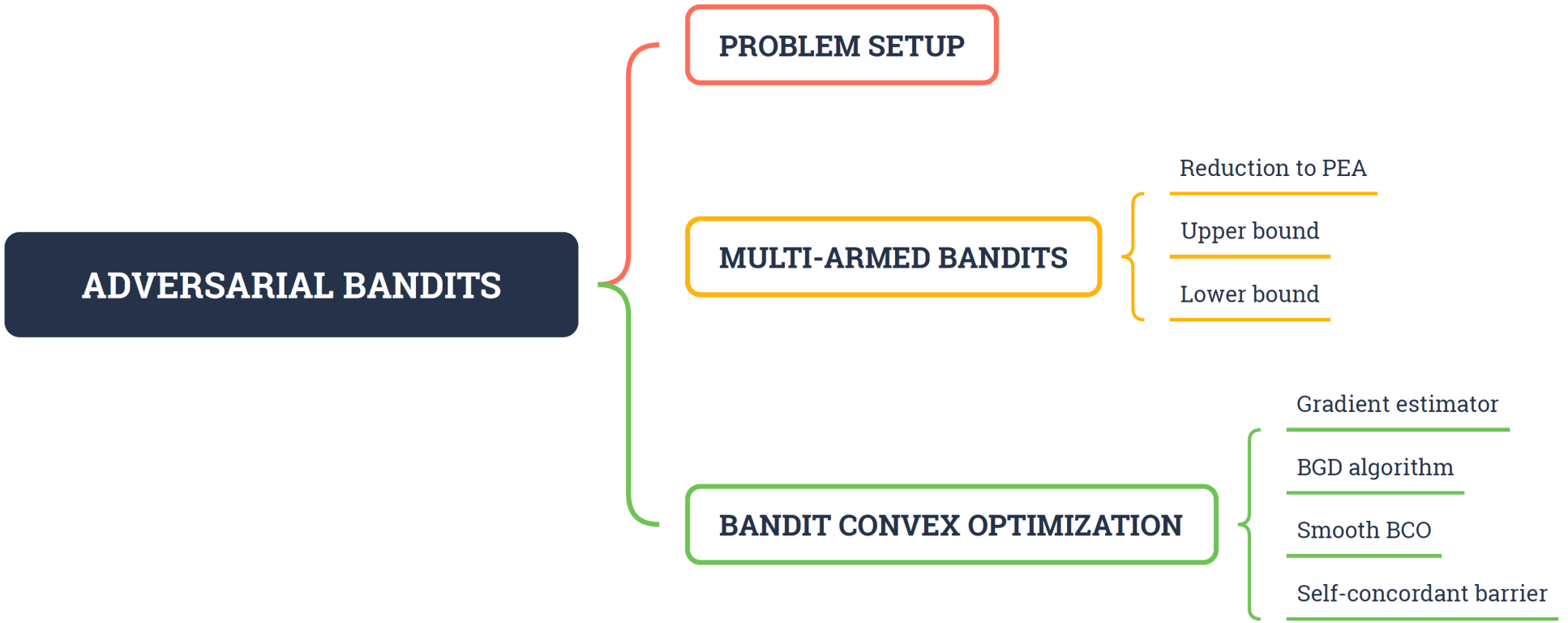
Online linear optimization has been often considered, yet primarily in the full-information setting where the learner sees all of $f_t(\cdot)$ rather than just $f_t(\mathbf{x}_t)$. In the full-information model, it has been known for some time that the optimal regret bound is $O(\sqrt{T})$, and it had been conjectured that the same should hold for the bandit setting as well. Nevertheless, several initially proposed algorithms were shown only

[1] In the case of online shortest path, the convex set can be represented as a set of vectors in $\mathbb{R}^{|E|}$. Hence, the dependence on number of paths in the graph can be circumvented.

COLT 2008
best paper award

Jacob Abernethy, Elad Hazan, and Alexander Rakhlin. Competing in the dark: An efficient algorithm for bandit linear optimization. COLT, 2008.

Summary



Q & A

Thanks!