Efficient Learning in Stochastic Bandits

Xiaotian Yu
Department of Computer Science and Engineering
The Chinese University of Hong Kong

Feb. 21, 2019

Outline

Outline

An Example: Clinical Treatment with Two Pills

(Thompson, 1933)

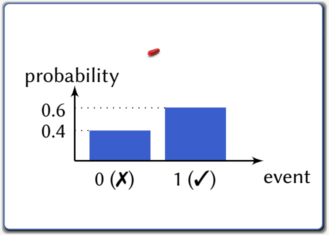
	1	2	3	4	5	6 ...	patient
	✓	✗			✓	...	
			✓	✓		✗ ...	

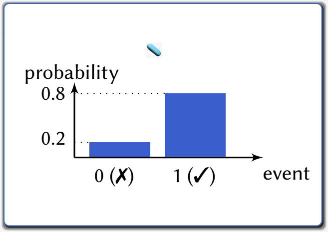
- Setting
  - A sequence of patients with the same symptoms
  - Two treatments with different performance
  - ✓: the patient is cured
  - ✗ : the patient is uncured
- Question: for the next patient $t \in \mathbb{N}^+$, which pill should be adopted?

An Example: Clinical Treatment with Two Pills

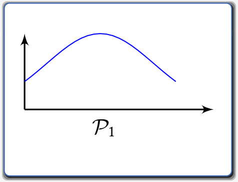
(Thompson, 1933)

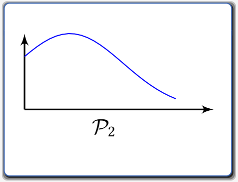
- Model: Bernoulli distributions (for stochastic feedback)



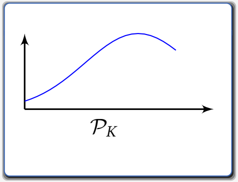


- Challenge: exploration and exploitation
- Extension: multi-armed bandits (Robbins, 1952)





. . .



Multi-Armed Bandits (MAB)

- Scenario: K arms

   $\cdots$ 

- Model: sequential decisions to *maximize cumulative rewards*

> **input:** the number of arms K, and the number of rounds $T \geq K$
> **for** $t = 1, \cdots, T$ **do**
>   select an arm $x_t \in \{1, \cdots, K\}$
>   observe a stochastic reward of arm x_t which is $y_t(x_t) \sim \mathcal{P}_{x_t}$
> **end for**

- Alias
  - Stochastic MAB
  - Online learning with bandit feedback
  - A simplified version of reinforcement learning

Multi-Armed Bandits (MAB)

Empirical average: a four-arm case with Bernoulli distributions

- An experiment

round	arm 1	arm 2	arm 3	arm 4	strategy
1-4	$\frac{1.0}{1} = 1$	$\frac{1.0}{1} = 1$	$\frac{1.0}{1} = 1$	$\frac{0.0}{1} = 0$	play each arm
5	$\frac{0.0+1.0}{2} = 0.5$	1	1	0	break ties randomly
6	0.5	$\frac{0.0+1.0}{2} = 0.5$	1	0	break ties randomly
7	0.5	0.5	$\frac{1.0+1.0}{2} = 1.0$	0	play the best arm
8	0.5	0.5	$\frac{0.0+2.0}{3} = \frac{2}{3}$	0	play the best arm
			$\vdots$		

- Issue: *arm 4 has never been explored*

Multi-Armed Bandits (MAB)

Empirical average + standard deviation

- An experiment
  - Standard deviation of estimate: $1 \rightarrow 0.7 \rightarrow 0.6 \rightarrow \cdots \rightarrow 0$

round	arm 1	arm 2	arm 3	arm 4
1-4	$\frac{1.0}{1} + 1 = 2$	$\frac{1.0}{1} + 1 = 2$	$\frac{1.0}{1} + 1 = 2$	$\frac{0.0}{1} + 1 = 1$
5	$\frac{0.0+1.0}{2} + 0.7 = 1.2$	2	2	1
6	1.2	$\frac{0.0+1.0}{2} + 0.7 = 1.2$	2	1
7	1.2	1.2	$\frac{0.0+1.0}{2} + 0.7 = 1.2$	1
8	1.2	1.2	$\frac{0.0+1.0}{3} + 0.6 = 0.9$	1
		$\vdots$		

- Standard deviation works like a confidence bound in (Robbins, 1952)
  - Standard deviation controls the quality of estimate

Efficient Learning in Stochastic Bandits

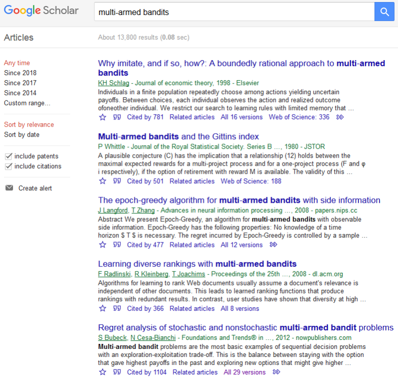
- Our general problem

> *How to make decisions based on stochastic feedback?*

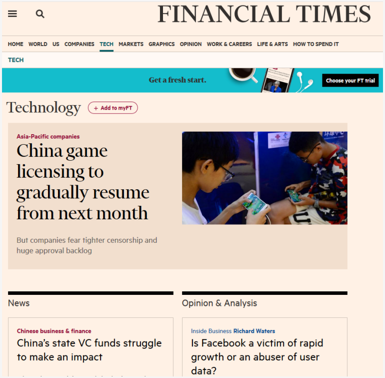
- Two general goals
  - To develop realizable and practical bandit algorithms
  - To derive theoretical guarantees for bandit algorithms
- Motivating examples
  - Clinical trials
  - Online personalized recommendations
  - Network routing
  - Online resource allocation
  - ...

Online Personalized Recommendations

- Recommendation with item information ⇒ contextual bandits



sponsored web search (Lu et al., 2010)



news recommendation (Li et al., 2011)

Online Resource Allocation

(Huo & Fu, 2017)

- A continuous arm set $\Rightarrow$ bandit optimization
  - Sequential investments with M units of money

target 1: $w_1 \in \mathbb{R}$ | target 2: $w_2 \in \mathbb{R}$ | $\cdots$ | target d: $w_d \in \mathbb{R}$

$$\Rightarrow \mathbf{w} = [w_1, \cdots, w_d] \text{ and } \sum_{i=1}^{d} w_i = 1 \textit{ with } w_i > 0$$

  - Goal: to maximize cumulative rewards with the assumption of $f(\mathbf{w})$

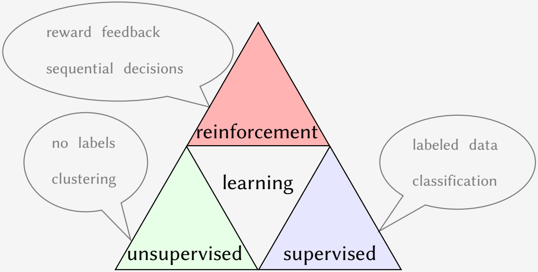
$$\Rightarrow \max \sum_{t=1}^{T} f(\mathbf{w}_t)$$

Outline

Stochastic Bandits in Machine Learning

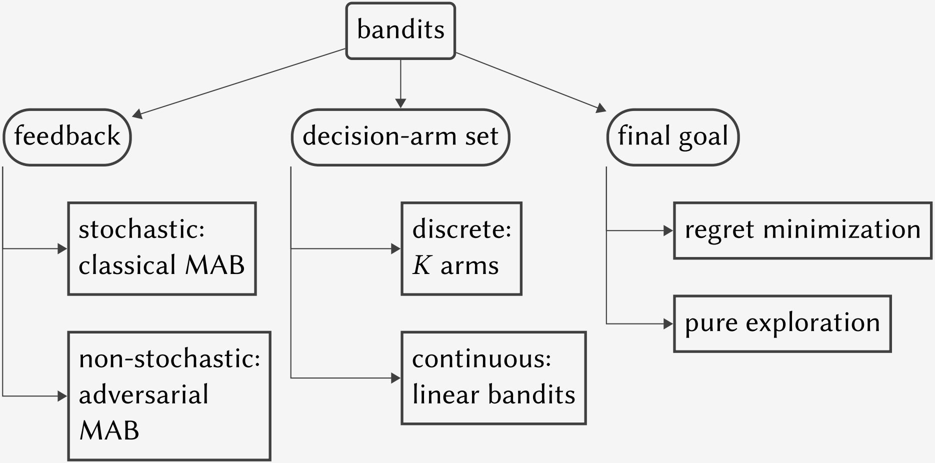
Reinforcement learning and zeroth-order optimization

- Three paradigms (Jordan and Mitchell, 2015)



- Optimization
  - Zeroth-order optimization ⇔ Bandit optimization
  - First-order optimization
  - Second-order optimization

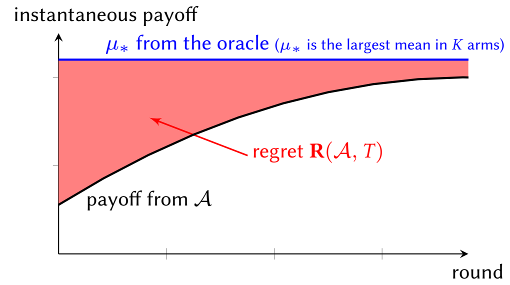
A Taxonomy


bandits
feedback
decision-arm set
final goal
stochastic: classical MAB
non-stochastic: adversarial MAB
discrete: K arms
continuous: linear bandits
regret minimization
pure exploration

Goal and Metric for Algorithm $\mathcal{A}$

- Regret minimization: $\min \mathbf{R}(\mathcal{A}, T)$

$$\mathbf{R}(\mathcal{A}, T) \triangleq \underbrace{\max_{i=1,\cdots,K} \mathbb{E}\left[\sum_{t=1}^{T} y_t(i)\right]}_{\text{an oracle}} - \underbrace{\mathbb{E}\left[\sum_{t=1}^{T} y_t(x_t)\right]}_{\text{decisions by } \mathcal{A}}. \tag{1}$$



Goal and Metric for Algorithm $\mathcal{A}$

- A different view

> What if we care more about the final decision at T?

- Pure exploration (or best arm identification): $\min \mathbb{P}[x_T \neq \text{Opt}]$
  - x_T is the output of $\mathcal{A}$ at time T, and Opt is the true optimal arm
  - To solve $\mathbb{P}[x_T = Opt] \geq 1 - \delta$ for $\delta \in (0, 1)$
  - Two settings: fixed confidence and fixed budget

fixed confidence

Given δ, what is the smallest T?

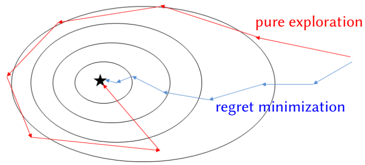
fixed budget

Given T, what is the smallest δ?

- Theoretical guarantees
  - T: sample complexity for fixed confidence
  - δ: probability of error for fixed budget

Regret Minimization versus Pure Exploration

- Application
  - Regret minimization: online advertising for news (Li et al., 2010)
  - Pure exploration: marketing for cosmetic products (Bubeck et al., 2009)
- Focus
  - Regret minimization: all decisions
  - Pure exploration: the final decision
- Hardness (Bubeck et al., 2011)
  - Regret minimization is at least as hard as pure exploration



convexity of $f(x)$

we define $\hat{x}_T \triangleq \frac{x_1+x_2+\cdots+x_T}{T}$

$$\Rightarrow f(\hat{x}_T) \leq \frac{f(x_1)+f(x_2)+\cdots+f(x_T)}{T}$$

$$\Rightarrow f(\hat{x}_T) - f(\mathrm{Opt}) \leq \frac{\mathbf{R}(\mathcal{A},T)}{T}$$

Theoretical Advancements of Regret Minimization

original problem (Thompson, 1933)

↓

first theoretical results (Lai & Robbins, 1985)

$T \to \infty$ (asymptotic bounds)

lower bound: $\frac{\mathbf{R}(\mathcal{A},T)}{\log(T)} = \sum_i \frac{\Delta_i}{KL(\theta_i,\theta_*)}$

upper bound: $\frac{\mathbf{R}(\mathrm{UCB},T)}{\log(T)} = \sum_i \frac{\Delta_i}{KL(\theta_i,\theta_*)}$

↓

finite time analysis of bandits (Agrawal, 1995)

T is finite and known (finite-time bound)

upper bound: $\mathbf{R}(\mathrm{SM}, T) = O\left(\sum_i \frac{\Delta_i \log(T)}{KL(\theta_i,\theta_*)}\right)$

↓

efficient algorithms for MAB (Auer et al, 2002)

T is finite and known

upper bound: $\mathbf{R}(\mathrm{UCB1}, T) = O\left(\sum_i \frac{\log(T)}{\Delta_i}\right)$

↓

⋮

Theoretical Advancements of Regret Minimization

⋮

improved analysis (Auer & Ortner, 2010)

T is finite
upper bound for known T:
$\mathbf{R}(\text{IUCB}, T) = O\left(\sum_i \frac{\log(T\Delta_i^2)}{\Delta_i}\right)$
upper bound for unknown T:
$\mathbf{R}(\text{IUCB}, T) = O\left(\sum_i \frac{\log(T\Delta_i^2)}{\Delta_i} + \max(\Delta_i T)\right)$

a new lower bound for MAB (Bubeck, 2013)

T is finite
lower bound: $\mathbf{R}(\mathcal{A}, T) = \Omega\left(\sum_i \frac{\log(T\Delta_i^2)}{\Delta_i}\right)$

Thompson sampling for MAB (Kaufmann et al., 2012)

T is finite and unknown
upper bound: $\mathbf{R}(\text{TS}, T) = O\left(\sum_i \Delta_i \log(T)\right)$
upper bound: $\mathbf{R}(\text{TS}, T) = O\left(\sqrt{T \log(T)}\right)$

Theoretical Advancements of Pure Exploration

Fixed confidence

PAC learning (Valiant, 1984)

↓

PAC bounds in MAB (Even-Dar et al., 2002)

bounded payoffs in $[0, 1]$

SE: $\mathbb{P}\left[T \leq \sum_{k=1}^{K} \Delta_k^{-2} \log\left(\frac{K}{\delta \Delta_k}\right)\right] \geq 1 - \delta$

ME: $\mathbb{P}\left[T \leq \frac{K}{\epsilon^2} \log\left(\frac{1}{\delta}\right)\right] \geq 1 - \delta$

↓

improved PAC bounds for bandits (Karnin et al., 2013)

bounded payoffs in $[0, 1]$

EGE:

$\mathbb{P}\left[T \leq \sum_{k=1}^{K} \Delta_k^{-2} \log\left(\frac{1}{\delta} \log\left(\frac{1}{\Delta_k}\right)\right)\right] \geq 1 - \delta$

↓

bandits with sub-Gaussian noises (Jamieson et al., 2014)

sub-Gaussian noises

LILUCB:

$\mathbb{P}\left[T \leq H_1 \log\left(\frac{1}{\delta}\right) + H_3\right] \geq 1 - 4\sqrt{c\delta} - 4c\delta$

↓

two-armed Gaussian bandits (Kaufmann et al., 2016)

two-armed Gaussian bandits

α-E: $\mathbb{P}\left[T \leq \frac{(\sigma_1+\sigma_2)^2}{(\mu_1-\mu_2)^2} \log\left(\frac{1}{\delta}\right)\right] \geq 1 - \delta$

Theoretical Advancements of Pure Exploration

Fixed budget

best arm identification of bandits (Bubeck, 2009)

↓

first results in best arm identificaiton (Audibert & Bubeck, 2010)

bounded payoffs in $[0, 1]$

UCB-E: $\mathbb{P}[\text{Out} \neq \text{Opt}] \leq TK \exp\left(-\frac{T-K}{H_1}\right)$

SR: $\mathbb{P}[\text{Out} \neq \text{Opt}] \leq K(K-1) \exp\left(-\frac{T-K}{\overline{\log}(K) H_2}\right)$

↓

ϵ-optimal learning (Gabillon et al., 2012)

bounded payoffs in $[0, b]$

UGapEb: $\mathbb{P}[\mu_{\text{Out}} - \mu_{\text{Opt}} \geq \epsilon] \leq TK \exp\left(-\frac{T-K}{H_\epsilon}\right)$

↓

improved algorithm and analysis (Karnin et al., 2013)

bounded payoffs in $[0, 1]$

SH: $\mathbb{P}[\text{Out} \neq \text{Opt}] \leq \log(K) \exp\left(-\frac{T}{\log(K) H_2}\right)$

↓

two-armed Gaussian bandits (Kaufmann et al., 2016)

two-armed Gaussian bandits

SS: $\mathbb{P}[\text{Out} \neq \text{Opt}] \leq \exp\left(-\frac{(\mu_1-\mu_2)^2 T}{2(\sigma_1+\sigma_2)^2}\right)$

Methodology for Stochastic Bandits

- Setting: K independent arms with different means $\{\mu_1, \cdots, \mu_K\}$
- Frequentist approach
  - Unknown fixed parameters: $\{\mu_1, \cdots, \mu_K\}$
  - Observed rewards: conditionally independent
  - Tool: empirical average and confidence interval
- Bayesian approach
  - Each parameter follows a distribution: $\mu_k \sim \mathcal{P}_k, \forall k \in [K]$
  - $\mathcal{P}_k$ is a prior
  - Observed rewards: conditionally independent
  - Tool: sampling from posterior, e.g., Thompson sampling

Outline

Existing Problems in Learning of Stochastic Bandits

- Sub-Gaussian noises in rewards
  - Bounded rewards
  - Rewards following Bernoulli distributions
  - Rewards following Gaussian distributions

Can rewards be more general?

$\Rightarrow$ Yes, such as heavy-tailed rewards (Bubeck et al., 2013)

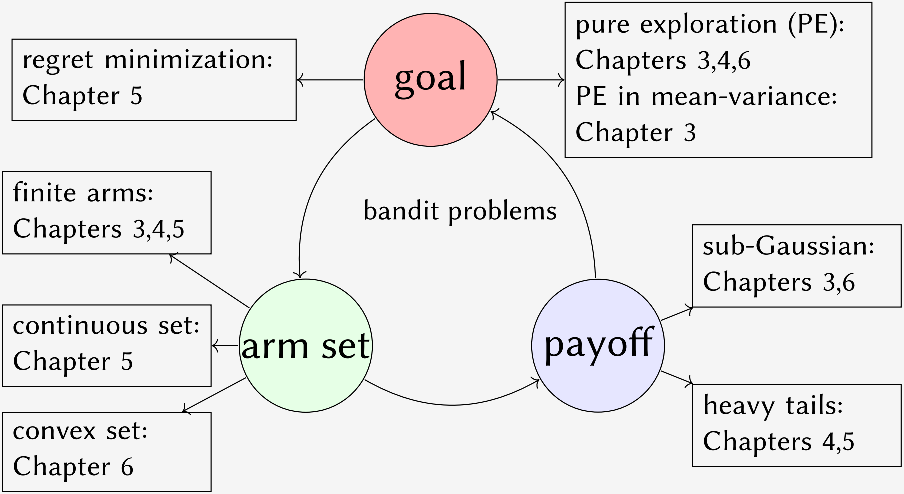
- Discrete arm sets and linear reward mapping
  - Finite arms (corresponding to vertex in a polytope)
  - Linear reward mapping

Can arm sets be continuous?

Can rewards come from nonlinear mappings?

$\Rightarrow$ Yes, such as bandit convex optimization (Hazan & Levy, 2014)

Roadmap


regret minimization:
Chapter 5
goal
pure exploration (PE):
Chapters 3,4,6
PE in mean-variance:
Chapter 3
finite arms:
Chapters 3,4,5
bandit problems
sub-Gaussian:
Chapters 3,6
continuous set:
Chapter 5
arm set
payoff
convex set:
Chapter 6
heavy tails:
Chapters 4,5

Pure Exploration of MAB

Previous work: mean information

. . .

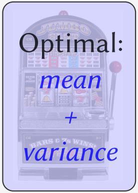
prior work

work	theoretical guarantee
(Even-Dar et al., 2006)	lower bound of probability of error
(Audibert & Bubeck, 2010)	$\mathbb{P}[\text{error}] \leq A\exp(-aT)$
(Gabillon et al., 2012)	a unified model
(Jamieson et al., 2014)	lower bound of sample complexity

*error: it denotes that the output is not the true optimal arm

Pure Exploration of MAB

Our work (new task): mean-variance



...

- Motivations
  - Clinical trial with additional risk
  - Financial investments in markets
- Setting
  - Metric: mean-variance as $\omega = \sigma^2 - \kappa\mu$ with a known $\kappa > 0$
  - Goal: identify the optimal arm with *the minimum mean-variance*
- Effect of κ
  - κ is small enough or even $\kappa = 0$: $\omega = \sigma^2$
  - κ is large enough: $\min \omega \Leftrightarrow \max \mu$

Pure Exploration of Mean-Variance (PEMV)

Fixed budget

- Problem

> Given κ and T, what is the optimal arm of ω among K arms?

- Challenges
  - What is *the error of the mean-variance estimate*?
  - How to design *a selection strategy*?
  - What is *the probability of error* for the final selected arm?

Pure Exploration of Mean-Variance (PEMV)

Technical contributions

- New metric for the optimal arm
  - Prior: empirical average ⇒ mean (sub-Gaussian estimate errors)
  - Ours: empirical mean-variance ⇒ mean-variance?
    ⇒ Yes. We prove *sub-gamma estimate errors*
- Intuitive understanding of algorithms

Confidence Bound (CB)

- empirical mean-variance
- a CB term for mean-variance estimate
- trade-off between the estimate and CB

halving technique

- binary search
- estimate error
- probability of error

Our Algorithms

PEMV.CB

input: T, K, R, $\mathbf{H}_1$, $\mathbf{H}_3$, κ
$\delta = \min\left(\frac{25(T-2K)}{576(96R^2+\kappa^2)R^2\mathbf{H}_1}, \frac{5(T-2K)}{96R^2\mathbf{H}_3}\right)$
play each arm twice and observe payoffs
for $t = 1, 2, \cdots, T$ **do**
    for $k \in [K]$ **do**
        $\hat{\omega}_t(k) = \hat{\sigma}_t^2(k) - \kappa\hat{\mu}_t(k)$
        $CB_t(k) = \sqrt{\frac{128R^4(s_t(k)+1)\delta}{(s_t(k)-1)^2} + \frac{4\kappa^2R^2\delta}{s_t(k)} + \frac{8R^2\delta}{(s_t(k)-1)}}$
        $p_t(k) = \hat{\omega}_t(k) - \mathrm{CB}_t(k)$
    end for
    $x_t = \arg\min_{k\in[K]}\ p_t(k)$ ▷ *break ties arbitrarily*
    observe a payoff $y_t(x_t)$ and save information
end for
return $x_T = \arg\min_{k\in[K]}\ \hat{\omega}_t(k)$

Our Algorithms

PEMV.HALVING

input T, K, κ
construct a decision-arm set $\mathcal{D}_1 = [K]$, $t = 0$
for $k = 1, \cdots, \lceil \log_2(K) \rceil$ **do**
  $T_k = \left\lfloor \frac{T}{|\mathcal{D}_k| \lceil \log_2(K) \rceil} \right\rfloor$
  for $a \in \mathcal{D}_k$ **do**
    for $j = 1, \cdots, T_k$ **do**
      $t = t + 1$
      select a and observe $y_j(a)$
    end for
  end for
  if $|\mathcal{D}_k| > 1$ **then**
    for $j = 1, \cdots, \left\lfloor \frac{|\mathcal{D}_k|}{2} \right\rfloor$ **do**
      select an arm $x_j = \arg\max_{a \in \mathcal{D}_k} \hat{\omega}_k(a)$
      $\mathcal{D}_k = \mathcal{D}_k \backslash x_j$ ▷ *delete an arm*
    end for
  end if
  $\mathcal{D}_{k+1} = \mathcal{D}_k$
end for
return $x_T = \mathcal{D}_{\lceil \log_2(K) \rceil + 1}$

Theoretical Results

- Estimate error: $\rho_t(a) \triangleq \hat{\omega}_t(a) - \omega(a)$ for $a \in [K]$

In Theorem 3.3 on Page 47, we prove

$$\mathbb{E}[\exp(\lambda \rho_t(a))] \leq \exp\left(\frac{\lambda^2 v}{2(1 - c\lambda)}\right), \tag{2}$$

where $\lambda \in (0, 1/c)$, $c > 0$, $v > 0$
See the definition of sub-gamma distributions in (Boucheron & Lugosi, 2013)

proof sketch: Moment Generating Function (MGF)

Step 1. calculate the MGF of empirical average
Step 2. calculate the MGF of empirical variance
Step 3. take the trade-off of the above two terms to obtain Eq. (2)

Theoretical Results

- Probability of error for PEMV.CB

Theorem 3.1 on Page 46 in the thesis

$$\mathbb{P}[x_T \neq \mathrm{Opt}] = O\left(\exp\left(-\frac{(T-2K)}{\min(\mathbf{H}_1, \mathbf{H}_3)}\right)\right) \tag{3}$$

*$\mathbf{H}_1$ and $\mathbf{H}_3$ are required in the algorithm

- Probability of error for PEMV.HALVING

Theorem 3.2 on Page 47 in the thesis

$$\mathbb{P}[x_T \neq \mathrm{Opt}] = O\left(\exp\left(-\frac{T}{\min(\mathbf{H}_4, 3\mathbf{H}_2)}\right)\right) \tag{4}$$

*$\mathbf{H}_1$-$\mathbf{H}_4$ denote problem hardness on Page 40 in the thesis

Experiments

- Settings
  - Synthetic dataset for *pure exploration* of mean-variance
  - Real financial dataset for *risk control* of investments
  - Baselines: UCBE and CuRisk
  - Metric: probability of error and cumulative returns
- Datasets
  - Statistics of synthetic datasets

dataset	#arm	$\{\mu(y)\}$	$\{\sigma^2(y)\}$
S1	20	$[1.0, 2.9]$ with a uniform gap	$\sigma^2(11)\sim\sigma^2(15) = 0.6$, $\sigma^2(20) = 0.6$, others 0.3
S2	10	random value in $[0.0, 1.0]$	random value in $[1.0, 2.0]$
S3	30	$\mu(1) = 1.0$, $\mu(y) = 1 - \frac{1.0}{2y^2}$	$\sigma^2(1) = 1.0$, $\sigma^2(y) = 2.0 - \frac{1.0}{2y^2}$

  - Historical returns on *stocks*, *bonds* and *bills*

http://pages.stern.nyu.edu/~adamodar/New_Home_Page/datafile/

Results for Synthetic Data

- Probability of error with $\kappa = 1.0$ and $T = 1000$

algorithm	S1	S2	S3
UCBE	0.63 ± 0.12	0.95 ± 0.04	0.95 ± 0.03
CuRisk	0.43 ± 0.06	0.63 ± 0.11	0.38 ± 0.10
PEMV.CB	0.19 ± 0.10	0.55 ± 0.08	$\mathbf{0.17} \pm 0.06$
PEMV.HALVING	$\mathbf{0.05} \pm 0.01$	$\mathbf{0.40} \pm 0.12$	0.23 ± 0.09

- Probability of error with $\kappa = 10.0$ and $T = 1000$

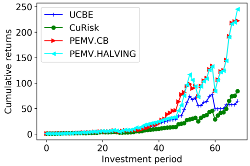
algorithm	S1	S2	S3
UCBE	0.32 ± 0.04	0.52 ± 0.10	0.47 ± 0.23
CuRisk	0.56 ± 0.12	0.67 ± 0.11	0.52 ± 0.12
PEMV.CB	0.47 ± 0.17	0.62 ± 0.09	$\mathbf{0.24} \pm 0.03$
PEMV.HALVING	$\mathbf{0.08} \pm 0.05$	$\mathbf{0.47} \pm 0.10$	0.31 ± 0.10

*More results can be found on Page 62-64 in the thesis

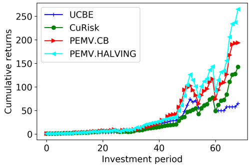
Results for Financial Data

Sharp ratio: UCBE (-0.23), CuRisk (-5.14), PEMV.CB (0.59), *PEMV.HALVING (0.72)*

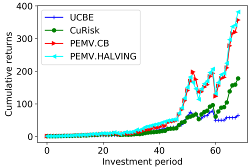
$\kappa = 1.0$ and $W = 20$



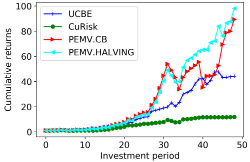
$\kappa = 1.5$ and $W = 20$



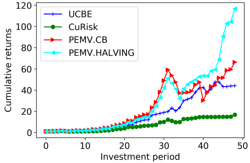
$\kappa = 2.0$ and $W = 20$



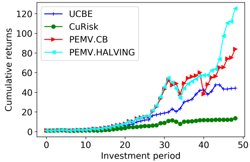
$\kappa = 1.0$ and $W = 40$



$\kappa = 1.5$ and $W = 40$



$\kappa = 2.0$ and $W = 40$



Summary

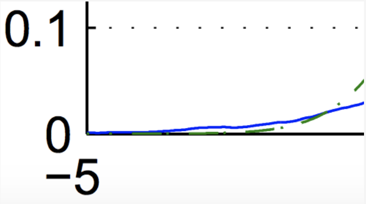
- Study the task of *pure exploration of mean-variance*
- Prove the *sub-gamma estimation error* in pure exploration of mean-variance
- Design *two algorithms* for pure exploration of mean-variance
- Prove the *probability of error* for pure exploration of mean-variance

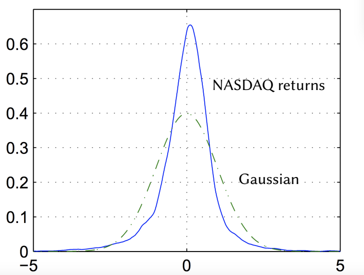
$$\boxed{\mathbb{P}[\text{error}] \leq A \exp(-aT)}$$

*The results were published in ICDM
(*Yu X.*, King I. and Lyu M. R., 2017)

What Is A Heavy-Tailed Distribution?

- Noises of rewards are not sub-Gaussian
- High-probability extreme returns in financial markets



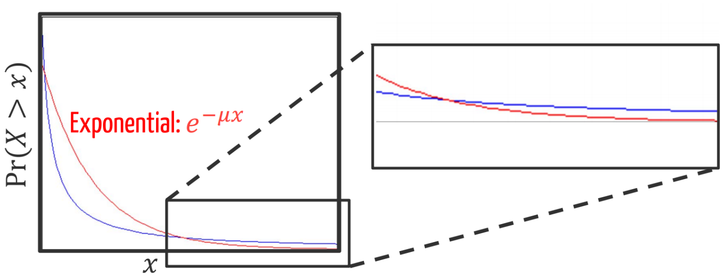


- Many other real cases
  - Delays in communication networks (Liebeherr et al., 2012)
  - Analysis of biological data (Burnecki et al., 2015)
  - ..

Heavy-Tailed Distributions

Intuition and definition

- A distribution with a “tail” that is “heavier” than an exponential decay



Ref: http://users.cms.caltech.edu/~adamw/papers/2013-SIGMETRICS-heavytails.pdf

- Mathematically, a random variable X is said to be heavy-tailed if $\lim_{x\to\infty} e^{\phi x}\mathbb{P}[|X| > x] = \infty$ *for all* $\phi > 0$ (Nolan, 2003)

Heavy-Tailed Distributions in Bandits

- Heavy-tailed distributions in bandits (Bubeck et al., 2013)

$$\mathbb{E}[X^p] < +\infty, \tag{5}$$

where X is a stochastic observation/noise, and $p \in (1, 2]$

- Remarks
  - Eq. (5) is *a subcase* of the general definition of heavy tails
  - In previous work, payoffs are assumed to have *sub-Gaussian* noises, i.e.,

$$\mathbb{E}[e^{\lambda X}] \leq \exp\left(\frac{\lambda^2 R^2}{2}\right), \tag{6}$$

    for all $\lambda \in \mathbb{R}$ and $R > 0$
  - Payoffs with sub-Gaussian noises are *light-tailed* with finite variance
    $\Rightarrow$ There is a connection between sub-Gaussian noises and heavy-tailed noises with $p = 2$

Weaker Assumption: Bounded p-th Moments

Examples

- Standard Student's t-Distribution with 3 degrees of freedom
  - The 2-nd *central moment* is bounded by 3
  - The 2-nd *raw moment* of signal (with a constant shift a) under noises following Standard Student's t-Distribution is bounded by $3 + a^2$, where $a \in \mathbb{R}$
  - The p-th moments satisfy the above properties with $p \in (1, 2]$ (*Jensen's inequality*)
- Pareto distribution with shape parameter α and scale parameter x_m
  - The p-th raw moments are bounded by $\alpha x_m^p / (\alpha - p)$, for all $p \in (1, \alpha)$
  - The p-th central moments are not directly available

Pure Exploration with Heavy Tails

- Settings
  - New task: identify the optimal arm with the largest mean under heavy tails
  - Input parameter: fixed budget or fixed confidence
- Challenges
  - What is *tail probability* of empirical average?
  - How to design *new tools* for decisions with heavy tails?
  - What is the *theoretical guarantee* for the new tool?

Pure Exploration with Heavy Tails

Fixed budget and fixed confidence

- Intuitive understanding
  - Truncation helps in extreme values

Where should we truncate?

- Technical contributions
  - Analyze *tail probability* of empirical average and truncated empirical average
  - Develop *two bandit algorithms* for pure exploration of heavy tails
  - Derive *theoretical guarantees* for the two algorithms

Our Algorithms

successive elimination-δ (SE-δ(TEA)) for fixed confidence

input: δ, K, p, B
initialization: $\hat{\mu}_1^{\dagger}(x) \leftarrow 0$ for any arm $x \in [K]$, $S_1 \leftarrow [K]$, and $b_1 \leftarrow 0$
$t \leftarrow 1$ ▷ *begin to explore arms in* $[K]$
while $|S_t| > 1$ **do**
  $c_t \leftarrow 5B^{\frac{1}{p}} \left(\frac{\log(2K/\delta)}{t}\right)^{\frac{p-1}{p}}$ ▷ *update confidence bound*
  $b_t \leftarrow \left(\frac{Bt}{\log(2K/\delta)}\right)^{\frac{1}{p}}$ ▷ *update truncating parameter*
  for $x \in S_t$ **do**
    play arm x and observe a payoff $\pi_t(x)$
    $\hat{\mu}_t^{\dagger}(x) \leftarrow \frac{1}{t}\sum_i^t \pi_i(x)\mathbb{1}_{[|\pi_i(x)| \leq b_i]}$ ▷ *calculate TEA*
  end for
  $x_t \leftarrow \arg\max_{x \in [K]} \hat{\mu}_t^{\dagger}(x)$ ▷ *choose the best arm at t*
  $S_{t+1} \leftarrow \emptyset$ ▷ *create a new arm set for* $t+1$
  for $x \in S_t$ **do**
    if $\hat{\mu}_t^{\dagger}(x_t) - \hat{\mu}_t^{\dagger}(x) \leq 2c_t$ **then**
      $S_{t+1} \leftarrow S_{t+1} + \{x\}$ ▷ *add arm x to* S_{t+1}
    end if
  end for
  $t \leftarrow t+1$ ▷ *update time index*
end while
Out $\leftarrow S_t[0]$ ▷ *assign the first entry of* S_t *to Out*
return: Out

Our Algorithms

successive rejects-T (SR-T(TEA)) for fixed budget

input $T, K, p, B, \underline{\Delta} > 0$
initialization: $\hat{\mu}^{\dagger}(x) \leftarrow 0$ for any arm $x \in [K]$, $S_1 \leftarrow [K]$, $n_0 \leftarrow 0$, $b \leftarrow 0$ and
$\bar{K} \leftarrow \sum_{i=1}^{K} \frac{1}{i}$, $b \leftarrow \left(\frac{3Bp}{\underline{\Delta}}\right)^{\frac{1}{p-1}}$ ▷ *calculate truncating parameter*
$\Phi(x) \leftarrow \emptyset$ for all $x \in S_1$ ▷ *construct sets to store time index*
for $k \in [K-1]$ **do**
  $n_k \leftarrow \left\lceil \frac{T-K}{\bar{K}(K+1-k)} \right\rceil$ ▷ *calculate n_k at stage k*
  $n \leftarrow n_k - n_{k-1}$ ▷ *calculate the number of times to pull arms*
  for $x \in S_k$ **do**
    for $i \in [n]$ **do**
      $t \leftarrow t + 1$
      play arm x, and observe a payoff $\pi_t(x)$
      $\Phi(x) \leftarrow \Phi(x) + \{t\}$ ▷ *store time index for arm x*
    end for
    $\hat{\mu}_k^{\dagger}(x) \leftarrow \frac{1}{|\Phi(x)|} \sum_{i \in \Phi(x)} \pi_i(x) \mathbb{1}_{[|\pi_i(x)| \leq b]}$
  end for
  $x_k \leftarrow \arg\min_{x \in S_k} \hat{\mu}_t^{\dagger}(x)$ ▷ *choose the worst arm at k*
  $S_{k+1} \leftarrow S_k - \{x_k\}$ ▷ *successively reject arm x_k*
end for
Out $\leftarrow S_K[0]$ ▷ *assign the first entry of S_K to Out*
return: Out

Theoretical Results

- Fixed confidence

$1 < p \leq 2$

For SE-δ(EA), we have $T = O\left(\left(\frac{1}{\delta}\right)^{\frac{1}{p-1}}\right)$

For SE-δ(TEA), we have $T = O\left(\log\left(\frac{1}{\delta}\right)\right)$

- Remarks
    - SE-δ(TEA) has an improvement *in terms of δ*
    - For sub-Gaussian noises, we have
      $T = O\left(\log\left(\frac{1}{\delta}\right)\right)$ (see Page 77 in the thesis)
      $\Rightarrow$ *SE-δ(TEA) recovers the sub-Gaussian results*
    - To have the results when $p \geq 2$ (see Page 85 in the thesis)

Theoretical Results

- Fixed budget

> $1 < p \leq 2$
>
> For SR-T(EA), we have $\mathbb{P}[\text{Out} \neq \text{Opt}] = O\left(\left(\frac{1}{T}\right)^{p-1}\right)$
> For SR-T(TEA), we have $\mathbb{P}[\text{Out} \neq \text{Opt}] = O\left(\exp\left(-T\right)\right)$

- Remarks
  - For sub-Gaussian noises, we have $\mathbb{P}[\text{error}] \leq A \exp(-aT)$ (see Page 78 in the thesis)
    $\Rightarrow$ *SR-T(TEA) recovers the sub-Gaussian results*
  - To have the results when $p \geq 2$ (see Page 87 in the thesis)

Experiments

- Setting
  - Synthetic dataset for *pure exploration of heavy tails*
  - Real-world datasets in *cryptocurrency*
  - Metric: sample complexity and probability of error
- Datasets
  - Statistics of synthetic datasets

dataset	#arms	$\{\mu(x)\}$	heavy-tailed $\{p, B, C\}$
S1	10	one arm is 2.0 and nine arms are over [0.7, 1.5] with a uniform gap	{2, 7, 3}
S2	10	one arm is 2.0 and nine arms are over [1.0, 1.8] with a uniform gap	{2, 7, 3}

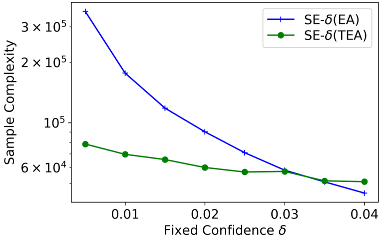
  - Top ten cryptocurrency in terms of market value

https://www.cryptocompare.com/

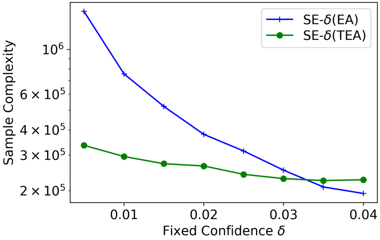
Results for Synthetic Data

Fixed confidence

S1 {10}



S2 {10}

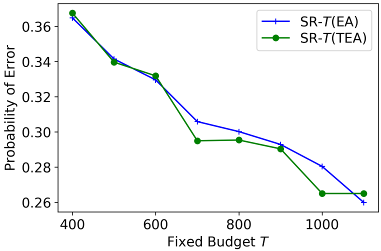


- SE-δ(TEA) outperforms SE-δ(EA) with small δ for S1 and S2
- The crossover point occurs when δ is large

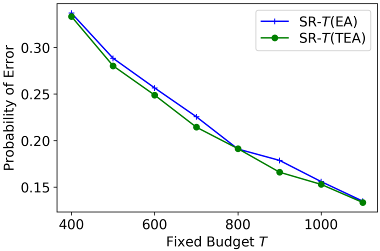
Results for Synthetic Data

Fixed budget

S1 {10}



S2 {10}



- SR-T(TEA) is comparable to SR-T(EA) for S1 and S2
- The constant factors in the theoretical results matter

Results for Financial Data

- Ten selected cryptocurrencies in experiments

full name	symbol	market value in April 2018 (unit: billion US dollar)
Bitcoin	BTC	155
Ethereum Classic	ETC	66
Ripple	XRP	32
Bitcoin Cash	BCH	23
EOS	EOS	15
Litecoin	LTC	8
Cardano	ADA	8
Stellar	XLM	7
IOTA	IOT	5
NEO	NEO	5

Results for Financial Data

- Statistical property of ten selected cryptocurrencies with hourly returns from Feb. 3rd, 2018 to Apr. 27th, 2018 (KS-test1 denotes Kolmogrov-Smirnov (KS) test with a null hypothesis that real data follow a Gaussian distribution, and KS-test2 denotes KS test with a null hypothesis that real data follow a *Student's t-distribution*)

symbol	empirical statistics (mean$\times 10^3$, variance$\times 10^3$)	KS-test1 (statistic, $\bar{p}$-value)	KS-test2 (statistic, $\bar{p}$-value)
BTC	$(0.36, 0.54)$	$(0.08, 0.005)$	$(0.05, 0.20)$
ETC	$(0.29, 1.03)$	$(0.07, 0.02)$	$(0.03, 0.89)$
XRP	$(0.33, 0.94)$	$(0.09, 0.0004)$	$(0.03, 0.61)$
BCH	$(0.78, 0.92)$	$(0.08, 0.001)$	$(0.03, 0.64)$
EOS	$(\mathbf{1.56}, 1.18)$	$(0.09, 0.0002)$	$(0.03, 0.88)$
LTC	$(0.68, 0.86)$	$(0.10, 0.0002)$	$(0.04, 0.49)$
ADA	$(0.02, 1.22)$	$(0.07, 0.03)$	$(0.02, 0.99)$
XLM	$(0.62, 0.12)$	$(0.07, 0.02)$	$(0.03, 0.80)$
IOT	$(0.68, 0.11)$	$(0.07, 0.02)$	$(0.04, 0.57)$
NEO	$(-0.31, 1.26)$	$(0.10, 0.0002)$	$(0.04, 0.53)$

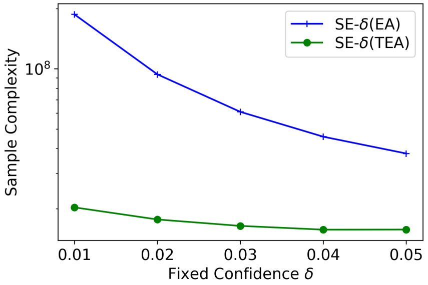
Results for Financial Data

- Estimated parameters for ten cryptocurrencies

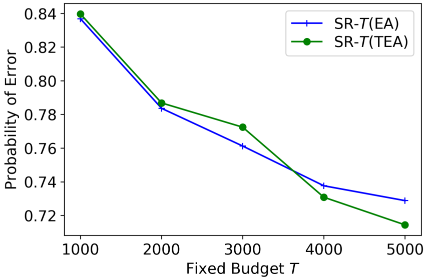
symbol	degree of freedom	(p, B, C) in experiments
BTC	3.50	
ETC	3.81	
XRP	2.53	
BCH	3.00	
EOS	2.90	
LTC	2.75	$(2, 1.577 \times 10^{-3}, 1.575 \times 10^{-3})$
ADA	3.55	
XLM	3.81	
IOT	4.66	
NEO	3.13	

Results for Finanical Data

fixed confidence



fixed budget



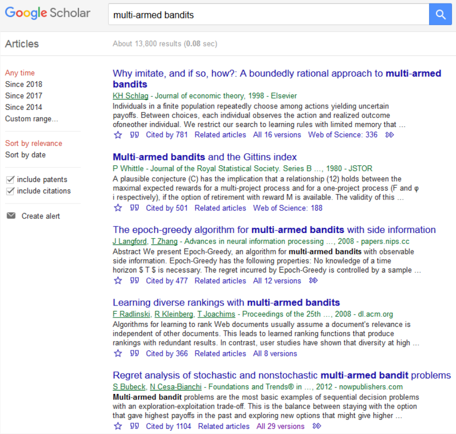
- SE-δ(TEA) and SR-T(TEA) perform better

Summary

- Study *pure exploration of bandits with heavy tails*
- Derive *tail probability* of empirical average and truncated empirical average
- Design *two algorithms* for pure exploration of bandits with heavy tails
- Derive *theoretical guarantees* of the two bandit algorithms

*The results were published in UAI
(*Yu X.*, Shao H., Lyu M. R. and King I., 2018)

Linear Stochastic Bandits (LinSB)



sponsored web search
(Lu et al., 2010)

- Arm space:
  d-dimensional space
- Reward function:
  a linear mapping
- Tool:
  least square estimate
- Regret for sub-Gaussian noises:
  $\tilde{O}(\sqrt{T})$
  $^*\tilde{O}(\cdot)$ omits the logarithmic factors of T

LinSB with Heavy-Tailed Payoffs

- Scenario

- Setting
  - At t, an algorithm is given $D_t \subset \mathbb{R}^d$ with $\theta_* \in \mathbb{R}^d$
  - Select an arm $x_t \in D_t$, and observe $y_t(x_t) = \langle x_t, \theta_* \rangle + \eta_t$
  - The goal is to maximize $\sum_{t=1}^{T} y_t(x_t)$
- Assumption: $y_t(x_t)$ or η_t is *heavy-tailed conditional on $\mathcal{F}_{t-1}$*

Problem Definition

Linear stochastic Bandits with hEavy-Tailed payoffs (LinBET)

LinBET

Given a decision set D_t for time step $t = 1, \cdots, T$, an algorithm $\mathcal{A}$, of which the goal is to maximize cumulative payoffs over T rounds, chooses an arm $x_t \in D_t$. With $\mathcal{F}_{t-1}$, the observed stochastic payoff $y_t(x_t)$ is conditionally heavy-tailed, i.e., $\mathbb{E}\left[|y_t|^p|\mathcal{F}_{t-1}\right] \leq b$ or $\mathbb{E}\left[|y_t - \langle x_t, \theta_* \rangle|^p|\mathcal{F}_{t-1}\right] \leq c$, where $p \in (1, 2]$, and $b, c \in (0, +\infty)$.

Challenges and Contributions

- Challenges
  - The *lower bound* of LinBET
  - How to develop *a robust estimator* of the parameter for LinBET and *bandit algorithms*
  - Results in previous work (Medina & Yang, 2016) are *far from optimal*
    - Sub-Gaussian: regret is $\tilde{O}(\sqrt{T})$
    - Prior results when $p = 2$: regret is $\tilde{O}(T^{\frac{3}{4}})$

  $\Rightarrow$ How to develop results when $p = 2$ *recovering the regret with sub-Gaussian noises*?
- Contributions
  - The first to provide the *lower bound* for LinBET
  - Develop *two novel bandit algorithms* to solve LinBET
  - Two algorithms are *optimal* up to logarithmic factors

Lower Bound of LinBET

- Setting
  Assume $d \geq 2$ is even. For $D_t \in \mathbb{R}^d$, we fix the decision set as $D_t = D_{(d)}$, where $D_{(d)} \triangleq \{(x_1, \cdots, x_d) \in \mathbb{R}_+^d : x_1 + x_2 = \cdots = x_{d-1} + x_d = 1\}$. Let $S_d \triangleq \{(\theta_1, \cdots, \theta_d) : \forall i \in [d/2], (\theta_{2i-1}, \theta_{2i}) \in \{(2\Delta, \Delta), (\Delta, 2\Delta)\}\}$ with $\Delta \in (0, 1/d]$. Payoffs are in $\{0, (1/\Delta)^{\frac{1}{p-1}}\}$ such that, for every $x \in D_{(d)}$, the expected payoff is $\theta_*^\top x$.
- Result (Theorem 5.1 on Page 107 in the thesis)

lower bound

$$\mathbb{E}\left[\mathbf{R}(\mathcal{A}, T)\right] = \Omega(T^{\frac{1}{p}})$$

*Sub-Gaussian noises: regret lower bound $\Omega(\sqrt{T})$

Existing Problems in Prior Work

Regret

- Least square estimate: $\hat{\theta}_t = (\mathbf{I}_d + X_t X_t^T)^{-1} X_t Y_t$
  where $X_t = (x_1, \cdots, x_t)$ and $Y_t = (y_1, \cdots, y_t)^T$
- Considering $V_t = \mathbf{I}_d + X_t X_t^T$, we have

$$\mathbf{R}(\mathcal{A}, T) = \sum_{t=1}^{T} r_t = \sum_{t=1}^{T} \langle \theta_*, x_* \rangle - \langle \theta_*, x_t \rangle \leq \sum_{t=1}^{T} \langle \hat{\theta}_t, x_* \rangle - \langle \theta_*, x_t \rangle$$

$$\leq \sum_{t=1}^{T} \underbrace{\left(\|\hat{\theta}_t - \hat{\theta}_{t-1}\|_{V_{t-1}} - \|\hat{\theta}_{t-1} - \hat{\theta}_*\|_{V_{t-1}} \right)}_{\text{an ellipsoid: } \{\theta | \|\hat{\theta}_t - \theta\|_{V_t} \leq \beta_t\}} \|x_t\|_{V_{t-1}^{-1}}$$

- Sub-Gaussian noises: $\beta_t = O(\sqrt{\log(t)})$
  $\Rightarrow \mathbf{R}(\mathcal{A}, T) = \tilde{O}(\max_{t \in [T]} \beta_{t-1} \sqrt{T})$
- (Medina & Yang, 2016): $\mathbf{R}(\mathcal{A}, T) = \tilde{O}\left(T^{\frac{3}{4}}\right)$ when $p = 2$

Algorithms: MEdian of meaNs under OFU (MENU)

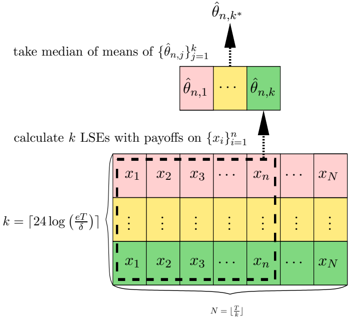
MENU

input d, c, p, δ, λ, S, T, $\{D_n\}_{n=1}^{N}$
initialization: $k = \lceil 24 \log\left(\frac{eT}{\delta}\right) \rceil$, $N = \lfloor \frac{T}{k} \rfloor$, $V_0 = \lambda I_d$, $C_0 = \mathbb{B}(\mathbf{0}, S)$
for $n = 1, 2, \cdots, N$ **do**
    $(x_n, \tilde{\theta}_n) = \arg\max_{(x,\theta)\in D_n \times C_{n-1}} \langle x, \theta \rangle$
    Play x_n with k times and observe payoffs $y_{n,1}, y_{n,2}, \cdots, y_{n,k}$
    $V_n = V_{n-1} + x_n x_n^\top$
    For $j \in [k]$, $\hat{\theta}_{n,j} = V_n^{-1} \sum_{i=1}^{n} y_{i,j} x_i$
    For $j \in [k]$, *let* r_j *be the median of* $\{\|\hat{\theta}_{n,j} - \hat{\theta}_{n,s}\|_{V_n} : s \in [k] \backslash j\}$
    $k^* = \arg\min_{j\in[k]} r_j$
    $\beta_n = 3\left((9dc)^{\frac{1}{p}} n^{\frac{2-p}{2p}} + \lambda^{\frac{1}{2}} S\right)$
    $C_n = \{\theta : \|\theta - \hat{\theta}_{n,k^*}\|_{V_n} \le \beta_n\}$
end for

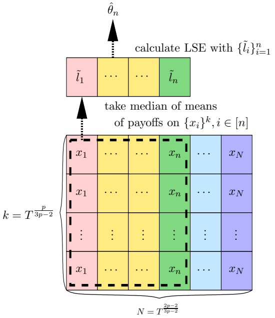
Understanding of MENU

Framework comparison

MENU



MoM by (Medina & Yang, 2016)



Upper Bound Analysis: MENU

Results

- Intuitive idea
  - $\mathbb{P}\left(\|\hat{\theta}_n - \theta_*\|_{V_n} \leq (9dc)^{\frac{1}{p}} n^{\frac{2-p}{2p}} + \lambda^{\frac{1}{2}} S\right) \geq \frac{3}{4}$
  - With probability at least $1 - e^{-\frac{k}{24}}$, $\|\hat{\theta}_{n,k^*} - \theta_*\|_{V_n} \leq 3\gamma$
  - $\mathbf{R}(\mathcal{A}, T) \leq \sum_{t=1}^{T} \left(\|\hat{\theta}_t - \hat{\theta}_{t-1}\|_{V_{t-1}} - \|\hat{\theta}_{t-1} - \hat{\theta}_*\|_{V_{t-1}}\right) \|x_t\|_{V_{t-1}^{-1}}$
- Our result (Theorem 5.2 on Page 111 in the thesis)

$$\boxed{\mathbf{R}(\text{MENU}, T) = \widetilde{O}(T^{\frac{1}{p}})}$$

Algorithms: Truncation under OFU (TOFU)

TOFU

input d, b, p, δ, λ, T, $\{D_t\}_{t=1}^{T}$
initialization: $V_0 = \lambda I_d$, $C_0 = \mathbb{B}(\mathbf{0}, S)$
for $t = 1, 2, \cdots, T$ **do**
  $b_t = \left(\frac{b}{\log(\frac{2T}{\delta})}\right)^{\frac{1}{p-1}} t^{\frac{2-p}{2p}}$
  $(x_t, \tilde{\theta}_t) = \arg\max_{(x,\theta)\in D_t \times C_{t-1}} \langle x, \theta \rangle$
  Play x_t and observe a payoff y_t
  $V_t = V_{t-1} + x_t x_t^\top$ and $X_t^\top = [x_1, \cdots, x_t]$
  $[u_1, \cdots, u_d]^\top = V_t^{-1/2} X_t^\top$
  for $i = 1, \cdots, d$ **do**
    $Y_i^\dagger = (y_1 \mathbb{1}_{u_{i,1} y_1 \le b_t}, \cdots, y_t \mathbb{1}_{u_{i,t} y_t \le b_t})$
  end for
  $\theta_t^\dagger = V_t^{-1/2}(u_1^\top Y_1^\dagger, \cdots, u_d^\top Y_d^\dagger)$
  $\beta_t = 4\sqrt{d}\, b^{\frac{1}{p}} \left(\log\left(\frac{2dT}{\delta}\right)\right)^{\frac{p-1}{p}} t^{\frac{2-p}{2p}} + \lambda^{\frac{1}{2}} S$
  Update $C_t = \{\theta : \|\theta - \theta_t^\dagger\|_{V_t} \le \beta_t\}$
end for

Understanding of TOFU

Framework comparison

- For TOFU, at time t, all of the history payoffs are truncated by b_t for each u_i
  - $Y_i^\dagger = (y_1 \mathbb{1}_{u_{i,1} y_1 \le b_t}, \cdots, y_t \mathbb{1}_{u_{i,t} y_t \le b_t})$
  - $\theta_t^\dagger = V_t^{-1/2}(u_1^\top Y_1^\dagger, \cdots, u_d^\top Y_d^\dagger)$
- For CRT in (Medina & Yang, 2016), the payoff at time t is truncated by α_t
  - $y_t^\dagger = y_t \mathbb{1}_{y_t \le \alpha_t}$

Upper Bound Analysis: TOFU

Results

- Intuitive idea
  - Trade-off between truncation error and bounded payoffs
  - Truncation parameter related to historical information
  - CRT in (Medina & Yang, 2016) only cares about time step
- Our result (Theorem 5.2 on Page 113 in the thesis)

$$\boxed{\mathbf{R}(\text{TOFU}, T) = \widetilde{O}(T^{\frac{1}{p}})}$$

Experimental Results

- Datasets
  - Four synthetic datasets
  - Metric: cumulative payoffs
  - Baselines: MoM and CRT (Medina & Yang, 2016)
- Settings
  - Run independently ten times for each experiment
  - Show cumulative payoffs with one standard variance

Experimental Results

Synthetic datasets

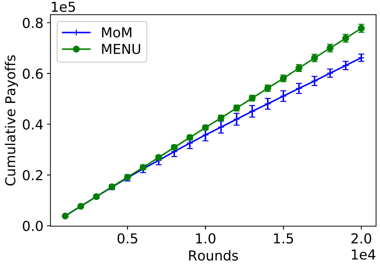
statistics

dataset	{#arms,#dim}	distribution {parameters}	$\{\epsilon, b, c\}$	optimal arm
S1	{20,10}	Student's t-distribution $\{\nu = 3, l_p = 0, s_p = 1\}$	{1.00, NA, 3.00}	4.00
S2	{100,20}	Student's t-distribution $\{\nu = 3, l_p = 0, s_p = 1\}$	{1.00, NA, 3.00}	7.40
S3	{20,10}	Pareto distribution $\{\alpha = 2, s_m = \frac{x_t^\top \theta_*}{2}\}$	{0.50, 7.72, NA}	3.10
S4	{100,20}	Pareto distribution $\{\alpha = 2, s_m = \frac{x_t^\top \theta_*}{2}\}$	{0.50, 54.37, NA}	11.39

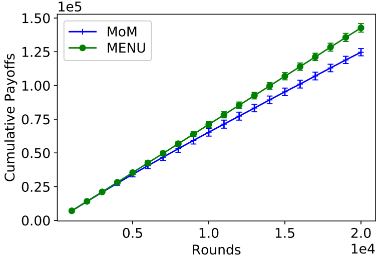
Experimental Results

Central moments

S1: {20,10}



S2: {100,20}

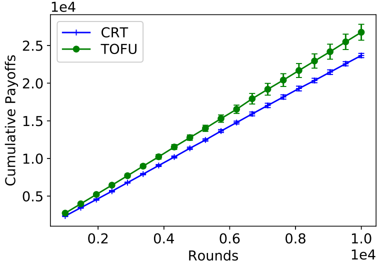


- Our algorithm MENU outperforms MoM in (Medina & Yang, 2016)

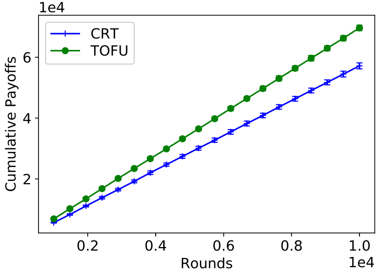
Experimental Results

Raw moments

S3: {20,10}



S4: {100,20}



- Our algorithm TOFU outperforms CRT in (Medina & Yang, 2016)

Summary

- Contributions
  - Derive *lower bound* for LinBET
  - Develop *two novel bandit algorithms* to solve LinBET
  - Theoretical results are *optimal up to logarithmic factors*

improvements: almost matching the lower bound $\Omega(T^{\frac{1}{p}})$

algorithm	MoM	**MENU**	CRT	**TOFU**
regret	$\widetilde{O}(T^{\frac{2p-1}{3p-2}})$	$\widetilde{O}(T^{\frac{1}{p}})$	$\widetilde{O}(T^{\frac{1}{2}+\frac{1}{2p}})$	$\widetilde{O}(T^{\frac{1}{p}})$
complexity	$O(T)$	$O(T\log T)$	$O(T)$	$O(T^2)$
storage	$O(1)$	$O(\log T)$	$O(1)$	$O(T)$

*The results were published in NIPS
(Shao H., *Yu X.*, King I. and Lyu M. R., 2018)

Nonlinear Stochastic Bandits

- Reward function: non-linear
- Settings: convex and non-convex (a discussion)

Stochastic Zeroth-order Convex Optimization (SZCO)

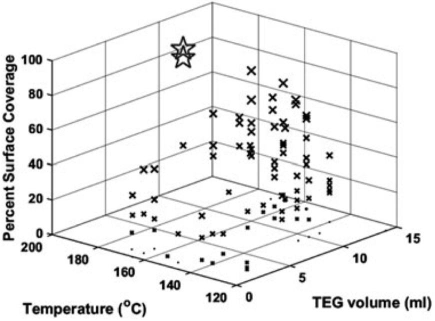
Practical scenarios

- White-box optimization
  - Linear regression
  - Logistic loss for binary classification
  - Convex optimization
- Stochastic black-box optimization
  - Unknown objective functions
  - Noisy feedbacks
- Many real cases
  1. Online advertisement selections (Wibisono et al., 2012)
  2. Stochastic structured predictions (Sokolov et al., 2016)
  3. Optimization in biological experiments (Nakamura et al., 2017)
  4. ...

Stochastic Zeroth-order Convex Optimization (SZCO)

Practical scenarios

Plot of real experimental output in (Nakamura et al., 2017) for an industrial device with different input parameters, i.e., Temperature and Tetraethylene Glycol (TEG)



Stochastic Zeroth-order Convex Optimization (SZCO)

Motivation

- How to determine the optimal parameter in a convex and compact set
    - A lot of real experiments
    - A statistical analysis (with a convexity assumption)
- Drawbacks of previous work
    - Time consuming for experiments
    - Expensive
- Settings of our work
    - Convex objective functions $\Leftrightarrow$ Concave reward functions
    - Noisy feedbacks
    - Unknown objective functions

Stochastic Zeroth-order Convex Optimization (SZCO)

Definition

- $f(\mathbf{x};\xi)$ is the convex model in learning problems
  - $\mathbf{x}$ is the parameter to be learned with $\mathbf{x} \in \mathbb{R}^d$
  - ξ is the samples with noises
- The goal is to solve

$$\min_{\mathbf{x}\in\Omega} f(\mathbf{x}) \triangleq \mathbb{E}_{\xi}[f(\mathbf{x};\xi)] \tag{7}$$

- ϵ-optimal solution
  An ϵ-optimal solution $\widehat{\mathbf{x}}$ satisfies the following condition:
  $\mathbb{E}[f(\widehat{\mathbf{x}};\xi) - \min_{\mathbf{x}\in\Omega} f(\mathbf{x};\xi)] \leq \epsilon$
- Theoretical guarantee
  How many samples do we need in order to get $\widehat{\mathbf{x}}$? (iteration complexity)

Two Settings in SZCO

- One-Point Evaluation (OPE)
  - For each round, one noisy observation is revealed
  - Noisy gradient estimator (Flaxman et al., 2005)

$$\mathbf{g}_t^{\mathrm{f}} = \frac{d}{\delta} f(\mathbf{x}_t + \delta \mathbf{u}_t; \xi_t)\mathbf{u}_t, \tag{8}$$

  where $\mathbf{u}_t \sim \mathbb{B}(\mathbf{0}, 1)$ and $\delta > 0$.

- Two-Point Evaluation (TPE)
  - Noisy gradient estimator (Agarwal et al., 2010)

$$\mathbf{g}_t^{\mathrm{a}} = \frac{d}{2\delta}\big(f(\mathbf{x}_t + \delta \mathbf{u}_t; \xi_t) - f(\mathbf{x}_t - \delta \mathbf{u}_t; \xi_t)\big)\mathbf{u}_t \tag{9}$$

- Solver: stochastic gradient descent

Previous Work

setting	algorithm	assumption	iteration complexity	h.p. or exp.
OPE	(Flaxman et al., 2005)	LC	$O\left(\frac{d^2}{\epsilon^4}\right)$	exp.
	(Agarwal et al., 2010)	LC + SC	$\widetilde{O}\left(\frac{d^2}{\epsilon^3}\right)$	exp.
		LC + SC + SM	$\widetilde{O}\left(\frac{d^2}{\epsilon^2}\right)$	exp.
TPE	(Agarwal et al., 2010)	LC	$O\left(\frac{d^2}{\epsilon^2}\right)$	h.p.
		LC + SC	$\widetilde{O}\left(\frac{d^2}{\epsilon}\right)$	h.p.
	(Nesterov, 2017)	LC	$\widetilde{O}\left(\frac{d^2}{\epsilon^2}\right)$	exp.
		LC + SM	$O\left(\frac{d}{\epsilon^2}\right)$	exp.
	(Duchi et al., 2015)	LC	$\widetilde{O}\left(\frac{d \log d}{\epsilon^2}\right)$	exp.
		LC + SM	$O\left(\frac{d}{\epsilon^2}\right)$	exp.
	(Shamir, 2017)	LC	$O\left(\frac{d}{\epsilon^2}\right)$	exp.

LC: Lipschitz Continuous, SC: Strong Convexity, and SM: SMoothness

Local Error Bound (LEB)

- LEB works for first-order optimization: acceleration
  - Previous work (Yang et al., 2015; Bolte et al., 2015; Xu et al., 2017)

A problem of Eq. (7) satisfies the LEB condition on a compact set Ω if there exist $\theta \in (0, 1]$ and $c > 0$ such that for any $\mathbf{x} \in \Omega$

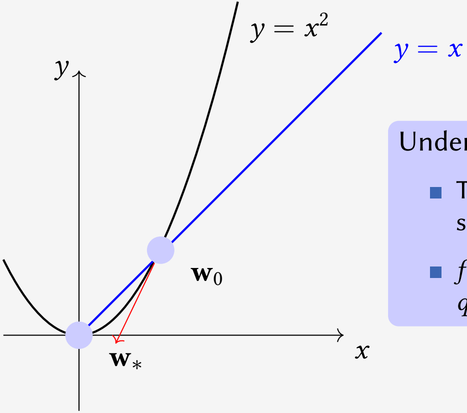
$$\mathrm{dist}(\mathbf{x}, \Omega_*) \le c(f(\mathbf{x}) - \min_{\mathbf{x}\in\Omega} f(\mathbf{x}))^\theta, \tag{10}$$

where $\mathrm{dist}(\mathbf{x}, \Omega_*) \triangleq \min_{\mathbf{v}\in\Omega_*} \|\mathbf{v} - \mathbf{x}\|_2$

- How can we apply LEB into SZCO?
  - To improve the iteration complexity of SZCO

Local Error Bound (LEB)

- An example: quadratic condition with $\theta = 1/2$



Understanding:

- The quadratic function has a sharper slope
- $f(\mathbf{w}_1) - f(\mathbf{w}_2) \geq q\|\mathbf{w}_1 - \mathbf{w}_2\|_2^2$, with $q > 0$

Local Error Bound (LEB)

Examples

- Example 1
  When $f(\mathbf{x};\xi) = \mathbf{x}^\top \xi$ is a linear function and Ω is a polyhedral set (e.g., hypercube), then the problem of Eq. (7) satisfies the LEB with $\theta = 1$. These functions are considered in online bandit linear optimization. More generally, if $f(\mathbf{x})$ is a polyhedral function and Ω is a polyhedral set, then LEB with $\theta = 1$ holds. For instance, $f(\mathbf{x}) = \sum_{i=1}^n |\mathbf{a}_i^\top \mathbf{x} - b_i|/n$ and $\Omega = \{\|\mathbf{x}\|_1 \leq s\}$.
- Example 2
  When $f(\mathbf{x})$ is strongly convex, then the LEB condition holds with $\theta = 1/2$
- Example 3
  Even when $f(\mathbf{x})$ is not strongly convex, the LEB condition with $\theta = 1/2$ may still hold, such as $f(\mathbf{x}) = \sum_{i=1}^n (\mathbf{a}_i^\top \mathbf{x} - b_i)^2/n$ and Ω is a polyhedral set.

Algorithm: A Generic Approach for Accelerating SZCO

Algorithm 1

initialization $\mathbf{x}_0, K, \eta_1, \delta_1, D_1$

for $k = 1, \cdots, K$ **do**

    $\mathbf{x}_k^1 = \mathbf{x}_{k-1}, \mathbb{D}_k = \Omega \cap \mathbb{B}(\mathbf{x}_k^1, D_k)$

    for $\tau = 1, \cdots, t$ **do**

        compute a gradient estimator in light of Eq. (8) or Eq. (9)

        compute $\mathbf{x}_k^\tau$ *according to stochastic gradient descent (under domain strinkage) with a step size* η_k*, a parameter* δ_k*, and a domain* $\mathbb{D}_k$

    end for

    let $\mathbf{x}_k = \sum_{\tau=1}^{t} \mathbf{x}_k^\tau / t$

    update δ_{k+1}, D_{k+1} *and* η_{k+1}

end for

return $\mathbf{x}_K$

Our Results: OPE

setting	algorithm	assumption	iteration complexity	h.p. or exp.
OPE	(Flaxman et al., 2005)	LC	$O\left(\frac{d^2}{\epsilon^4}\right)$	exp.
	(Agarwal et al., 2010)	LC + SC	$\widetilde{O}\left(\frac{d^2}{\epsilon^3}\right)$	exp.
		LC + SC + SM	$\widetilde{O}\left(\frac{d^2}{\epsilon^2}\right)$	exp.
	our work	LC + LEB	$\widetilde{O}\left(\frac{d^2}{\epsilon^{2(2-\theta)}}\right), \theta \in (0, \frac{1}{2}]$	exp.
			$\widetilde{O}\left(\frac{d^2}{\epsilon^{2(2-\theta)}}\right), \theta \in (0, 1]$	h.p.
	our work	LC + LEB + SM	$\widetilde{O}\left(\frac{d^2}{\epsilon^{3-2\theta}}\right), \theta \in (0, \frac{1}{2}]$	exp.
			$\widetilde{O}\left(\frac{d^2}{\epsilon^{3-2\theta}}\right), \theta \in (0, 1]$	h.p.

LC: Lipschitz Continuous, SC: Strong Convexity, SM: SMoothness, and LEB: Local Error Bound

- An order improvement in convergence rate

Our Results: TPE

setting	algorithm	assumption	iteration complexity	h.p. or exp.
TPE	(Agarwal et al., 2010)	LC	$O\left(\frac{d^2}{\epsilon^2}\right)$	h.p.
		LC + SC	$\widetilde{O}\left(\frac{d^2}{\epsilon}\right)$	h.p.
	(Nesterov, 2017)	LC	$\widetilde{O}\left(\frac{d^2}{\epsilon^2}\right)$	exp.
		LC + SM	$O\left(\frac{d}{\epsilon^2}\right)$	exp.
	(Duchi et al., 2015)	LC	$\widetilde{O}\left(\frac{d \log d}{\epsilon^2}\right)$	exp.
		LC + SM	$O\left(\frac{d}{\epsilon^2}\right)$	exp.
	(Shamir, 2017)	LC	$O\left(\frac{d}{\epsilon^2}\right)$	exp.
	our work	LC + LEB	$\widetilde{O}\left(\frac{d^2}{\epsilon^{2(1-\theta)}}\right), \theta \in (0, 1]$	h.p.
	our work	LC + LEB	$\widetilde{O}\left(\frac{d}{\epsilon^{2(1-\theta)}}\right), \theta \in (0, \frac{1}{2}]$	exp.

LC: Lipschitz Continuous, SC: Strong Convexity, SM: SMoothness, and LEB: Local Error Bound

Experimental Results

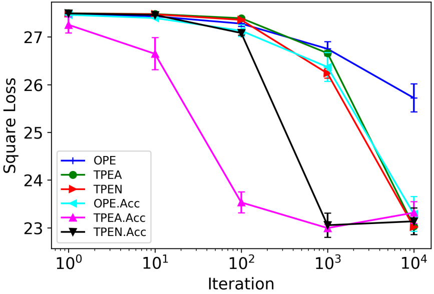
- Datasets
  - Two real-world datasets
    - Music recommendation competition data
    - Industrial data on ceramic thin films
  - Metric
    - Iteration complexity with respect to objectives
- Setting
  - Three baselines and add '.Acc' for each baseline as the method based on Algorithm 1
  - Run experiments in a personal computer with Intel CPU@3.70GHz and 16 GB memory
  - Independent ten times for each epoch

Experimental Results

Music recommendation competition data (KDD 2011)

- KDD competition: suppose we have multiple models to conduct score prediction, how to determine the weight of each model?
  $\Rightarrow$ Online resource allocation

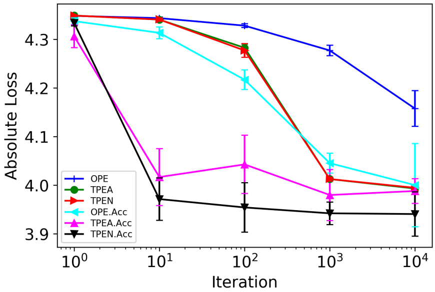
$$f(\mathbf{x}) = \frac{\sum_{i=1}^{N}(\mathbf{w}_i^\top \mathbf{x} - r_i)^2}{N}, \theta = 0.5, \text{ and } T = 10^4$$



Experimental Results

Music recommendation competition data (KDD 2011)

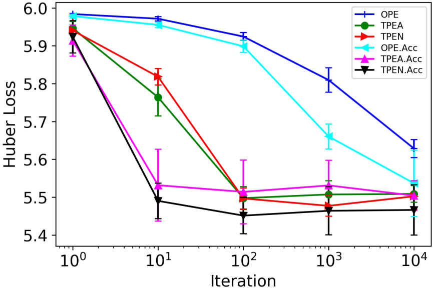
$$f(\mathbf{x}) = \frac{\sum_{i=1}^{N} |\mathbf{w}_i^\top \mathbf{x} - r_i|}{N}, \theta = 1 \text{ and } T = 10^4$$



Experimental Results

Music recommendation competition data (KDD 2011)

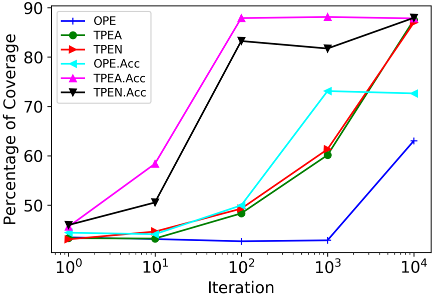
$f(\mathbf{x})$ is averaged Huber loss, $\theta = 1$ and $T = 10^4$



Experimental Results

Industrial data on ceramic thin films

Growth of ceramic thin films with $T = 10^4$



Summary

- Contributions
  - Design *a generic framework* for SZCO with LEB
  - Derive *iteration complexity* of the generic framework
  - Theoretical guarantees *beat* the state-of-the-art results
  - The results can be extent into non-convex cases (feed-forward networks)

*The results were published in IJCAI
(*Yu X.*, King I., Lyu M. R. and Yang T., 2018)

Outline

Conclusion

- Contributions
  - Goals: pure exploration and regret minimization
  - Settings: mean-variance, heavy tails, nonlinear payoffs
  - Output: algorithms with theoretical guarantees

task	pure exploration	regret minimization
mean-variance	*(Yu et al., 2017) in ICDM*	(Audibert et al., 2010)
MAB with heavy tails	*(Yu et al., 2018) in UAI*	(Bubeck et al., 2013)
linear bandits with heavy tails	(Hsu & Sabato, 2016)	(Medina & Yang, 2016) *(Shao et al., 2018) in NIPS*
nonlinear payoffs	(Flaxman et al., 2005; Agarwal et al., 2010) *(Yu et al., 2018) in IJCAI*	(Hazan & Levy, 2014; Bubeck et al., 2016)

Conclusion

Future work

- Adaptive learning in bandits
- Learning in bandits with dependent arms

List of Publications

[1] **Xiaotian Yu**, Haiqin Yang, Irwin King and Michael R. Lyu. Online non-negative dictionary learning via moment information for sparse poisson coding. In *Proceedings of IEEE International Joint Conference on Neural Networks (IJCNN)*, pages 5094–5101, 2016.

[2] **Xiaotian Yu**, Michael R. Lyu and Irwin King. CBRAP: contextual bandits with random projection. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence (AAAI)*, pages 2859–2866, 2017.

[3] **Xiaotian Yu**, Irwin King and Michael R. Lyu. Risk control of best arm identification in multi-armed bandits via successive rejects. In *Proceedings of IEEE International Conference on Data Mining (ICDM)*, pages 1147–1152, 2017. *(Chapter 3)*

[4] **Xiaotian Yu**, Irwin King, Michael R. Lyu and Tianbao Yang. A generic approach for accelerating stochastic zeroth-order convex optimization. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence (IJCAI)*, pages 3040–3046, 2018. *(Chapter 6)*

[5] **Xiaotian Yu**, Han Shao, Michael R. Lyu and Irwin King. Pure exploration of multi-armed bandits with heavy-tailed payoffs. In *Proceedings of the Thirty-Fourth Conference on Uncertainty in Artificial Intelligence (UAI)*, pages 937–946, 2018. *(Chapter 4)*

[6] Han Shao, **Xiaotian Yu**, Irwin King and Michael R. Lyu. Almost optimal algorithms for linear stochastic bandits with heavy-tailed payoffs. In *Proceedings of Advances in Neural Information Processing Systems (NIPS)*, pages 8430–8439, 2018. **Spotlight presentation**. *(Chapter 5)*

[7] **Xiaotian Yu**, Irwin King and Michael R. Lyu. Fixed-budget pure exploration of multi-armed bandits with second-order information. Submitted to IEEE Transactions on Knowledge and Data Engineering (TKDE)

End

Chapter 3: Theoretical Results

Theorem (Estimate error for mean-variance)

For pure exploration of mean-variance in MAB with K arms, suppose Assumptions 3.1-3.3 are satisfied. We define a random variable as $\rho_t(a) \triangleq \hat{\omega}_t(a) - \omega(a)$ for any $a \in [K]$. Then, we have $\rho_t(a)$ is sub-gamma on the right tail, implying

$$\mathbb{E}[\exp(\lambda\rho_t(a))] \leq \exp\left(\frac{\lambda^2 v}{2(1 - c\lambda)}\right), \tag{11}$$

where $\lambda \in (0, 1/c)$, $c = 8R^2$, $v = (192R^2 + \kappa^2)R^2$ for any $a \in [K]$ and $t \in [T]$.

Proof of Theorem 3.3 on Page 47 in the thesis

Chapter 3: Theoretical Results

Theorem (Probability of error for PEMV.CB)

For pure exploration of mean-variance with K-arm MAB, suppose Assumptions 3.1-3.3 are satisfied. If PEMV.CB is run with a fixed budget TK, we have the upper bound of the probability of error for PEMV.CB as

$$\mathbb{P}[x_T \neq Opt] \leq 2TK \exp\left(-\frac{\delta}{5}\right), \tag{12}$$

where $\delta \in \left(0, \min\left(\frac{25(T-2K)}{576(96R^2+\kappa^2)R^2\mathbf{H}_1}, \frac{5(T-2K)}{96R^2\mathbf{H}_3}\right)\right]$.

Proof of Theorem 3.1 on Page 53 in the thesis

Chapter 3: Theoretical Results

Theorem (Probability of error for PEMV.HALVING)

For pure exploration of mean-variance with K-arm MAB, suppose Assumptions 3.1-3.3 are satisfied. If PEMV.HALVING is run with a fixed budget T, we have the upper bound of the probability of error for PEMV.HALVING as

$$\mathbb{P}[x_T \neq \mathit{Opt}] \leq 2K \exp\left(-\frac{T}{\log_2(K)\mathbf{H}}\right), \tag{13}$$

where $\mathbf{H} = 12(96R^2 + \kappa^2)R^2 \min(\mathbf{H}_4, 3\mathbf{H}_2)$.

Proof of Theorem 3.2 on Page 58 in the thesis

Chapter 4: Theoretical Results

Theorem (Sample complexity of SE-δ)

For pure exploration in MAB with K arms, with probability at least $1-\delta$, SE-δ identifies the optimal arm Opt with sample complexity as

- *for SE-δ(EA)*

$$T \leq \sum_{x=1}^{K} \left(\frac{2^{2p+1} K C}{\Delta_x^p \delta} \right)^{\frac{1}{p-1}};$$

- *for SE-δ(TEA)*

$$T \leq \sum_{x=1}^{K} \left(\frac{20 B^{\frac{1}{p}}}{\Delta_x} \right)^{\frac{p}{p-1}} \log\left(\frac{2K}{\delta} \right),$$

where $p \in (1, 2]$.

Proof of Theorem 4.1 on Page 88 in the thesis

Chapter 4: Theoretical Results

Theorem (Probability of error for SR-T)

For pure exploration in MAB with K arms, if Algorithm SR-T is run with a fixed budget T, we have probability of error for $p \in (1, 2]$ as

- *for SR-T(EA)*

$$\mathbb{P}[Out \neq Opt] \leq 2^{p+1} CK(K-1)H_2^p \left(\frac{\bar{K}}{T-K}\right)^{p-1};$$

- *for SR-T(TEA)*

$$\mathbb{P}[Out \neq Opt] \leq 2K(K-1)\exp\left(-\frac{(T-K)\bar{B}_1}{\bar{K}K\underline{\Delta}^{p/(1-p)}}\right),$$

where $\bar{B}_1 = \frac{p-1}{4(2^p 3Bp^p)^{\frac{1}{p-1}}}$.

Proof of Theorem 4.2 on Page 90 in the thesis

Chapter 5: Lower Bound of LinBET

- Setting
  Assume $d \geq 2$ is even. For $D_t \in \mathbb{R}^d$, we fix the decision set as $D_t = D_{(d)}$, where $D_{(d)} \triangleq \{(x_1, \cdots, x_d) \in \mathbb{R}_+^d : x_1 + x_2 = \cdots = x_{d-1} + x_d = 1\}$. Let $S_d \triangleq \{(\theta_1, \cdots, \theta_d) : \forall i \in [d/2], (\theta_{2i-1}, \theta_{2i}) \in \{(2\Delta, \Delta), (\Delta, 2\Delta)\}\}$ with $\Delta \in (0, 1/d]$. Payoffs are in $\{0, (1/\Delta)^{\frac{1}{p-1}}\}$ such that, for every $x \in D_{(d)}$, the expected payoff is $\theta_*^\top x$.
- Result

Theorem (Lower bound of LinBET)

If θ_* is chosen uniformly at random from S_d, and the payoff for each $x \in D_{(d)}$ is in $\{0, (1/\Delta)^{\frac{1}{p-1}}\}$ with mean $\theta_*^\top x$, then for any algorithm $\mathcal{A}$ and every $T \geq (d/12)^{\frac{p-1}{p}}$, we have

$$\mathbb{E}\left[\mathbf{R}(\mathcal{A}, T)\right] \geq \frac{d}{192} T^{\frac{1}{p}}.$$

Chapter 5: Lower Bound of LinBET

$d = 2$ and $\mathbb{E}\left[|y_t|^p | \mathcal{F}_{t-1}\right] \leq d$ case

- Decision set: $D_{(2)} \triangleq \{(x_1, x_2) \in \mathbb{R}_+^2 : x_1 + x_2 = 1\}$
- Payoff function of x:

$$y(x) = \begin{cases} \left(\frac{1}{\Delta}\right)^{\frac{1}{p-1}} & \text{with a probability of } \Delta^{\frac{1}{p-1}} \theta_*^\top x, \\ 0 & \text{with a probability of } 1 - \Delta^{\frac{1}{p-1}} \theta_*^\top x \end{cases}$$

- θ_* is chosen uniformly at random from $\{\mu_1, \mu_2\}$, where $\mu_1 = (2\Delta, \Delta)$ and $\mu_2 = (\Delta, 2\Delta)$
- Change of measure (through $\mu_0 = (\Delta, \Delta)$)
- Set $\Delta = T^{-\frac{p-1}{p}}/12$
- $\mathbb{E}\left[\mathbf{R}(\mathcal{A}, T)\right] \geq \frac{1}{96} T^{\frac{1}{p}}$
- Extend it to $d > 2$

Chapter 5: Algorithm for Linear Stochastic Bandits

OFUL (Abbasi-Yadkori et al., 2011)

- At time t, select arm x_t by
  - $(x_t, \tilde{\theta}_t) = \arg\max_{(x,\theta)\in D_t \times C_{t-1}} \langle x, \theta \rangle$
  - $C_t = \{\theta : \|\theta - \hat{\theta}_{t,k^*}\|_{V_t} \leq \beta_t\}$, $V_t = \lambda I + \sum_{\tau=1}^{t} x_\tau x_\tau^\top$
- $\beta_0 = \sqrt{\lambda}\|\theta_*\|_2$
- For sub-Gaussian case, $\beta_t = O\left(\sqrt{\log t}\right)$
- The regret is bounded by $\widetilde{O}\left(\max_{t\in[T]} \beta_{t-1}\sqrt{T}\right)$

Chapter 5: Upper Bound Analysis: MENU

Results

Theorem

Assume that for all t and $x_t \in D_t$ with $\|x_t\|_2 \le D$, $\|\theta_*\|_2 \le S$, $|x_t^\top \theta_*| \le L$ and $\mathbb{E}[|\eta_t|^p|\mathcal{F}_{t-1}] \le c$. Then, with probability at least $1-\delta$, for every $T \ge 256 + 24\log(e/\delta)$, the regret of the MENU algorithm satisfies

$$\mathbf{R}(MENU, T) \le \widetilde{O}(c^{\frac{1}{p}} d^{\frac{1}{2}+\frac{1}{p}} T^{\frac{1}{p}}).$$

Proof of Theorem 5.2 on Page 118 in thesis

Chapter 5: Upper Bound Analysis: MENU

Proof sketch

- Lemma 1 (Confidence Ellipsoid of LSE)
  Let $\hat{\theta}_n$ denote the LSE of θ_* with the sequence of decisions $x_1, \cdots, x_n$ and observed payoffs $y_1, \cdots, y_n$. Assume that for all $\tau \in [n]$ and all $x_\tau \in D_\tau \subseteq \mathbb{R}^d$, $\mathbb{E}[|\eta_\tau|^p | \mathcal{F}_{\tau-1}] \leq c$ and $\|\theta_*\|_2 \leq S$. Then $\hat{\theta}_n$ satisfies

$$\mathbb{P}\left(\|\hat{\theta}_n - \theta_*\|_{V_n} \leq (9dc)^{\frac{1}{p}} n^{\frac{2-p}{2p}} + \lambda^{\frac{1}{2}} S\right) \geq \frac{3}{4},$$

- Lemma 2
  Recall $\hat{\theta}_{n,j}$, $\hat{\theta}_{n,k^*}$ and V_n in MENU. If there exists a $\gamma > 0$ such that $\Pr\left(\|\hat{\theta}_{n,j} - \theta_*\|_{V_n} \leq \gamma\right) \geq \frac{3}{4}$ holds for all $j \in [k]$ with $k \geq 1$, then with probability at least $1 - e^{-\frac{k}{24}}$, $\|\hat{\theta}_{n,k^*} - \theta_*\|_{V_n} \leq 3\gamma$.

Chapter 5: Upper Bound Analysis: MENU

Proof sketch of Lemma 1

- Let u_i denote the i-th row of $V_t^{-1/2} X_t^\top$
- $\|\hat{\theta}_n - \theta_*\|_{V_n} \le \sqrt{\sum_{i=1}^d \left(u_i^\top (Y_n - X_n\theta_*)\right)^2} + \lambda \|\theta_*\|_{V_n^{-1}}$
- Union bound

$$
\begin{aligned}
&\mathbb{P}\left(\sum_{i=1}^{d}\left(\sum_{\tau=1}^{n} u_{i,\tau}\eta_\tau\right)^2 > \gamma^2\right) \\
&\le \mathbb{P}\left(\exists i, \tau, |u_{i,\tau}\eta_\tau| > \gamma\right) + \mathbb{P}\left(\sum_{i=1}^{d}\left(\sum_{\tau=1}^{n} u_{i,\tau}\eta_\tau \mathbb{1}_{|u_{i,\tau}\eta_\tau| \le \gamma}\right)^2 > \gamma^2\right),
\end{aligned}
$$

where $\mathbb{1}_{\{\cdot\}}$ is the indicator function

- Both terms could be bounded by Markov's inequality
- Set $\gamma = (9dc)^{\frac{1}{p}} n^{\frac{2-p}{2p}}$

Chapter 5: Upper Bound Analysis: MENU

Proof sketch of Lemma 2

- By Azuma-Hoeffding's inequality, we have with prob. at least $1 - e^{-\frac{k}{24}}$, more than 2/3 of $\{\hat{\theta}_{n,1}, \cdots, \hat{\theta}_{n,k}\}$ are contained in $\mathbb{B}_{V_n}(\theta_*, \gamma) \triangleq \{\theta : \|\theta - \theta_*\|_{V_n} \leq \gamma\}$
- r_j be the median of $\{\|\hat{\theta}_{n,j} - \hat{\theta}_{n,s}\|_{V_n} : s \in [k] \backslash j\}$
- Select arm $\arg\min_{j \in [k]} r_j$
  - If $\hat{\theta}_{n,j} \in \mathbb{B}_{V_n}(\theta_*, \gamma)$, $\|\hat{\theta}_{n,j} - \hat{\theta}_{n,s}\|_{V_n} \leq 2\gamma$ for all $\hat{\theta}_{n,s} \in \mathbb{B}_{V_n}(\theta_*, \gamma)$ by triangle inequality. Therefore, $r_j \leq 2\gamma$
  - If $\hat{\theta}_{n,j} \notin \mathbb{B}_{V_n}(\theta_*, 3\gamma)$, $\|\hat{\theta}_{n,j} - \hat{\theta}_{n,s}\|_{V_n} > 2\gamma$ for all $\hat{\theta}_{n,s} \in \mathbb{B}_{V_n}(\theta_*, \gamma)$ by triangle inequality. Therefore, $r_j > 2\gamma$

Chapter 5: Upper Bound Analysis: TOFU

Results

Theorem

Assume that for all t and $x_t \in D_t$ with $\|x_t\|_2 \leq D$, $\|\theta_*\|_2 \leq S$, $|x_t^\top \theta_*| \leq L$ and $\mathbb{E}[|y_t|^p|\mathcal{F}_{t-1}] \leq b$. Then, with probability at least $1-\delta$, for every $T \geq 1$, the regret of the TOFU algorithm satisfies

$$\mathbf{R}(\mathit{TOFU}, T) \leq \widetilde{O}(b^{\frac{1}{p}} d T^{\frac{1}{p}}).$$

Proof of Theorem 5.3 on Page 122 in the thesis

Chapter 5: Upper Bound Analysis: TOFU

Lemma 3. [Confidence Ellipsoid of Truncated Estimate] With the sequence of decisions $x_1, \cdots, x_t$, the truncated payoffs $\{Y_i^\dagger\}_{i=1}^d$ and the parameter estimate $\theta_t^\dagger$ are defined in TOFU. Assume that for all $\tau \in [t]$ and all $x_\tau \in D_\tau \subseteq \mathbb{R}^d$, $\mathbb{E}[|y_\tau|^p|\mathcal{F}_{\tau-1}] \leq b$ and $\|\theta_*\|_2 \leq S$. With probability at least $1-\delta$, we have

$$\|\theta_t^\dagger - \theta_*\|_{V_t} \leq 4\sqrt{d} b^{\frac{1}{p}} \left(\log\left(\frac{2d}{\delta}\right)\right)^{\frac{p-1}{p}} t^{\frac{2-p}{2p}} + \lambda^{\frac{1}{2}} S, \tag{14}$$

where $\lambda > 0$ is a regularization parameter and $V_t = \lambda I_d + \sum_{\tau=1}^t x_\tau x_\tau^\top$.

Chapter 5: Upper Bound Analysis: TOFU

Proof sketch of Lemma 3

- Like before,

$$\|\theta_t^\dagger - \theta_*\|_{V_t} \leq \sqrt{\sum_{i=1}^{d} \left(u_i^\top (Y_i^\dagger - X_t\theta_*)\right)^2} + \lambda\|\theta_*\|_{V_n^{-1}}$$

- For each i

$$u_i^\top \left(Y_i^\dagger - X_t\theta_*\right) = \sum_{\tau=1}^{t} u_{i,\tau} \left(Y_{i,\tau}^\dagger - \mathbb{E}[Y_{i,\tau}|\mathcal{F}_{\tau-1}]\right)$$
$$\leq \left|\sum_{\tau=1}^{t} u_{i,\tau}(Y_{i,\tau}^\dagger - \mathbb{E}[Y_{i,\tau}^\dagger|\mathcal{F}_{\tau-1}])\right| + \left|\sum_{\tau=1}^{t} u_{i,\tau}\mathbb{E}[Y_{i,\tau}\mathbb{1}_{|u_{i,\tau}Y_{i,\tau}|>b_t}|\mathcal{F}_{\tau-1}]\right|$$

- The first term is bounded by Bernstein's inequality
- Set $b_t = (b/\log(2d/\delta))^{\frac{1}{p}}\ t^{\frac{2-p}{2p}}$

Chapter 6: Lemmas

- Lemma 1 (Flaxman et al., 2005)
  Given $\mathbf{u} \sim \mathbb{B}(\mathbf{0}, 1)$, we have $\mathbb{E}_{\mathbf{u}}[\mathbf{g}_t^{\mathrm{f}}] = \nabla \hat{f}(\mathbf{x}_t; \xi_t)$, and $\|\mathbf{g}_t^{\mathrm{f}}\|_2 \leq dB/\delta$. If $f(\mathbf{x}; \xi)$ is G-Lipschitz continuous, we have $|f(\mathbf{x}; \xi) - \hat{f}(\mathbf{x}; \xi)| \leq G\delta$. If $f(\mathbf{x}; \xi)$ is L-smooth, we have $|f(\mathbf{x}; \xi) - \hat{f}(\mathbf{x}; \xi)| \leq L\delta^2/2$.
- Lemma 2 (Agarwal et al., 2010)
  Given $\mathbf{u} \sim \mathbb{B}(\mathbf{0}, 1)$, we have $\mathbb{E}_{\mathbf{u}}[\mathbf{g}_t^{\mathrm{a}}] = \nabla \hat{f}(\mathbf{x}_t; \xi_t)$. If $f(\mathbf{x}; \xi)$ is G-Lipschitz continuous, we have $\|\mathbf{g}_t^{\mathrm{a}}\|_2 \leq Gd$, $\mathbb{E}_{\mathbf{u}}[\|\mathbf{g}_t^{\mathrm{a}}\|_2^2] \leq db^2G^2C$, and $|f(\mathbf{x}; \xi) - \hat{f}(\mathbf{x}; \xi)| \leq G\delta$, where C is a universal constant and b is a constant such that $(\mathbb{E}[\|\mathbf{u}\|_2^4])^{1/4} \leq b$. If $f(\mathbf{x}; \xi)$ is L-smooth, we have $|f(\mathbf{x}; \xi) - \hat{f}(\mathbf{x}; \xi)| \leq L\delta^2/2$.

Chapter 6: Lemmas

- Lemma 3 (Nesterov et al., 2017)
  Considering $\mathbf{u} \sim \mathcal{N}(0,1)$, we have $\mathbb{E}_{\mathbf{u}}[\mathbf{g}_t^{\mathsf{n}}] = \nabla \hat{f}(\mathbf{x}_t; \xi_t)$. If $f(\mathbf{x}; \xi)$ is G-Lipschitz continuous, we have $\mathbb{E}_{\mathbf{u}}[\|\mathbf{g}_t^{\mathsf{n}}\|_2^2] \le G^2(d+4)^2$, and $|f(\mathbf{x}; \xi_t) - \hat{f}(\mathbf{x}; \xi_t)| \le \delta G d^{1/2}$. If $f(\mathbf{x}; \xi)$ is G-Lipschitz continuous and L-smooth, we have $\mathbb{E}_{\mathbf{u}}[\|\mathbf{g}_t^{\mathsf{n}}\|_2^2] \le \delta^2 (d+6)^3 L^2/2 + 2(d+4)G^2$, and $|f(\mathbf{x}; \xi) - \hat{f}(\mathbf{x}; \xi)| \le \delta^2 L d/2$.

$$\mathbf{g}_t^{\mathsf{n}} = \frac{1}{\delta}(f(\mathbf{x}_t + \delta \mathbf{u}_t; \xi_t) - f(\mathbf{x}_t; \xi_t))\mathbf{u}_t. \tag{15}$$

Chapter 6: Proof Sketch of Results in Expectation (OPE)

- Cumulative errors of $\forall \mathbf{x} \in \Omega$

$$\sum_{t=1}^{T} f(\mathbf{x}_t; \xi_t) - f(\mathbf{x}; \xi_t) \leq 2TG\delta + \frac{\eta T d^2 B^2}{2\delta^2} + \frac{\|\mathbf{x}_1 - \mathbf{x}\|_2^2}{2\eta} + \sum_{t=1}^{T} (\nabla \hat{f}(\mathbf{x}_t; \xi_t) - \mathbf{g}_t^{\mathrm{f}})^\top (\mathbf{x}_t - \mathbf{x}).$$

- At the k-th stage

$$\mathbb{E}[f(\mathbf{x}_k) - f(\mathbf{x})] \leq \frac{\mathbb{E}[\|\mathbf{x}_{k-1} - \mathbf{x}\|_2^2]}{2\eta_k t} + \frac{\eta_k d^2 B^2}{2\delta_k^2} + 2G\delta_k,$$

- By induction, we prove $\mathbb{E}[f(\mathbf{x}_k) - f_*] \leq \epsilon_k$

$\mathbb{E}[f(\mathbf{x}_k) - f_*] \leq \epsilon_k$ (OPE)

$$
\begin{aligned}
&\mathbb{E}[f(\mathbf{x}_k) - f(\mathbf{x}_{k-1,*})] \\
&\leq \frac{\mathbb{E}[\|\mathbf{x}_{k-1} - \mathbf{x}_{k-1,*}\|_2^2]}{2\eta_k t} + \frac{\eta_k d^2 B^2}{2\delta_k^2} + 2G\delta_k \\
&\leq \frac{c(\mathbb{E}[f(\mathbf{x}_{k-1}) - f(\mathbf{x}_{k-1,*})])^{2\theta}}{2\eta_k t} + \frac{\eta_k d^2 B^2}{2\delta_k^2} + 2G\delta_k \\
&\leq \frac{c\epsilon_{k-1}^{2\theta}}{2\eta_k t} + \frac{\eta_k d^2 B^2}{2\delta_k^2} + 2G\delta_k,
\end{aligned}
$$

$$
\begin{aligned}
&\frac{c^2\epsilon_{k-1}^{2\theta}}{2\eta_k t} \leq \frac{\epsilon_{k-1}}{6} \Rightarrow t \geq \frac{1296 d^2 B^2 G^2 c^2}{\epsilon_{k-1}^{2(2-\theta)}}, \\
&\frac{\eta_k d^2 B^2}{2\delta_k^2} \leq \frac{\epsilon_k}{3} \Rightarrow \eta_k \leq \frac{\epsilon_k^3}{54 G^2 d^2 B^2}, \\
&2G\delta_k \leq \frac{\epsilon_k}{3} \Rightarrow \delta_k \leq \frac{\epsilon_k}{6G}.
\end{aligned}
$$

Chapter 6: Proof Sketch of Results with High Probability (OPE)

- High probability error

$$\hat{f}(\hat{\mathbf{x}}_T) - \hat{f}(\mathbf{x}) \leq \frac{\|\mathbf{x}_1 - \mathbf{x}\|_2^2}{2\eta T} + \frac{\eta d^2 B^2}{2\delta^2} + \frac{4dBD\sqrt{3\log(\frac{1}{\tilde{p}})}}{\sqrt{T}\delta},$$

- By induction, we prove $f(\mathbf{x}_k) - f_* \leq \epsilon_k$

$$f(\mathbf{x}_k) - f(\mathbf{x}_{k-1,*}) \leq \frac{c^2\epsilon_{k-1}^{2\theta}}{2\eta_k t} + \frac{\eta_k d^2 B^2}{2\delta_k^2} + \frac{4dBc\epsilon_{k-1}^{\theta}\sqrt{3\log(\frac{1}{\tilde{p}})}}{\sqrt{t}\delta_k} + 2G\delta_k$$