NJUAI 南京大学 人工智能学院
SCHOOL OF ARTIFICIAL INTELLIGENCE, NANJING UNIVERSITY

Lecture 6. Online Mirror Descent

Advanced Optimization (Fall 2024)

Peng Zhao

zhaop@lamda.nju.edu.cn

Nanjing University

Outline

- Prediction with Expert Advice
- Online Mirror Descent
- Follow-the-Regularized Leader

Part 1. Prediction with Expert Advice

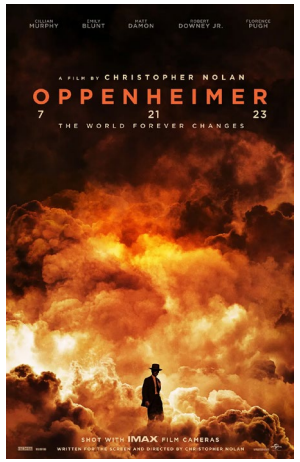
- Problem Setup
- Algorithms
- Regret Analysis

Motivation

- Consider that we are making predictions based on external experts.

A Chinese Odyssey Part Two - Cinderella

豆 9.2/10 | RT 87% | IMDb 7.8/10



Oppenheimer

豆 8.8/10 | RT 93% | IMDb 8.5/10

Titanic

豆 9.5/10 | RT 88% | IMDb 7.9/10

Prediction with Expert Advice

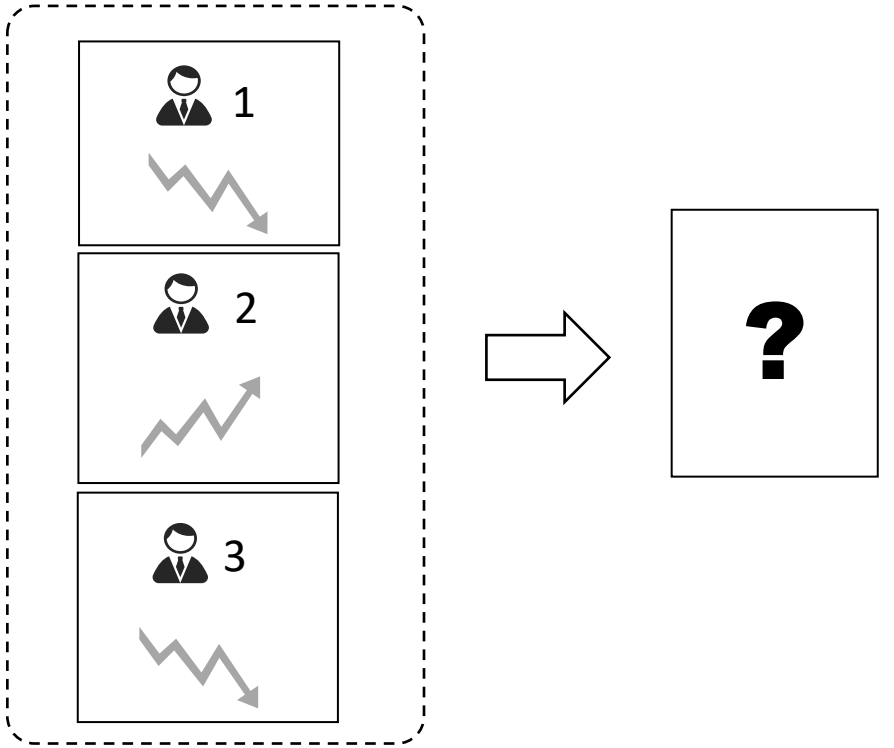
- Another Example: Universal Portfolio Selection

- Universal Portfolio Selection 
  - a total of d stocks in the stock market.
  - each round, the player chooses stocks by a distribution $\mathbf{x}_t \in \Delta_d$.
  - the market returns the price ratio $\boldsymbol{\theta}_t$ between iter t and $t+1$,

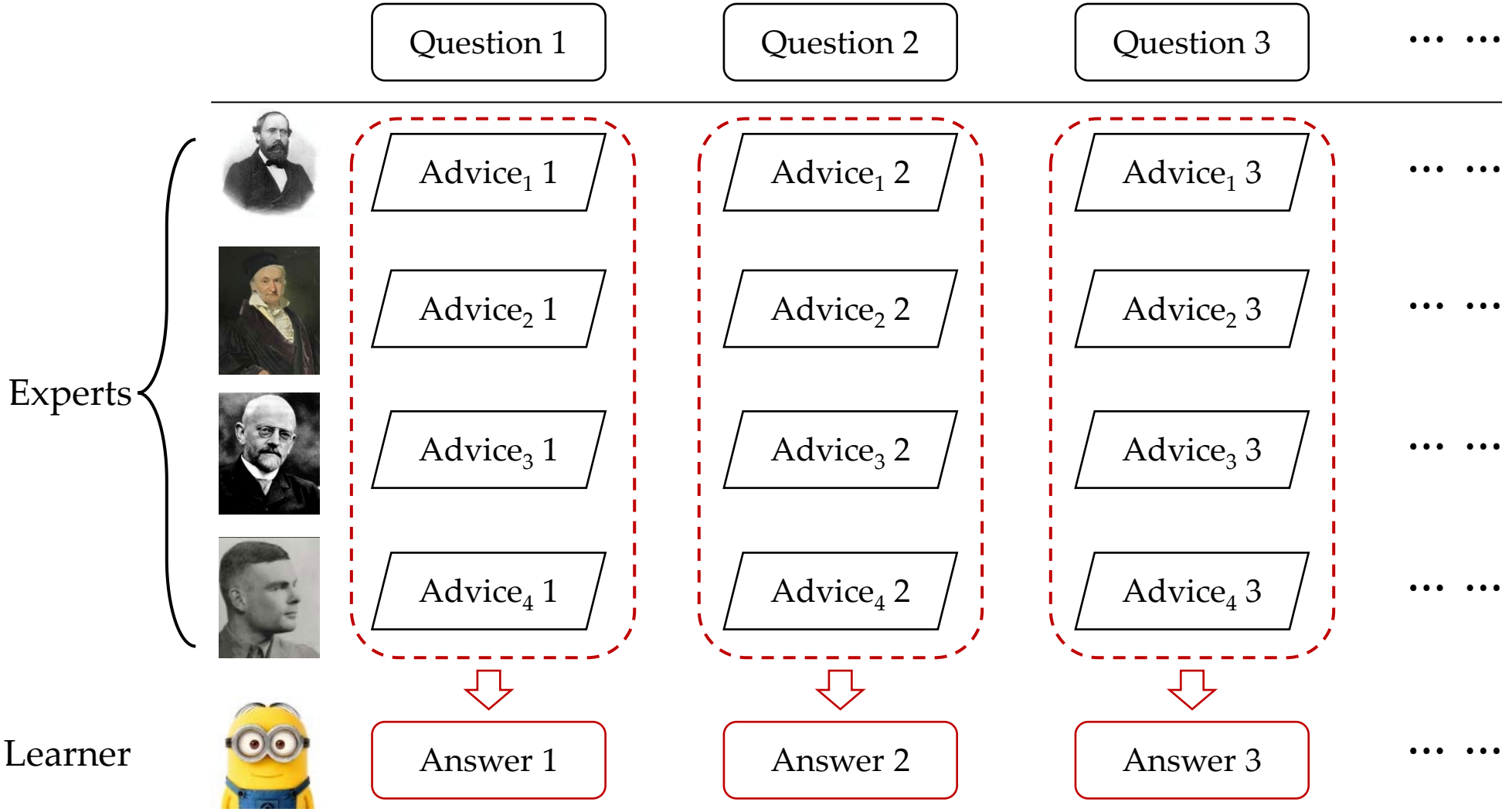
$$\boldsymbol{\theta}_t(i) = \frac{\text{price of stock}_i \text{ at time } t+1}{\text{price of stock}_i \text{ at time } t}$$

which means that our final wealth W_T will be: $W_T = W_1 \cdot \prod_{t=1}^{T} \boldsymbol{\theta}_t^\top \mathbf{x}_t$

⇨ Our goal is to maximize our wealth at time T.



PEA Problem Setup



PEA: Formulization

- The online learner (player) aims to make the prediction based by combining N experts' advice.

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

(1) the player first picks a weight $\boldsymbol{p}_t$ from a simplex Δ_N;

(2) and simultaneously environments pick a loss vector $\boldsymbol{\ell}_t \in \mathbb{R}^N$;

(3) the player suffers loss $f_t(\boldsymbol{p}_t) \triangleq \langle \boldsymbol{p}_t, \boldsymbol{\ell}_t \rangle$, observes $\boldsymbol{\ell}_t$ and updates the model.

The feasible domain is the $(N-1)$-dim simplex $\Delta_N = \left\{ \boldsymbol{p} \in \mathbb{R}^N \mid p_i \geq 0, \sum_{i=1}^N p_i = 1 \right\}$.

We typically assume that $0 \leq \ell_{t,i} \leq 1$ holds for all $t \in [T]$ and $i \in [N]$.

PEA: Formulization

- The online learner (player) aims to make the prediction based by combining N experts' advice.

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

(1) the player first picks a weight $\boldsymbol{p}_t$ from a simplex Δ_N;

(2) and simultaneously environments pick a loss vector $\boldsymbol{\ell}_t \in \mathbb{R}^N$;

(3) the player suffers loss $f_t(\boldsymbol{p}_t) \triangleq \langle \boldsymbol{p}_t, \boldsymbol{\ell}_t \rangle$, observes $\boldsymbol{\ell}_t$ and updates the model.

- The goal is to minimize the regret with respect to the ***best expert***:

$$\text{Regret}_T \triangleq \sum_{t=1}^{T} \langle \boldsymbol{p}_t, \boldsymbol{\ell}_t \rangle - \min_{\boldsymbol{p} \in \Delta_N} \sum_{t=1}^{T} \langle \boldsymbol{p}, \boldsymbol{\ell}_t \rangle = \sum_{t=1}^{T} \langle \boldsymbol{p}_t, \boldsymbol{\ell}_t \rangle - \min_{i \in [N]} \sum_{t=1}^{T} \ell_{t,i}$$

A Natural Solution

- **Follow the Leader (FTL)**

Select the expert that ***performs best so far***, specifically,

$$\boldsymbol{p}_t^{\text{FTL}} = \arg\min_{\boldsymbol{p}\in\Delta_N} \langle \boldsymbol{p}, L_{t-1}\rangle = \operatorname*{argmin}_{i\in[N]} L_{t-1,i}$$

where $L_{t-1} \in \mathbb{R}^N$ is the cumulative loss vector with $L_{t-1,i} \triangleq \sum_{s=1}^{t-1} \ell_{s,i}$.

$\ell_{1,1} = 0.49$ ⇨ $\ell_{2,1} = 1$ ⇨ $\ell_{3,1} = 0$ ⇨ ⋯ ⋯

$\ell_{1,2} = 0.51$ ⇨ $\ell_{2,2} = 0$ ⇨ $\ell_{3,2} = 1$ ⇨ ⋯ ⋯

$$\text{Reg}_T = \sum_{t=1}^{T} \langle \boldsymbol{p}_t, \boldsymbol{\ell}_t\rangle - \min_{i\in[N]} \sum_{t=1}^{T} \ell_{t,i}$$

$$= T - \frac{T}{2} = \mathcal{O}(T)$$

FTL achieves ***linear regret*** in the worst case!

A Natural Solution

- **Follow the Leader (FTL)**

  Select the expert that ***performs best so far***, specifically,

$$\boldsymbol{p}_t^{\mathrm{FTL}} = \arg\min_{\boldsymbol{p}\in\Delta_N} \langle \boldsymbol{p}, \boldsymbol{L}_{t-1} \rangle = \operatorname*{argmin}_{i\in[N]} L_{t-1,i}$$

  where $\boldsymbol{L}_{t-1} \in \mathbb{R}^N$ is the cumulative loss vector with $L_{t-1,i} \triangleq \sum_{s=1}^{t-1} \ell_{s,i}$.

⇨ Pitfall: online decision is made ***blindly*** based on the historical performance!

⇨ Replacing the "max" operation in FTL by "***softmax***".

Hedge: Algorithm

- Hedge: replacing the "*max*" operation in FTL by "***softmax***".

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

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

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

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

FTL update

$$\boldsymbol{p}_t^{\text{FTL}} = \arg\max_{\boldsymbol{p} \in \Delta_N} \langle \boldsymbol{p}, -\boldsymbol{L}_{t-1} \rangle$$

Hedge update

$$p_{t,i} \propto \exp\left(-\eta L_{t-1,i}\right), \forall i \in [N]$$

Lazy and Greedy Updates

- Hedge algorithm

lazy update

$$p_{t+1,i} \propto \exp\left(-\eta L_{t,i}\right), \forall i \in [N] \qquad L_{t,i} = \sum_{s=1}^{t} \ell_{s,i}, \ \forall i \in [N]$$

- Another equivalent update (when the learning rate η is ***fixed***)

greedy update

$$p_{t+1,i} \propto p_{t,i} \exp\left(-\eta \ell_{t,i}\right), \forall i \in [N]$$

where we set the uniform initialization as $p_{0,i} = 1/N, \forall i \in [N]$.

⇨ But the two updates can be ***significantly different when learning rate is changing***.

Hedge: Regret Bound

Theorem 1. *Suppose that* $\forall t \in [T]$ *and* $i \in [N], 0 \leq \ell_{t,i} \leq 1$*, then Hedge with learning rate* η *guarantees*

$$\mathrm{Regret}_T \leq \frac{\ln N}{\eta} + \eta T = \mathcal{O}\big(\sqrt{T \log N}\big),$$

where the last equality is by setting η *optimally as* $\sqrt{(\ln N)/T}$*.*

Proof. We present a 'potential-based' proof here, where the potential is defined as

$$\Phi_t \triangleq \frac{1}{\eta} \ln \left(\sum_{i=1}^{N} \exp\left(-\eta L_{t,i}\right) \right).$$

Proof of Hedge Regret Bound

Proof.

$$\Phi_t - \Phi_{t-1} = \frac{1}{\eta} \ln \left(\frac{\sum_{i=1}^N \exp\left(-\eta L_{t,i}\right)}{\sum_{i=1}^N \exp\left(-\eta L_{t-1,i}\right)} \right) \qquad \Phi_t \triangleq \frac{1}{\eta} \ln \left(\sum_{i=1}^N \exp\left(-\eta L_{t,i}\right) \right)$$

$$= \frac{1}{\eta} \ln \left(\sum_{i=1}^N \left(\frac{\exp\left(-\eta L_{t-1,i}\right)}{\sum_{i=1}^N \exp\left(-\eta L_{t-1,i}\right)} \exp(-\eta \ell_{t,i}) \right) \right)$$

$$= \frac{1}{\eta} \ln \left(\sum_{i=1}^N p_{t,i} \exp\left(-\eta \ell_{t,i}\right) \right) \qquad \text{(update step of } \boldsymbol{p}_t\text{)}$$

$$\leq \frac{1}{\eta} \ln \left(\sum_{i=1}^N p_{t,i} \left(1 - \eta \ell_{t,i} + \eta^2 \ell_{t,i}^2\right) \right) \qquad (\forall x \geq 0, e^{-x} \leq 1 - x + x^2)$$

$$= \frac{1}{\eta} \ln \left(1 - \eta \left\langle \boldsymbol{p}_t, \boldsymbol{\ell}_t \right\rangle + \eta^2 \sum_{i=1}^N p_{t,i} \ell_{t,i}^2 \right)$$

Proof of Hedge Regret Bound

Proof. $$\Phi_t - \Phi_{t-1} = \frac{1}{\eta} \ln \left(\frac{\sum_{i=1}^N \exp\left(-\eta L_{t,i}\right)}{\sum_{i=1}^N \exp\left(-\eta L_{t-1,i}\right)} \right)$$

$$\leq -\left\langle \boldsymbol{p}_t, \boldsymbol{\ell}_t \right\rangle + \eta \sum_{i=1}^N p_{t,i} \ell_{t,i}^2 \qquad (\ln(1+x) \leq x)$$

Summing over t, we have

$$\sum_{t=1}^T \left\langle \boldsymbol{p}_t, \boldsymbol{\ell}_t \right\rangle \leq \Phi_0 - \Phi_T + \eta \sum_{t=1}^T \sum_{i=1}^N p_{t,i} \ell_{t,i}^2 \qquad \Phi_t \triangleq \frac{1}{\eta} \ln \left(\sum_{i=1}^N \exp\left(-\eta L_{t,i}\right) \right)$$

$$\leq \frac{\ln N}{\eta} - \frac{1}{\eta} \ln \left(\exp\left(-\eta L_{T,i^\star}\right) \right) + \eta \sum_{t=1}^T \sum_{i=1}^N p_{t,i} \ell_{t,i}^2$$

$$\leq \frac{\ln N}{\eta} + L_{T,i^\star} + \eta \sum_{t=1}^T \sum_{i=1}^N p_{t,i} \ell_{t,i}^2$$

Proof of Hedge Regret Bound

Proof.

$$\sum_{t=1}^{T} \langle \boldsymbol{p}_t, \boldsymbol{\ell}_t \rangle \leq \frac{\ln N}{\eta} + L_{T,i^\star} + \eta \sum_{t=1}^{T} \sum_{i=1}^{N} p_{t,i} \ell_{t,i}^2$$

Rearranging the term gives

$$\begin{aligned} \sum_{t=1}^{T} \langle \boldsymbol{p}_t, \boldsymbol{\ell}_t \rangle - L_{T,i^\star} &\leq \frac{\ln N}{\eta} + \eta \sum_{t=1}^{T} \sum_{i=1}^{N} p_{t,i} \ell_{t,i}^2 \\ &\leq \frac{\ln N}{\eta} + \eta T \qquad (\ell_{t,i} \leq 1) \end{aligned}$$

Thus, setting $\eta = \sqrt{\ln N / T}$ yields

$$\mathrm{Regret}_T \leq \frac{\ln N}{\eta} + \eta T = 2\sqrt{T \ln N}.$$

□

Lower bound of PEA

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

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

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

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

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

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

Lower bound of PEA

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

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

Proof. We construct the 'hard' instance by randomization. Let $\mathcal{D}$ be the uniform distribution over $\{0, 1\}$. We have

$$\begin{aligned}
\max_{\boldsymbol{\ell}_1,\ldots,\boldsymbol{\ell}_T} \mathrm{Regret}_T &\geq \mathbb{E}_{\boldsymbol{\ell}_1,\ldots,\boldsymbol{\ell}_T \overset{\text{i.i.d.}}{\sim} \mathcal{D}^N} [\mathrm{Regret}_T] \\
&\qquad \text{(conditional expectation decomposition)} \\
&= \sum_{t=1}^{T} \mathbb{E}_{\boldsymbol{\ell}_1,\ldots,\boldsymbol{\ell}_{t-1}} \mathbb{E}_{\boldsymbol{\ell}_t} [\langle p_t, \boldsymbol{\ell}_t \rangle \mid \boldsymbol{\ell}_{t-1}, \ldots, \boldsymbol{\ell}_1] - \mathbb{E}_{\boldsymbol{\ell}_1,\ldots,\boldsymbol{\ell}_T} \left[\min_{i\in[N]} \sum_{t=1}^{T} \ell_{t,i} \right] \\
&= \sum_{t=1}^{T} \mathbb{E}_{\boldsymbol{\ell}_1,\ldots,\boldsymbol{\ell}_{t-1}} \langle p_t, \mathbb{E}_{\boldsymbol{\ell}_t} [\boldsymbol{\ell}_t \mid \boldsymbol{\ell}_{t-1}, \ldots, \boldsymbol{\ell}_1] \rangle - \mathbb{E}_{\boldsymbol{\ell}_1,\ldots,\boldsymbol{\ell}_T} \left[\min_{i\in[N]} \sum_{t=1}^{T} \ell_{t,i} \right]
\end{aligned}$$

Lower bound of PEA

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

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

Proof. $$\max_{\boldsymbol{\ell}_1,\ldots,\boldsymbol{\ell}_T} \mathrm{Regret}_T \geq \sum_{t=1}^{T} \mathbb{E}_{\boldsymbol{\ell}_1,\ldots,\boldsymbol{\ell}_{t-1}} \left\langle \boldsymbol{p}_t, \mathbb{E}_{\boldsymbol{\ell}_t} \left[\boldsymbol{\ell}_t \mid \boldsymbol{\ell}_{t-1}, \ldots, \boldsymbol{\ell}_1 \right] \right\rangle - \mathbb{E}_{\boldsymbol{\ell}_1,\ldots,\boldsymbol{\ell}_T} \left[\min_{i\in[N]} \sum_{t=1}^{T} \ell_{t,i} \right]$$

$$= T/2 - \mathbb{E}_{\boldsymbol{\ell}_1,\ldots,\boldsymbol{\ell}_T} \left[\min_{i\in[N]} \sum_{t=1}^{T} \ell_{t,i} \right] = \mathbb{E}_{\boldsymbol{\ell}_1,\ldots,\boldsymbol{\ell}_T} \left[\max_{i\in[N]} \sum_{t=1}^{T} \left(\frac{1}{2} - \ell_{t,i} \right) \right]$$

$$= \frac{1}{2} \mathbb{E}_{\sigma_1,\ldots,\sigma_T} \left[\max_{i\in[N]} \sum_{t=1}^{T} \sigma_{t,i} \right],$$

($\boldsymbol{\ell}_t \overset{\text{i.i.d.}}{\sim} \mathcal{D}$ with $\mathcal{D}$ be the uniform distribution over $\{0,1\}$)

(σ_t for $i \in [N], t \in [T]$ are i.i.d. Rademacher random variables)

Lower bound of PEA

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

$$\sup_{T,N} \max_{\ell_1,\ldots,\ell_T} \frac{\mathrm{Regret}_T}{\sqrt{T \ln N}} \geq \frac{1}{\sqrt{2}}.$$

Proof.

$$\max_{\ell_1,\ldots,\ell_T} \mathrm{Regret}_T \geq \frac{1}{2} \mathbb{E}_{\sigma_1,\ldots,\sigma_T} \left[\max_{i \in [N]} \sum_{t=1}^{T} \sigma_{t,i} \right]$$

($\sigma_{t,i}$ for $i \in [N], t \in [T]$ are i.i.d. Rademacher random variables)

Using the result from probability theory (*Prediction, Learning, and Games*, Chapter 3.7) of Rademacher variables,

$$\Rightarrow \quad \lim_{T \to \infty} \lim_{N \to \infty} \frac{\mathbb{E}_{\sigma_1,\ldots,\sigma_T} \left[\max_{i \in [N]} \sum_{t=1}^{T} \sigma_{t,i} \right]}{\sqrt{T \ln N}} = \sqrt{2}. \qquad \square$$

Upper Bound and Lower Bound

Theorem 1. *Suppose that* $\forall t \in [T]$ *and* $i \in [N], 0 \leq \boldsymbol{\ell}_{t,i} \leq 1$*, then Hedge with learning rate* η *guarantees*

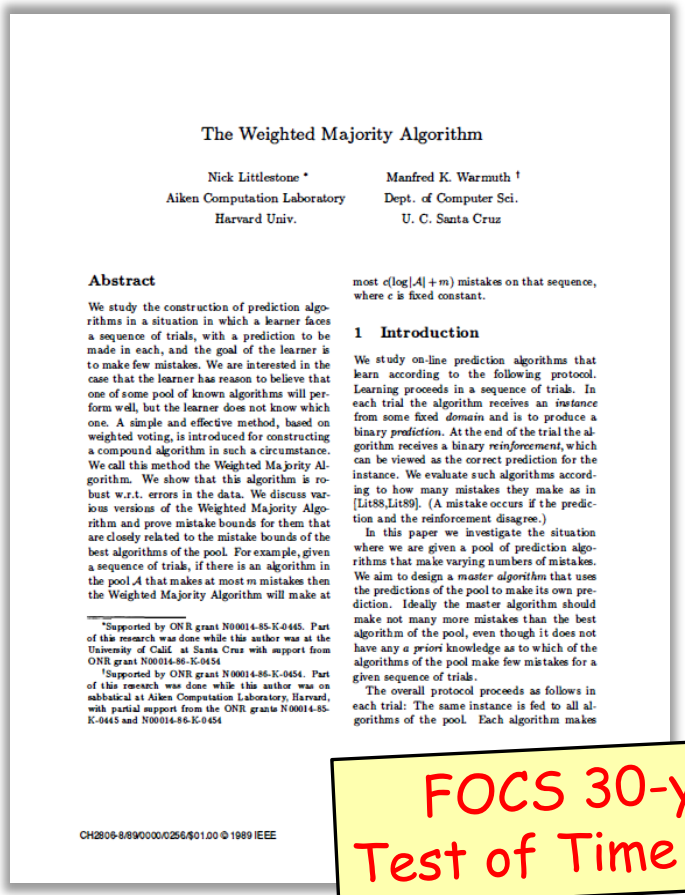
$$\mathrm{Regret}_T \leq \frac{\ln N}{\eta} + \eta T = \mathcal{O}\big(\sqrt{T \log N}\big),$$

where the last equality is by setting η *optimally as* $\sqrt{(\ln N)/T}$*.*

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

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

Prediction with Expert Advice: history bits



The Weighted Majority Algorithm

Nick Littlestone *
Aiken Computation Laboratory
Harvard Univ.

Manfred K. Warmuth †
Dept. of Computer Sci.
U. C. Santa Cruz

Abstract

We study the construction of prediction algorithms in a situation in which a learner faces a sequence of trials, with a prediction to be made in each, and the goal of the learner is to make few mistakes. We are interested in the case that the learner has reason to believe that one of some pool of known algorithms will perform well, but the learner does not know which one. A simple and effective method, based on weighted voting, is introduced for constructing a compound algorithm in such a circumstance. We call this method the Weighted Majority Algorithm. We show that this algorithm is robust w.r.t. errors in the data. We discuss various versions of the Weighted Majority Algorithm and prove mistake bounds for them that are closely related to the mistake bounds of the best algorithms of the pool. For example, given a sequence of trials, if there is an algorithm in the pool $\mathcal{A}$ that makes at most m mistakes then the Weighted Majority Algorithm will make at most $c(\log|\mathcal{A}|+m)$ mistakes on that sequence, where c is fixed constant.

*Supported by ONR grant N00014-85-K-0445. Part of this research was done while this author was at the University of Calif. at Santa Cruz with support from ONR grant N00014-86-K-0454
†Supported by ONR grant N00014-86-K-0454. Part of this research was done while this author was on sabbatical at Aiken Computation Laboratory, Harvard, with partial support from the ONR grants N00014-85-K-0445 and N00014-86-K-0454

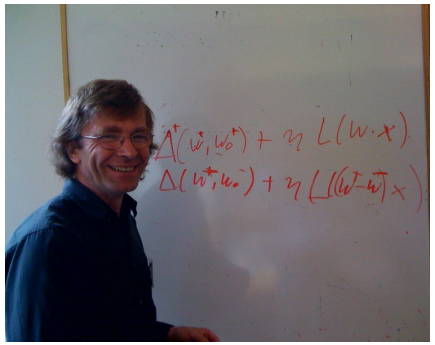
1 Introduction

We study on-line prediction algorithms that learn according to the following protocol. Learning proceeds in a sequence of trials. In each trial the algorithm receives an *instance* from some fixed *domain* and is to produce a binary *prediction*. At the end of the trial the algorithm receives a binary *reinforcement*, which can be viewed as the correct prediction for the instance. We evaluate such algorithms according to how many mistakes they make as in [Lit88,Lit89]. (A mistake occurs if the prediction and the reinforcement disagree.)

In this paper we investigate the situation where we are given a pool of prediction algorithms that make varying numbers of mistakes. We aim to design a *master algorithm* that uses the predictions of the pool to make its own prediction. Ideally the master algorithm should make not many more mistakes than the best algorithm of the pool, even though it does not have any *a priori* knowledge as to which of the algorithms of the pool make few mistakes for a given sequence of trials.

The overall protocol proceeds as follows in each trial: The same instance is fed to all algorithms of the pool. Each algorithm makes

CH2808-8/89/0000/0256/$01.00 © 1989 IEEE

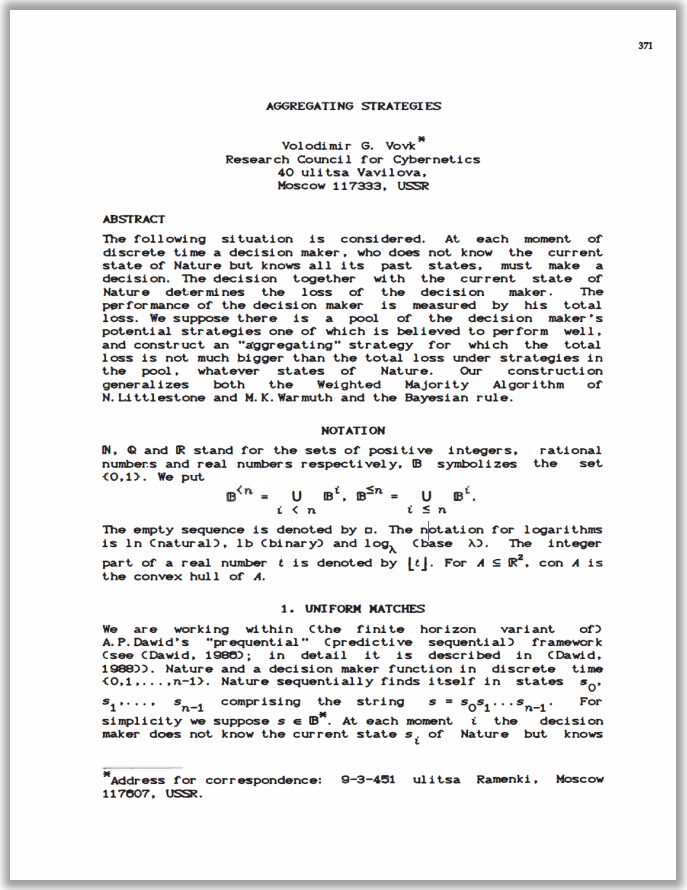


Manfred Warmuth

UC Santa Cruz

FOCS 30-year Test of Time Award!

Nick Littlestone and Manfred K. Warmuth.
"The Weighted Majority Algorithm." FOCS 1989: 256-261.



AGGREGATING STRATEGIES

Volodimir G. Vovk*
Research Council for Cybernetics
40 ulitsa Vavilova,
Moscow 117333, USSR

ABSTRACT

The following situation is considered. At each moment of discrete time a decision maker, who does not know the current state of Nature but knows all its past states, must make a decision. The decision together with the current state of Nature determines the loss of the decision maker. The performance of the decision maker is measured by his total loss. We suppose there is a pool of the decision maker's potential strategies one of which is believed to perform well, and construct an "aggregating" strategy for which the total loss is not much bigger than the total loss under strategies in the pool, whatever states of Nature. Our construction generalizes both the Weighted Majority Algorithm of N. Littlestone and M. K. Warmuth and the Bayesian rule.

NOTATION

$\mathbb{N}$, $\mathbb{Q}$ and $\mathbb{R}$ stand for the sets of positive integers, rational numbers and real numbers respectively, $\mathbb{B}$ symbolizes the set $\langle 0,1\rangle$. We put

$$\mathbb{B}^{<n} = \bigcup_{i<n} \mathbb{B}^i, \quad \mathbb{B}^{\le n} = \bigcup_{i\le n} \mathbb{B}^i.$$

The empty sequence is denoted by $\square$. The notation for logarithms is ln (natural), lb (binary) and $\log_\lambda$ (base λ). The integer part of a real number t is denoted by $\lfloor t \rfloor$. For $A \subseteq \mathbb{R}^2$, con A is the convex hull of A.

1. UNIFORM MATCHES

We are working within (the finite horizon variant of) A. P. Dawid's "prequential" (predictive sequential) framework (see (Dawid, 1985); in detail it is described in (Dawid, 1988)). Nature and a decision maker function in discrete time $\langle 0,1,\ldots,n-1\rangle$. Nature sequentially finds itself in states $s_0, s_1, \ldots, s_{n-1}$ comprising the string $s = s_0 s_1 \ldots s_{n-1}$. For simplicity we suppose $s \in \mathbb{B}^n$. At each moment i the decision maker does not know the current state s_i of Nature but knows

*Address for correspondence: 9-3-451 ulitsa Ramenki, Moscow 117607, USSR.

Volodimir G. Vovk. "Aggregating Strategies."
COLT 1990: 371-383.

Volodimir G. Vovk
Royal Holloway,
University of London

Prediction with Expert Advice: history bits

Yoav Freund

Robert Schapire

Goldel Prize 2003

This paper introduced AdaBoost, an adaptive algorithm to improve the accuracy of hypotheses in machine learning. The algorithm demonstrated novel possibilities in analyzing data and is a permanent contribution to science even beyond computer science.

JOURNAL OF COMPUTER AND SYSTEM SCIENCES **55**, 119–139 (1997)
ARTICLE NO. SS971504

A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting*

Yoav Freund and Robert E. Schapire†

AT&T Labs, 180 Park Avenue, Florham Park, New Jersey 07932

Received December 19, 1996

In the first part of the paper we consider the problem of dynamically apportioning resources among a set of options in a worst-case on-line framework. The model we study can be interpreted as a broad, abstract extension of the well-studied on-line prediction model to a general decision-theoretic setting. We show that the multiplicative weight-update Littlestone–Warmuth rule can be adapted to this model, yielding bounds that are slightly weaker in some cases, but applicable to a considerably more general class of learning problems. We show how the resulting learning algorithm can be applied to a variety of problems, including gambling, multiple-outcome prediction, repeated games, and prediction of points in $\mathbb{R}^n$. In the second part of the paper we apply the multiplicative weight-update technique to derive a new boosting algorithm. This boosting algorithm does not require any prior knowledge about the performance of the weak learning algorithm. We also study generalizations of the new boosting algorithm to the problem of learning functions whose range, rather than being binary, is an arbitrary finite set or a bounded segment of the real line.

converting a "weak" PAC learning algorithm that performs just slightly better than random guessing into one with arbitrarily high accuracy.

We formalize our *on-line allocation model* as follows. The allocation agent A has N options or *strategies* to choose from; we number these using the integers $1, \ldots, N$. At each time step $t = 1, 2, \ldots, T$, the allocator A decides on a distribution $\mathbf{p}^t$ over the strategies; that is $p_i^t \geqslant 0$ is the amount allocated to strategy i, and $\sum_{i=1}^N p_i^t = 1$. Each strategy i then suffers some *loss* ℓ_i^t which is determined by the (possibly adversarial) "environment." The loss suffered by A is then $\sum_{i=1}^n p_i^t \ell_i^t = \mathbf{p}^t \cdot \boldsymbol{\ell}^t$, i.e., the average loss of the strategies with respect to A's chosen allocation rule. We call this loss function the *mixture loss*.

In this paper, we always assume that the loss suffered by any strategy is bounded so that, without loss of generality, $\ell_i^t \in [0, 1]$. Besides this condition, we make no assumptions

Reference: Y. Freund and R. Schapire. A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting. JCSS 1997.

Prediction with Expert Advice: history bits

Yoav Freund

Robert Schapire

Goldel Prize 2003

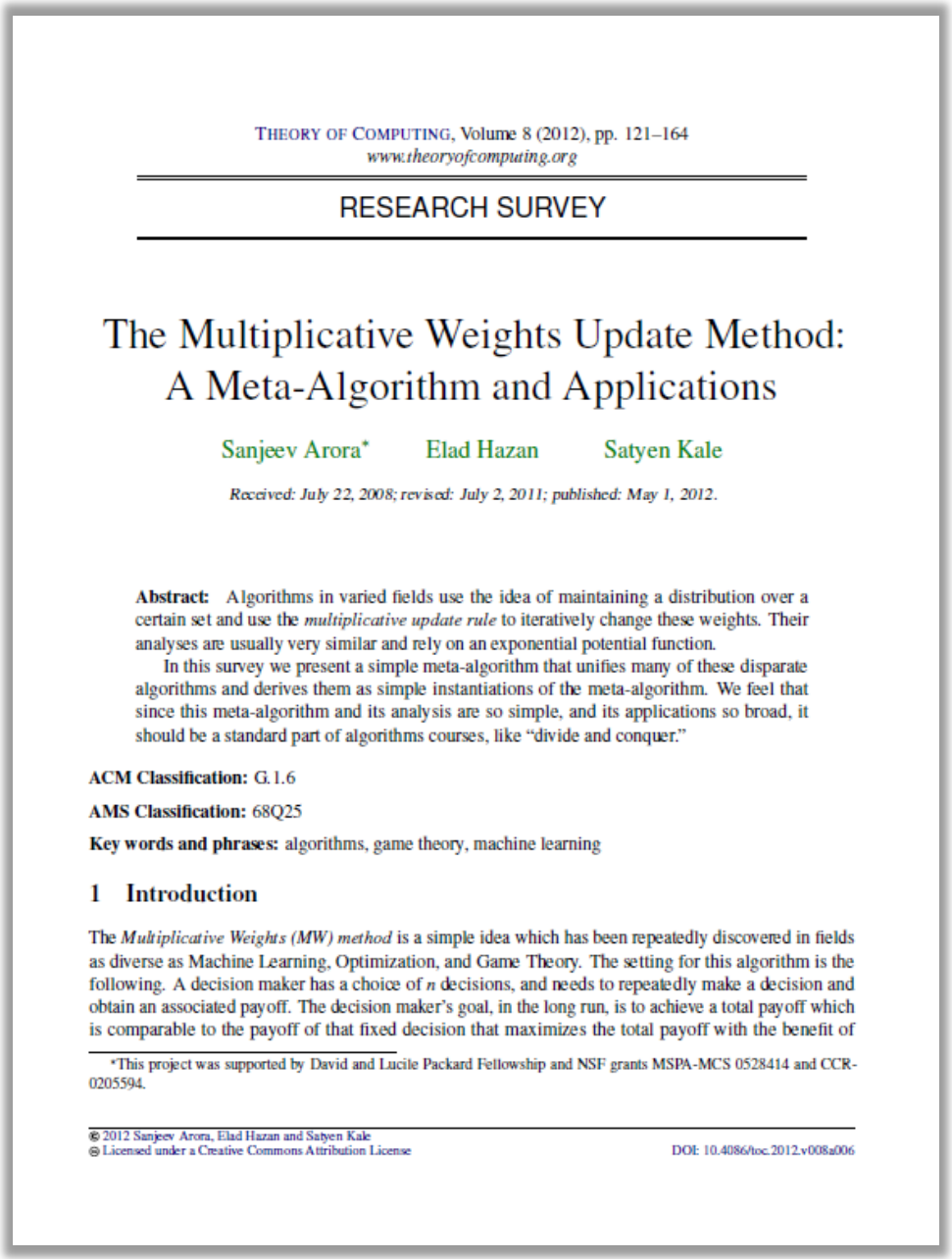
This paper introduced AdaBoost, an adaptive algorithm to improve the accuracy of hypotheses in machine learning. The algorithm demonstrated novel possibilities in analyzing data and is a permanent contribution to science even beyond computer science.

Photo@ICML'24（维也纳，July 22, 2024）

Why is PEA useful?

- Prediction with Expert Advice is essentially a **meta-algorithm** for combining different experts, and the "expert" can be interpreted as any learning model with a particular kind of expertise.

- It is used in a variety of algorithmic design (see HW1 for example).



THEORY OF COMPUTING, Volume 8 (2012), pp. 121–164
www.theoryofcomputing.org

RESEARCH SURVEY

The Multiplicative Weights Update Method: A Meta-Algorithm and Applications

Sanjeev Arora* Elad Hazan Satyen Kale

Received: July 22, 2008; revised: July 2, 2011; published: May 1, 2012.

Abstract: Algorithms in varied fields use the idea of maintaining a distribution over a certain set and use the *multiplicative update rule* to iteratively change these weights. Their analyses are usually very similar and rely on an exponential potential function.

In this survey we present a simple meta-algorithm that unifies many of these disparate algorithms and derives them as simple instantiations of the meta-algorithm. We feel that since this meta-algorithm and its analysis are so simple, and its applications so broad, it should be a standard part of algorithms courses, like "divide and conquer."

ACM Classification: G.1.6

AMS Classification: 68Q25

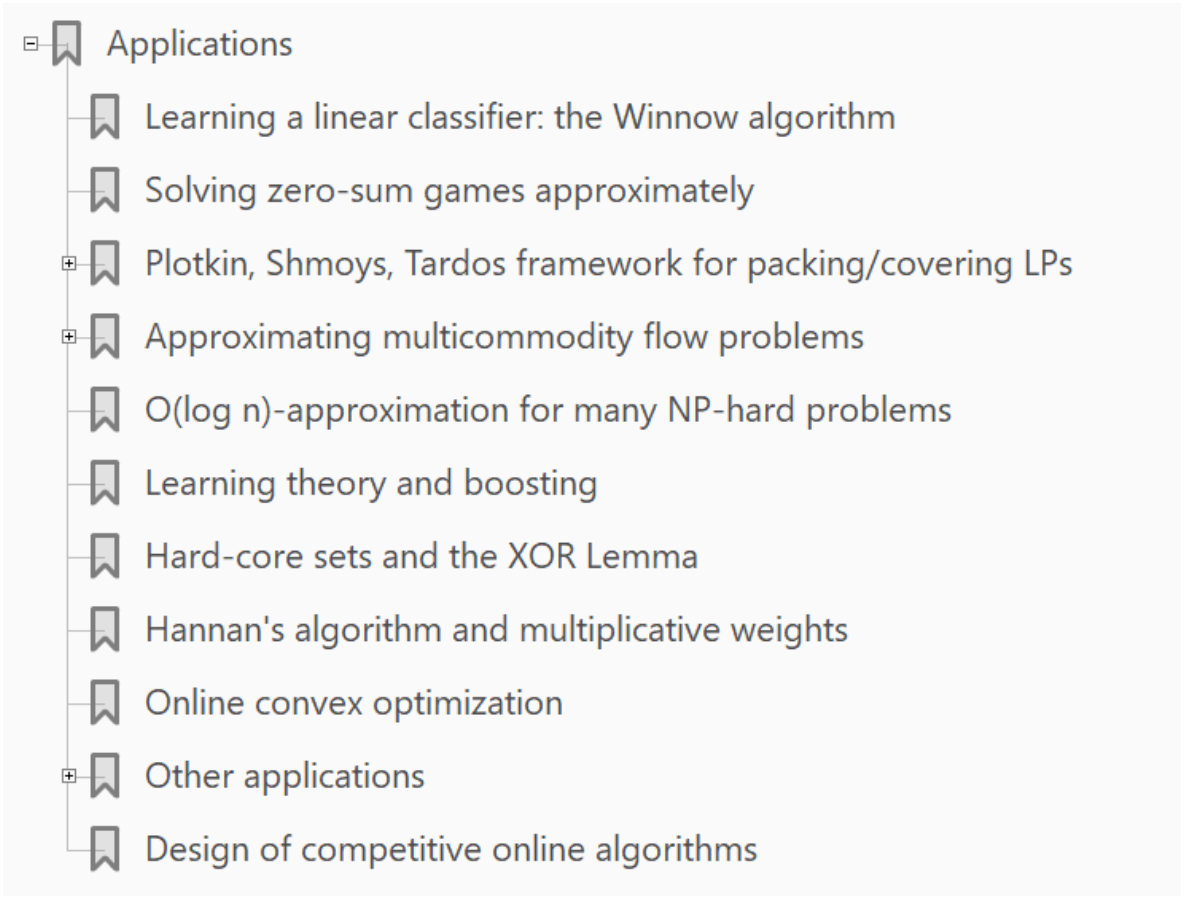
Key words and phrases: algorithms, game theory, machine learning

1 Introduction

The *Multiplicative Weights (MW) method* is a simple idea which has been repeatedly discovered in fields as diverse as Machine Learning, Optimization, and Game Theory. The setting for this algorithm is the following. A decision maker has a choice of n decisions, and needs to repeatedly make a decision and obtain an associated payoff. The decision maker's goal, in the long run, is to achieve a total payoff which is comparable to the payoff of that fixed decision that maximizes the total payoff with the benefit of

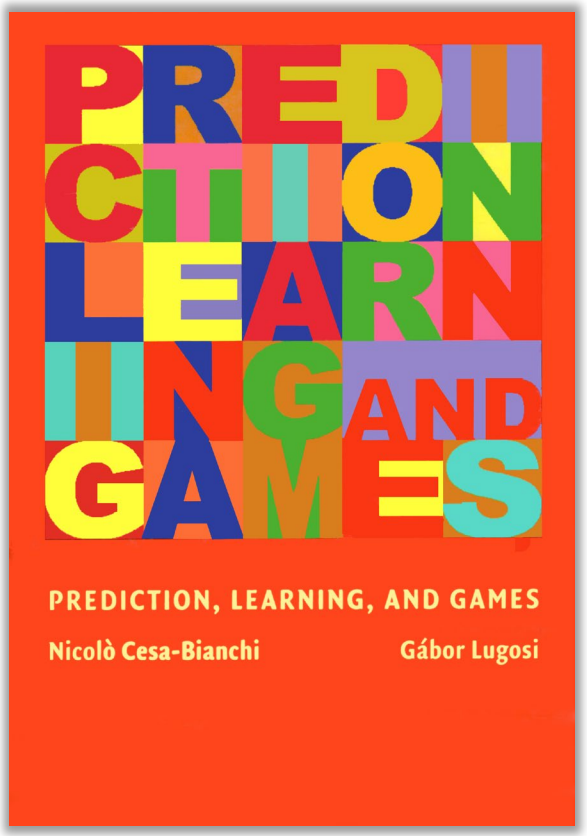
*This project was supported by David and Lucile Packard Fellowship and NSF grants MSPA-MCS 0528414 and CCR-0205594.

DOI: 10.4086/toc.2012.v008a006

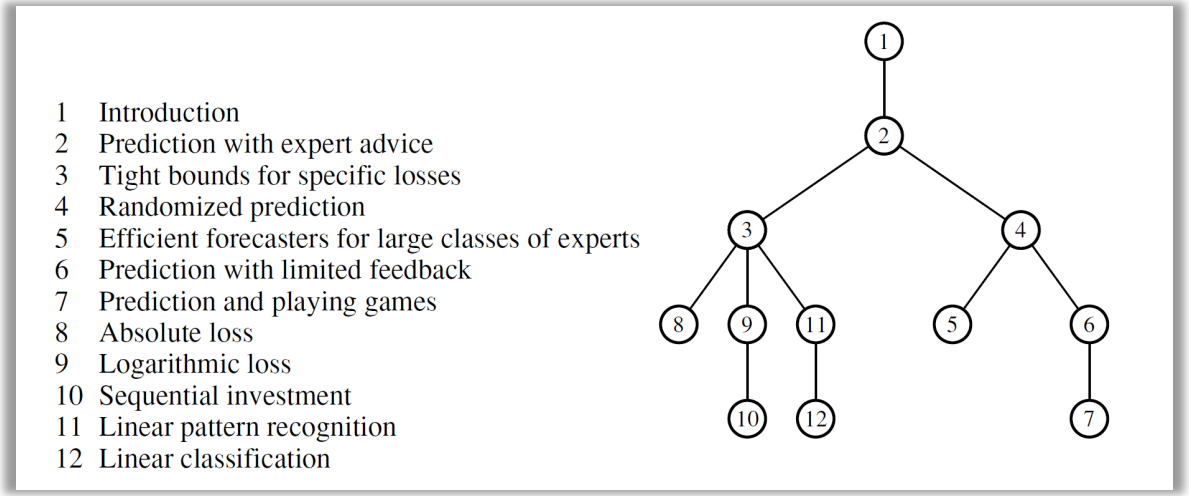


The multiplicative weights update method: a meta-algorithm and applications. S Arora, E Hazan, S Kale. Theory of Computing, 2012

More Results on PEA



Prediction, Learning and Games.
Nicolò Cesa-Bianchi and Gabor Lugosi.
Cambridge University Press, 2006.



Nicolò Cesa-Bianchi

Gabor Lugosi

Part 2. OMD Framework

- Algorithmic Framework
- Regret Analysis
- Interpretation from Primal-Dual View

PEA vs. OCO

At each round $t = 1, 2, \cdots$ **Prediction with Expert Advice**

(1) the player first picks a weight $\boldsymbol{p}_t$ from a simplex Δ_N;

(2) and simultaneously environments pick an loss vector $\boldsymbol{\ell}_t \in \mathbb{R}^N$;

(3) the player suffers loss $f_t(\boldsymbol{p}_t) \triangleq \langle \boldsymbol{p}_t, \boldsymbol{\ell}_t \rangle$, observes $\boldsymbol{\ell}_t$ and updates the model.

require domain to be a simplex $\mathcal{X} = \Delta_N$ ⇧ linear loss $f_t(\mathbf{x}) \triangleq \langle \mathbf{x}, \boldsymbol{\ell}_t \rangle$

PEA is a ***special case*** of OCO!

At each round $t = 1, 2, \cdots$ **Online Convex Optimization**

(1) the player first picks a model $\mathbf{x}_t \in \mathcal{X}$;

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

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

Deploying OGD to PEA

- PEA is a special case of OCO:

  Why not directly deploy OGD (proposed in last lecture) to address PEA?

Theorem 4 (Regret bound for OGD). *Under Assumption 1, 2 and 3, online gradient descent (OGD) with step sizes* $\eta_t = \frac{D}{G\sqrt{t}}$ *for* $t \in [T]$ *guarantees:*

$$\text{Regret}_T = \sum_{t=1}^{T} f_t(\mathbf{x}_t) - \min_{\mathbf{x}\in\mathcal{X}} \sum_{t=1}^{T} f_t(\mathbf{x}) \leq \frac{3}{2} GD\sqrt{T}.$$

Regret guarantee: $D = \max_{\mathbf{x},\mathbf{y}\in\Delta_N} \|\mathbf{x}-\mathbf{y}\|_2 = \sqrt{2}$ $\quad G = \max_{\boldsymbol{\ell}_t\in\mathbb{R}^N} \|\boldsymbol{\ell}_t\|_2 = \sqrt{N}$

$$\Longrightarrow \text{Regret}_T = \sum_{t=1}^{T} \langle \boldsymbol{p}_t, \boldsymbol{\ell}_t \rangle - \min_{\boldsymbol{p}\in\Delta_N} \sum_{t=1}^{T} \langle \boldsymbol{p}, \boldsymbol{\ell}_t \rangle \leq \mathcal{O}(\sqrt{TN})$$

Deploying OGD to PEA

- OGD for PEA Problem:

$$D = \max_{\mathbf{x},\mathbf{y}\in\Delta_N} \|\mathbf{x}-\mathbf{y}\|_2 = \sqrt{2} \qquad G = \max_{\boldsymbol{\ell}_t\in\mathbb{R}^N} \|\boldsymbol{\ell}_t\|_2 = \sqrt{N}$$

$$\Rightarrow \quad \text{Regret}_T = \sum_{t=1}^{T} \langle \boldsymbol{p}_t, \boldsymbol{\ell}_t \rangle - \min_{\boldsymbol{p}\in\Delta_N} \sum_{t=1}^{T} \langle \boldsymbol{p}, \boldsymbol{\ell}_t \rangle \le \mathcal{O}(\sqrt{TN})$$

- A natural question: is the $\mathcal{O}(\sqrt{TN})$ regret bound tight enough?
  - recall that the lower bound of PEA is $\Omega(\sqrt{T\log N})$
  - OGD is not optimal with respect to N (number of experts)

Deploying OGD to PEA

- PEA is a special case of OCO:

  Why not directly deploy OGD (proposed in last lecture) to address PEA?

Theorem 4 (Regret bound for OGD). *Under Assumption 1, 2 and 3, online gradient descent (OGD) with step sizes* $\eta_t = \frac{D}{G\sqrt{t}}$ *for* $t \in [T]$ *guarantees:*

$$\text{Regret}_T = \sum_{t=1}^{T} f_t(\mathbf{x}_t) - \min_{\mathbf{x}\in\mathcal{X}} \sum_{t=1}^{T} f_t(\mathbf{x}) \leq \frac{3}{2} GD\sqrt{T}.$$

Regret guarantee: $D = \max_{\mathbf{x},\mathbf{y}\in\Delta_N} \|\mathbf{x}-\mathbf{y}\|_2 = \sqrt{2}$ $\qquad$ $G = \max_{\boldsymbol{\ell}_t\in\mathbb{R}^N} \|\boldsymbol{\ell}_t\|_2 = \sqrt{N}$

$$\Longrightarrow \quad \text{Regret}_T = \sum_{t=1}^{T} \langle \boldsymbol{p}_t, \boldsymbol{\ell}_t \rangle - \min_{\boldsymbol{p}\in\Delta_N} \sum_{t=1}^{T} \langle \boldsymbol{p}, \boldsymbol{\ell}_t \rangle \leq \mathcal{O}(\sqrt{TN})$$

Why OGD Fails for PEA

- PEA has a special structure whereas general OCO doesn't have.

Convex Problem	**PEA Problem**
Domain: convex set $\mathcal{X}$	Domain: simplex $\mathcal{X} = \Delta_N$
Online function: convex function f_t	Online function: linear $f_t(\boldsymbol{p}) \triangleq \langle \boldsymbol{p}, \boldsymbol{\ell}_t \rangle$
Lower Bound: $\Omega(GD\sqrt{T})$	Lower Bound: $\Omega(\sqrt{T \log N})$

Why OGD Fails for PEA

- Remember that for the general OCO, we linearized the function to analyze the first gradient descent lemma:

$$
\begin{aligned}
\|\mathbf{x}_{t+1} - \mathbf{x}^\star\|^2 &= \|\Pi_{\mathcal{X}}[\mathbf{x}_t - \eta_t \nabla f(\mathbf{x}_t)] - \mathbf{x}^\star\|^2 \quad \text{(GD)} \\
&\leq \|\mathbf{x}_t - \eta_t \nabla f(\mathbf{x}_t) - \mathbf{x}^\star\|^2 \quad \text{(Pythagoras Theorem)} \\
&= \|\mathbf{x}_t - \mathbf{x}^\star\|^2 - 2\eta_t \langle \nabla f(\mathbf{x}_t), \mathbf{x}_t - \mathbf{x}^\star \rangle + \eta_t^2 \|\nabla f(\mathbf{x}_t)\|^2 \\
&\leq \|\mathbf{x}_t - \mathbf{x}^\star\|^2 - 2\eta_t (f(\mathbf{x}_t) - f^\star) + \eta_t^2 \|\nabla f(\mathbf{x}_t)\|^2 \\
&\quad \text{(convexity: } f(\mathbf{x}_t) - f^\star = f(\mathbf{x}_t) - f(\mathbf{x}^\star) \leq \langle \nabla f(\mathbf{x}_t), \mathbf{x}_t - \mathbf{x}^\star \rangle\text{)}
\end{aligned}
$$

- So, linearized loss is not the essence, but the ***simplex domain*** of the PEA problem is worthy specifically considering.

Why OGD Fails for PEA?

- Recall that for general OCO, we update the model as follows:

General Online Convex Optimization

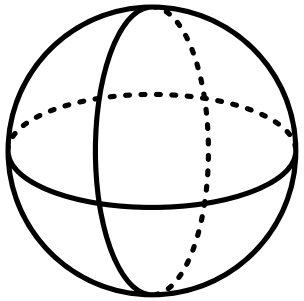
OGD:

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

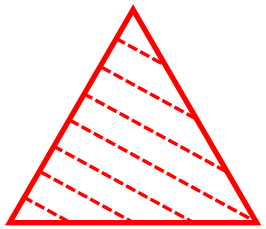
Proximal type update:

$$\mathbf{x}_{t+1} = \arg\min_{\mathbf{x} \in \mathcal{X}} \left\{ \langle \mathbf{x}, \eta_t \nabla f_t(\mathbf{x}_t) \rangle + \frac{1}{2} \|\mathbf{x} - \mathbf{x}_t\|_2^2 \right\}$$

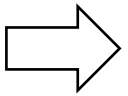
- In PEA, is it proper to use 2-norm (ball) to measure distance?

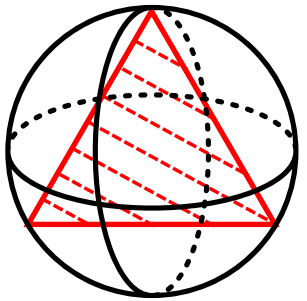


Ball



Simplex





A ball is too pessimistic (loose) to measure a simplex!

Proximal View

- Recall that for general OCO, we update the model as follows:

General Online Convex Optimization

OGD:

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

Proximal type update:

$$\mathbf{x}_{t+1} = \arg\min_{\mathbf{x} \in \mathcal{X}} \left\{ \langle \mathbf{x}, \eta_t \nabla f_t(\mathbf{x}_t) \rangle + \frac{1}{2} \|\mathbf{x} - \mathbf{x}_t\|_2^2 \right\}$$

- In PEA, is it proper to use 2-norm (ball) to measure distance?

⇨ We need to find an alternative distance measure for the ***special structure*** in PEA.

Proximal View

⇨ We need to find an alternative distance measure for the ***special structure*** in PEA.

- Intuitively, for Euclidean space, 2-norm is the most natural measure:

$$\|\mathbf{x} - \mathbf{y}\|_2^2$$

- For PEA problem
  - the decision can be viewed as a distribution within the simplex
  - for two distributions P and Q, KL divergence is a natural measure:

$$\mathrm{KL}(P\|Q) = \sum_{x \in \mathcal{X}} P(x) \log \left(\frac{P(x)}{Q(x)} \right)$$

Reinvent Hedge Algorithm

Theorem 3. *Consider* $f_t(\boldsymbol{p}) = \langle \boldsymbol{p}, \boldsymbol{\ell}_t \rangle$. *An online learning algorithm that updates the model following*

$$\boldsymbol{p}_{t+1} = \arg\min_{\boldsymbol{p} \in \Delta_N} \left\{ \eta \langle \boldsymbol{p}, \nabla f_t(\boldsymbol{p}_t) \rangle + \mathrm{KL}(\boldsymbol{p} \| \boldsymbol{p}_t) \right\}$$

is equal to Hedge update, i.e.,

$$p_{t+1,i} \propto p_{t,i} \exp\left(-\eta \ell_{t,i}\right) \text{ for all } i \in [N].$$

Proof.

$$\begin{aligned} \boldsymbol{p}_{t+1} &= \arg\min_{\boldsymbol{p} \in \Delta_N} \eta \langle \boldsymbol{p}, \nabla f_t(\boldsymbol{p}_t) \rangle + \mathrm{KL}(\boldsymbol{p} \| \boldsymbol{p}_t) \\ &= \arg\min_{\boldsymbol{p} \in \Delta_N} \underbrace{\eta \langle \boldsymbol{p}, \nabla f_t(\boldsymbol{p}_t) \rangle - \sum_{i=1}^{N} p_i \ln\left(\frac{p_{t,i}}{p_i}\right)}_{F(\boldsymbol{p})} \qquad \text{(definition of KL divergence)} \end{aligned}$$

Reinvent Hedge Algorithm

- Proximal update rule for OGD:

$$\mathbf{x}_{t+1} = \arg\min_{\mathbf{x}\in\mathcal{X}} \left\{ \eta_t \langle \mathbf{x}, \nabla f_t(\mathbf{x}_t) \rangle + \frac{1}{2} \|\mathbf{x} - \mathbf{x}_t\|_2^2 \right\}$$

- Proximal update rule for Hedge:

$$\mathbf{x}_{t+1} = \arg\min_{\mathbf{x}\in\mathcal{X}} \left\{ \eta_t \langle \mathbf{x}, \nabla f_t(\mathbf{x}_t) \rangle + \mathrm{KL}(\mathbf{x} \| \mathbf{x}_t) \right\}$$

- More possibility: changing the distance measure to a more general form using ***Bregman divergence***

$$\mathbf{x}_{t+1} = \arg\min_{\mathbf{x}\in\mathcal{X}} \left\{ \eta_t \langle \mathbf{x}, \nabla f_t(\mathbf{x}_t) \rangle + \mathcal{D}_\psi(\mathbf{x}, \mathbf{x}_t) \right\}$$

Bregman Divergence

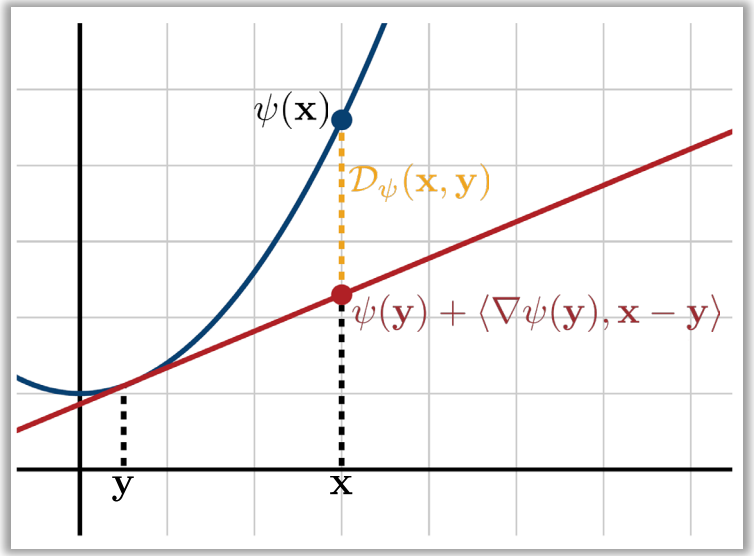
Definition 1 (Bregman Divergence). Let ψ be a strongly convex and differentiable function over a convex set $\mathcal{X}$, then for any $\mathbf{x}, \mathbf{y} \in \mathcal{X}$, the bregman divergence $\mathcal{D}_\psi$ associated to ψ is defined as

$$\mathcal{D}_\psi(\mathbf{x}, \mathbf{y}) = \psi(\mathbf{x}) - \psi(\mathbf{y}) - \langle \nabla\psi(\mathbf{y}), \mathbf{x} - \mathbf{y} \rangle.$$

- Bregman divergence measures the difference of a function and its linear approximation.
- Using second-order Taylor expansion, we know

$$\mathcal{D}_\psi(\mathbf{x}, \mathbf{y}) = \frac{1}{2}\|\mathbf{x} - \mathbf{y}\|^2_{\nabla^2\psi(\boldsymbol{\xi})}$$

for some $\boldsymbol{\xi} \in [\mathbf{x}, \mathbf{y}]$.



Bregman Divergence

Definition 1 (Bregman Divergence). Let ψ be a strongly convex and differentiable function over a convex set $\mathcal{X}$, then for any $\mathbf{x}, \mathbf{y} \in \mathcal{X}$, the bregman divergence $\mathcal{D}_\psi$ associated to ψ is defined as

$$\mathcal{D}_\psi(\mathbf{x}, \mathbf{y}) = \psi(\mathbf{x}) - \psi(\mathbf{y}) - \langle \nabla\psi(\mathbf{y}), \mathbf{x} - \mathbf{y} \rangle.$$

Table 1: Choice of $\psi(\cdot)$ and the corresponding Bregman divergence.

	$\psi(\mathbf{x})$	$\mathcal{D}_\psi(\mathbf{x}, \mathbf{y})$
Squared L_2-distance	$\|\mathbf{x}\|_2^2$	$\|\mathbf{x} - \mathbf{y}\|_2^2$
Mahalanobis distance	$\|\mathbf{x}\|_A^2$	$\|\mathbf{x} - \mathbf{y}\|_A^2$
Negative entropy	$\sum_i x_i \log x_i$	$\mathrm{KL}(\mathbf{x}\|\mathbf{y})$

Online Mirror Descent

Online Mirror Descent

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

$$\mathbf{x}_{t+1} = \arg\min_{\mathbf{x} \in \mathcal{X}} \left\{ \eta_t \langle \mathbf{x}, \nabla f_t(\mathbf{x}_t) \rangle + \mathcal{D}_\psi(\mathbf{x}, \mathbf{x}_t) \right\}$$

where $\mathcal{D}_\psi(\mathbf{x}, \mathbf{y}) = \psi(\mathbf{x}) - \psi(\mathbf{y}) - \langle \nabla \psi(\mathbf{y}), \mathbf{x} - \mathbf{y} \rangle$ is the Bregman divergence.

- $\psi(\cdot)$ is a required to be strongly convex and differentiable over a convex set $\mathcal{X}$.
- Strong convexity of ψ will ensure the uniqueness of the minimization problem, and actually we further need some analytical assumptions (see later mirror map defintion) to ensure the solutions' feasibility in domain $\mathcal{X}$.

Online Mirror Descent

- So we can unify OGD and Hedge from the same view of OMD.

$$\mathbf{x}_{t+1} = \arg\min_{\mathbf{x}\in\mathcal{X}} \left\{ \eta_t \langle \mathbf{x}, \nabla f_t(\mathbf{x}_t) \rangle + \mathcal{D}_\psi(\mathbf{x}, \mathbf{x}_t) \right\}$$

Algo.	OMD/proximal form	$\psi(\cdot)$	η_t	Regret_T
OGD	$\mathbf{x}_{t+1} = \arg\min_{\mathbf{x}\in\mathcal{X}} \left\{ \eta_t \langle \mathbf{x}, \nabla f_t(\mathbf{x}_t) \rangle + \frac{1}{2} \|\mathbf{x} - \mathbf{x}_t\|_2^2 \right\}$	$\|\mathbf{x}\|_2^2$	$\frac{1}{\sqrt{t}}$	$\mathcal{O}(\sqrt{T})$
Hedge	$\mathbf{x}_{t+1} = \arg\min_{\mathbf{x}\in\Delta_N} \left\{ \eta_t \langle \mathbf{x}, \nabla f_t(\mathbf{x}_t) \rangle + \text{KL}(\mathbf{x} \| \mathbf{x}_t) \right\}$	$\sum_{i=1}^N x_i \log x_i$	$\sqrt{\frac{\ln N}{T}}$	$\mathcal{O}(\sqrt{T \log N})$

- We also learn ONS for exp-concave functions, can it be included?

Recap: ONS in a view of Proximal Gradient

Convex Problem

Property: $f_t(\mathbf{x}) \geq f_t(\mathbf{y}) + \nabla f_t(\mathbf{y})^\top (\mathbf{x} - \mathbf{y})$

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

Proximal type update:

$$\mathbf{x}_{t+1} = \arg\min_{\mathbf{x} \in \mathcal{X}} \langle \mathbf{x}, \nabla f_t(\mathbf{x}_t) \rangle + \frac{1}{2\eta_t} \|\mathbf{x} - \mathbf{x}_t\|_2^2$$

Exp-concave Problem

Property: $f_t(\mathbf{x}) \geq f_t(\mathbf{y}) + \nabla f_t(\mathbf{y})^\top (\mathbf{x} - \mathbf{y}) + \frac{\gamma}{2} \|\mathbf{x} - \mathbf{y}\|^2_{\nabla f_t(\mathbf{y}) \nabla f_t(\mathbf{y})^\top}$

ONS: $A_t = A_{t-1} + \nabla f_t(\mathbf{x}_t) \nabla f_t(\mathbf{x}_t)^\top$

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

Proximal type update:

$$\mathbf{x}_{t+1} = \arg\min_{\mathbf{x} \in \mathcal{X}} \langle \mathbf{x}, \nabla f_t(\mathbf{x}_t) \rangle + \frac{\gamma}{2} \|\mathbf{x} - \mathbf{x}_t\|^2_{A_t}$$

Online Mirror Descent

- Our previous mentioned algorithms can all be covered by OMD.

Algo.	OMD/proximal form	$\psi(\cdot)$	η_t	Regret_T
OGD for convex	$\mathbf{x}_{t+1} = \arg\min_{\mathbf{x}\in\mathcal{X}} \eta_t \langle \mathbf{x}, \nabla f_t(\mathbf{x}_t)\rangle + \frac{1}{2}\|\mathbf{x}-\mathbf{x}_t\|_2^2$	$\|\mathbf{x}\|_2^2$	$\frac{1}{\sqrt{t}}$	$\mathcal{O}(\sqrt{T})$
OGD for strongly c.	$\mathbf{x}_{t+1} = \arg\min_{\mathbf{x}\in\mathcal{X}} \eta_t \langle \mathbf{x}, \nabla f_t(\mathbf{x}_t)\rangle + \frac{1}{2}\|\mathbf{x}-\mathbf{x}_t\|_2^2$	$\|\mathbf{x}\|_2^2$	$\frac{1}{\sigma t}$	$\mathcal{O}(\frac{1}{\sigma}\log T)$
ONS for exp-concave	$\mathbf{x}_{t+1} = \arg\min_{\mathbf{x}\in\mathcal{X}} \eta_t \langle \mathbf{x}, \nabla f_t(\mathbf{x}_t)\rangle + \frac{1}{2}\|\mathbf{x}-\mathbf{x}_t\|_{A_t}^2$	$\|\mathbf{x}\|_{A_t}^2$	$\frac{1}{\gamma}$	$\mathcal{O}(\frac{d}{\gamma}\log T)$
Hedge for PEA	$\mathbf{x}_{t+1} = \arg\min_{\mathbf{x}\in\Delta_N} \eta_t \langle \mathbf{x}, \nabla f_t(\mathbf{x}_t)\rangle + \text{KL}(\mathbf{x}\|\mathbf{x}_t)$	$\sum_{i=1}^N x_i \log x_i$	$\sqrt{\frac{\ln N}{T}}$	$\mathcal{O}(\sqrt{T\log N})$

General Regret Analysis for OMD

OMD update:

$$\mathbf{x}_{t+1} = \arg\min_{\mathbf{x}\in\mathcal{X}} \left\{ \eta_t \langle \mathbf{x}, \nabla f_t(\mathbf{x}_t) \rangle + \mathcal{D}_\psi(\mathbf{x}, \mathbf{x}_t) \right\}$$

Theorem 4 (General Regret Bound for OMD). *Assume ψ is λ-strongly convex w.r.t. $\|\cdot\|$ and $\eta_t = \eta, \forall t \in [T]$. Then, for all $\mathbf{u} \in \mathcal{X}$, the following regret bound holds*

$$\sum_{t=1}^{T} f_t(\mathbf{x}_t) - \sum_{t=1}^{T} f_t(\mathbf{u}) \leq \frac{\mathcal{D}_\psi(\mathbf{u}, \mathbf{x}_1)}{\eta} + \frac{\eta}{\lambda} \sum_{t=1}^{T} \|\nabla f_t(\mathbf{x}_t)\|_\star^2 - \frac{1}{\eta} \sum_{t=1}^{T} \mathcal{D}_\psi(\mathbf{x}_{t+1}, \mathbf{x}_t)$$

General Regret Analysis for OMD

Lemma 1 (Mirror Descent Lemma)**.** *Let $\mathcal{D}_\psi$ be the Bregman divergence w.r.t. $\psi : \mathcal{X} \to \mathbb{R}$ and assume ψ to be λ-strongly convex with respect to a norm $\|\cdot\|$. Then, $\forall \mathbf{u} \in \mathcal{X}$, the following inequality holds*

$$f_t(\mathbf{x}_t) - f_t(\mathbf{u}) \leq \frac{1}{\eta_t}\left(\mathcal{D}_\psi(\mathbf{u}, \mathbf{x}_t) - \mathcal{D}_\psi(\mathbf{u}, \mathbf{x}_{t+1})\right) + \frac{\eta_t}{\lambda}\|\nabla f_t(\mathbf{x}_t)\|_\star^2 - \frac{1}{\eta_t}\mathcal{D}_\psi(\mathbf{x}_{t+1}, \mathbf{x}_t)$$

Proof of Mirror Descent Lemma

Lemma 1 (Mirror Descent Lemma). *Let $\mathcal{D}_\psi$ be the Bregman divergence w.r.t. $\psi : \mathcal{X} \to \mathbb{R}$ and assume ψ to be λ-strongly convex with respect to a norm $\|\cdot\|$. Then, $\forall \mathbf{u} \in \mathcal{X}$, the following inequality holds*

$$f_t(\mathbf{x}_t) - f_t(\mathbf{u}) \le \frac{1}{\eta_t}\left(\mathcal{D}_\psi(\mathbf{u}, \mathbf{x}_t) - \mathcal{D}_\psi(\mathbf{u}, \mathbf{x}_{t+1})\right) + \frac{\eta_t}{\lambda}\|\nabla f_t(\mathbf{x}_t)\|_\star^2 - \frac{1}{\eta_t}\mathcal{D}_\psi(\mathbf{x}_{t+1}, \mathbf{x}_t)$$

bias term (range term) *variance term (stability term)* *negative term*

Proof. $f_t(\mathbf{x}_t) - f_t(\mathbf{u}) \le \langle \nabla f_t(\mathbf{x}_t), \mathbf{x}_t - \mathbf{u} \rangle$

$$\le \underbrace{\langle \nabla f_t(\mathbf{x}_t), \mathbf{x}_t - \mathbf{x}_{t+1} \rangle}_{\text{term (a)}} + \underbrace{\langle \nabla f_t(\mathbf{x}_t), \mathbf{x}_{t+1} - \mathbf{u} \rangle}_{\text{term (b)}}$$

We use ***stability lemma*** to analyze term (a), and use ***Bregman proximal inequality*** to analyze term (b).

Proof of Mirror Descent Lemma

Proof. $$f_t(\mathbf{x}_t) - f_t(\mathbf{u}) \leq \underbrace{\langle \nabla f_t(\mathbf{x}_t), \mathbf{x}_t - \mathbf{x}_{t+1} \rangle}_{\text{term (a)}} + \underbrace{\langle \nabla f_t(\mathbf{x}_t), \mathbf{x}_{t+1} - \mathbf{u} \rangle}_{\text{term (b)}}$$

We introduce the following stability lemma to analyze term (a):

Lemma 2 (Stability Lemma). *Consider the following updates:*

$$\begin{cases} \mathbf{x} = \arg\min_{\mathbf{x} \in \mathcal{X}} \langle \mathbf{g}, \mathbf{x} \rangle + \mathcal{D}_\psi(\mathbf{x}, \mathbf{c}) \\ \mathbf{x}' = \arg\min_{\mathbf{x} \in \mathcal{X}} \langle \mathbf{g}', \mathbf{x} \rangle + \mathcal{D}_\psi(\mathbf{x}, \mathbf{c}) \end{cases}$$

When the regularizer $\psi : \mathcal{X} \mapsto \mathbb{R}$ is a λ-strongly convex function with respect to norm $\|\cdot\|$, we have

$$\lambda \|\mathbf{x} - \mathbf{x}'\| \leq \|\mathbf{g} - \mathbf{g}'\|_\star .$$

Stability Lemma

Lemma 2 (Stability Lemma). *Consider the following updates:*

$$\begin{cases} \mathbf{x} = \arg\min_{\mathbf{x}\in\mathcal{X}} \langle \mathbf{g}, \mathbf{x}\rangle + \mathcal{D}_\psi(\mathbf{x}, \mathbf{c}) \\ \mathbf{x}' = \arg\min_{\mathbf{x}\in\mathcal{X}} \langle \mathbf{g}', \mathbf{x}\rangle + \mathcal{D}_\psi(\mathbf{x}, \mathbf{c}) \end{cases}$$

When the regularizer $\psi : \mathcal{X} \mapsto \mathbb{R}$ is a λ-strongly convex function with respect to norm $\|\cdot\|$, we have

$$\lambda \|\mathbf{x} - \mathbf{x}'\| \leq \|\mathbf{g} - \mathbf{g}'\|_\star .$$

Proof. For any convex function f, we have the first-order optimality condition:

$$f(\mathbf{x}) \leq f(\mathbf{y}) \quad \forall \mathbf{y} \in \mathcal{X} \Longleftrightarrow \nabla f(\mathbf{x})^\top (\mathbf{y} - \mathbf{x}) \geq 0 \quad \forall \mathbf{y} \in \mathcal{X}$$

Therefore, for $\mathbf{x}' = \arg\min_{\mathbf{x}\in\mathcal{X}} \{\langle \mathbf{g}', \mathbf{x}\rangle + \mathcal{D}_\psi(\mathbf{x}, \mathbf{c})\}$, we have

$$\langle \mathbf{g}' + \nabla\psi(\mathbf{x}') - \nabla\psi(\mathbf{c}), \mathbf{u} - \mathbf{x}'\rangle \geq 0 \text{ holds for } \forall \mathbf{u} \in \mathcal{X}.$$

Stability Lemma

Lemma 2 (Stability Lemma). *Consider the following updates:*

$$\begin{cases} \mathbf{x} = \arg\min_{\mathbf{x}\in\mathcal{X}} \langle \mathbf{g}, \mathbf{x}\rangle + \mathcal{D}_\psi(\mathbf{x}, \mathbf{c}) \\ \mathbf{x}' = \arg\min_{\mathbf{x}\in\mathcal{X}} \langle \mathbf{g}', \mathbf{x}\rangle + \mathcal{D}_\psi(\mathbf{x}, \mathbf{c}) \end{cases}$$

When the regularizer $\psi : \mathcal{X} \mapsto \mathbb{R}$ is a λ-strongly convex function with respect to norm $\|\cdot\|$, we have

$$\lambda \|\mathbf{x} - \mathbf{x}'\| \leq \|\mathbf{g} - \mathbf{g}'\|_\star .$$

Proof. $\langle \mathbf{g}' + \nabla\psi(\mathbf{x}') - \nabla\psi(\mathbf{c}), \mathbf{u} - \mathbf{x}'\rangle \geq 0$ holds for $\forall \mathbf{u} \in \mathcal{X}$.

By the first-order optimality conditions of $\mathbf{x}$ and $\mathbf{x}'$,

$$\langle \nabla\psi(\mathbf{x}) - \nabla\psi(\mathbf{c}) + \mathbf{g}, \mathbf{x}' - \mathbf{x}\rangle \geq 0$$
$$\langle \nabla\psi(\mathbf{x}') - \nabla\psi(\mathbf{c}) + \mathbf{g}', \mathbf{x} - \mathbf{x}'\rangle \geq 0$$

$$\Rightarrow \quad \langle \mathbf{x}' - \mathbf{x}, \mathbf{g} - \mathbf{g}'\rangle \geq \langle \nabla\psi(\mathbf{x}) - \nabla\psi(\mathbf{x}'), \mathbf{x} - \mathbf{x}'\rangle \quad (1)$$

Stability Lemma

Lemma 2 (Stability Lemma). *Consider the following updates:*

$$\begin{cases} \mathbf{x} = \arg\min_{\mathbf{x}\in\mathcal{X}} \langle \mathbf{g}, \mathbf{x} \rangle + \mathcal{D}_\psi(\mathbf{x}, \mathbf{c}) \\ \mathbf{x}' = \arg\min_{\mathbf{x}\in\mathcal{X}} \langle \mathbf{g}', \mathbf{x} \rangle + \mathcal{D}_\psi(\mathbf{x}, \mathbf{c}) \end{cases}$$

When the regularizer $\psi : \mathcal{X} \mapsto \mathbb{R}$ is a λ-strongly convex function with respect to norm $\|\cdot\|$, we have

$$\lambda \|\mathbf{x} - \mathbf{x}'\| \leq \|\mathbf{g} - \mathbf{g}'\|_\star .$$

Proof. Besides, by the strong convexity of ψ, we have

$$\langle \nabla\psi(\mathbf{x}), \mathbf{x} - \mathbf{x}' \rangle \geq \psi(\mathbf{x}) - \psi(\mathbf{x}') + \frac{\lambda}{2} \|\mathbf{x} - \mathbf{x}'\|^2$$

$$\langle \nabla\psi(\mathbf{x}'), \mathbf{x}' - \mathbf{x} \rangle \geq \psi(\mathbf{x}') - \psi(\mathbf{x}) + \frac{\lambda}{2} \|\mathbf{x} - \mathbf{x}'\|^2$$

Summing them up, we get

$$\langle \nabla\psi(\mathbf{x}) - \nabla\psi(\mathbf{x}'), \mathbf{x} - \mathbf{x}' \rangle \geq \lambda \|\mathbf{x} - \mathbf{x}'\|^2 \qquad (2)$$

Stability Lemma

Lemma 2 (Stability Lemma). *Consider the following updates:*

$$\begin{cases} \mathbf{x} = \arg\min_{\mathbf{x}\in\mathcal{X}} \langle \mathbf{g}, \mathbf{x} \rangle + \mathcal{D}_\psi(\mathbf{x}, \mathbf{c}) \\ \mathbf{x}' = \arg\min_{\mathbf{x}\in\mathcal{X}} \langle \mathbf{g}', \mathbf{x} \rangle + \mathcal{D}_\psi(\mathbf{x}, \mathbf{c}) \end{cases}$$

When the regularizer $\psi : \mathcal{X} \mapsto \mathbb{R}$ is a λ-strongly convex function with respect to norm $\|\cdot\|$, we have

$$\lambda \|\mathbf{x} - \mathbf{x}'\| \leq \|\mathbf{g} - \mathbf{g}'\|_\star .$$

Proof.

$$\langle \mathbf{x}' - \mathbf{x}, \mathbf{g} - \mathbf{g}' \rangle \geq \langle \nabla\psi(\mathbf{x}) - \nabla\psi(\mathbf{x}'), \mathbf{x} - \mathbf{x}' \rangle \quad (1)$$

$$\langle \nabla\psi(\mathbf{x}) - \nabla\psi(\mathbf{x}'), \mathbf{x} - \mathbf{x}' \rangle \geq \lambda \|\mathbf{x} - \mathbf{x}'\|^2 \quad (2)$$

$$\Rightarrow \quad \lambda \|\mathbf{x} - \mathbf{x}'\|^2 \leq \langle \nabla\psi(\mathbf{x}) - \nabla\psi(\mathbf{x}'), \mathbf{x} - \mathbf{x}' \rangle \leq \langle \mathbf{x}' - \mathbf{x}, \mathbf{g} - \mathbf{g}' \rangle$$

$$\leq \|\mathbf{x} - \mathbf{x}'\| \, \|\mathbf{g} - \mathbf{g}'\|_\star \quad \text{(Hölder's inequality)}$$

$$\Rightarrow \quad \lambda \|\mathbf{x} - \mathbf{x}'\| \leq \|\mathbf{g} - \mathbf{g}'\|_\star \qquad \square$$

Proof of Mirror Descent Lemma

Proof. $f_t(\mathbf{x}_t) - f_t(\mathbf{u}) \leq \underbrace{\langle \nabla f_t(\mathbf{x}_t), \mathbf{x}_t - \mathbf{x}_{t+1} \rangle}_{\text{term (a)}} + \underbrace{\langle \nabla f_t(\mathbf{x}_t), \mathbf{x}_{t+1} - \mathbf{u} \rangle}_{\text{term (b)}}$

We further introduce following lemma to analyze term (b).

Lemma 3 (Bregman Proximal Inequality). *Let $\mathcal{X}$ be a convex set in a Banach space $\mathcal{B}$. Let $f : \mathcal{X} \mapsto \mathbb{R}$ be a closed proper convex function on $\mathcal{X}$. Given a convex regularizer $\psi : \mathcal{X} \mapsto \mathbb{R}$, we denote its induced Bregman divergence by $\mathcal{D}_\psi(\cdot, \cdot)$. Then, any update of the form*

$$\mathbf{x}_{t+1} = \arg\min_{\mathbf{x} \in \mathcal{X}} \{\langle \mathbf{g}_t, \mathbf{x} \rangle + \mathcal{D}_\psi(\mathbf{x}, \mathbf{x}_t)\}$$

satisfies the following inequality for any $\mathbf{u} \in \mathcal{X}$:

$$\langle \mathbf{g}_t, \mathbf{x}_{t+1} - \mathbf{u} \rangle \leq \mathcal{D}_\psi(\mathbf{u}, \mathbf{x}_t) - \mathcal{D}_\psi(\mathbf{u}, \mathbf{x}_{t+1}) - \mathcal{D}_\psi(\mathbf{x}_{t+1}, \mathbf{x}_t).$$

Crucial for analysis of first-order optimization methods based on Bregman divergence.

Bregman Proximal Inequality

Lemma 3 (Bregman Proximal Inequality). *The Bregman proximal update in the form of* $\mathbf{x}_{t+1} = \arg\min_{\mathbf{x}\in\mathcal{X}} \{\langle \mathbf{g}_t, \mathbf{x}\rangle + \mathcal{D}_\psi(\mathbf{x}, \mathbf{x}_t)\}$ *satisfies*

$$\langle \mathbf{g}_t, \mathbf{x}_{t+1} - \mathbf{u}\rangle \le \mathcal{D}_\psi(\mathbf{u}, \mathbf{x}_t) - \mathcal{D}_\psi(\mathbf{u}, \mathbf{x}_{t+1}) - \mathcal{D}_\psi(\mathbf{x}_{t+1}, \mathbf{x}_t).$$

Proof. Recall that for any convex function f, we have the following first-order optimality condition:

$$f(\mathbf{x}) \le f(\mathbf{y}) \quad \forall \mathbf{y} \in \mathcal{X} \iff \nabla f(\mathbf{x})^\top (\mathbf{y} - \mathbf{x}) \ge 0 \quad \forall \mathbf{y} \in \mathcal{X}$$

Therefore, for $\mathbf{x}_{t+1} = \arg\min_{\mathbf{x}\in\mathcal{X}} \{\langle \mathbf{g}_t, \mathbf{x}\rangle + \mathcal{D}_\psi(\mathbf{x}, \mathbf{x}_t)\}$, we have

$$\langle \mathbf{g}_t + \nabla\psi(\mathbf{x}_{t+1}) - \nabla\psi(\mathbf{x}_t), \mathbf{u} - \mathbf{x}_{t+1}\rangle \ge 0 \text{ holds for any } \mathbf{u} \in \mathcal{X}.$$

Bregman Proximal Inequality

Lemma 3 (Bregman Proximal Inequality). *The Bregman proximal update in the form of* $\mathbf{x}_{t+1} = \arg\min_{\mathbf{x}\in\mathcal{X}} \{\langle \mathbf{g}_t, \mathbf{x}\rangle + \mathcal{D}_\psi(\mathbf{x}, \mathbf{x}_t)\}$ *satisfies*

$$\langle \mathbf{g}_t, \mathbf{x}_{t+1} - \mathbf{u}\rangle \le \mathcal{D}_\psi(\mathbf{u}, \mathbf{x}_t) - \mathcal{D}_\psi(\mathbf{u}, \mathbf{x}_{t+1}) - \mathcal{D}_\psi(\mathbf{x}_{t+1}, \mathbf{x}_t).$$

Proof. $\langle \mathbf{g}_t + \nabla\psi(\mathbf{x}_{t+1}) - \nabla\psi(\mathbf{x}_t), \mathbf{u} - \mathbf{x}_{t+1}\rangle \ge 0$ holds for any $\mathbf{u} \in \mathcal{X}$.

On the other hand, the right side of Lemma 3 is:

$$\begin{aligned}
&\mathcal{D}_\psi(\mathbf{u}, \mathbf{x}_t) - \mathcal{D}_\psi(\mathbf{u}, \mathbf{x}_{t+1}) - \mathcal{D}_\psi(\mathbf{x}_{t+1}, \mathbf{x}_t) \\
&= \cancel{\psi(\mathbf{u})} - \cancel{\psi(\mathbf{x}_t)} - \langle \nabla\psi(\mathbf{x}_t), \mathbf{u} - \cancel{\mathbf{x}_t}\rangle - \cancel{\psi(\mathbf{u})} + \cancel{\psi(\mathbf{x}_{t+1})} + \langle \nabla\psi(\mathbf{x}_{t+1}), \mathbf{u} - \mathbf{x}_{t+1}\rangle \\
&\qquad - \cancel{\psi(\mathbf{x}_{t+1})} + \cancel{\psi(\mathbf{x}_t)} + \langle \nabla\psi(\mathbf{x}_t), \mathbf{x}_{t+1} - \cancel{\mathbf{x}_t}\rangle \\
&= \langle \nabla\psi(\mathbf{x}_{t+1}) - \nabla\psi(\mathbf{x}_t), \mathbf{u} - \mathbf{x}_{t+1}\rangle .
\end{aligned}$$

Rearranging the terms can finish the proof. □

Proof of Mirror Descent Lemma

Proof. $f_t(\mathbf{x}_t) - f_t(\mathbf{u}) \le \underbrace{\langle \nabla f_t(\mathbf{x}_t), \mathbf{x}_t - \mathbf{x}_{t+1} \rangle}_{\text{term (a)}} + \underbrace{\langle \nabla f_t(\mathbf{x}_t), \mathbf{x}_{t+1} - \mathbf{u} \rangle}_{\text{term (b)}}$

Lemma 2 (Stability Lemma)**.**

$$\lambda \|\mathbf{x}_1 - \mathbf{x}_2\| \le \|\mathbf{g}_1 - \mathbf{g}_2\|_\star$$

⇨ $\text{term (a)} = \langle \nabla f_t(\mathbf{x}_t), \mathbf{x}_t - \mathbf{x}_{t+1} \rangle \le \frac{\eta_t}{\lambda} \|\nabla f_t(\mathbf{x}_t)\|_\star^2$

(think of two updates: one for $\mathbf{x}_{t+1}$ with $\nabla f_t(\mathbf{x}_t)$ and another one for $\mathbf{x}_t$ with $\mathbf{0}$)

Lemma 3 (Bregman Proximal Inequality)**.**

$$\langle \mathbf{g}_t, \mathbf{x}_{t+1} - \mathbf{u} \rangle \le \mathcal{D}_\psi(\mathbf{u}, \mathbf{x}_t) - \mathcal{D}_\psi(\mathbf{u}, \mathbf{x}_{t+1}) - \mathcal{D}_\psi(\mathbf{x}_{t+1}, \mathbf{x}_t)$$

⇨ $\text{term (b)} \le \frac{1}{\eta_t} \left(\mathcal{D}_\psi(\mathbf{u}, \mathbf{x}_t) - \mathcal{D}_\psi(\mathbf{u}, \mathbf{x}_{t+1}) - \mathcal{D}_\psi(\mathbf{x}_{t+1}, \mathbf{x}_t) \right)$

(negative term, usually dropped; but sometimes highly useful)

⇨ $f_t(\mathbf{x}_t) - f_t(\mathbf{u}) \le \frac{1}{\eta_t} \left(\mathcal{D}_\psi(\mathbf{u}, \mathbf{x}_t) - \mathcal{D}_\psi(\mathbf{u}, \mathbf{x}_{t+1}) \right) + \frac{\eta_t}{\lambda} \|\nabla f_t(\mathbf{x}_t)\|_\star^2 - \frac{1}{\eta_t} \mathcal{D}_\psi(\mathbf{x}_{t+1}, \mathbf{x}_t)$ □

General Regret Analysis for OMD

Lemma 1 (Mirror Descent Lemma). *Let $\mathcal{D}_\psi$ be the Bregman divergence w.r.t. $\psi : \mathcal{X} \to \mathbb{R}$ and assume ψ to be λ-strongly convex with respect to a norm $\|\cdot\|$. Then, $\forall \mathbf{u} \in \mathcal{X}$, the following inequality holds*

$$f_t(\mathbf{x}_t) - f_t(\mathbf{u}) \leq \frac{1}{\eta_t}\big(\mathcal{D}_\psi(\mathbf{u}, \mathbf{x}_t) - \mathcal{D}_\psi(\mathbf{u}, \mathbf{x}_{t+1})\big) + \frac{\eta_t}{\lambda}\|\nabla f_t(\mathbf{x}_t)\|_\star^2 - \frac{1}{\eta_t}\mathcal{D}_\psi(\mathbf{x}_{t+1}, \mathbf{x}_t)$$

Using Lemma 1, we can easily prove the following cumulative regret bound for OMD.

Theorem 4 (General Regret Bound for OMD). *Assume ψ is λ-strongly convex w.r.t. $\|\cdot\|$ and $\eta_t = \eta, \forall t \in [T]$. Then, for all $\mathbf{u} \in \mathcal{X}$, the following regret bound holds*

$$\sum_{t=1}^{T} f_t(\mathbf{x}_t) - \sum_{t=1}^{T} f_t(\mathbf{u}) \leq \frac{\mathcal{D}_\psi(\mathbf{u}, \mathbf{x}_1)}{\eta} + \frac{\eta}{\lambda}\sum_{t=1}^{T}\|\nabla f_t(\mathbf{x}_t)\|_\star^2 - \frac{1}{\eta}\sum_{t=1}^{T}\mathcal{D}_\psi(\mathbf{x}_{t+1}, \mathbf{x}_t)$$

General Regret Analysis for OMD

Theorem 4 (General Regret Bound for OMD). *Assume ψ is λ-strongly convex w.r.t. $\|\cdot\|$ and $\eta_t = \eta, \forall t \in [T]$. Then, for all $\mathbf{u} \in \mathcal{X}$, the following regret bound holds*

$$\sum_{t=1}^{T} f_t(\mathbf{x}_t) - \sum_{t=1}^{T} f_t(\mathbf{u}) \leq \frac{\mathcal{D}_\psi(\mathbf{u}, \mathbf{x}_1)}{\eta} + \frac{\eta}{\lambda} \sum_{t=1}^{T} \|\nabla f_t(\mathbf{x}_t)\|_\star^2 - \frac{1}{\eta} \sum_{t=1}^{T} \mathcal{D}_\psi(\mathbf{x}_{t+1}, \mathbf{x}_t)$$

Proof.

$$\begin{aligned}
\sum_{t=1}^{T} f_t(\mathbf{x}_t) - \sum_{t=1}^{T} f_t(\mathbf{u}) &\leq \sum_{t=1}^{T} \left(\frac{1}{\eta_t}\mathcal{D}_\psi(\mathbf{u}, \mathbf{x}_t) - \frac{1}{\eta_t}\mathcal{D}_\psi(\mathbf{u}, \mathbf{x}_{t+1})\right) + \sum_{t=1}^{T} \frac{\eta_t}{\lambda} \|\nabla f_t(\mathbf{x}_t)\|_\star^2 - \sum_{t=1}^{T} \frac{1}{\eta_t}\mathcal{D}_\psi(\mathbf{x}_{t+1}, \mathbf{x}_t) \\
&= \frac{1}{\eta_1}\mathcal{D}_\psi(\mathbf{u}, \mathbf{x}_1) - \frac{1}{\eta_T}\mathcal{D}_\psi(\mathbf{u}, \mathbf{x}_{T+1}) + \sum_{t=2}^{T} \left(\frac{1}{\eta_t} - \frac{1}{\eta_{t-1}}\right) \mathcal{D}_\psi(\mathbf{u}, \mathbf{x}_t) + \sum_{t=1}^{T} \frac{\eta_t}{\lambda} \|f_t(\mathbf{x}_t)\|_\star^2 - \sum_{t=1}^{T} \frac{1}{\eta_t}\mathcal{D}_\psi(\mathbf{x}_{t+1}, \mathbf{x}_t) \\
&\leq \frac{\mathcal{D}_\psi(\mathbf{u}, \mathbf{x}_1)}{\eta} + \frac{\eta}{\lambda} \sum_{t=1}^{T} \|\nabla f_t(\mathbf{x}_t)\|_\star^2 - \frac{1}{\eta} \sum_{t=1}^{T} \mathcal{D}_\psi(\mathbf{x}_{t+1}, \mathbf{x}_t) \qquad (\eta_t = \eta_{t-1})
\end{aligned}$$

□

General Regret Analysis for OMD

Theorem 4 (General Regret Bound for OMD). *Assume ψ is λ-strongly convex w.r.t. $\|\cdot\|$ and $\eta_t = \eta, \forall t \in [T]$. Then, for all $\mathbf{u} \in \mathcal{X}$, the following regret bound holds*

$$\sum_{t=1}^{T} f_t(\mathbf{x}_t) - \sum_{t=1}^{T} f_t(\mathbf{u}) \leq \frac{\mathcal{D}_\psi(\mathbf{u}, \mathbf{x}_1)}{\eta} + \frac{\eta}{\lambda} \sum_{t=1}^{T} \|\nabla f_t(\mathbf{x}_t)\|_\star^2 - \frac{1}{\eta} \sum_{t=1}^{T} \mathcal{D}_\psi(\mathbf{x}_{t+1}, \mathbf{x}_t)$$

With this general regret bound for OMD, it will become straightforward to analyze OGD/Hedge/ONS ***in a unified way***, which we previously analyzed specifically for each algorithm.

OMD Implication: Recovering OGD

Algorithm. With Theorem 3, it is straightforward to recover OGD:

OGD for convex	$\mathbf{x}_{t+1} = \arg\min_{\mathbf{x}\in\mathcal{X}} \left\{ \langle \mathbf{x}, \frac{1}{\sqrt{t}} \nabla f_t(\mathbf{x}_t) \rangle + \frac{1}{2} \|\mathbf{x} - \mathbf{x}_t\|_2^2 \right\}$

- $\psi(\mathbf{x}) = \frac{1}{2} \|\mathbf{x}\|_2^2$ is 1-strongly convex w.r.t. $\|\cdot\|_2$
- The dual norm of $\|\cdot\|_2$ is still $\|\cdot\|_2$

Regret Analysis.

$$\Longrightarrow \quad \sum_{t=1}^{T} f_t(\mathbf{x}_t) - \sum_{t=1}^{T} f_t(\mathbf{u}) \le \sum_{t=1}^{T} \left(\frac{1}{\eta_t} \|\mathbf{u} - \mathbf{x}_t\|_2^2 - \frac{1}{\eta_t} \|\mathbf{u} - \mathbf{x}_{t+1}\|_2^2 \right) + \sum_{t=1}^{T} \eta_t \|\nabla f_t(\mathbf{x}_t)\|_2^2$$

$$= \frac{1}{\eta_1} \|\mathbf{u} - \mathbf{x}_1\|_2^2 - \frac{1}{\eta_T} \|\mathbf{u} - \mathbf{x}_{T+1}\|_2^2 + \sum_{t=2}^{T} \left(\frac{1}{\eta_t} - \frac{1}{\eta_{t-1}} \right) \|\mathbf{u} - \mathbf{x}_t\|_2^2 + \sum_{t=1}^{T} \frac{\eta_t}{\lambda} \|f_t(\mathbf{x}_t)\|_2^2$$

$$\le \frac{D^2}{\eta_1} + \frac{D^2}{\eta_T} + \sum_{t=1}^{T} \eta_t G^2 \quad \le \quad 3DG\sqrt{T} \qquad \left(\eta_t = \tfrac{D}{G\sqrt{t}} \text{ and } \textstyle\sum_{t=1}^{T} \tfrac{1}{\sqrt{t}} \le 2\sqrt{T}\right) \qquad \square$$

OMD Implication: Recovering Hedge

Algorithm. With Theorem 3, it is straightforward to recover Hegde:

Hedge for PEA	$\mathbf{x}_{t+1} = \arg\min_{\mathbf{x}\in\mathcal{X}} \left\{ \langle \mathbf{x}, \eta \nabla f_t(\mathbf{x}_t) \rangle + \mathrm{KL}(\mathbf{x} \| \mathbf{x}_t) \right\}$

- Negative entropy is 1-strongly convex w.r.t. $\|\cdot\|_1$
- The dual norm of $\|\cdot\|_1$ is $\|\cdot\|_\infty$
- We initialize the initial prediction $\mathbf{x}_1 = \{\frac{1}{N}, \ldots, \frac{1}{N}\}$

Regret Analysis.

$$\Longrightarrow \quad \sum_{t=1}^{T} f_t(\mathbf{x}_t) - \sum_{t=1}^{T} f_t(\mathbf{u}) \leq \frac{\mathrm{KL}(\mathbf{u}\|\mathbf{x}_1)}{\eta} + \eta \sum_{t=1}^{T} \|\boldsymbol{\ell}_t\|_\infty^2 \leq \frac{\ln N}{\eta} + \eta T$$

$(\mathrm{KL}(\mathbf{u}\|\mathbf{x}_1) \leq \ln N, \forall \mathbf{u})$ $(\ell_t(i) \leq 1, \forall i \in [N])$

□

OMD Implication: Recovering ONS

Algorithm. With Theorem 3, it is straightforward to recover ONS:

ONS for exp-concave	$\mathbf{x}_{t+1} = \arg\min_{\mathbf{x}\in\mathcal{X}} \left\{ \langle \mathbf{x}, \frac{1}{\gamma}\nabla f_t(\mathbf{x}_t)\rangle + \frac{1}{2}\|\mathbf{x}-\mathbf{x}_t\|_{A_t}^2 \right\}$

- $\psi_t(\mathbf{x}) = \frac{1}{2}\|\mathbf{x}\|_{A_t}^2$ is 1-strongly convex w.r.t. $\|\cdot\|_{A_t}$ with $A_t = \varepsilon I + \sum_{s=1}^{t} \nabla f_t(\mathbf{x}_t)\nabla f_t(\mathbf{x}_t)^\top$
- The dual norm of $\|\cdot\|_{A_t}$ is $\|\cdot\|_{A_t^{-1}}$

Regret Analysis.

$$\Longrightarrow \quad \sum_{t=1}^{T} f_t(\mathbf{x}_t) - \sum_{t=1}^{T} f_t(\mathbf{u}) \le \frac{\gamma}{2}\sum_{t=1}^{T}\left(\|\mathbf{u}-\mathbf{x}_t\|_{A_t}^2 - \|\mathbf{u}-\mathbf{x}_{t+1}\|_{A_t}^2 - \|\mathbf{u}-\mathbf{x}_t\|_{\nabla f_t(\mathbf{x}_t)\nabla f_t(\mathbf{x}_t)^\top}^2\right) + \frac{1}{2\gamma}\sum_{t=1}^{T}\|\nabla f_t(\mathbf{x}_t)\|_{A_t^{-1}}^2$$

(exp-concavity)

$$= \frac{\gamma}{2}\sum_{t=1}^{T}\left(\|\mathbf{u}-\mathbf{x}_t\|_{A_{t-1}}^2 - \|\mathbf{u}-\mathbf{x}_{t+1}\|_{A_t}^2\right) + \frac{1}{2\gamma}\sum_{t=1}^{T}\|\nabla f_t(\mathbf{x}_t)\|_{A_t^{-1}}^2$$

(telescope)

$$\le \frac{\gamma}{2}\|\mathbf{u}-\mathbf{x}_1\|_{A_0}^2 + \frac{1}{2\gamma}\sum_{t=1}^{T}\|\nabla f_t(\mathbf{x}_t)\|_{A_t^{-1}}^2$$

□

OMD Implication: Recovering OGD for S.C.

Algorithm. With Theorem 3, we can recover OGD for strongly convex function:

OGD for strongly convex	$\mathbf{x}_{t+1} = \arg\min_{\mathbf{x}\in\mathcal{X}} \left\{ \langle \mathbf{x}, \frac{1}{\sigma t}\nabla f_t(\mathbf{x}_t)\rangle + \frac{1}{2}\left\|\mathbf{x}-\mathbf{x}_t\right\|_2^2 \right\}$

- $\psi_t(\mathbf{x}) = \frac{1}{2}\left\|\mathbf{x}\right\|_2^2$ is 1-strongly convex w.r.t. $\|\cdot\|_2$
- The dual norm of $\|\cdot\|_2$ is $\|\cdot\|_2$

Regret Analysis.

$$\Longrightarrow \quad \sum_{t=1}^T f_t(\mathbf{x}_t) - \sum_{t=1}^T f_t(\mathbf{u}) \le \frac{1}{2}\sum_{t=1}^T \left(\frac{1}{\eta_t}\left\|\mathbf{u}-\mathbf{x}_t\right\|_2^2 - \frac{1}{\eta_t}\left\|\mathbf{u}-\mathbf{x}_{t+1}\right\|_2^2 - \sigma\left\|\mathbf{u}-\mathbf{x}_t\right\|_2^2 \right) + \frac{1}{2}\sum_{t=1}^T \eta_t \left\|\nabla f_t(\mathbf{x}_t)\right\|_2^2$$

(strong convexity)

$$= \frac{1}{2}\sum_{t=1}^T \left(\frac{1}{\eta_t} - \frac{1}{\eta_{t-1}} - \sigma \right)\left\|\mathbf{u}-\mathbf{x}_t\right\|_2^2 + \frac{1}{2}\sum_{t=1}^T \eta_t G^2$$

$$= 0 + \frac{1}{2}\sum_{t=1}^T \frac{G^2}{\sigma t}$$

□

A Summary of OMD Deployment

- Our previous mentioned algorithms can all be covered by OMD.

Algo.	OMD/proximal form	$\psi(\cdot)$	η_t	Regret_T
OGD for convex	$\mathbf{x}_{t+1} = \arg\min_{\mathbf{x}\in\mathcal{X}} \eta_t \langle \mathbf{x}, \nabla f_t(\mathbf{x}_t)\rangle + \frac{1}{2}\|\mathbf{x}-\mathbf{x}_t\|_2^2$	$\|\mathbf{x}\|_2^2$	$\frac{1}{\sqrt{t}}$	$\mathcal{O}(\sqrt{T})$
OGD for strongly c.	$\mathbf{x}_{t+1} = \arg\min_{\mathbf{x}\in\mathcal{X}} \eta_t \langle \mathbf{x}, \nabla f_t(\mathbf{x}_t)\rangle + \frac{1}{2}\|\mathbf{x}-\mathbf{x}_t\|_2^2$	$\|\mathbf{x}\|_2^2$	$\frac{1}{\sigma t}$	$\mathcal{O}(\frac{1}{\sigma}\log T)$
ONS for exp-concave	$\mathbf{x}_{t+1} = \arg\min_{\mathbf{x}\in\mathcal{X}} \eta_t \langle \mathbf{x}, \nabla f_t(\mathbf{x}_t)\rangle + \frac{1}{2}\|\mathbf{x}-\mathbf{x}_t\|_{A_t}^2$	$\|\mathbf{x}\|_{A_t}^2$	$\frac{1}{\gamma}$	$\mathcal{O}(\frac{d}{\gamma}\log T)$
Hedge for PEA	$\mathbf{x}_{t+1} = \arg\min_{\mathbf{x}\in\Delta_N} \eta_t \langle \mathbf{x}, \nabla f_t(\mathbf{x}_t)\rangle + \text{KL}(\mathbf{x}\|\mathbf{x}_t)$	$\sum_{i=1}^N x_i \log x_i$	$\sqrt{\frac{\ln N}{T}}$	$\mathcal{O}(\sqrt{T\log N})$

Another View for Mirror Descent

Theorem 5. *The OMD update form*

$$\mathbf{x}_{t+1} = \arg\min_{\mathbf{x}\in\mathcal{X}} \left\{\eta_t \langle \mathbf{x}, \nabla f_t(\mathbf{x}_t)\rangle + \mathcal{D}_\psi(\mathbf{x}, \mathbf{x}_t)\right\} \qquad (\star)$$

is equivalent to the following two-step updates:

$$\begin{cases} \nabla\psi(\mathbf{y}_{t+1}) = \nabla\psi(\mathbf{x}_t) - \eta_t \nabla f_t(\mathbf{x}_t) \\ \mathbf{x}_{t+1} = \arg\min_{\mathbf{x}\in\mathcal{X}} \mathcal{D}_\psi(\mathbf{x}, \mathbf{y}_{t+1}) \end{cases} \qquad (\diamond)$$

Proof. $(\diamond)$

$$\begin{aligned} \mathbf{x}_{t+1} &= \arg\min_{\mathbf{x}\in\mathcal{X}} \mathcal{D}_\psi(\mathbf{x}, \mathbf{y}_{t+1}) \\ &= \arg\min_{\mathbf{x}\in\mathcal{X}} \psi(\mathbf{x}) - \psi(\mathbf{y}_{t+1}) - \langle \nabla\psi(\mathbf{y}_{t+1}), \mathbf{x} - \mathbf{y}_{t+1}\rangle \\ &= \arg\min_{\mathbf{x}\in\mathcal{X}} \psi(\mathbf{x}) - \langle \nabla\psi(\mathbf{y}_{t+1}), \mathbf{x}\rangle \qquad \text{(definition of Bregman divergence)} \\ &= \arg\min_{\mathbf{x}\in\mathcal{X}} \psi(\mathbf{x}) - \langle \nabla\psi(\mathbf{x}_t) - \eta_t \nabla f_t(\mathbf{x}_t), \mathbf{x}\rangle \end{aligned}$$

Another View for Mirror Descent

Theorem 5. *The OMD update form*

$$\mathbf{x}_{t+1} = \arg\min_{\mathbf{x}\in\mathcal{X}} \left\{\eta_t \langle \mathbf{x}, \nabla f_t(\mathbf{x}_t)\rangle + \mathcal{D}_\psi(\mathbf{x}, \mathbf{x}_t)\right\} \qquad (\star)$$

is equivalent to the following two-step updates:

$$\begin{cases} \nabla\psi(\mathbf{y}_{t+1}) = \nabla\psi(\mathbf{x}_t) - \eta_t \nabla f_t(\mathbf{x}_t) \\ \mathbf{x}_{t+1} = \arg\min_{\mathbf{x}\in\mathcal{X}} \mathcal{D}_\psi(\mathbf{x}, \mathbf{y}_{t+1}) \end{cases} \qquad (\diamond)$$

Proof. $(\star)$ $\mathbf{x}_{t+1} = \arg\min_{\mathbf{x}\in\mathcal{X}} \left\{\eta_t \langle \nabla f_t(\mathbf{x}_t), \mathbf{x}\rangle + \mathcal{D}_\psi(\mathbf{x}, \mathbf{x}_t)\right\}$

$$= \arg\min_{\mathbf{x}\in\mathcal{X}} \left\{\eta_t \langle \nabla f_t(\mathbf{x}_t), \mathbf{x}\rangle + \psi(\mathbf{x}) - \psi(\mathbf{x}_t) - \langle \nabla\psi(\mathbf{x}_t), \mathbf{x} - \mathbf{x}_t\rangle \right\}$$

(definition of Bregman divergence)

$$= \arg\min_{\mathbf{x}\in\mathcal{X}} \left\{\eta_t \langle \nabla f_t(\mathbf{x}_t), \mathbf{x}\rangle + \psi(\mathbf{x}) - \langle \nabla\psi(\mathbf{x}_t), \mathbf{x}\rangle \right\}$$

□

Another View for Mirror Descent

- A two-step update for mirror descent

$$\begin{cases} \nabla\psi(\mathbf{y}_{t+1}) = \nabla\psi(\mathbf{x}_t) - \eta_t \nabla f_t(\mathbf{x}_t) \\ \\ \mathbf{x}_{t+1} = \arg\min_{\mathbf{x}\in\mathcal{X}} \mathcal{D}_\psi(\mathbf{x}, \mathbf{y}_{t+1}) \end{cases}$$

  - The first step is somehow similar to a "***gradient descent***" step;
  - The second step looks like a "***projection***" step.

⇨ Key role in mirror descent: the operator $\nabla\psi(\cdot)$

Primal-Dual View for Mirror Descent

- Recall the gradient descent update

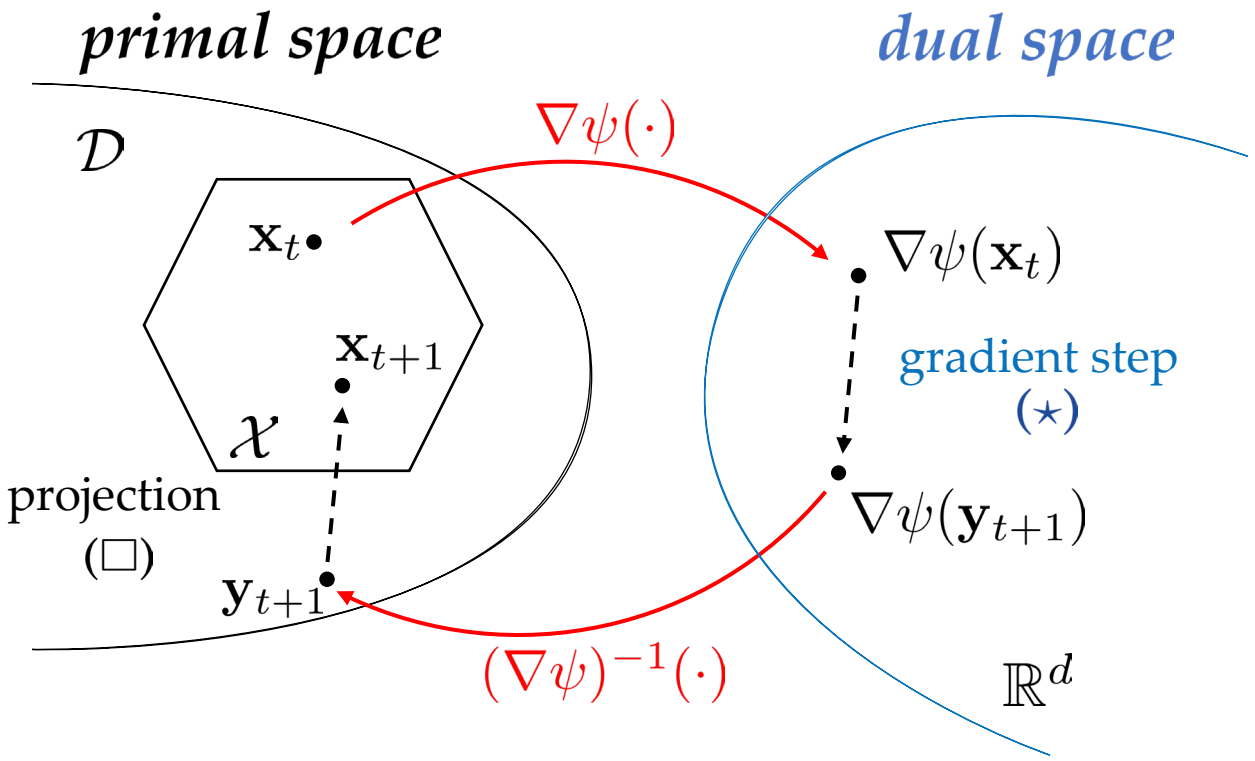
$$\mathbf{x} - \eta \nabla f(\mathbf{x})$$

but this simply ***does not make sense*** for general non-Euclidean space…

- Bits in convex analysis
  - consider a Banach space $\mathcal{B}$, whose dual space is denoted by $\mathcal{B}^*$
  - $\mathbf{x}$ is in the primal space $\mathcal{B}$, and $\nabla f(\mathbf{x})$ is in the dual space $\mathcal{B}^*$

$\Rightarrow$ a simple intuition: $f(\mathbf{x} + \Delta\mathbf{x}) \approx \langle \nabla f(\mathbf{x}), \Delta\mathbf{x} \rangle$

Primal-Dual View for Mirror Descent



$$(\star)\ \nabla\psi(\mathbf{y}_{t+1}) = \nabla\psi(\mathbf{x}_t) - \eta\nabla f(\mathbf{x}_t)$$

$$(\square)\ \ \mathbf{x}_{t+1} \in \Pi_{\mathcal{X}}^{\psi}[\mathbf{y}_{t+1}]$$

$$(\Pi_{\mathcal{X}}^{\psi}[\mathbf{y}] = \arg\min_{\mathbf{x}\in\mathcal{X}\cap\mathcal{D}} \mathcal{D}_{\psi}(\mathbf{x},\mathbf{y}))$$

$\nabla\psi(\cdot)$ is the mirror map to link two spaces

Mirror Map

Definition 2 (Mirror Map). Let $\mathcal{D} \subset \mathbb{R}^n$ be a convex open set such that $\mathcal{X}$ is included in its closure, that is $\mathcal{X} \subset \overline{\mathcal{D}}$, and $\mathcal{X} \cap \mathcal{D} \neq \emptyset$. We say that $\psi : \mathcal{D} \to \mathbb{R}$ is a mirror map if it safisfies the following properties:

(i) ψ is strictly convex and differentiable;

(ii) The gradient of ψ takes all possible values, that is $\nabla\psi(\mathcal{D}) = \mathbb{R}^n$;

(iii) The gradient of ψ diverges on the boundary of $\mathcal{D}$, that is

$$\lim_{\mathbf{x} \to \partial\mathcal{D}} \|\nabla\psi(\mathbf{x})\| = +\infty$$

See Chapter 4.1 of Bubeck's book for rigorous discussions.

Mirror Map Calculation

$$\nabla\psi(\mathbf{y}_{t+1}) = \nabla\psi(\mathbf{x}_t) - \eta_t \nabla f_t(\mathbf{x}_t)$$
$$\mathbf{x}_{t+1} = \arg\min_{\mathbf{x}\in\mathcal{X}} \mathcal{D}_\psi(\mathbf{x}, \mathbf{y}_{t+1})$$

equivalent $\Longleftrightarrow$

$$\mathbf{y}_{t+1} = \nabla\psi^\star\left(\nabla\psi\left(\mathbf{x}_t\right) - \eta_t \nabla f_t(\mathbf{x}_t)\right)$$
$$\mathbf{x}_{t+1} = \arg\min_{\mathbf{x}\in\mathcal{X}} \mathcal{D}_\psi(\mathbf{x}, \mathbf{y}_{t+1})$$

- Here, $\nabla\psi^\star(\cdot)$ is the ***Fenchel Conjugate*** of $\nabla\psi(\cdot)$.

Definition 3 (Fenchel Conjugate). For a function $f : \mathbb{R}^d \to [-\infty, \infty]$, we define its Fenchel conjugate $f^\star : \mathbb{R}^d \to [-\infty, \infty]$ as

$$f^\star(\mathbf{g}) = \sup_{\mathbf{y}\in\mathbb{R}^d} \left\{\langle \mathbf{g}, \mathbf{y}\rangle - f(\mathbf{y})\right\}.$$

Mirror Map Calculation

Proof. We first show for any convex and closed f, $\mathbf{g} = \nabla f(\mathbf{x}) \iff \mathbf{x} = \nabla f^\star(\mathbf{g})$.

By the convexity of f ($f(\mathbf{y}) \geq f(\mathbf{x}) + \langle \mathbf{g}, \mathbf{y} - \mathbf{x} \rangle, \forall \mathbf{y}$):

$$\langle \mathbf{g}, \mathbf{x} \rangle - f(\mathbf{x}) \geq \langle \mathbf{g}, \mathbf{y} \rangle - f(\mathbf{y}), \forall \mathbf{y}$$

which means $\langle \mathbf{g}, \mathbf{y} \rangle - f(\mathbf{y})$ achieves its supremum in $\mathbf{y}$ at $\mathbf{y} = \mathbf{x}$. Thus, by the definition of Fenchel Conjugate:

$$f^\star(\mathbf{g}) = \sup_{\mathbf{y} \in \mathbb{R}^d} \langle \mathbf{g}, \mathbf{y} \rangle - f(\mathbf{y}) = \langle \mathbf{g}, \mathbf{x} \rangle - f(\mathbf{x})$$

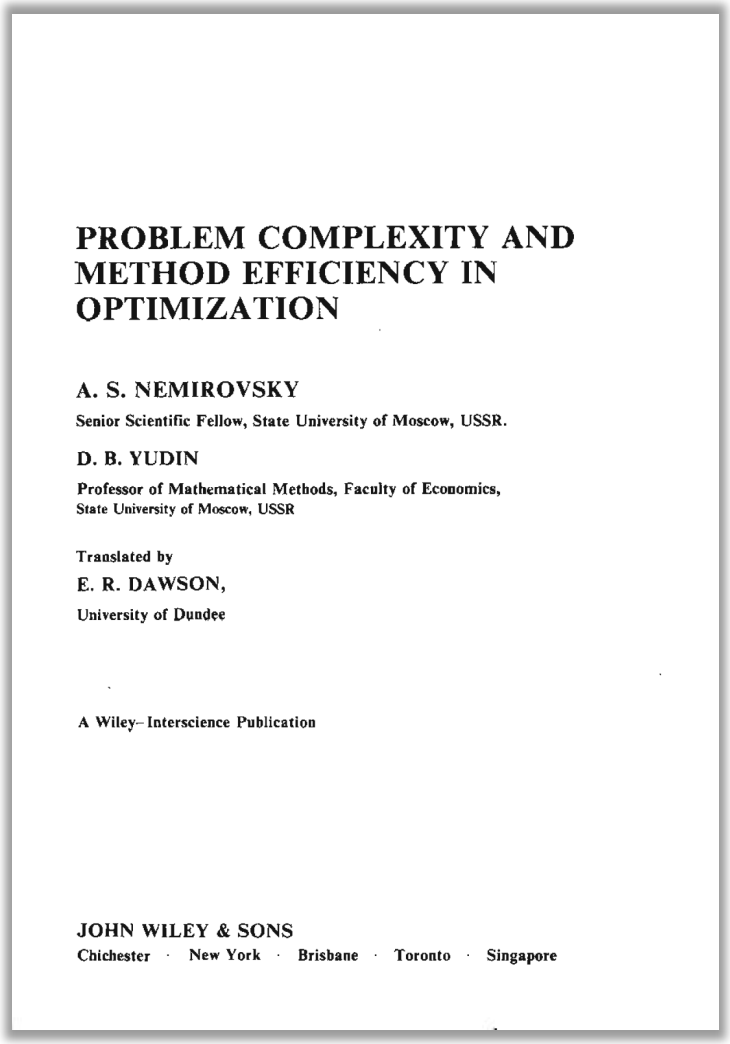
By taking the gradient w.r.t. $\mathbf{g}$ at both sides:

$$\nabla f^\star(\mathbf{g}) = \mathbf{x}$$

Therefore we have proved that $\mathbf{g} = \nabla f(\mathbf{x}) \iff \mathbf{x} = \nabla f^\star(\mathbf{g})$.

By setting $f(\cdot) = \psi(\cdot)$ and $\mathbf{x} = \mathbf{y}_{t+1}$, we finish the proof. □

Mirror Descent: history bits



PROBLEM COMPLEXITY AND METHOD EFFICIENCY IN OPTIMIZATION

A. S. NEMIROVSKY
Senior Scientific Fellow, State University of Moscow, USSR.

D. B. YUDIN
Professor of Mathematical Methods, Faculty of Economics, State University of Moscow, USSR

Translated by
E. R. DAWSON,
University of Dundee

A Wiley–Interscience Publication

JOHN WILEY & SONS
Chichester · New York · Brisbane · Toronto · Singapore

A. S. Nemirovski (1947 -

D. B. Yudin (1919 - 2006)

A.S. Nemirovski, D.B. Yudin, **Problem Complexity and Method Efficiency in Optimization**. Wiley-Interscience Series in Discrete Mathematics (A Wiley-Interscience Publication/Wiley, New York, 1983)

23. Nemirovskiy, A. S., and Yudin, D. B. (1979). Efficient methods of solving convex-programming problems of high dimensionality. *Ekonomika i matem. metody*, **XV**, No. 1. (In Russian.)

Part 3. Follow-the-Regularized Leader

- Algorithmic Framework
- Regret Analysis
- Interpretation from Primal-Dual View

Another OCO Framework: FTRL

- Recall: **Follow the Leader (FTL)**

Select the expert that *performs best so far*, specifically,

$$\boldsymbol{p}_t^{\text{FTL}} = \arg\min_{\boldsymbol{p}\in\Delta_N} \langle \boldsymbol{p}, \boldsymbol{L}_{t-1} \rangle$$

where $\boldsymbol{L}_{t-1} \triangleq \sum_{s=1}^{t-1} \boldsymbol{\ell}_s \in \mathbb{R}^N$ is the cumulative loss vector.

$\ell_{1,1} = 0.49$ ⇨ $\ell_{2,1} = 1$ ⇨ $\ell_{3,1} = 0$ ⇨ ⋯ ⋯

$\ell_{1,2} = 0.51$ ⇨ $\ell_{2,2} = 0$ ⇨ $\ell_{3,2} = 1$ ⇨ ⋯ ⋯

$$\text{Reg}_T = \sum_{t=1}^{T} \langle \boldsymbol{p}_t, \boldsymbol{\ell}_t \rangle - \min_{i\in[N]} \sum_{t=1}^{T} \ell_{t,i}$$

$$= T - \frac{T}{2} = \mathcal{O}(T)$$

FTL achieves *linear regret* in the worst case!

Another OCO Framework: FTRL

- Recall: **Follow the Leader (FTL)**

> Select the expert that ***performs best so far***, specifically,
>
> $$\boldsymbol{p}_t^{\mathrm{FTL}} = \arg\min_{\boldsymbol{p}\in\Delta_N} \langle \boldsymbol{p}, \boldsymbol{L}_{t-1} \rangle$$
>
> where $\boldsymbol{L}_{t-1} \triangleq \sum_{s=1}^{t-1} \boldsymbol{\ell}_s \in \mathbb{R}^N$ is the cumulative loss vector.

- As mentioned, FTL is ***sub-optimal*** due to its ***unstable*** nature.

⇨ a natural idea: adding regularizers to stabilize the algorithm.

Another OCO Framework: FTRL

Follow The Regularized Leader (FTRL)

$$\mathbf{x}_{t+1} = \arg\min_{\mathbf{x}\in\mathcal{X}} \left\{ \sum_{s=1}^{t} f_s(\mathbf{x}) + \psi_{t+1}(\mathbf{x}) \right\},$$

where $\psi_{t+1} : \mathcal{X} \mapsto \mathbb{R}$ is the regularizer at round $t+1$ update.

FTRL: essentially adding regularizer to stabilize the FTL algorithm.

We use time-varying regularizer to encode the potentially changing step sizes.

FTRL vs. OMD: Update Styles

- OMD update style:

$$\mathbf{x}_{t+1} = \arg\min_{\mathbf{x}\in\mathcal{X}} \left\{ \langle \mathbf{x}, \eta_t \nabla f_t(\mathbf{x}_t) \rangle + \mathcal{D}_\psi(\mathbf{x}, \mathbf{x}_t) \right\}$$

- FTRL update style:

$$\mathbf{x}_{t+1} = \arg\min_{\mathbf{x}\in\mathcal{X}} \left\{ \sum_{s=1}^{t} f_s(\mathbf{x}) + \psi_{t+1}(\mathbf{x}) \right\}$$

Comparison:

– in OMD, $\mathbf{x}_{t+1}$ depends on $\mathbf{x}_t$ and $f_t(\cdot)$;

– in FTRL, $\mathbf{x}_{t+1}$ depends on entire history $\{f_s(\cdot)\}_{s=1}^t$ and regularizer ψ_{t+1}.

Linearization in FTRL

- FTRL update requires to store all the historical online functions.

$$\mathbf{x}_{t+1} = \arg\min_{\mathbf{x}\in\mathcal{X}} \left\{ \sum_{s=1}^{t} f_s(\mathbf{x}) + \psi_{t+1}(\mathbf{x}) \right\}$$

- ***Surrogate optimization***: maintain regret while achieving one-pass update

$$f_t(\mathbf{x}_t) - f_t(\mathbf{u}) \leq \langle \nabla f_t(\mathbf{x}_t), \mathbf{x}_t - \mathbf{u} \rangle \triangleq \ell_t(\mathbf{x}_t) - \ell_t(\mathbf{u})$$

where we define the linear surrogate loss as $\ell_t(\mathbf{x}) \triangleq \langle \nabla f_t(\mathbf{x}_t), \mathbf{x} \rangle$.

surrogate ⇒

$$\mathbf{x}_{t+1} = \arg\min_{\mathbf{x}\in\mathcal{X}} \left\{ \sum_{s=1}^{t} \ell_s(\mathbf{x}) + \psi_{t+1}(\mathbf{x}) \right\}$$

$$= \arg\min_{\mathbf{x}\in\mathcal{X}} \left\{ \sum_{s=1}^{t} \langle \nabla f_s(\mathbf{x}_s), \mathbf{x} \rangle + \psi_{t+1}(\mathbf{x}) \right\}$$

It suffices to store gradient vectors only.

General Analysis of FTRL

Lemma 4 (FTRL Regret). *We denote that* $F_t(\mathbf{x}) \triangleq \psi_t(\mathbf{x}) + \sum_{s=1}^{t-1} f_s(\mathbf{x})$. *Thus, the FTRL algorithm runs* $\mathbf{x}_t = \arg\min_{\mathbf{x}\in\mathcal{X}} F_t(\mathbf{x})$. *Then, for any* $\mathbf{u} \in X$, *we have*

$$\sum_{t=1}^{T} f_t(\mathbf{x}_t) - \sum_{t=1}^{T} f_t(\mathbf{u}) = \psi_{T+1}(\mathbf{u}) - \min_{\mathbf{x}\in\mathcal{X}} \psi_1(\mathbf{x}) \quad \text{(range term)}$$

$$+ \sum_{t=1}^{T} \Big(F_t(\mathbf{x}_t) - F_{t+1}(\mathbf{x}_{t+1}) + f_t(\mathbf{x}_t) \Big) \quad \text{(stability term)}$$

$$+ F_{T+1}(\mathbf{x}_{T+1}) - F_{T+1}(\mathbf{u}) \quad (\mathbf{x}_{T+1} = \arg\min_{\mathbf{x}} F_{T+1}(\mathbf{x}), \text{ thus} \le 0)$$

General Analysis of FTRL

$$F_t(\mathbf{x}) \triangleq \psi_t(\mathbf{x}) + \sum_{s=1}^{t-1} f_s(\mathbf{x})$$

Lemma 4.

$$\sum_{t=1}^{T} f_t(\mathbf{x}_t) - \sum_{t=1}^{T} f_t(\mathbf{u}) = \psi_{T+1}(\mathbf{u}) - \min_{\mathbf{x}\in\mathcal{X}} \psi_1(\mathbf{x}) \quad \text{(range term)}$$

$$+ \sum_{t=1}^{T} \Big(F_t(\mathbf{x}_t) - F_{t+1}(\mathbf{x}_{t+1}) + f_t(\mathbf{x}_t) \Big) \quad \text{(stability term)}$$

$$+ F_{T+1}(\mathbf{x}_{T+1}) - F_{T+1}(\mathbf{u}) \quad \text{(negative term)}$$

Proof. The term $\sum_{t=1}^{T} f_t(\mathbf{x}_t)$ appears at both side of the equality, thus we verify

$$-\sum_{t=1}^{T} f_t(\mathbf{u}) = \psi_{T+1}(\mathbf{u}) - \min_{\mathbf{x}\in\mathcal{X}} \psi_1(\mathbf{x}) + \sum_{t=1}^{T} \Big(F_t(\mathbf{x}_t) - F_{t+1}(\mathbf{x}_{t+1}) \Big) + F_{T+1}(\mathbf{x}_{T+1}) - F_{T+1}(\mathbf{u}).$$

General Analysis of FTRL

$$F_t(\mathbf{x}) \triangleq \psi_t(\mathbf{x}) + \sum_{s=1}^{t-1} f_s(\mathbf{x})$$

Proof. The term $\sum_{t=1}^{T} f_t(\mathbf{x}_t)$ appears at both side of the equality, thus we verify

$$-\sum_{t=1}^{T} f_t(\mathbf{u}) = \psi_{T+1}(\mathbf{u}) - \min_{\mathbf{x}\in\mathcal{X}} \psi_1(\mathbf{x}) + \sum_{t=1}^{T} \Big(F_t(\mathbf{x}_t) - F_{t+1}(\mathbf{x}_{t+1}) \Big) + F_{T+1}(\mathbf{x}_{T+1}) - F_{T+1}(\mathbf{u}).$$

Recall that $F_1(\mathbf{x}_1) = \min_{\mathbf{x}\in\mathcal{X}} \psi_1(\mathbf{x})$, telescoping over $\sum_{t=1}^{T} \Big(F_t(\mathbf{x}_t) - F_{t+1}(\mathbf{x}_{t+1}) \Big)$

$$\sum_{t=1}^{T} \Big(F_t(\mathbf{x}_t) - F_{t+1}(\mathbf{x}_{t+1}) \Big) = F_1(\mathbf{x}_1) - F_{T+1}(\mathbf{x}_{T+1})$$

$$\Rightarrow -\sum_{t=1}^{T} f_t(\mathbf{u}) = \psi_{T+1}(\mathbf{u}) - F_1(\mathbf{x}_1) + F_1(\mathbf{x}_1) - F_{T+1}(\mathbf{x}_{T+1}) + F_{T+1}(\mathbf{x}_{T+1}) - F_{T+1}(\mathbf{u})$$

$$= \psi_{T+1}(\mathbf{u}) - F_{T+1}(\mathbf{u}),$$

which is true by the definition of $F_{T+1}(\mathbf{x}) \triangleq \psi_{T+1}(\mathbf{x}) + \sum_{s=1}^{T} f_s(\mathbf{x})$. □

General Analysis of FTRL

Lemma 4 (FTRL Regret). *We denote that* $F_t(\mathbf{x}) \triangleq \psi_t(\mathbf{x}) + \sum_{s=1}^{t-1} f_s(\mathbf{x})$. *Thus, the FTRL algorithm runs* $\mathbf{x}_t = \arg\min_{\mathbf{x}\in\mathcal{X}} F_t(\mathbf{x})$. *Then, for any* $\mathbf{u} \in X$, *we have*

$$\sum_{t=1}^{T} f_t(\mathbf{x}_t) - \sum_{t=1}^{T} f_t(\mathbf{u}) = \psi_{T+1}(\mathbf{u}) - \min_{\mathbf{x}\in\mathcal{X}} \psi_1(\mathbf{x}) \quad \text{(range term)}$$

$$+ \sum_{t=1}^{T} \Big(F_t(\mathbf{x}_t) - F_{t+1}(\mathbf{x}_{t+1}) + f_t(\mathbf{x}_t) \Big) \quad \text{(stability term)}$$

$$+ F_{T+1}(\mathbf{x}_{T+1}) - F_{T+1}(\mathbf{u}) \quad (\mathbf{x}_{T+1} = \arg\min_{\mathbf{x}} F_{T+1}(\mathbf{x}), \text{ thus} \leq 0)$$

- The first and third terms are similar to those in OMD regret analysis.
- The second term is the stability term, which is crucial for the regret analysis, and we will explain why it's called stability later.

FTRL Stability

$$F_t(\mathbf{x}) \triangleq \psi_t(\mathbf{x}) + \sum_{s=1}^{t-1} f_s(\mathbf{x})$$

Lemma 5 (FTRL Stability). *Assume that ψ_t is λ_t-strongly convex w.r.t. $\|\cdot\|$. Then, the FTRL update satisfies*

$$F_t(\mathbf{x}_t) - F_{t+1}(\mathbf{x}_{t+1}) + f_t(\mathbf{x}_t) \le \frac{\|\nabla f_t(\mathbf{x}_t)\|_*^2}{\lambda_t} + \psi_t(\mathbf{x}_{t+1}) - \psi_{t+1}(\mathbf{x}_{t+1}).$$

Proof. $F_t(\mathbf{x}_t) - F_{t+1}(\mathbf{x}_{t+1}) + f_t(\mathbf{x}_t)$

$$= F_t(\mathbf{x}_t) + f_t(\mathbf{x}_t) - (F_t(\mathbf{x}_{t+1}) + f_t(\mathbf{x}_{t+1})) + \psi_t(\mathbf{x}_{t+1}) - \psi_{t+1}(\mathbf{x}_{t+1})$$

$$\le \langle \nabla F_t(\mathbf{x}_t) + \nabla f_t(\mathbf{x}_t), \mathbf{x}_t - \mathbf{x}_{t+1} \rangle - \frac{\lambda_t}{2} \|\mathbf{x}_t - \mathbf{x}_{t+1}\|^2 + \psi_t(\mathbf{x}_{t+1}) - \psi_{t+1}(\mathbf{x}_{t+1}) \quad \text{(strong convexity)}$$

$$\le \langle \nabla f_t(\mathbf{x}_t), \mathbf{x}_t - \mathbf{x}_{t+1} \rangle - \frac{\lambda_t}{2} \|\mathbf{x}_t - \mathbf{x}_{t+1}\|^2 + \psi_t(\mathbf{x}_{t+1}) - \psi_{t+1}(\mathbf{x}_{t+1}) \quad (\mathbf{x}_t = \arg\min_{\mathbf{x}\in\mathcal{X}} F_t(\mathbf{x}))$$

FTRL Stability

$$F_t(\mathbf{x}) \triangleq \psi_t(\mathbf{x}) + \sum_{s=1}^{t-1} f_s(\mathbf{x})$$

Lemma 5 (FTRL Stability). *Assume that ψ_t is λ_t-strongly convex w.r.t. $\|\cdot\|$. Then, the FTRL update satisfies*

$$F_t(\mathbf{x}_t) - F_{t+1}(\mathbf{x}_{t+1}) + f_t(\mathbf{x}_t) \le \frac{\|\nabla f_t(\mathbf{x}_t)\|_*^2}{\lambda_t} + \psi_t(\mathbf{x}_{t+1}) - \psi_{t+1}(\mathbf{x}_{t+1}).$$

Proof. $F_t(\mathbf{x}_t) - F_{t+1}(\mathbf{x}_{t+1}) + f_t(\mathbf{x}_t)$

$$\le \langle \nabla f_t(\mathbf{x}_t), \mathbf{x}_t - \mathbf{x}_{t+1} \rangle - \frac{\lambda_t}{2} \|\mathbf{x}_t - \mathbf{x}_{t+1}\|^2 + \psi_t(\mathbf{x}_{t+1}) - \psi_{t+1}(\mathbf{x}_{t+1})$$

$$\le \|\nabla f_t(\mathbf{x}_t)\|_* \cdot \|\mathbf{x}_t - \mathbf{x}_{t+1}\| - \frac{\lambda_t}{2} \|\mathbf{x}_t - \mathbf{x}_{t+1}\|^2 + \psi_t(\mathbf{x}_{t+1}) - \psi_{t+1}(\mathbf{x}_{t+1}) \quad \text{(Hölder's inequality)}$$

$$\le \frac{1}{\lambda_t} \|\nabla f_t(\mathbf{x}_t)\|_*^2 - \frac{\lambda_t}{4} \|\mathbf{x}_t - \mathbf{x}_{t+1}\|^2 + \psi_t(\mathbf{x}_{t+1}) - \psi_{t+1}(\mathbf{x}_{t+1}) \qquad \square \quad (ab \le \tfrac{a^2}{\lambda} + \tfrac{\lambda}{4} b^2)$$

Regret Bound for FTRL

$$F_t(\mathbf{x}) \triangleq \psi_t(\mathbf{x}) + \sum_{s=1}^{t-1} f_s(\mathbf{x})$$

Theorem 6 (Regret Bound for FTRL). *Assume $\psi_t(\mathbf{x})$ is λ_t-strongly convex on domain $\mathcal{X}$ w.r.t. $\|\cdot\|$. We further assume that $\psi_t(\mathbf{x}) \le \psi_{t+1}(\mathbf{x})$ for $t \in [T]$.Then, for FTRL satisfies*

$$\sum_{t=1}^{T} f_t(\mathbf{x}_t) - \sum_{t=1}^{T} f_t(\mathbf{u}) \le \psi_{T+1}(\mathbf{u}) + \sum_{t=1}^{T} \frac{1}{\lambda_t} \|\nabla f_t(\mathbf{x}_t)\|_*^2 - \sum_{t=2}^{T} \frac{\lambda_t}{4} \|\mathbf{x}_t - \mathbf{x}_{t-1}\|^2.$$

Proof.

$$\sum_{t=1}^{T} (f_t(\mathbf{x}_t) - f_t(\mathbf{u})) = \psi_{T+1}(\mathbf{u}) - \min_{\mathbf{x}\in\mathcal{X}} \psi_1(\mathbf{x}) \quad \text{(range term)}$$

$$+ \sum_{t=1}^{T} \Big(F_t(\mathbf{x}_t) - F_{t+1}(\mathbf{x}_{t+1}) + f_t(\mathbf{x}_t) \Big) \quad \text{(stability term)}$$

$$+ F_{T+1}(\mathbf{x}_{T+1}) - F_{T+1}(\mathbf{u}) \quad (\mathbf{x}_{T+1} = \arg\min_{\mathbf{x}} F_{T+1}(\mathbf{x}), \text{ thus} \le 0)$$

Regret Bound for FTRL

$$F_t(\mathbf{x}) \triangleq \psi_t(\mathbf{x}) + \sum_{s=1}^{t-1} f_s(\mathbf{x})$$

Theorem 6 (Regret Bound for FTRL). *Assume $\psi_t(\mathbf{x})$ is λ_t-strongly convex on domain $\mathcal{X}$ w.r.t. $\|\cdot\|$. We further assume that $\psi_t(\mathbf{x}) \le \psi_{t+1}(\mathbf{x})$ for $t \in [T]$.Then, for FTRL satisfies*

$$\sum_{t=1}^{T} f_t(\mathbf{x}_t) - \sum_{t=1}^{T} f_t(\mathbf{u}) \le \psi_{T+1}(\mathbf{u}) + \sum_{t=1}^{T} \frac{1}{\lambda_t} \|\nabla f_t(\mathbf{x}_t)\|_*^2 - \sum_{t=2}^{T} \frac{\lambda_t}{4} \|\mathbf{x}_t - \mathbf{x}_{t-1}\|^2.$$

Proof.

$$\sum_{t=1}^{T} (f_t(\mathbf{x}_t) - f_t(\mathbf{u})) \overset{\text{(stability)}}{\le} \psi_{T+1}(\mathbf{u}) + \sum_{t=1}^{T} \left(\frac{\|\nabla f_t(\mathbf{x}_t)\|_*^2}{\lambda_t} + \psi_t(\mathbf{x}_{t+1}) - \psi_{t+1}(\mathbf{x}_{t+1}) \right)$$

$$\le \psi_{T+1}(\mathbf{u}) + \sum_{t=1}^{T} \frac{1}{\lambda_t} \|\nabla f_t(\mathbf{x}_t)\|_*^2 - \sum_{t=2}^{T} \frac{\lambda_t}{4} \|\mathbf{x}_t - \mathbf{x}_{t-1}\|^2$$

□

FTRL can be equivalent to OMD

Claim 1. Under online linear optimization (OLO) setting, with the same constant step size $\eta > 0$ and the same regularizer ψ (which is required to be *strongly convex* and a *barrier* function over $\mathcal{X}$), the OMD and FTRL algorithms share the same output:

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

and

$$\mathbf{x}_t = \arg\min_{\mathbf{x}\in\mathcal{X}} \left\{ \langle \eta \mathbf{g}_{t-1}, \mathbf{x} \rangle + \mathcal{D}_\psi(\mathbf{x}, \mathbf{x}_{t-1}) \right\}.$$

FTRL vs. OMD: Equivalence Condition

Proof. For OMD, taking the gradient and setting it to 0 will lead to:

$$\eta \mathbf{g}_{t-1} + \nabla \psi(\mathbf{x}_t) - \nabla \psi(\mathbf{x}_{t-1}) = 0 \quad \text{(due to the barrier property of } \psi\text{)}$$

Telescoping from 1 to $t-1$, and define $\mathbf{x}_0 \triangleq \arg\min_{\mathbf{x} \in \mathcal{X}} \psi(\mathbf{x})$,

$$\nabla \psi(\mathbf{x}_t) = -\eta \sum_{s=1}^{t-1} \mathbf{g}_s$$

On the other hand, for FTRL, setting the gradient to zero will lead to:

$$\nabla \psi(\mathbf{x}_t) = -\eta \sum_{s=1}^{t-1} \mathbf{g}_s \qquad \square$$

FTRL as Dual Averaging

- Mirror Descent

$$\nabla \psi_t \left(\mathbf{y}_{t+1}\right) = \nabla \psi_t(\mathbf{x}_t) - \eta_t \nabla f_t(\mathbf{x}_t)$$
$$\mathbf{x}_{t+1} = \arg\min_{\mathbf{x} \in \mathcal{X}} \mathcal{D}_\psi(\mathbf{x}, \mathbf{y}_{t+1})$$

- Dual Averaging (**lazy** mirror descent)

$$\nabla \psi_t(\mathbf{y}_{t+1}) = \nabla \psi_t(\mathbf{y}_t) - \eta_t \nabla f_t(\mathbf{x}_t)$$

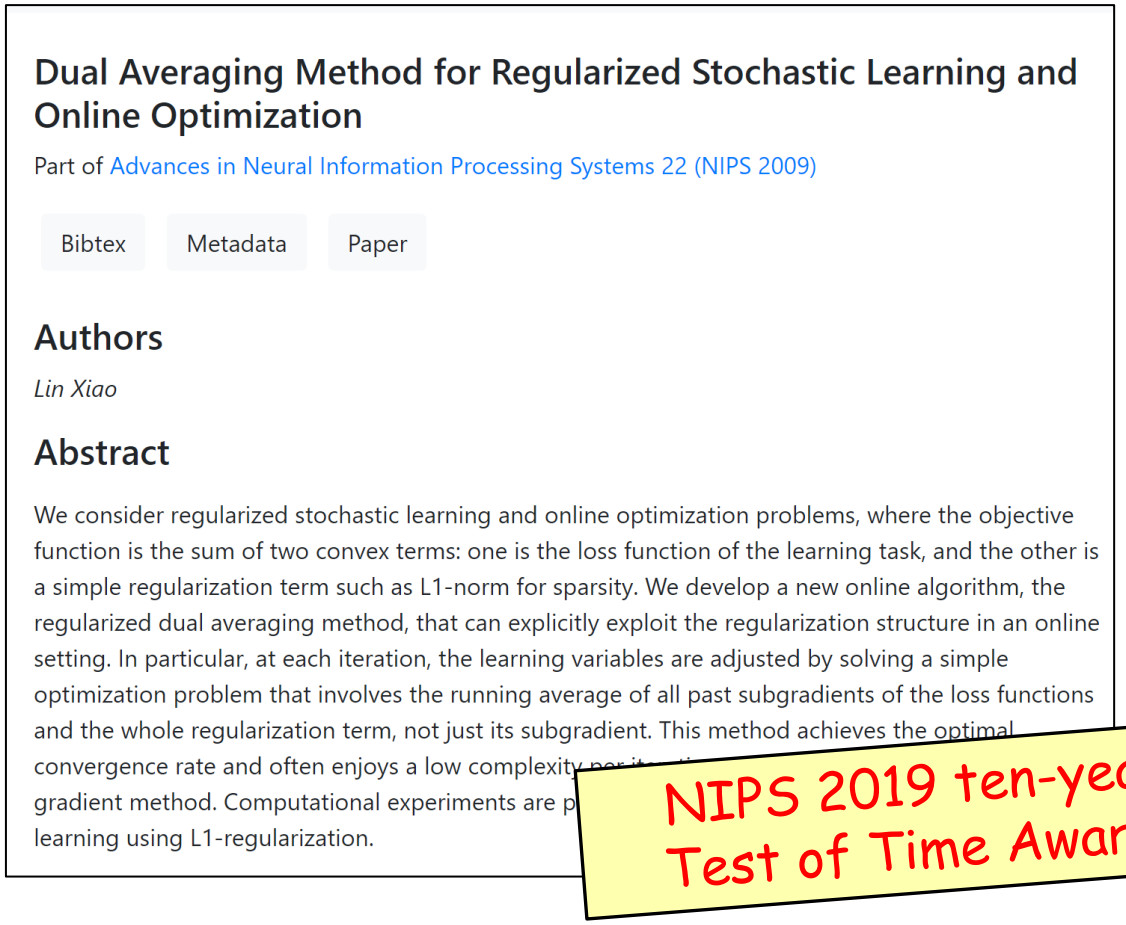
averaging updates in dual space

$$\mathbf{x}_{t+1} = \arg\min_{\mathbf{x} \in \mathcal{X}} \mathcal{D}_\psi(\mathbf{x}, \mathbf{y}_{t+1})$$

$$\Rightarrow \quad \mathbf{x}_{t+1} = \arg\min_{\mathbf{x} \in \mathcal{X}} \left\{ \eta \sum_{s=1}^{t-1} \langle \nabla f_s(\mathbf{x}_s), \mathbf{x} \rangle + \psi(\mathbf{x}) \right\}$$

this is FTRL update (consider fixed step size for simplicity)

FTRL as Dual Averaging



Dual Averaging Method for Regularized Stochastic Learning and Online Optimization

Part of Advances in Neural Information Processing Systems 22 (NIPS 2009)

Bibtex Metadata Paper

Authors

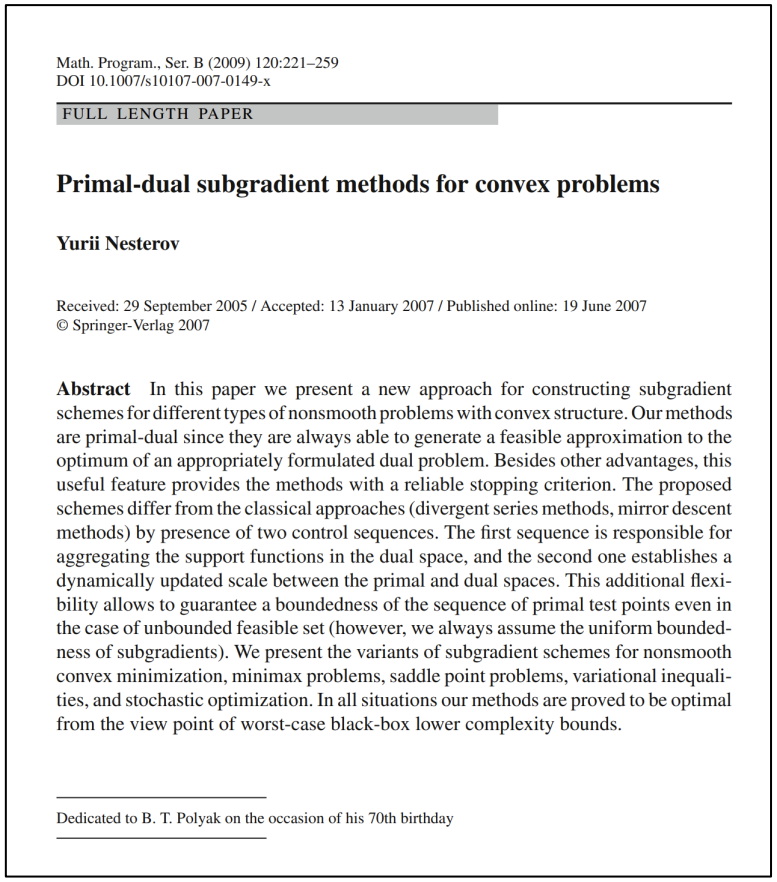
Lin Xiao

Abstract

We consider regularized stochastic learning and online optimization problems, where the objective function is the sum of two convex terms: one is the loss function of the learning task, and the other is a simple regularization term such as L1-norm for sparsity. We develop a new online algorithm, the regularized dual averaging method, that can explicitly exploit the regularization structure in an online setting. In particular, at each iteration, the learning variables are adjusted by solving a simple optimization problem that involves the running average of all past subgradients of the loss functions and the whole regularization term, not just its subgradient. This method achieves the optimal convergence rate and often enjoys a low complexity per
gradient method. Computational experiments are p
learning using L1-regularization.

NIPS 2019 ten-year Test of Time Award!

Lin Xiao. Dual Averaging Method for Regularized Stochastic Learning and Online Optimization. NIPS 2009.



Math. Program., Ser. B (2009) 120:221–259
DOI 10.1007/s10107-007-0149-x

FULL LENGTH PAPER

Primal-dual subgradient methods for convex problems

Yurii Nesterov

Received: 29 September 2005 / Accepted: 13 January 2007 / Published online: 19 June 2007
© Springer-Verlag 2007

Abstract In this paper we present a new approach for constructing subgradient schemes for different types of nonsmooth problems with convex structure. Our methods are primal-dual since they are always able to generate a feasible approximation to the optimum of an appropriately formulated dual problem. Besides other advantages, this useful feature provides the methods with a reliable stopping criterion. The proposed schemes differ from the classical approaches (divergent series methods, mirror descent methods) by presence of two control sequences. The first sequence is responsible for aggregating the support functions in the dual space, and the second one establishes a dynamically updated scale between the primal and dual spaces. This additional flexibility allows to guarantee a boundedness of the sequence of primal test points even in the case of unbounded feasible set (however, we always assume the uniform boundedness of subgradients). We present the variants of subgradient schemes for nonsmooth convex minimization, minimax problems, saddle point problems, variational inequalities, and stochastic optimization. In all situations our methods are proved to be optimal from the view point of worst-case black-box lower complexity bounds.

Dedicated to B. T. Polyak on the occasion of his 70th birthday

Y. Nesterov. Primal-dual subgradient methods for convex problems, 2005.

1 Introduction

1.1 Prehistory

The results presented in this paper are not very new. Most of them were obtained by the author in 2001–2002. However, a further purification of the developed framework led to rather surprising results related to the smoothing technique. Namely, in [11] it was shown that many nonsmooth convex minimization problems with an appropriate

At that moment of time, the author got an illusion that the importance of black-box approach in Convex Optimization will be irreversibly vanishing, and, finally, this approach will be completely replaced by other ones based on a clever use of problem's structure (interior-point methods, smoothing, etc.). This explains why the results included in this paper were not published at time. However, the developments of the last years clearly demonstrated that in some situations the black-box methods are irreplaceable. Indeed, the structure of a convex problem may be too complex for constructing a good self-concordant barrier or for applying a smoothing technique. Note also, that optimization schemes sometimes are employed for modelling certain *adjustment processes* in real-life systems. In this situation, we are not free in selecting the type of optimization scheme and in the choice of its parameters. However, the results on convergence and the rate of convergence of corresponding methods remain interesting.

Yurii Nesterov
1956 –
UCLouvain, Belgium

FTRL vs. OMD

- FTRL and OMD frameworks can recover different OCO methods.

- They share many similarities in both algorithm and regret, but they are ***fundamentally different*** in essence, especially when the step size scheduling is time-varying.

- The dynamics of FTRL and OMD also exhibits great difference when considering beyond static regret minimization, such as in dynamic regret minimization, or repeated game convergence.

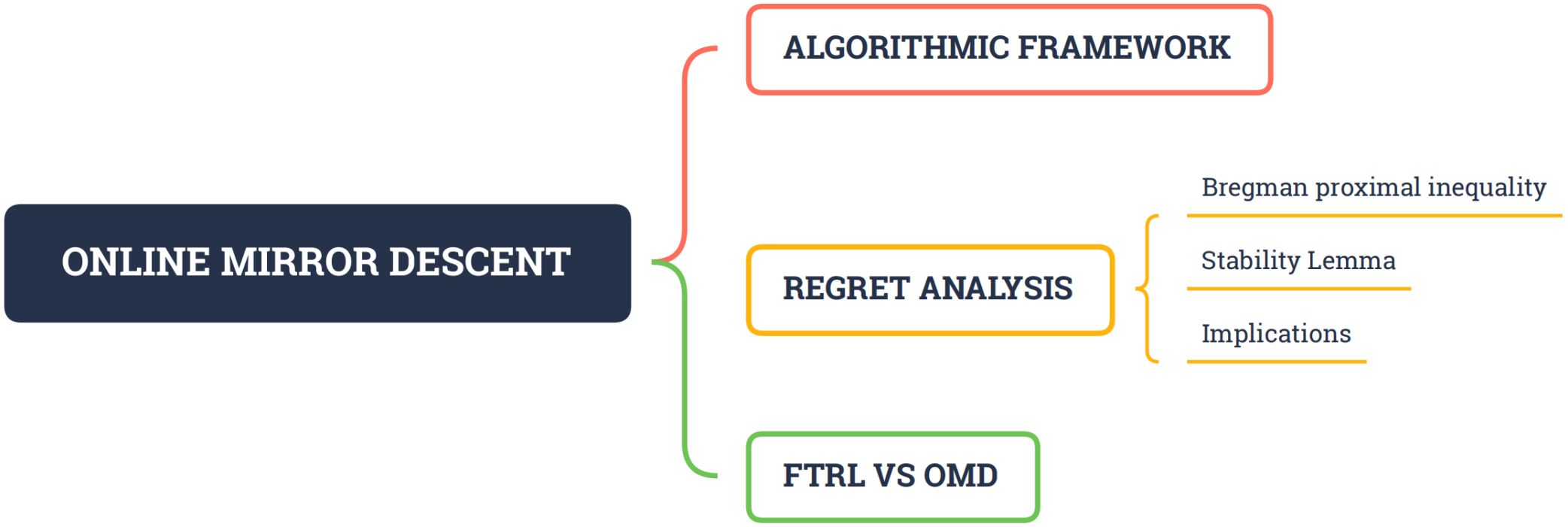
Congrats to Nemirovski and Nesterov

Congrats to WLA Prize (actually)

2023年顶科协奖“智能科学或数学奖”

2023年11月6日

Summary



Q & A

Thanks!