# Online Learning

## — Online-to-Batch Conversion

Joseph Chuang-Chieh Lin

Department of Computer Science & Information Engineering,
Tamkang University

Spring 2023

## Credits for the resource

The slides are based on the lectures of Prof. Luca Trevisan:
https://lucatrevisan.github.io/40391/index.html

the lectures of Prof. Shipra Agrawal:
https://ieor8100.github.io/mab/

the lectures of Prof. Francesco Orabona:
https://parameterfree.com/lecture-notes-on-online-learning/
the monograph: https://arxiv.org/abs/1912.13213

and also Elad Hazan's textbook:
*Introduction to Online Convex Optimization, 2nd Edition.*

# Outline

1. Stochastic Optimization $\Rightarrow$ OCO

2. Example: Binary Classification

# Goal of this Subject

- Reduce stochastic optimization of convex functions to online convex optimization (OCO).

# The Main Theorem

## Theorem (3.1 in the monograph by Prof. Orabona)

- Assume that we are given $F(\mathbf{x}) = \mathbf{E}_{\mathbf{z}\sim\rho(V)}[h(\mathbf{x},\mathbf{z})]$, such that
  - $\mathbf{z}$ is drawn from $\rho$ over a vector space $V$.
  - $h : \mathbb{R}^d \times V \mapsto \mathbb{R}$ is convex w.r.t. the first argument.
- Drawn $T$ samples $\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_T$ i.i.d. from $\rho$ and receive the sequence of losses $f_t(\mathbf{x}) := \alpha_t h(\mathbf{x}, \mathbf{z}_t)$, where $\alpha_t > 0$ are deterministic.
- Run any OCO algorithm over the losses $f_t$ to construct the sequence of predictions $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_{T+1}$.

Then, we have

$$\mathbf{E}\left[F\left(\frac{1}{\sum_{t=1}^{T}\alpha_t}\sum_{t=1}^{T}\alpha_t\mathbf{x}_t\right)\right] \le F(\mathbf{u}) + \frac{\mathbf{E}[\mathrm{Regret}_T(\mathbf{u})]}{\sum_{t=1}^{T}\alpha_t},$$

for any $\mathbf{u} \in \mathbb{R}^d$.

# Proof of the Theorem (1/4)

- Note that we already have $f_t(\mathbf{x}) := \alpha_t h(\mathbf{x}, \mathbf{z}_t)$ and $F(\mathbf{x}) = \mathbf{E}[h(\mathbf{x}, \mathbf{z})]$.

First, we claim that

$$\mathbf{E}\left[\sum_{t=1}^{T} \alpha_t F(\mathbf{x}_t)\right] = \mathbf{E}\left[\sum_{t=1}^{T} f_t(\mathbf{x}_t)\right].$$

# Proof of the Theorem (1/4)

- Note that we already have $f_t(\mathbf{x}) := \alpha_t h(\mathbf{x}, \mathbf{z}_t)$ and $F(\mathbf{x}) = \mathbf{E}[h(\mathbf{x}, \mathbf{z})]$.

First, we claim that

$$\mathbf{E}\left[\sum_{t=1}^{T} \alpha_t F(\mathbf{x}_t)\right] = \mathbf{E}\left[\sum_{t=1}^{T} f_t(\mathbf{x}_t)\right].$$

- By the linearity of expectation:

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

# Proof of the Theorem (1/4)

- Note that we already have $f_t(\mathbf{x}) := \alpha_t h(\mathbf{x}, \mathbf{z}_t)$ and $F(\mathbf{x}) = \mathbf{E}[h(\mathbf{x}, \mathbf{z})]$.

First, we claim that

$$\mathbf{E}\left[\sum_{t=1}^{T} \alpha_t F(\mathbf{x}_t)\right] = \mathbf{E}\left[\sum_{t=1}^{T} f_t(\mathbf{x}_t)\right].$$

- By the linearity of expectation:

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

# Proof of the Theorem (2/4)

- From the law of total expectation, we have

$$\mathbf{E}[f_t(\mathbf{x}_t)] = \mathbf{E}[\mathbf{E}[f_t(\mathbf{x}_t) \mid \mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_{t-1}]]$$

# Proof of the Theorem (2/4)

- From the law of total expectation, we have

$$\begin{aligned}\mathbf{E}[f_t(\mathbf{x}_t)] &= \mathbf{E}[\mathbf{E}[f_t(\mathbf{x}_t) \mid \mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_{t-1}]] \\ &= \mathbf{E}[\mathbf{E}[\alpha_t h(\mathbf{x}_t, \mathbf{z}_t) \mid \mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_{t-1}]]\end{aligned}$$

# Proof of the Theorem (2/4)

- From the law of total expectation, we have

$$\begin{aligned}\mathbf{E}[f_t(\mathbf{x}_t)] &= \mathbf{E}[\mathbf{E}[f_t(\mathbf{x}_t) \mid \mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_{t-1}]] \\ &= \mathbf{E}[\mathbf{E}[\alpha_t h(\mathbf{x}_t, \mathbf{z}_t) \mid \mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_{t-1}]] \\ &= \mathbf{E}[\alpha_t F(\mathbf{x}_t)].\end{aligned}$$

  - Note: $\mathbf{x}_t$ only depends on $\mathbf{z}_1, \ldots, \mathbf{z}_{t-1}$.

# Proof of the Theorem (2/4)

- From the law of total expectation, we have

$$\begin{aligned}\mathbf{E}[f_t(\mathbf{x}_t)] &= \mathbf{E}[\mathbf{E}[f_t(\mathbf{x}_t) \mid \mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_{t-1}]] \\ &= \mathbf{E}[\mathbf{E}[\alpha_t h(\mathbf{x}_t, \mathbf{z}_t) \mid \mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_{t-1}]] \\ &= \mathbf{E}[\alpha_t F(\mathbf{x}_t)].\end{aligned}$$

  - Note: $\mathbf{x}_t$ only depends on $\mathbf{z}_1, \ldots, \mathbf{z}_{t-1}$.

- Thus, we have

$$\mathbf{E}\left[\sum_{t=1}^{T} \alpha_t F(\mathbf{x}_t)\right] = \mathbf{E}\left[\sum_{t=1}^{T} f_t(\mathbf{x}_t)\right].$$

# Proof of the Theorem (3/4)

By Jensen's inequality and the fact that $f$ is convex,

$$F\left(\frac{1}{\sum_{t=1}^{T}\alpha_t}\sum_{t=1}^{T}\alpha_t \mathbf{x}_t\right) \leq$$

# Proof of the Theorem (3/4)

By Jensen's inequality and the fact that $f$ is convex,

$$F\left(\frac{1}{\sum_{t=1}^{T}\alpha_t}\sum_{t=1}^{T}\alpha_t\mathbf{x}_t\right) \le \frac{1}{\sum_{t=1}^{T}\alpha_t}\sum_{t=1}^{T}\alpha_t F(\mathbf{x}_t).$$

# Proof of the Theorem (3/4)

By Jensen's inequality and the fact that $f$ is convex,

$$F\left(\frac{1}{\sum_{t=1}^{T}\alpha_t}\sum_{t=1}^{T}\alpha_t\mathbf{x}_t\right) \le \frac{1}{\sum_{t=1}^{T}\alpha_t}\sum_{t=1}^{T}\alpha_t F(\mathbf{x}_t).$$

So

$$\begin{aligned}
\mathbf{E}\left[F\left(\frac{1}{\sum_{t=1}^{T}\alpha_t}\sum_{t=1}^{T}\alpha_t\mathbf{x}_t\right)\right] &\le \mathbf{E}\left[\frac{1}{\sum_{t=1}^{T}\alpha_t}\sum_{t=1}^{T}\alpha_t F(\mathbf{x}_t)\right] \\
&= \frac{1}{\sum_{t=1}^{T}\alpha_t}\mathbf{E}\left[\sum_{t=1}^{T}\alpha_t F(\mathbf{x}_t)\right] \\
&= \frac{1}{\sum_{t=1}^{T}\alpha_t}\mathbf{E}\left[\sum_{t=1}^{T} f_t(\mathbf{x}_t)\right].
\end{aligned}$$

# Proof of the Theorem (4/4)

To wrap up all of these:

$$\mathbf{E}\left[F\left(\frac{1}{\sum_{t=1}^{T}\alpha_t}\sum_{t=1}^{T}\alpha_t\mathbf{x}_t\right)\right] \leq \frac{1}{\sum_{t=1}^{T}\alpha_t}\mathbf{E}\left[\sum_{t=1}^{T}f_t(\mathbf{x}_t)\right].$$

# Proof of the Theorem (4/4)

To wrap up all of these:

$$\mathbf{E}\left[F\left(\frac{1}{\sum_{t=1}^{T}\alpha_t}\sum_{t=1}^{T}\alpha_t\mathbf{x}_t\right)\right] \leq \frac{1}{\sum_{t=1}^{T}\alpha_t}\mathbf{E}\left[\sum_{t=1}^{T} f_t(\mathbf{x}_t)\right].$$

Together with

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

and

# Proof of the Theorem (4/4)

To wrap up all of these:

$$\mathbf{E}\left[F\left(\frac{1}{\sum_{t=1}^{T}\alpha_t}\sum_{t=1}^{T}\alpha_t \mathbf{x}_t\right)\right] \le \frac{1}{\sum_{t=1}^{T}\alpha_t}\mathbf{E}\left[\sum_{t=1}^{T} f_t(\mathbf{x}_t)\right].$$

Together with

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

and $\mathbf{E}[f_t(\mathbf{u})] = (\sum_{t=1}^{T}\alpha_t)F(\mathbf{u})$, dividing $\sum_{t=1}^{T}\alpha_t$ we can finish the proof of the theorem.

# Outline

# Example: Binary Classification (1/2)

- The inputs: $\mathbf{z}_i \in \mathbb{R}^d$.
- The outputs: $y_i \in \{-1, 1\}$.
- The loss function (hinge loss): $f(\mathbf{x}, (\mathbf{z}, y)) = \max(1 - y\langle \mathbf{z}, \mathbf{x}\rangle, 0)$.
- **Our goal:** Minimize the training error over a training set of $N$ samples: $\{(\mathbf{z}_i, y_i)\}_{i=1}^N$.
  - That is,
$$\min_{\mathbf{x}} F(\mathbf{x}) := \frac{1}{N} \max(1 - y_i\langle \mathbf{z}_i, \mathbf{x}_i\rangle, 0).$$
    and let
$$\mathbf{x}^* = \arg\min_{\mathbf{x}} F(\mathbf{x}).$$
- Additional assumption: the maximum $L_2$-norm of the samples is $D$.

# Example: Binary Classification (2/2)

- Using the reduction in the theorem for $T$ iterations, and use OGD as the OCO algorithm.

# Example: Binary Classification (2/2)

- Using the reduction in the theorem for $T$ iterations, and use OGD as the OCO algorithm.
- Loss in each iteration: $f_t(\mathbf{x}) = \max(1 - y_t\langle \mathbf{z}_t, \mathbf{x}\rangle, 0)$, sampling a training point $(\mathbf{z}_t, y_t)$ uniformly at random from 1 to $N$.
    - Here, $\alpha_t = 1$ for each $t$.

## Example: Binary Classification (2/2)

- Using the reduction in the theorem for $T$ iterations, and use OGD as the OCO algorithm.
- Loss in each iteration: $f_t(\mathbf{x}) = \max(1 - y_t\langle \mathbf{z}_t, \mathbf{x}\rangle, 0)$, sampling a training point $(\mathbf{z}_t, y_t)$ uniformly at random from 1 to $N$.
  - Here, $\alpha_t = 1$ for each $t$.
- Set $\mathbf{x}_1 = \mathbf{0}$ and learning rate $\eta = \frac{1}{D\sqrt{T}}$. We have

$$\mathbf{E}\left[F\left(\frac{1}{T}\sum_{t=1}^{T}\mathbf{x}_t\right)\right] - F(\mathbf{x}^*) \leq D\frac{\|\mathbf{x}^*\|_2^2 + 1}{2\sqrt{T}}.$$

# Remark

- What does the previous example tell us?

# Remark

- What does the previous example tell us?
- We can use an online-convex optimization algorithm to stochastically optimize a function.

# Remark

- What does the previous example tell us?
- We can use an online-convex optimization algorithm to stochastically optimize a function.
  - The regret is transformed into a convergence rate.

# Discussions