Online Learning
Lecture 1

Nicolò Cesa-Bianchi

Università degli Studi di Milano

Contents

Contents

Contents

Contents

Contents

1. Online learning, online convex optimization, Follow-the-Leader (FTL)
2. Follow-the-Regularized-Leader (FTRL), Euclidean (OGD) and entropic (EG) regularization
3. FRTL analysis, regret bounds for OGD and EG
4. Experts, bandits, and feedback graphs
5. Additional topics (parameter-free algorithms, dynamic regret, . . .)

Contents

1. Online learning, online convex optimization, Follow-the-Leader (FTL)
2. Follow-the-Regularized-Leader (FTRL), Euclidean (OGD) and entropic (EG) regularization
3. FRTL analysis, regret bounds for OGD and EG
4. Experts, bandits, and feedback graphs
5. Additional topics (parameter-free algorithms, dynamic regret, . . .)

- We do some (short) proofs

Online learning

- Data streams are ubiquitous: sensors, markets, user interactions

Online learning

- Data streams are ubiquitous: sensors, markets, user interactions
- New data is being generated all the time

Online learning

- Data streams are ubiquitous: sensors, markets, user interactions
- New data is being generated all the time
- The train-test model of statistical learning is not well suited for learning on data streams

Online learning

- Data streams are ubiquitous: sensors, markets, user interactions
- New data is being generated all the time
- The train-test model of statistical learning is not well suited for learning on data streams
- Online learning algorithms incrementally adjust their models after observing each new data point

Some history


- Online learning model formalized by Nick Littlestone and Manfred Warmuth (Mistake bounds and logarithmic linear-threshold learning algorithms, 1989)

Some history


- Online learning model formalized by Nick Littlestone and Manfred Warmuth (Mistake bounds and logarithmic linear-threshold learning algorithms, 1989)
- Volodya Vovk independently develops a related framework (Aggregating strategies, 1990)

Some history

- Online learning model formalized by Nick Littlestone and Manfred Warmuth (Mistake bounds and logarithmic linear-threshold learning algorithms, 1989)
- Volodya Vovk independently develops a related framework (Aggregating strategies, 1990)
- Similar ideas also independently emerged in game theory and information theory:
  - Tom Cover
  - Adrew Barron
  - Rakesh Vohra and Dean Foster
  - Sergiu Hart and Andreu Mas-Colell

The online learning protocol

The algorithm starts with a default model $h_1 \in \mathcal{H}$

For $t = 1, 2, \ldots$

The online learning protocol

The algorithm starts with a default model $h_1 \in \mathcal{H}$

For $t = 1, 2, \dots$

1. The current model $h_t \in \mathcal{H}$ is tested on the next data point $(\boldsymbol{x}_t, y_t)$ in the stream

The online learning protocol

The algorithm starts with a default model $h_1 \in \mathcal{H}$

For $t = 1, 2, \ldots$

1. The current model $h_t \in \mathcal{H}$ is tested on the next data point $(\boldsymbol{x}_t, y_t)$ in the stream
2. A is charged with loss $\ell(y_t, h_t(\boldsymbol{x}_t))$

The online learning protocol

The algorithm starts with a default model $h_1 \in \mathcal{H}$

For $t = 1, 2, \dots$

1. The current model $h_t \in \mathcal{H}$ is tested on the next data point $(\boldsymbol{x}_t, y_t)$ in the stream
2. A is charged with loss $\ell(y_t, h_t(\boldsymbol{x}_t))$
3. $h_{t+1} \in \mathcal{H}$ is computed based on h_t and $(\boldsymbol{x}_t, y_t)$

The online learning protocol

The algorithm starts with a default model $h_1 \in \mathcal{H}$

For $t = 1, 2, \dots$

1. The current model $h_t \in \mathcal{H}$ is tested on the next data point $(\boldsymbol{x}_t, y_t)$ in the stream
2. A is charged with loss $\ell(y_t, h_t(\boldsymbol{x}_t))$
3. $h_{t+1} \in \mathcal{H}$ is computed based on h_t and $(\boldsymbol{x}_t, y_t)$

- Computation of h_{t+1} relies on local information

The online learning protocol

The algorithm starts with a default model $h_1 \in \mathcal{H}$

For $t = 1, 2, \dots$

1. The current model $h_t \in \mathcal{H}$ is tested on the next data point $(\boldsymbol{x}_t, y_t)$ in the stream
2. A is charged with loss $\ell(y_t, h_t(\boldsymbol{x}_t))$
3. $h_{t+1} \in \mathcal{H}$ is computed based on h_t and $(\boldsymbol{x}_t, y_t)$

- Computation of h_{t+1} relies on local information
- No stochastic assumptions on the generation of the data stream!

Regret

Sequential risk

Given a convex loss ℓ and a stream $(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \ldots$, the sequential risk of an online learner A generating models $h_1, h_2, \ldots$ is

$$\sum_{t=1}^{T} \ell(y_t, h_t(\boldsymbol{x}_t))$$

Regret

Sequential risk

Given a convex loss ℓ and a stream $(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \ldots$, the sequential risk of an online learner A generating models $h_1, h_2, \ldots$ is

$$\sum_{t=1}^{T} \ell(y_t, h_t(\boldsymbol{x}_t))$$

Regret: $$R_T = \sum_{t=1}^{T} \ell(y_t, h_t(\boldsymbol{x}_t)) - \inf_{h \in \mathcal{H}} \sum_{t=1}^{T} \ell(y_t, h(\boldsymbol{x}_t))$$

Regret

Sequential risk

Given a convex loss ℓ and a stream $(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \ldots$, the sequential risk of an online learner A generating models $h_1, h_2, \ldots$ is

$$\sum_{t=1}^{T} \ell(y_t, h_t(\boldsymbol{x}_t))$$

Regret: $$R_T = \sum_{t=1}^{T} \ell(y_t, h_t(\boldsymbol{x}_t)) - \inf_{h \in \mathcal{H}} \sum_{t=1}^{T} \ell(y_t, h(\boldsymbol{x}_t))$$

- A sequential counterpart to the estimation error in statistical learning

  $\ell_{\mathcal{D}}(h_S) - \inf_{h \in \mathcal{H}} \ell_{\mathcal{D}}(h)$ where $\ell_{\mathcal{D}}(h) = \mathbb{E}\Big[\ell(Y, h(X))\Big]$ is the statistical risk of h

Regret

Sequential risk

Given a convex loss ℓ and a stream $(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \ldots$, the sequential risk of an online learner A generating models $h_1, h_2, \ldots$ is

$$\sum_{t=1}^{T} \ell\big(y_t, h_t(\boldsymbol{x}_t)\big)$$

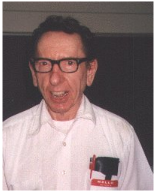
Regret: $$R_T = \sum_{t=1}^{T} \ell\big(y_t, h_t(\boldsymbol{x}_t)\big) - \inf_{h \in \mathcal{H}} \sum_{t=1}^{T} \ell\big(y_t, h(\boldsymbol{x}_t)\big)$$

- A sequential counterpart to the estimation error in statistical learning

  $\ell_{\mathcal{D}}(h_S) - \inf_{h \in \mathcal{H}} \ell_{\mathcal{D}}(h)$ where $\ell_{\mathcal{D}}(h) = \mathbb{E}\Big[\ell\big(Y, h(X)\big)\Big]$ is the statistical risk of h

- Can we ensure $\dfrac{R_T}{T} \to 0$ as $T \to \infty$ for all streams?

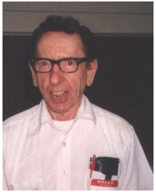
Online learning as a repeated game

Learning to play a game (1956)

- Theory of repeated games pioneered by James Hannan and David Blackwell

Online learning as a repeated game



Learning to play a game (1956)

- Theory of repeated games pioneered by James Hannan and David Blackwell
- Play a game repeatedly against a possibly suboptimal opponent (a.k.a. the data stream)

Zero-sum 2-person games played more than once

	1	2	$\dots$	M
1	$\ell(1,1)$	$\ell(1,2)$	$\dots$	
2	$\ell(2,1)$	$\ell(2,2)$	$\dots$	
$\vdots$	$\vdots$	$\vdots$	$\ddots$	
N				

$N \times M$ known loss matrix

- Row player (player) has N actions
- Column player (opponent) has M actions

Zero-sum 2-person games played more than once

	1	2	$\dots$	M
1	$\ell(1,1)$	$\ell(1,2)$	$\dots$	
2	$\ell(2,1)$	$\ell(2,2)$	$\dots$	
$\vdots$	$\vdots$	$\vdots$	$\ddots$	
N				

$N \times M$ known loss matrix

- Row player (player) has N actions
- Column player (opponent) has M actions

For each game round $t = 1, 2, \dots$

- Player chooses action i_t and opponent chooses action y_t

Zero-sum 2-person games played more than once

	1	2	$\dots$	M
1	$\ell(1,1)$	$\ell(1,2)$	$\dots$	
2	$\ell(2,1)$	$\ell(2,2)$	$\dots$	
$\vdots$	$\vdots$	$\vdots$	$\ddots$	
N				

$N \times M$ known loss matrix

- Row player (player) has N actions
- Column player (opponent) has M actions

For each game round $t = 1, 2, \dots$

- Player chooses action i_t and opponent chooses action y_t
- The player suffers loss $\ell(i_t, y_t)$ (= gain of opponent)

Zero-sum 2-person games played more than once

	1	2	$\dots$	M
1	$\ell(1,1)$	$\ell(1,2)$	$\dots$	
2	$\ell(2,1)$	$\ell(2,2)$	$\dots$	
$\vdots$	$\vdots$	$\vdots$	$\ddots$	
N				

$N \times M$ known loss matrix

- Row player (player) has N actions
- Column player (opponent) has M actions

For each game round $t = 1, 2, \dots$

- Player chooses action i_t and opponent chooses action y_t
- The player suffers loss $\ell(i_t, y_t)$ ($=$ gain of opponent)
- Player can learn from opponent's history of past choices $y_1, \dots, y_{t-1}$

Zero-sum 2-person games played more than once

	1	2	$\dots$	M
1	$\ell(1,1)$	$\ell(1,2)$	$\dots$	
2	$\ell(2,1)$	$\ell(2,2)$	$\dots$	
$\vdots$	$\vdots$	$\vdots$	$\ddots$	
N				

$N \times M$ known loss matrix

- Row player (player) has N actions
- Column player (opponent) has M actions

For each game round $t = 1, 2, \dots$

- Player chooses action i_t and opponent chooses action y_t
- The player suffers loss $\ell(i_t, y_t)$ (= gain of opponent)

- Player can learn from opponent's history of past choices $y_1, \dots, y_{t-1}$
- Replace opponent choices with sequence of loss functions, e.g., $\ell_t = \ell(y_t, \cdot)$

Online convex optimization

Model space $\mathbb{V} \subseteq \mathbb{R}^d$ convex, closed, and nonempty

For $t = 1, 2, \dots$

1. The current $h_t \in \mathcal{H}$ is tested on the next data point $(\boldsymbol{x}_t, y_t)$ in the stream
2. A is charged with loss $\ell(y_t, h_t(\boldsymbol{x}_t))$
3. h_{t+1} is computed based on h_t and $(\boldsymbol{x}_t, y_t)$

Online convex optimization

Model space $\mathbb{V} \subseteq \mathbb{R}^d$ convex, closed, and nonempty

For $t = 1, 2, \ldots$

1. The current $\boldsymbol{w}_t \in \mathbb{V}$ is tested on the next convex loss function ℓ_t in the stream
2. A is charged with loss $\ell(y_t, h_t(\boldsymbol{x}_t))$
3. h_{t+1} is computed based on h_t and $(\boldsymbol{x}_t, y_t)$

Online convex optimization

Model space $\mathbb{V} \subseteq \mathbb{R}^d$ convex, closed, and nonempty

For $t = 1, 2, \ldots$

1. The current $\boldsymbol{w}_t \in \mathbb{V}$ is tested on the next convex loss function ℓ_t in the stream
2. A is charged loss $\ell_t(\boldsymbol{w}_t)$
3. h_{t+1} is computed based on h_t and $(\boldsymbol{x}_t, y_t)$

Online convex optimization

Model space $\mathbb{V} \subseteq \mathbb{R}^d$ convex, closed, and nonempty

For $t = 1, 2, \ldots$

1. The current $\boldsymbol{w}_t \in \mathbb{V}$ is tested on the next convex loss function ℓ_t in the stream
2. A is charged loss $\ell_t(\boldsymbol{w}_t)$
3. $\boldsymbol{w}_{t+1}$ is computed based on $\boldsymbol{w}_t$ and feedback information (e.g., $\nabla \ell_t(\boldsymbol{w}_t)$, first-order oracle)

Online convex optimization

Model space $\mathbb{V} \subseteq \mathbb{R}^d$ convex, closed, and nonempty

For $t = 1, 2, \ldots$

1. The current $\boldsymbol{w}_t \in \mathbb{V}$ is tested on the next convex loss function ℓ_t in the stream
2. A is charged loss $\ell_t(\boldsymbol{w}_t)$
3. $\boldsymbol{w}_{t+1}$ is computed based on $\boldsymbol{w}_t$ and feedback information
   (e.g., $\nabla \ell_t(\boldsymbol{w}_t)$, first-order oracle)

Regret

$$R_T(\boldsymbol{u}) = \sum_{t=1}^{T} \ell_t(\boldsymbol{w}_t) - \sum_{t=1}^{T} \ell_t(\boldsymbol{u}) \qquad \boldsymbol{u} \in \mathbb{V}$$

Online convex optimization

Model space $\mathbb{V} \subseteq \mathbb{R}^d$ convex, closed, and nonempty

For $t = 1, 2, \ldots$

1. The current $\boldsymbol{w}_t \in \mathbb{V}$ is tested on the next convex loss function ℓ_t in the stream
2. A is charged loss $\ell_t(\boldsymbol{w}_t)$
3. $\boldsymbol{w}_{t+1}$ is computed based on $\boldsymbol{w}_t$ and feedback information
   (e.g., $\nabla \ell_t(\boldsymbol{w}_t)$, first-order oracle)

Regret

$$R_T = \sum_{t=1}^{T} \ell_t(\boldsymbol{w}_t) - \inf_{\boldsymbol{u} \in \mathbb{V}} \sum_{t=1}^{T} \ell_t(\boldsymbol{u})$$

Stochastic optimization

Online convex optimization can be used to minimize the training error

$$\inf_{\boldsymbol{w} \in \mathbb{V}} \sum_{i=1}^{m} \ell(\boldsymbol{w}, (\boldsymbol{x}_i, y_i))$$

$\ell(\boldsymbol{w}, (\boldsymbol{x}_i, y_i))$ measures the (convex) loss of $\boldsymbol{w}$ on the training example $(\boldsymbol{x}_i, y_i)$

Stochastic optimization

Online convex optimization can be used to minimize the training error

$$\inf_{\boldsymbol{w} \in \mathbb{V}} \sum_{i=1}^{m} \ell(\boldsymbol{w}, (\boldsymbol{x}_i, y_i))$$

$\ell(\boldsymbol{w}, (\boldsymbol{x}_i, y_i))$ measures the (convex) loss of $\boldsymbol{w}$ on the training example $(\boldsymbol{x}_i, y_i)$

- When m is large we cannot afford to spend more than constant time on each data point

Stochastic optimization

Online convex optimization can be used to minimize the training error

$$\inf_{\boldsymbol{w} \in \mathbb{V}} \sum_{i=1}^{m} \ell(\boldsymbol{w}, (\boldsymbol{x}_i, y_i))$$

$\ell(\boldsymbol{w}, (\boldsymbol{x}_i, y_i))$ measures the (convex) loss of $\boldsymbol{w}$ on the training example $(\boldsymbol{x}_i, y_i)$

- When m is large we cannot afford to spend more than constant time on each data point
- Stochastic optimization:

Stochastic optimization

Online convex optimization can be used to minimize the training error

$$\inf_{\boldsymbol{w}\in\mathbb{V}} \sum_{i=1}^{m} \ell(\boldsymbol{w},(\boldsymbol{x}_i,y_i))$$

$\ell(\boldsymbol{w},(\boldsymbol{x}_i,y_i))$ measures the (convex) loss of $\boldsymbol{w}$ on the training example $(\boldsymbol{x}_i,y_i)$

- When m is large we cannot afford to spend more than constant time on each data point
- Stochastic optimization:
  1. Draw $(\boldsymbol{X}_1,Y_1),(\boldsymbol{X}_2,Y_2)\dots$ uniformly i.i.d. from the training set

Stochastic optimization

Online convex optimization can be used to minimize the training error

$$\inf_{\boldsymbol{w} \in \mathbb{V}} \sum_{i=1}^{m} \ell(\boldsymbol{w}, (\boldsymbol{x}_i, y_i))$$

$\ell(\boldsymbol{w}, (\boldsymbol{x}_i, y_i))$ measures the (convex) loss of $\boldsymbol{w}$ on the training example $(\boldsymbol{x}_i, y_i)$

- When m is large we cannot afford to spend more than constant time on each data point
- Stochastic optimization:
  1. Draw $(\boldsymbol{X}_1, Y_1), (\boldsymbol{X}_2, Y_2) \dots$ uniformly i.i.d. from the training set
  2. Run online algorithm on the sequence of loss functions $\ell_t = \ell(\cdot, (\boldsymbol{X}_t, Y_t))$

Follow the Leader

- Predict using the best model on previous data: $\boldsymbol{w}_{t+1} = \operatorname*{argmin}_{\boldsymbol{w} \in \mathbb{V}} \sum_{s=1}^{t} \ell_s(\boldsymbol{w})$

Follow the Leader

- Predict using the best model on previous data: $\boldsymbol{w}_{t+1} = \operatorname*{argmin}_{\boldsymbol{w} \in \mathbb{V}} \sum_{s=1}^{t} \ell_s(\boldsymbol{w})$
- An online version of empirical risk minimization

Follow the Leader

- Predict using the best model on previous data: $$\boldsymbol{w}_{t+1} = \operatorname*{argmin}_{\boldsymbol{w}\in\mathbb{V}} \sum_{s=1}^{t} \ell_s(\boldsymbol{w})$$
- An online version of empirical risk minimization

FTL Lemma

$$R_T = \sum_{t=1}^{T} \ell_t(\boldsymbol{w}_t) - \inf_{\boldsymbol{w}\in\mathbb{V}} \sum_{t=1}^{T} \ell_t(\boldsymbol{w})$$

Follow the Leader

- Predict using the best model on previous data: $\boldsymbol{w}_{t+1} = \operatorname*{argmin}_{\boldsymbol{w}\in\mathbb{V}} \sum_{s=1}^{t} \ell_s(\boldsymbol{w})$
- An online version of empirical risk minimization

FTL Lemma

$$R_T = \sum_{t=1}^{T} \ell_t(\boldsymbol{w}_t) - \inf_{\boldsymbol{w}\in\mathbb{V}} \sum_{t=1}^{T} \ell_t(\boldsymbol{w})$$
$$= \sum_{t=1}^{T} \Big(\ell_t(\boldsymbol{w}_t) - \ell_t(\boldsymbol{w}_{T+1})\Big)$$

Follow the Leader

- Predict using the best model on previous data: $\boldsymbol{w}_{t+1} = \underset{\boldsymbol{w}\in\mathbb{V}}{\operatorname{argmin}} \sum_{s=1}^{t} \ell_s(\boldsymbol{w})$
- An online version of empirical risk minimization

FTL Lemma

$$
\begin{aligned}
R_T &= \sum_{t=1}^{T} \ell_t(\boldsymbol{w}_t) - \inf_{\boldsymbol{w}\in\mathbb{V}} \sum_{t=1}^{T} \ell_t(\boldsymbol{w}) \\
&= \sum_{t=1}^{T} \Big(\ell_t(\boldsymbol{w}_t) - \ell_t(\boldsymbol{w}_{T+1}) \Big) \\
&= \sum_{t=1}^{T} \Big(L_t(\boldsymbol{w}_t) - L_{t-1}(\boldsymbol{w}_t) \Big) - L_T(\boldsymbol{w}_{T+1}) \qquad (L_t = \ell_1 + \cdots + \ell_t,\ L_0 \equiv 0)
\end{aligned}
$$

Follow the Leader

- Predict using the best model on previous data: $\boldsymbol{w}_{t+1} = \operatorname*{argmin}_{\boldsymbol{w} \in \mathbb{V}} \sum_{s=1}^{t} \ell_s(\boldsymbol{w})$
- An online version of empirical risk minimization

FTL Lemma

$$
\begin{aligned}
R_T &= \sum_{t=1}^{T} \ell_t(\boldsymbol{w}_t) - \inf_{\boldsymbol{w} \in \mathbb{V}} \sum_{t=1}^{T} \ell_t(\boldsymbol{w}) \\
&= \sum_{t=1}^{T} \Big(\ell_t(\boldsymbol{w}_t) - \ell_t(\boldsymbol{w}_{T+1}) \Big) \\
&= \sum_{t=1}^{T} \Big(L_t(\boldsymbol{w}_t) - L_{t-1}(\boldsymbol{w}_t) \Big) - L_T(\boldsymbol{w}_{T+1}) \qquad (L_t = \ell_1 + \cdots + \ell_t,\ L_0 \equiv 0) \\
&= \sum_{t=1}^{T} \Big(L_t(\boldsymbol{w}_t) - L_t(\boldsymbol{w}_{t+1}) \Big)
\end{aligned}
$$

Strongly convex losses

- A differentiable $\ell : \mathbb{R}^d \to \mathbb{R}$ is μ-strongly convex on $\mathbb{V}$ with respect to $\|\cdot\|$ if
$$\ell(\boldsymbol{u}) \geq \ell(\boldsymbol{v}) + \nabla\ell(\boldsymbol{v})^\top(\boldsymbol{u}-\boldsymbol{v}) + \frac{\mu}{2}\|\boldsymbol{u}-\boldsymbol{v}\|^2 \qquad \boldsymbol{u},\boldsymbol{v} \in \mathbb{V}$$

Strongly convex losses

- A differentiable $\ell : \mathbb{R}^d \to \mathbb{R}$ is μ-strongly convex on $\mathbb{V}$ with respect to $\|\cdot\|$ if
$$\ell(\boldsymbol{u}) \geq \ell(\boldsymbol{v}) + \nabla \ell(\boldsymbol{v})^\top (\boldsymbol{u} - \boldsymbol{v}) + \frac{\mu}{2} \|\boldsymbol{u} - \boldsymbol{v}\|^2 \qquad \boldsymbol{u}, \boldsymbol{v} \in \mathbb{V}$$
- If ℓ is twice differentiable, then μ-strong convexity is equivalent to requiring that smallest eigenvalue of the Hessian matrix be at least μ

Strongly convex losses

- A differentiable $\ell : \mathbb{R}^d \to \mathbb{R}$ is μ-strongly convex on $\mathbb{V}$ with respect to $\|\cdot\|$ if

$$\ell(\boldsymbol{u}) \geq \ell(\boldsymbol{v}) + \nabla \ell(\boldsymbol{v})^\top (\boldsymbol{u} - \boldsymbol{v}) + \frac{\mu}{2} \|\boldsymbol{u} - \boldsymbol{v}\|^2 \qquad \boldsymbol{u}, \boldsymbol{v} \in \mathbb{V}$$

- If ℓ is twice differentiable, then μ-strong convexity is equivalent to requiring that smallest eigenvalue of the Hessian matrix be at least μ
- The squared Euclidean norm $\frac{1}{2}\|\cdot\|_2^2$ is 1-strongly convex w.r.t. $\|\cdot\|_2$

Strongly convex losses

- A differentiable $\ell : \mathbb{R}^d \to \mathbb{R}$ is μ-strongly convex on $\mathbb{V}$ with respect to $\|\cdot\|$ if
$$\ell(\boldsymbol{u}) \geq \ell(\boldsymbol{v}) + \nabla \ell(\boldsymbol{v})^\top (\boldsymbol{u} - \boldsymbol{v}) + \frac{\mu}{2} \|\boldsymbol{u} - \boldsymbol{v}\|^2 \qquad \boldsymbol{u}, \boldsymbol{v} \in \mathbb{V}$$
- If ℓ is twice differentiable, then μ-strong convexity is equivalent to requiring that smallest eigenvalue of the Hessian matrix be at least μ
- The squared Euclidean norm $\frac{1}{2}\|\cdot\|_2^2$ is 1-strongly convex w.r.t. $\|\cdot\|_2$
- The negative entropy $\sum_i p_i \ln p_i$ is 1-strongly convex w.r.t. $\|\cdot\|_1$ over the probability simplex

First-order optimality for convex functions

Let $f : \mathbb{V} \to \mathbb{R}$ be a differentiable convex function.

$$\boldsymbol{w}^* = \operatorname*{argmin}_{\boldsymbol{w}\in\mathbb{V}} f(\boldsymbol{w}) \quad \text{iff} \quad \nabla f(\boldsymbol{w}^*)^\top (\boldsymbol{w} - \boldsymbol{w}^*) \geq 0 \qquad \boldsymbol{w} \in \mathbb{V}$$

No descent direction inside $\mathbb{V}$

Stability of FTL with strongly convex losses

- For all $t \geq 1$, ℓ_t is μ-strongly convex and G-Lipschitz with respect to $\|\cdot\|$

Stability of FTL with strongly convex losses

- For all $t \geq 1$, ℓ_t is μ-strongly convex and G-Lipschitz with respect to $\|\cdot\|$
- $L_t = \ell_1 + \cdots + \ell_t$ is μt-strongly convex with respect to $\|\cdot\|$ for all $t = 1, \ldots, T$

Stability of FTL with strongly convex losses

- For all $t \geq 1$, ℓ_t is μ-strongly convex and G-Lipschitz with respect to $\|\cdot\|$
- $L_t = \ell_1 + \cdots + \ell_t$ is μt-strongly convex with respect to $\|\cdot\|$ for all $t = 1, \ldots, T$
- FTL prediction: $\boldsymbol{w}_{t+1} = \operatorname*{argmin}_{\boldsymbol{w} \in \mathbb{V}} L_t(\boldsymbol{w})$

Stability of FTL with strongly convex losses

- For all $t \geq 1$, ℓ_t is μ-strongly convex and G-Lipschitz with respect to $\|\cdot\|$
- $L_t = \ell_1 + \cdots + \ell_t$ is μt-strongly convex with respect to $\|\cdot\|$ for all $t = 1, \ldots, T$
- FTL prediction: $\boldsymbol{w}_{t+1} = \operatorname*{argmin}_{\boldsymbol{w} \in \mathbb{V}} L_t(\boldsymbol{w})$

$$L_t(\boldsymbol{w}_t) - L_t(\boldsymbol{w}_{t+1}) \geq \nabla L_t(\boldsymbol{w}_{t+1})^\top (\boldsymbol{w}_t - \boldsymbol{w}_{t+1}) + \frac{\mu t}{2} \|\boldsymbol{w}_t - \boldsymbol{w}_{t+1}\|^2 \geq \frac{\mu t}{2} \|\boldsymbol{w}_t - \boldsymbol{w}_{t+1}\|^2$$

Stability of FTL with strongly convex losses

- For all $t \geq 1$, ℓ_t is μ-strongly convex and G-Lipschitz with respect to $\|\cdot\|$
- $L_t = \ell_1 + \cdots + \ell_t$ is μt-strongly convex with respect to $\|\cdot\|$ for all $t = 1, \ldots, T$
- FTL prediction: $\boldsymbol{w}_{t+1} = \operatorname*{argmin}_{\boldsymbol{w} \in \mathbb{V}} L_t(\boldsymbol{w})$

$$L_t(\boldsymbol{w}_t) - L_t(\boldsymbol{w}_{t+1}) \geq \nabla L_t(\boldsymbol{w}_{t+1})^\top (\boldsymbol{w}_t - \boldsymbol{w}_{t+1}) + \frac{\mu t}{2} \|\boldsymbol{w}_t - \boldsymbol{w}_{t+1}\|^2 \geq \frac{\mu t}{2} \|\boldsymbol{w}_t - \boldsymbol{w}_{t+1}\|^2$$

$$L_t(\boldsymbol{w}_t) - L_t(\boldsymbol{w}_{t+1}) = L_{t-1}(\boldsymbol{w}_t) - L_{t-1}(\boldsymbol{w}_{t+1}) + \ell_t(\boldsymbol{w}_t) - \ell_t(\boldsymbol{w}_{t+1})$$

Stability of FTL with strongly convex losses

- For all $t \geq 1$, ℓ_t is μ-strongly convex and G-Lipschitz with respect to $\|\cdot\|$
- $L_t = \ell_1 + \cdots + \ell_t$ is μt-strongly convex with respect to $\|\cdot\|$ for all $t = 1, \dots, T$
- FTL prediction: $\boldsymbol{w}_{t+1} = \operatorname*{argmin}_{\boldsymbol{w} \in \mathbb{V}} L_t(\boldsymbol{w})$

$$L_t(\boldsymbol{w}_t) - L_t(\boldsymbol{w}_{t+1}) \geq \nabla L_t(\boldsymbol{w}_{t+1})^\top (\boldsymbol{w}_t - \boldsymbol{w}_{t+1}) + \frac{\mu t}{2} \|\boldsymbol{w}_t - \boldsymbol{w}_{t+1}\|^2 \geq \frac{\mu t}{2} \|\boldsymbol{w}_t - \boldsymbol{w}_{t+1}\|^2$$

$$\begin{aligned} L_t(\boldsymbol{w}_t) - L_t(\boldsymbol{w}_{t+1}) &= L_{t-1}(\boldsymbol{w}_t) - L_{t-1}(\boldsymbol{w}_{t+1}) + \ell_t(\boldsymbol{w}_t) - \ell_t(\boldsymbol{w}_{t+1}) \\ &\leq \ell_t(\boldsymbol{w}_t) - \ell_t(\boldsymbol{w}_{t+1}) \qquad \text{(because } \boldsymbol{w}_t \text{ minimizes } L_{t-1}\text{)} \end{aligned}$$

Stability of FTL with strongly convex losses

- For all $t \geq 1$, ℓ_t is μ-strongly convex and G-Lipschitz with respect to $\|\cdot\|$
- $L_t = \ell_1 + \cdots + \ell_t$ is μt-strongly convex with respect to $\|\cdot\|$ for all $t = 1, \ldots, T$
- FTL prediction: $\boldsymbol{w}_{t+1} = \operatorname*{argmin}_{\boldsymbol{w} \in \mathbb{V}} L_t(\boldsymbol{w})$

$$L_t(\boldsymbol{w}_t) - L_t(\boldsymbol{w}_{t+1}) \geq \nabla L_t(\boldsymbol{w}_{t+1})^\top (\boldsymbol{w}_t - \boldsymbol{w}_{t+1}) + \frac{\mu t}{2} \|\boldsymbol{w}_t - \boldsymbol{w}_{t+1}\|^2 \geq \frac{\mu t}{2} \|\boldsymbol{w}_t - \boldsymbol{w}_{t+1}\|^2$$

$$\begin{aligned} L_t(\boldsymbol{w}_t) - L_t(\boldsymbol{w}_{t+1}) &= L_{t-1}(\boldsymbol{w}_t) - L_{t-1}(\boldsymbol{w}_{t+1}) + \ell_t(\boldsymbol{w}_t) - \ell_t(\boldsymbol{w}_{t+1}) \\ &\leq \ell_t(\boldsymbol{w}_t) - \ell_t(\boldsymbol{w}_{t+1}) \qquad \text{(because } \boldsymbol{w}_t \text{ minimizes } L_{t-1}\text{)} \\ &\leq G \|\boldsymbol{w}_t - \boldsymbol{w}_{t+1}\| \end{aligned}$$

Stability of FTL with strongly convex losses

- For all $t \geq 1$, ℓ_t is μ-strongly convex and G-Lipschitz with respect to $\|\cdot\|$
- $L_t = \ell_1 + \cdots + \ell_t$ is μt-strongly convex with respect to $\|\cdot\|$ for all $t = 1, \ldots, T$
- FTL prediction: $\boldsymbol{w}_{t+1} = \operatorname*{argmin}_{\boldsymbol{w} \in \mathbb{V}} L_t(\boldsymbol{w})$

$$L_t(\boldsymbol{w}_t) - L_t(\boldsymbol{w}_{t+1}) \geq \nabla L_t(\boldsymbol{w}_{t+1})^\top (\boldsymbol{w}_t - \boldsymbol{w}_{t+1}) + \frac{\mu t}{2} \|\boldsymbol{w}_t - \boldsymbol{w}_{t+1}\|^2 \geq \frac{\mu t}{2} \|\boldsymbol{w}_t - \boldsymbol{w}_{t+1}\|^2$$

$$\begin{aligned} L_t(\boldsymbol{w}_t) - L_t(\boldsymbol{w}_{t+1}) &= L_{t-1}(\boldsymbol{w}_t) - L_{t-1}(\boldsymbol{w}_{t+1}) + \ell_t(\boldsymbol{w}_t) - \ell_t(\boldsymbol{w}_{t+1}) \\ &\leq \ell_t(\boldsymbol{w}_t) - \ell_t(\boldsymbol{w}_{t+1}) \qquad \text{(because } \boldsymbol{w}_t \text{ minimizes } L_{t-1}\text{)} \\ &\leq G \|\boldsymbol{w}_t - \boldsymbol{w}_{t+1}\| \end{aligned}$$

- Then we have $\|\boldsymbol{w}_t - \boldsymbol{w}_{t+1}\| \leq \dfrac{2G}{\mu t}$

Stability of FTL with strongly convex losses

- For all $t \geq 1$, ℓ_t is μ-strongly convex and G-Lipschitz with respect to $\|\cdot\|$
- $L_t = \ell_1 + \cdots + \ell_t$ is μt-strongly convex with respect to $\|\cdot\|$ for all $t = 1, \ldots, T$
- FTL prediction: $\boldsymbol{w}_{t+1} = \operatorname*{argmin}_{\boldsymbol{w} \in \mathbb{V}} L_t(\boldsymbol{w})$

$$L_t(\boldsymbol{w}_t) - L_t(\boldsymbol{w}_{t+1}) \geq \nabla L_t(\boldsymbol{w}_{t+1})^\top (\boldsymbol{w}_t - \boldsymbol{w}_{t+1}) + \frac{\mu t}{2} \|\boldsymbol{w}_t - \boldsymbol{w}_{t+1}\|^2 \geq \frac{\mu t}{2} \|\boldsymbol{w}_t - \boldsymbol{w}_{t+1}\|^2$$

$$\begin{aligned} L_t(\boldsymbol{w}_t) - L_t(\boldsymbol{w}_{t+1}) &= L_{t-1}(\boldsymbol{w}_t) - L_{t-1}(\boldsymbol{w}_{t+1}) + \ell_t(\boldsymbol{w}_t) - \ell_t(\boldsymbol{w}_{t+1}) \\ &\leq \ell_t(\boldsymbol{w}_t) - \ell_t(\boldsymbol{w}_{t+1}) \qquad \text{(because } \boldsymbol{w}_t \text{ minimizes } L_{t-1}\text{)} \\ &\leq G \|\boldsymbol{w}_t - \boldsymbol{w}_{t+1}\| \end{aligned}$$

- Then we have $\|\boldsymbol{w}_t - \boldsymbol{w}_{t+1}\| \leq \dfrac{2G}{\mu t}$
- Implying $L_t(\boldsymbol{w}_t) - L_t(\boldsymbol{w}_{t+1}) \leq \dfrac{2G^2}{\mu t}$

FTL regret bound

$$R_T = \sum_{t=1}^{T} \Big(L_t(\boldsymbol{w}_t) - L_t(\boldsymbol{w}_{t+1}) \Big)$$

FTL regret bound

$$
\begin{aligned}
R_T &= \sum_{t=1}^{T} \Big(L_t(\boldsymbol{w}_t) - L_t(\boldsymbol{w}_{t+1}) \Big) \\
&\le \sum_{t=1}^{T} \frac{2G^2}{\mu t}
\end{aligned}
$$

FTL regret bound

$$
\begin{aligned}
R_T &= \sum_{t=1}^{T} \Big(L_t(\boldsymbol{w}_t) - L_t(\boldsymbol{w}_{t+1}) \Big) \\
&\le \sum_{t=1}^{T} \frac{2G^2}{\mu t} \\
&\le \frac{2G^2}{\mu}(1 + \ln T)
\end{aligned}
$$

A lower bound for FTL

- What happens if losses have no curvature?

A lower bound for FTL

- What happens if losses have no curvature?
- $\mathbb{V} = [-1, 1]$

A lower bound for FTL

- What happens if losses have no curvature?
- $\mathbb{V} = [-1, 1]$
- $\ell_1(w) = \frac{w}{2}$

A lower bound for FTL

- What happens if losses have no curvature?
- $\mathbb{V} = [-1, 1]$
- $\ell_1(w) = \frac{w}{2}$
- for $t > 1$, $\quad \ell_t(w) = \begin{cases} w & t \text{ is odd} \\ -w & \text{otherwise} \end{cases}$

A lower bound for FTL

- What happens if losses have no curvature?
- $\mathbb{V} = [-1, 1]$
- $\ell_1(w) = \frac{w}{2}$
- for $t > 1$, $\quad \ell_t(w) = \begin{cases} w & t \text{ is odd} \\ -w & \text{otherwise} \end{cases}$
- $\displaystyle\sum_{s=1}^{t} \ell_s(w) = \begin{cases} w/2 & t \text{ is odd} \\ -w/2 & \text{otherwise} \end{cases}$

A lower bound for FTL

- What happens if losses have no curvature?
- $\mathbb{V} = [-1, 1]$
- $\ell_1(w) = \frac{w}{2}$
- for $t > 1$, $\quad \ell_t(w) = \begin{cases} w & t \text{ is odd} \\ -w & \text{otherwise} \end{cases}$
- $\displaystyle\sum_{s=1}^{t} \ell_s(w) = \begin{cases} w/2 & t \text{ is odd} \\ -w/2 & \text{otherwise} \end{cases}$
- FTL prediction at time $t+1$ is $w_{t+1} = \displaystyle\operatorname*{argmin}_{w \in [-1,1]} \sum_{s=1}^{t} \ell_s(w) = \begin{cases} -1 & t \text{ is odd} \\ 1 & \text{otherwise} \end{cases}$

A lower bound for FTL

- What happens if losses have no curvature?
- $\mathbb{V} = [-1, 1]$
- $\ell_1(w) = \frac{w}{2}$
- for $t > 1$, $\quad \ell_t(w) = \begin{cases} w & t \text{ is odd} \\ -w & \text{otherwise} \end{cases}$
- $\displaystyle\sum_{s=1}^{t} \ell_s(w) = \begin{cases} w/2 & t \text{ is odd} \\ -w/2 & \text{otherwise} \end{cases}$
- FTL prediction at time $t+1$ is $w_{t+1} = \displaystyle\operatorname*{argmin}_{w \in [-1,1]} \sum_{s=1}^{t} \ell_s(w) = \begin{cases} -1 & t \text{ is odd} \\ 1 & \text{otherwise} \end{cases}$
- $\ell_{t+1}(w_{t+1}) = 1$ for all $t > 1$, FTL regret grows linearly!

A lower bound for FTL

- What happens if losses have no curvature?
- $\mathbb{V} = [-1, 1]$
- $\ell_1(w) = \frac{w}{2}$
- for $t > 1$, $\quad \ell_t(w) = \begin{cases} w & t \text{ is odd} \\ -w & \text{otherwise} \end{cases}$
- $\displaystyle\sum_{s=1}^{t} \ell_s(w) = \begin{cases} w/2 & t \text{ is odd} \\ -w/2 & \text{otherwise} \end{cases}$
- FTL prediction at time $t+1$ is $w_{t+1} = \displaystyle\operatorname*{argmin}_{w \in [-1,1]} \sum_{s=1}^{t} \ell_s(w) = \begin{cases} -1 & t \text{ is odd} \\ 1 & \text{otherwise} \end{cases}$
- $\ell_{t+1}(w_{t+1}) = 1$ for all $t > 1$, FTL regret grows linearly!
- Best prediction is $w = 0$, zero loss

Follow the Regularized Leader

- If losses lack curvature, FTL is unstable

Follow the Regularized Leader

- If losses lack curvature, FTL is unstable
- We can introduce curvature using a regularizer $\psi : \mathbb{R}^d \to \mathbb{R}$

Follow the Regularized Leader

- If losses lack curvature, FTL is unstable
- We can introduce curvature using a regularizer $\psi : \mathbb{R}^d \to \mathbb{R}$
- $\boldsymbol{w}_{t+1} = \operatorname*{argmin}_{\boldsymbol{w} \in \mathbb{V}} \psi(\boldsymbol{w}) + \sum_{s=1}^{t} \ell_s(\boldsymbol{w})$

Follow the Regularized Leader

- If losses lack curvature, FTL is unstable
- We can introduce curvature using a regularizer $\psi : \mathbb{R}^d \to \mathbb{R}$
- $\boldsymbol{w}_{t+1} = \operatorname*{argmin}_{\boldsymbol{w} \in \mathbb{V}} \psi(\boldsymbol{w}) + \sum_{s=1}^{t} \ell_s(\boldsymbol{w})$
- Example: SVM objective function: $\operatorname*{argmin}_{\boldsymbol{w} \in \mathbb{V}} \frac{\lambda}{2} \|\boldsymbol{w}\|_2^2 + \frac{1}{m} \sum_{t=1}^{m} \ell_t(\boldsymbol{w})$

Follow the Regularized Leader

- If losses lack curvature, FTL is unstable
- We can introduce curvature using a regularizer $\psi : \mathbb{R}^d \to \mathbb{R}$
- $\boldsymbol{w}_{t+1} = \operatorname*{argmin}_{\boldsymbol{w} \in \mathbb{V}} \psi(\boldsymbol{w}) + \sum_{s=1}^{t} \ell_s(\boldsymbol{w})$
- Example: SVM objective function: $\operatorname*{argmin}_{\boldsymbol{w} \in \mathbb{V}} \frac{\lambda}{2} \|\boldsymbol{w}\|_2^2 + \frac{1}{m} \sum_{t=1}^{m} \ell_t(\boldsymbol{w})$
- If ℓ_t are all convex, this is equivalent to FTL over λ-strongly convex losses $\frac{\lambda}{2} \|\cdot\|_2^2 + \ell_t$

Follow the Regularized Leader

- If losses lack curvature, FTL is unstable
- We can introduce curvature using a regularizer $\psi : \mathbb{R}^d \to \mathbb{R}$
- $\boldsymbol{w}_{t+1} = \underset{\boldsymbol{w} \in \mathbb{V}}{\operatorname{argmin}} \, \psi(\boldsymbol{w}) + \sum_{s=1}^{t} \ell_s(\boldsymbol{w})$
- Example: SVM objective function: $\underset{\boldsymbol{w} \in \mathbb{V}}{\operatorname{argmin}} \, \frac{\lambda}{2} \|\boldsymbol{w}\|_2^2 + \frac{1}{m} \sum_{t=1}^{m} \ell_t(\boldsymbol{w})$
- If ℓ_t are all convex, this is equivalent to FTL over λ-strongly convex losses $\frac{\lambda}{2} \|\cdot\|_2^2 + \ell_t$
- How does the regularizer affect regret?