L03
Online Optimization and Learning: Basics

50.579 Optimization for Machine Learning

Ioannis Panageas

ISTD, SUTD

Playing the experts game

Definition. *For each day $t = 1...T$, you have to choose between alternatives A, B (e.g., rain or not rain).*

- *Choose A or B according to some rule.*
- *One of the alternatives realizes.*
- *If you choose correctly you are not penalized otherwise you lose one point.*
- *Imagine that there are n experts who on each day t, recommend either A or B.*

Playing the experts game

Definition. *For each day $t = 1...T$, you have to choose between alternatives A, B (e.g., rain or not rain).*

- *Choose A or B according to some rule.*
- *One of the alternatives realizes.*
- *If you choose correctly you are not penalized otherwise you lose one point.*
- *Imagine that there are n experts who on each day t, recommend either A or B.*

Can you be correct all the time? What is the "right" objective?

Playing the experts game

Definition. *For each day $t = 1...T$, you have to choose between alternatives A, B (e.g., rain or not rain).*

- *Choose A or B according to some rule.*
- *One of the alternatives realizes.*
- *If you choose correctly you are not penalized otherwise you lose one point.*
- *Imagine that there are n experts who on each day t, recommend either A or B.*

Can you be correct all the time? What is the "right" objective?

Perform close to best expert!

Playing the experts game

Algorithm (Weighted Majority). *We define the following algorithm:*

1. Initialize $w_i^0 = 1$ for all $i \in [n]$.
2. **For** t=1 ... T **do**
3. **If** $\sum_{i \text{ choose } A} w_i^{t-1} \geq \sum_{i \text{ choose } B} w_i^{t-1}$
4. **Choose** A, **otherwise** B.
5. **End If**
6. **For** expert i that made a mistake **do**
7. $w_i^t = (1-\epsilon) w_i^{t-1}$.
8. **End For**
9. **For** expert i that did not make a mistake **do**
10. $w_i^t = w_i^{t-1}$.
11. **End For**
12. **End For**

Remarks:

- ϵ is the stepsize (to be chosen later).
- Performs almost as good as ``best" expert (fewest mistakes)

Playing the experts game

Theorem (Weighted Majority). *Let M_T, M_T^B be the total number of mistakes the algorithm and best expert make until step T, respectively. It holds that*

$$M_T \leq 2(1+\epsilon)M_T^B + \frac{\log n}{\epsilon}.$$

Proof. Let's define the potential function $\phi_t = \sum_i w_i^t$.

Playing the experts game

Theorem (Weighted Majority)**.** *Let M_T, M_T^B be the total number of mistakes the algorithm and best expert make until step T, respectively. It holds that*

$$M_T \leq 2(1+\epsilon)M_T^B + \frac{\log n}{\epsilon}.$$

Proof. Let's define the potential function $\phi_t = \sum_i w_i^t$.

- $\phi_0 = n$.
- $\phi_{t+1} \leq \phi_t$ (why?).

Playing the experts game

Theorem (Weighted Majority). *Let M_T, M_T^B be the total number of mistakes the algorithm and best expert make until step T, respectively. It holds that*

$$M_T \leq 2(1+\epsilon)M_T^B + \frac{\log n}{\epsilon}.$$

Proof. Let's define the potential function $\phi_t = \sum_i w_i^t$.

- $\phi_0 = n$.
- $\phi_{t+1} \leq \phi_t$ (why?).

Observe that if we make a mistake at time t then the majority was wrong, that is at least $\frac{\phi_t}{2}$ will be multiplied by $(1-\epsilon)$.

Hence, if we make a mistake then $\phi_{t+1} \leq (1-\epsilon)\frac{\phi_t}{2} + \frac{\phi_t}{2} = (1-\frac{\epsilon}{2})\phi_t$

Playing the experts game

Theorem (Weighted Majority). *Let M_T, M_T^B be the total number of mistakes the algorithm and best expert make until step T, respectively. It holds that*

$$M_T \leq 2(1+\epsilon)M_T^B + \frac{\log n}{\epsilon}.$$

That is $\phi_{t+1} \leq (1-\frac{\epsilon}{2})\phi_t$ when we do a mistake, otherwise just $\phi_{t+1} \leq \phi_t$. Since we have M_T mistakes, then

$$\phi_T \leq \left(1-\frac{\epsilon}{2}\right)^{M_T} \phi_1.$$

Proof. Let's

- ϕ_0
- ϕ_t

Observe that … that is at least $\frac{\phi_t}{2}$ will be multiplied by $(1-\epsilon)$.

Hence, if we make a mistake then $\phi_{t+1} \leq (1-\epsilon)\frac{\phi_t}{2} + \frac{\phi_t}{2} = (1-\frac{\epsilon}{2})\phi_t$

Playing the experts game

Proof cont. Moreover, assuming the best expert (say i^*) did M_T^B mistakes, we have

$$\phi_T > w_{i^*}^T = (1-\epsilon)^{M_T^B}.$$

Playing the experts game

Proof cont. Moreover, assuming the best expert (say i^*) did M_T^B mistakes, we have

$$\phi_T > w_{i^*}^T = (1-\epsilon)^{M_T^B}.$$

We conclude that

$$(1-\epsilon)^{M_T^B} < \left(1 - \frac{\epsilon}{2}\right)^{M_T} n.$$

By taking the log, $M_T^B \log(1-\epsilon) < \log(1-\epsilon/2)M_T + \log n$.

Since $-x - x^2 < \log(1-x) < -x$, $\boxed{M_T^B(-\epsilon - \epsilon^2) < -M_T\epsilon/2 + \log n.}$

Playing the experts game (randomized)

Definition. *For each day $t = 1 \ldots T$, you have to choose between alternatives A, B (e.g., rain or not rain).*

- *Choose A or B with some probability.*
- *One of the alternatives realizes.*
- *If you choose correctly you are not penalized otherwise you lose one point.*
- *Imagine that there are n experts who on each day t, recommend either A or B.*

What is the “right” objective this time?

Playing the experts game (randomized)

Definition. *For each day $t = 1...T$, you have to choose between alternatives A, B (e.g., rain or not rain).*

- *Choose A or B with some probability.*
- *One of the alternatives realizes.*
- *If you choose correctly you are not penalized otherwise you lose one point.*
- *Imagine that there are n experts who on each day t, recommend either A or B.*

What is the "right" objective this time?

Perform in expectation close to best expert!

Playing the experts game (randomized)

Algorithm (Randomized Weighted Majority). *We define the following algorithm:*

1. Initialize $w_i^0 = 1$ for all $i \in [n]$.
2. **For** t=1 ... T **do**
3. **Choose** expert's i recommendation with probability proportional to w_i^{t-1}.
4. **For** expert i that made a mistake **do**
5. $w_i^t = (1-\epsilon) w_i^{t-1}$.
6. **End For**
7. **For** expert i that did not make a mistake **do**
8. $w_i^t = w_i^{t-1}$.
9. **End For**
10. **End For**

Remarks:

- ϵ is the stepsize (to be chosen later).
- Performs almost as good as ``best” expert (fewest mistakes).
- We choose i with probability $p_i^t = \frac{w_i^{t-1}}{\sum_j w_j^{t-1}}$.
- The algorithm is also called Multiplicative Weights Update!

Playing the experts game

Theorem (Weighted Majority)**.** *Let M_T, M_T^B be the total number of mistakes the algorithm and best expert make until step T, respectively. It holds that*

$$\mathbb{E}[M_T] \leq (1+\epsilon)M_T^B + \frac{\log n}{\epsilon}.$$

Proof. Let's define the potential function $\phi_t = \sum_i w_i^t$.

Using the exact same argument, if the best expert (say i^*) did M_T^B mistakes, we have

$$\phi_T > w_{i^*}^T = (1-\epsilon)^{M_T^B}.$$

$$\text{Now } \phi_{t+1} = \sum w_i^{t+1} = \sum w_i^t (1 - \epsilon \mathbf{1}_{i \text{ wrong at } t})$$

Playing the experts game

Theorem (Weighted Majority)**.** *Let M_T, M_T^B be the total number of mistakes the algorithm and best expert make until step T, respectively. It holds that*

$$\mathbb{E}[M_T] \leq (1+\epsilon)M_T^B + \frac{\log n}{\epsilon}.$$

Proof. Let's define the potential function $\phi_t = \sum_i w_i^t$.

Using the exact same argument, if the best expert (say i^*) did M_T^B mistakes, we have

$$\phi_T > w_{i^*}^T = (1-\epsilon)^{M_T^B}.$$

$$\text{Now } \phi_{t+1} = \sum w_i^{t+1} = \sum w_i^t (1 - \epsilon \mathbf{1}_{i \text{ wrong at } t})$$

$$= \sum \phi_t p_i^{t+1} (1 - \epsilon \mathbf{1}_{i \text{ wrong at } t})$$

Playing the experts game

Theorem (Weighted Majority). *Let M_T, M_T^B be the total number of mistakes the algorithm and best expert make until step T, respectively. It holds that*

$$\mathbb{E}[M_T] \leq (1+\epsilon)M_T^B + \frac{\log n}{\epsilon}.$$

Proof. Let's define the potential function $\phi_t = \sum_i w_i^t$.

Using the exact same argument, if the best expert (say i^*) did M_T^B mistakes, we have

$$\phi_T > w_{i^*}^T = (1-\epsilon)^{M_T^B}.$$

$$\text{Now } \phi_{t+1} = \sum w_i^{t+1} = \sum w_i^t (1 - \epsilon \mathbf{1}_{i \text{ wrong at } t})$$

$$= \sum \phi_t p_i^{t+1} (1 - \epsilon \mathbf{1}_{i \text{ wrong at } t})$$

$$= \phi_t \sum p_i^{t+1} (1 - \epsilon \mathbf{1}_{i \text{ wrong at } t})$$

Playing the experts game

Proof cont. Therefore

$$\phi_{t+1} = \phi_t \left(1 - \epsilon \sum_i p_i^{t+1} \mathbf{1}_{i \text{ wrong at } t} \right)$$

$$= \phi_t \sum p_i^{t+1} (1 - \epsilon \mathbf{1}_{i \text{ wrong}})$$

Playing the experts game

Proof cont. Therefore

$$\phi_{t+1} = \phi_t \left(1 - \epsilon \sum_i p_i^{t+1} \mathbf{1}_{i \text{ wrong at } t} \right)$$

$$= \phi_t \sum p_i^{t+1} (1 - \epsilon \mathbf{1}_{i \text{ wrong}})$$

$$= \phi_t (1 - \epsilon \mathbb{E}[\mathbf{1}_{\text{ we made mistake at } t}])$$

Playing the experts game

Proof cont. Therefore

$$\phi_{t+1} = \phi_t \left(1 - \epsilon \sum_i p_i^{t+1} \mathbf{1}_{i \text{ wrong at } t}\right)$$

$$= \phi_t \sum p_i^{t+1} (1 - \epsilon \mathbf{1}_{i \text{ wrong}})$$

$$= \phi_t (1 - \epsilon \mathbb{E}[\mathbf{1}_{\text{ we made mistake at } t}])$$

$$\leq \phi_t e^{-\epsilon \mathbb{E}[\mathbf{1}_{\text{ we made mistake at } t}]}$$

Telescopic product gives

$$\phi_T \leq \phi_1 e^{-\epsilon \mathbb{E}[M_T]}.$$

Playing the experts game

Proof cont. Therefore

$$\phi_{t+1} = \phi_t \left(1 - \epsilon \sum_i p_i^{t+1} \mathbf{1}_{i \text{ wrong at } t}\right)$$

$$= \phi_t \sum p_i^{t+1} (1 - \epsilon \mathbf{1}_{i \text{ wrong}})$$

$$= \phi_t (1 - \epsilon \mathbb{E}[\mathbf{1}_{\text{ we made mistake at } t}])$$

$$\leq \phi_t e^{-\epsilon \mathbb{E}[\mathbf{1}_{\text{ we made mistake at } t}]}$$

Telescopic product gives

$$\phi_T \leq \phi_1 e^{-\epsilon \mathbb{E}[M_T]}.$$

Therefore $(1-\epsilon)^{M_T^B} \leq e^{-\epsilon \mathbb{E}[M_T]} n$, or $M_T^B(-\epsilon - \epsilon^2) \leq \log n - \epsilon \mathbb{E}[M_T]$.

The general setting

Definition. *At each time step* $t = 1 \ldots T$.

- *Player chooses* $x_t \in \mathcal{K} \subset \mathbb{R}^n$ *(some closed convex set).*
- *Adversary chooses* $\ell_t \in \mathcal{F}$ *(set of convex functions).*
- *Player suffers loss* $\ell_t(x_t)$ *and observes feedback.*

The general setting

Definition. *At each time step $t = 1...T$.*

- *Player chooses $x_t \in \mathcal{K} \subset \mathbb{R}^n$ (some closed convex set).*
- *Adversary chooses $\ell_t \in \mathcal{F}$ (set of convex functions).*
- *Player suffers loss $\ell_t(x_t)$ and observes feedback.*

Player's goal is to minimize the (time average) Regret, that is:

$$\frac{1}{T}\left[\sum_{t=1}^{T}\ell_t(x_t) - \min_{u\in\mathcal{K}}\sum_{t=1}^{T}\ell_t(u)\right].$$

If Regret $\rightarrow 0$ as $T \rightarrow \infty$, the algorithm is called no-regret.

Convex optimization as special case

Definition. *At each time step $t = 1 ... T$.*

- *Player chooses $x_t \in \mathcal{K} \subset \mathbb{R}^n$ (some closed convex set).*
- *Adversary chooses same ℓ (convex function).*
- *Player suffers loss $\ell(x_t)$ and observes feedback.*

Player's goal is to minimize the (time average) Regret, that is:

$$\frac{1}{T}\left[\sum_{t=1}^{T} \ell(x_t) - \min_{u \in \mathcal{K}} \sum_{t=1}^{T} \ell(u)\right] \geq \ell\left(\frac{1}{T}\sum_{t=1}^{T} x_t\right) - \ell(x^*).$$

Regret for Experts problem

Player's goal is to minimize the (time average) Regret, that is:

$$\frac{(\mathbb{E}[M_T] - \#\text{mistakes best expert})}{T}.$$

Regret for Experts problem

Player's goal is to minimize the (time average) Regret, that is:

$$\frac{(\mathbb{E}[M_T] - \#\text{mistakes best expert})}{T}.$$

Explanation: We chose x_t the probability distribution at time t over experts and ℓ_t is the probability to do a mistake.

Regret for Experts problem

Player's goal is to minimize the (time average) Regret, that is:

$$\frac{(\mathbb{E}[M_T] - \#\text{mistakes best expert})}{T}.$$

Explanation: We chose x_t the probability distribution at time t over experts and ℓ_t is the probability to do a mistake.

Recall that,

$$\mathbb{E}[M_T] \leq (1+\epsilon)M_T^B + \frac{\log n}{\epsilon}.$$

Choosing $\epsilon = \sqrt{\frac{\log n}{T}}$ gives average regret $2\sqrt{\frac{\log n}{T}}$!

Regret for Experts problem

Player's goal is to minimize the (time average) Regret, that is:

$$\frac{(\mathbb{E}[M_T] - \#\text{mistakes best expert})}{T}.$$

Explanation: We chose x_t the probability distribution at time t over experts and ℓ_t is the probability to do a mistake.

Recall that,

$$\mathbb{E}[M_T] \leq (1+\epsilon)M_T^B + \frac{\log n}{\epsilon}.$$

Choosing $\epsilon = \sqrt{\frac{\log n}{T}}$ gives average regret $2\sqrt{\frac{\log n}{T}}$!

Can we do better?

Regret for Experts problem

Consider just two experts that choose one A and B respectively at all times. The adversary chooses uniformaly at random A or B.

The expected number of mistakes of an online algorithm is $\frac{T}{2}$.

One of the two fixed strategies will have with high probability (say 99%)

$$\frac{T}{2} - \Theta(\sqrt{T}) \text{ mistakes.}$$

Online Gradient Descent

Definition (Online Gradient Descent)**.** *Let $f : \mathbb{R}^n \to \mathbb{R}$ be convex function, differentiable and L-Lipschitz in some compact convex set $\mathcal{X}$ of diameter D. Online GD is defined:*

Initialize at some x_0.
For t:=1 to T do

1. Choose x_t and observe $\ell_t(x_t)$.
2. $y_t = x_t - \alpha_t \nabla \ell_t(x_t)$.
3. $x_{t+1} = \Pi_{\mathcal{X}}(y_t)$.

Regret: $\frac{1}{T}\left(\sum_{t=1}^{T} \ell_t(x_t) - \min_x \sum_{t=1}^{T} \ell_t(x)\right).$

Analysis of Online GD for L-Lipschitz

Theorem (Online Gradient Descent). *Let $f : \mathbb{R}^n \to \mathbb{R}$ be convex function, differentiable and L-Lipschitz in some compact convex set $\mathcal{X}$ of diameter D. It holds*

$$\left(\frac{1}{T}\sum_{t=1}^{T}\ell_t(x_t) - \min_x \sum_{t=1}^{T}\ell_t(x)\right) \le \frac{3}{2}\frac{LD}{\sqrt{T}},$$

with appropriately choosing $\alpha = \frac{D}{L\sqrt{t}}$.

Remarks:

- If we want error ϵ, we need $T = \Theta\left(\frac{L^2D^2}{\epsilon^2}\right)$ iterations (same as GD for L-Lipschitz).

Analysis of Online GD for L-Lipschitz

Proof. Let x^* be the argmin of $\sum \ell_t(x)$.

$$
\begin{aligned}
\ell_t(x_t) - \ell_t(x^*) &\le \nabla \ell_t(x_t)^\top (x_t - x^*) \text{ convexity,} \\
&= \frac{1}{\alpha_t}(x_t - y_t)^\top (x_t - x^*) \text{ definition of GD,}
\end{aligned}
$$

Analysis of Online GD for L-Lipschitz

Proof. Let x^* be the argmin of $\sum \ell_t(x)$.

$$
\begin{aligned}
\ell_t(x_t) - \ell_t(x^*) &\le \nabla \ell_t(x_t)^\top (x_t - x^*) \text{ convexity,} \\
&= \frac{1}{\alpha_t}(x_t - y_t)^\top (x_t - x^*) \text{ definition of GD,} \\
&= \frac{1}{2\alpha_t}\left(\|x_t - x^*\|_2^2 + \|x_t - y_t\|_2^2 - \|y_t - x^*\|_2^2\right) \text{ law of Cosines,} \\
&= \frac{1}{2\alpha_t}\left(\|x_t - x^*\|_2^2 - \|y_t - x^*\|_2^2\right) + \frac{\alpha_t}{2}\|\nabla \ell_t(x_t)\|_2^2 \text{ Def. of } y_t,
\end{aligned}
$$

Analysis of Online GD for L-Lipschitz

Proof. Let x^* be the argmin of $\sum \ell_t(x)$.

$$
\begin{aligned}
\ell_t(x_t) - \ell_t(x^*) &\leq \nabla \ell_t(x_t)^\top (x_t - x^*) \text{ convexity,} \\
&= \frac{1}{\alpha_t}(x_t - y_t)^\top (x_t - x^*) \text{ definition of GD,} \\
&= \frac{1}{2\alpha_t}\left(\|x_t - x^*\|_2^2 + \|x_t - y_t\|_2^2 - \|y_t - x^*\|_2^2\right) \text{ law of Cosines,} \\
&= \frac{1}{2\alpha_t}\left(\|x_t - x^*\|_2^2 - \|y_t - x^*\|_2^2\right) + \frac{\alpha_t}{2}\|\nabla \ell_t(x_t)\|_2^2 \text{ Def. of } y_t, \\
&\leq \frac{1}{2\alpha_t}\left(\|x_t - x^*\|_2^2 - \|y_t - x^*\|_2^2\right) + \frac{\alpha_t L^2}{2} \text{ Lipschitz,} \\
&\leq \frac{1}{2\alpha_t}\left(\|x_t - x^*\|_2^2 - \|x_{t+1} - x^*\|_2^2\right) + \frac{\alpha_t L^2}{2} \text{ projection.}
\end{aligned}
$$

Analysis of Online GD for L-Lipschitz

Proof cont. Since

$$\ell_t(x_t) - \ell_t(x^*) \leq \frac{1}{2\alpha_t}\left(\|x_t - x^*\|_2^2 - \|x_{t+1} - x^*\|_2^2\right) + \frac{\alpha_t L^2}{2},$$

taking the telescopic sum we have

$$\begin{aligned}\sum_{t=1}^{T}(\ell_t(x_t) - \ell_t(x^*)) &\leq \sum_{t=1}^{T}\|x_t - x^*\|_2^2\left(\frac{1}{2\alpha_t} - \frac{1}{2\alpha_{t-1}}\right) + \frac{L^2}{2}\sum_{t=1}^{T}\alpha_t. \\ &\leq \frac{D^2}{2}\sum_{t=1}^{T}\left(\frac{1}{\alpha_t} - \frac{1}{\alpha_{t-1}}\right) + \frac{L^2}{2}\sum_{t=1}^{T}\alpha_t.\end{aligned}$$

Analysis of Online GD for L-Lipschitz

Proof cont. Since

$$\ell_t(x_t) - \ell_t(x^*) \leq \frac{1}{2\alpha_t}\left(\|x_t - x^*\|_2^2 - \|x_{t+1} - x^*\|_2^2\right) + \frac{\alpha_t L^2}{2},$$

taking the telescopic sum we have

$$\begin{aligned}\sum_{t=1}^T (\ell_t(x_t) - \ell_t(x^*)) &\leq \sum_{t=1}^T \|x_t - x^*\|_2^2 \left(\frac{1}{2\alpha_t} - \frac{1}{2\alpha_{t-1}}\right) + \frac{L^2}{2}\sum_{t=1}^T \alpha_t. \\ &\leq \frac{D^2}{2}\sum_{t=1}^T \left(\frac{1}{\alpha_t} - \frac{1}{\alpha_{t-1}}\right) + \frac{L^2}{2}\sum_{t=1}^T \alpha_t. \\ &\leq \frac{D^2}{2\alpha_T} + \frac{L^2}{2}\sum_{t=1}^T \alpha_t \leq \frac{LD}{2}\sqrt{T} + 2\sqrt{T}\frac{LD}{2}.\end{aligned}$$

where we used the fact $\sum \frac{1}{\sqrt{t}} \leq 2\sqrt{T}$ and $\alpha_t = \frac{D}{\sqrt{t}L}$.

Conclusion

- Introduction to Online Optimization and Learning.
  - Experts problem and MWUA.
  - Online GD has rate of convergence $O\left(\frac{1}{\epsilon^2}\right)$ for L-Lipschitz.
  - Next Lecture we will see more about online learning.
- Next week we will talk about non-convex optimization!