Online Learning

— Follow The Regularized Leader (FTRL)

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. Follow The Regularized Leader (FTRL)
   - MWU Revisited
   - FTRL with 2-norm regularizer

Outline

1. Follow The Regularized Leader (FTRL)
   - MWU Revisited
   - FTRL with 2-norm regularizer

Introducing REGULARIZATION

- You might have already been using regularization for quite a long time.

Introducing REGULARIZATION

```
from keras import regularizers
model.add(Dense(64, input_dim=64,
                kernel_regularizer=regularizers.l2(0.01)
```

Introducing REGULARIZATION

```
# l1 data (only 5 informative features)
X_1, y_1 = datasets.make_classification(n_samples=n_samples,
                                        n_features=n_features, n_informative=5,
                                        random_state=1)

# l2 data: non sparse, but less features
y_2 = np.sign(.5 - rnd.rand(n_samples))
X_2 = rnd.randn(n_samples, n_features // 5) + y_2[:, np.newaxis]
X_2 += 5 * rnd.randn(n_samples, n_features // 5)

clf_sets = [(LinearSVC(penalty='l1', loss='squared_hinge', dual=False,
                       tol=1e-3),
             np.logspace(-2.3, -1.3, 10), X_1, y_1),
            (LinearSVC(penalty='l2', loss='squared_hinge', dual=True),
             np.logspace(-4.5, -2, 10), X_2, y_2)]
```

The regularizer

At each step, we compute the solution

$$\boldsymbol{x}_t := \arg\min_{\boldsymbol{x}\in\mathcal{K}} \left(R(\boldsymbol{x}) + \sum_{k=1}^{t-1} f_k(\boldsymbol{x}) \right).$$

This is called Follow the Regularized Leader (FTRL).

In short,

$$\text{FTRL} = \text{FTL} + \text{Regularizer}.$$

Analysis of FTRL

Theorem 3 (Analysis of FTRL)

For

- every sequence of cost function $\{f_t(\cdot)\}_{t\geq 1}$ and
- every regularizer function $R(\cdot)$,

for every $\boldsymbol{x}$, the regret with respect to $\boldsymbol{x}$ after T steps of the FTRL algorithm is bounded as

$$\text{regret}_T(\boldsymbol{x}) \leq \left(\sum_{t=1}^{T} f_t(\boldsymbol{x}_t) - f_t(\boldsymbol{x}_{t+1}) \right) + R(\boldsymbol{x}) - R(\boldsymbol{x}_1),$$

where $\text{regret}_T(\boldsymbol{x}) := \sum_{t=1}^{T} (f_t(\boldsymbol{x}_t) - f_t(\boldsymbol{x}))$.

Proof of Theorem 3

- Consider a *mental* experiment:

Proof of Theorem 3

- Consider a *mental* experiment:
  - We run the FTL algorithm for $T + 1$ steps.
  - The sequence of cost functions: R, f_1, f_2, . . . , f_T.
    - Use $\boldsymbol{x}_1$ as the first solution.
  - The solutions: $\boldsymbol{x}_1$, $\boldsymbol{x}_1$, $\boldsymbol{x}_2$, . . . , $\boldsymbol{x}_T$.

Proof of Theorem 3

- Consider a *mental* experiment:
  - We run the FTL algorithm for $T+1$ steps.
  - The sequence of cost functions: R, f_1, f_2, ..., f_T.
    - Use $\boldsymbol{x}_1$ as the first solution.
  - The solutions: $\boldsymbol{x}_1$, $\boldsymbol{x}_1$, $\boldsymbol{x}_2$, ..., $\boldsymbol{x}_T$.
- The regret:

$$R(\boldsymbol{x}_1) - R(\boldsymbol{x}) + \sum_{t=1}^{T} (f_t(\boldsymbol{x}_t) - f_t(\boldsymbol{x}))$$

Proof of Theorem 3

- Consider a *mental* experiment:
    - We run the FTL algorithm for $T+1$ steps.
    - The sequence of cost functions: $R, f_1, f_2, \ldots, f_T$.
        - Use $\boldsymbol{x}_1$ as the first solution.
    - The solutions: $\boldsymbol{x}_1, \boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_T$.
- The regret:

$$R(\boldsymbol{x}_1) - R(\boldsymbol{x}) + \sum_{t=1}^{T}(f_t(\boldsymbol{x}_t) - f_t(\boldsymbol{x})) \leq R(\boldsymbol{x}_1) - R(\boldsymbol{x}_1) + \sum_{t=1}^{T}(f_t(\boldsymbol{x}_t) - f_t(\boldsymbol{x}_{t+1}))$$

minimizer of $R(\cdot)$

Proof of Theorem 3

- Consider a *mental* experiment:
  - We run the FTL algorithm for $T+1$ steps.
  - The sequence of cost functions: $R, f_1, f_2, \ldots, f_T$.
    - Use $\boldsymbol{x}_1$ as the first solution.
  - The solutions: $\boldsymbol{x}_1, \boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_T$.
- The regret:

$$R(\boldsymbol{x}_1) - R(\boldsymbol{x}) + \sum_{t=1}^{T}(f_t(\boldsymbol{x}_t) - f_t(\boldsymbol{x})) \leq R(\boldsymbol{x}_1) - R(\boldsymbol{x}_1) + \sum_{t=1}^{T}(f_t(\boldsymbol{x}_t) - f_t(\boldsymbol{x}_{t+1}))$$

output of FTRL at $t+1$

Outline

1. Follow The Regularized Leader (FTRL)
   - MWU Revisited
   - FTRL with 2-norm regularizer

Using negative-entropy regularization

- We have seen an example that FTL tends to put all probability mass on one expert (it's bad!)

Using negative-entropy regularization

- We have seen an example that FTL tends to put all probability mass on one expert (it's bad!)
- **Idea:** penalize over "concentralized" distributions.
  - *negative*-entropy: a good measure of how centralized a distribution is.

Using negative-entropy regularization

- We have seen an example that FTL tends to put all probability mass on one expert (it's bad!)
- **Idea:** penalize over "concentralized" distributions.
    - *negative*-entropy: a good measure of how centralized a distribution is.

$$R(\boldsymbol{x}) := c \cdot \sum_{i=1}^{n} \boldsymbol{x}(i) \ln \boldsymbol{x}(i).$$

Using negative-entropy regularization

- We have seen an example that FTL tends to put all probability mass on one expert (it's bad!)
- **Idea:** penalize over "concentralized" distributions.
  - *negative*-entropy: a good measure of how centralized a distribution is.

$$R(\boldsymbol{x}) := c \cdot \sum_{i=1}^{n} \boldsymbol{x}(i) \ln \boldsymbol{x}(i).$$

- So our FTRL gives

$$\boldsymbol{x}_t = \arg\min_{\boldsymbol{x} \in \Delta} \left(\sum_{k=1}^{t-1} \langle \ell_k, \boldsymbol{x} \rangle + c \cdot \sum_{i=1}^{n} \boldsymbol{x}(i) \ln \boldsymbol{x}(i) \right).$$

Using negative entropy regularization

$$\boldsymbol{x}_t = \arg\min_{\boldsymbol{x}\in\Delta} \left(\sum_{k=1}^{t-1} \langle \ell_k, \boldsymbol{x} \rangle + c \cdot \sum_{i=1}^{n} \boldsymbol{x}(i) \ln \boldsymbol{x}(i) \right).$$

- The constraint $\boldsymbol{x} \in \Delta \Rightarrow \sum_i \boldsymbol{x}_i = 1$.
- So we use Lagrange multiplier to solve

$$\mathcal{L} = \left(\sum_{k=1}^{t-1} \langle \ell_k, \boldsymbol{x} \rangle \right) + c \cdot \left(\sum_{i=1}^{n} \boldsymbol{x}(i) \ln \boldsymbol{x}(i) \right) + \lambda \cdot (\langle \boldsymbol{x}, \mathbf{1} \rangle - 1).$$

Using negative entropy regularization

$$\mathbf{x}_t = \arg\min_{\mathbf{x}\in\Delta} \left(\sum_{k=1}^{t-1} \langle \ell_k, \mathbf{x} \rangle + c \cdot \sum_{i=1}^{n} \mathbf{x}(i) \ln \mathbf{x}(i) \right).$$

- The constraint $\mathbf{x} \in \Delta \Rightarrow \sum_i \mathbf{x}_i = 1$.
- So we use Lagrange multiplier to solve

$$\mathcal{L} = \left(\sum_{k=1}^{t-1} \langle \ell_k, \mathbf{x} \rangle \right) + c \cdot \left(\sum_{i=1}^{n} \mathbf{x}(i) \ln \mathbf{x}(i) \right) + \lambda \cdot (\langle \mathbf{x}, \mathbf{1} \rangle - 1).$$

- The partial derivative $\frac{\partial \mathcal{L}}{\partial \mathbf{x}(i)}$:

$$\left(\sum_{k=1}^{t-1} \ell_k(i) \right) + c \cdot (1 + \ln \mathbf{x}_i) + \lambda$$

Rediscover MWU?

$$\frac{\partial \mathcal{L}}{\partial \boldsymbol{x}(i)} = 0 \quad \Rightarrow \quad \boldsymbol{x}(i) = \exp\left(-1 - \frac{\lambda}{c} - \frac{1}{c}\sum_{k=1}^{t-1} \ell_k(i)\right)$$

Rediscover MWU?

$$\frac{\partial \mathcal{L}}{\partial \boldsymbol{x}(i)} = 0 \quad \Rightarrow \quad \boldsymbol{x}(i) = \exp\left(-1 - \frac{\lambda}{c} - \frac{1}{c}\sum_{k=1}^{t-1} \ell_k(i)\right)$$

Take the value of λ to make the solution a probability distribution. Thus,

Rediscover MWU?

$$\frac{\partial \mathcal{L}}{\partial \boldsymbol{x}(i)} = 0 \quad \Rightarrow \quad \boldsymbol{x}(i) = \exp\left(-1 - \frac{\lambda}{c} - \frac{1}{c}\sum_{k=1}^{t-1} \ell_k(i)\right)$$

Take the value of λ to make the solution a probability distribution. Thus,

$$\boldsymbol{x}(i) = \frac{\exp\left(-\frac{1}{c}\sum_{k=1}^{t-1} \ell_k(i)\right)}{\sum_j \exp\left(-\frac{1}{c}\sum_{k=1}^{t-1} \ell_k(j)\right)}.$$

Rediscover MWU?

$$\frac{\partial \mathcal{L}}{\partial \boldsymbol{x}(i)} = 0 \quad \Rightarrow \quad \boldsymbol{x}(i) = \exp\left(-1 - \frac{\lambda}{c} - \frac{1}{c}\sum_{k=1}^{t-1} \ell_k(i)\right)$$

Take the value of λ to make the solution a probability distribution. Thus,

$$\boldsymbol{x}(i) = \frac{\exp\left(-\frac{1}{c}\sum_{k=1}^{t-1} \ell_k(i)\right)}{\sum_j \exp\left(-\frac{1}{c}\sum_{k=1}^{t-1} \ell_k(j)\right)}.$$

Exactly the solution of MWU if we take $c = 1/\beta$!

Rediscover MWU?

$$\frac{\partial \mathcal{L}}{\partial \boldsymbol{x}(i)} = 0 \quad \Rightarrow \quad \boldsymbol{x}(i) = \exp\left(-1 - \frac{\lambda}{c} - \frac{1}{c}\sum_{k=1}^{t-1} \ell_k(i)\right)$$

Take the value of λ to make the solution a probability distribution. Thus,

$$\boldsymbol{x}(i) = \frac{\exp\left(-\frac{1}{c}\sum_{k=1}^{t-1} \ell_k(i)\right)}{\sum_j \exp\left(-\frac{1}{c}\sum_{k=1}^{t-1} \ell_k(j)\right)}.$$

Exactly the solution of MWU if we take $c = 1/\beta$!

- Now it remains to bound the deviation of each step.

Regret of FTRL + Negative-Entropy Regularization

- At each step,
  $f_t(\boldsymbol{x}_t) - f_t(\boldsymbol{x}_{t+1}) = \langle \ell_t, \boldsymbol{x}_t - \boldsymbol{x}_{t+1} \rangle$.
- Let's go back to use the notation of MWU.
  - $\boldsymbol{w}_1(i) = 1$ (initialization).
  - $\boldsymbol{w}_{t+1}(i) = \boldsymbol{w}_t(i) \cdot e^{-\ell_t(i)/c}$.

Regret of FTRL + Negative-Entropy Regularization

- At each step,
  $f_t(\boldsymbol{x}_t) - f_t(\boldsymbol{x}_{t+1}) = \langle \boldsymbol{\ell}_t, \boldsymbol{x}_t - \boldsymbol{x}_{t+1} \rangle$.
- Let's go back to use the notation of MWU.
  - $\boldsymbol{w}_1(i) = 1$ (initialization).
  - $\boldsymbol{w}_{t+1}(i) = \boldsymbol{w}_t(i) \cdot e^{-\ell_t(i)/c}$.
- So, $\boldsymbol{x}_t = \frac{\boldsymbol{w}_t(i)}{\sum_j \boldsymbol{w}_t(j)}$.
- Then,

$$\begin{aligned} \boldsymbol{x}_{t+1}(i) &= \frac{\boldsymbol{w}_{t+1}(i)}{\sum_j \boldsymbol{w}_{t+1}(j)} = \frac{\boldsymbol{w}_t(i)e^{-\ell_t(i)/c}}{\sum_j \boldsymbol{w}_{t+1}(j)} \geq \frac{\boldsymbol{w}_t(i)e^{-\ell_t(i)/c}}{\sum_j \boldsymbol{w}_t(j)} \\ &\geq \boldsymbol{x}_t(i) \cdot e^{-1/c} \geq (1 - 1/c)\boldsymbol{x}_t(i). \end{aligned}$$

Regret of FTRL + Negative-Entropy Regularization

- At each step,
  $f_t(\boldsymbol{x}_t) - f_t(\boldsymbol{x}_{t+1}) = \langle \ell_t, \boldsymbol{x}_t - \boldsymbol{x}_{t+1} \rangle$.
- Let's go back to use the notation of MWU.
  - $\boldsymbol{w}_1(i) = 1$ (initialization).
  - $\boldsymbol{w}_{t+1}(i) = \boldsymbol{w}_t(i) \cdot e^{-\ell_t(i)/c}$.
- So, $\boldsymbol{x}_t = \frac{\boldsymbol{w}_t(i)}{\sum_j \boldsymbol{w}_t(j)}$.
- Then,

$$\begin{aligned} \boldsymbol{x}_{t+1}(i) &= \frac{\boldsymbol{w}_{t+1}(i)}{\sum_j \boldsymbol{w}_{t+1}(j)} = \frac{\boldsymbol{w}_t(i) e^{-\ell_t(i)/c}}{\sum_j \boldsymbol{w}_{t+1}(j)} \geq \frac{\boldsymbol{w}_t(i) e^{-\ell_t(i)/c}}{\sum_j \boldsymbol{w}_t(j)} \\ &\geq \boldsymbol{x}_t(i) \cdot e^{-1/c} \geq (1 - 1/c)\boldsymbol{x}_t(i). \end{aligned}$$

$\because$ weights are non-increasing

Regret of FTRL + Negative-Entropy Regularization

- At each step,
  $f_t(\boldsymbol{x}_t) - f_t(\boldsymbol{x}_{t+1}) = \langle \ell_t, \boldsymbol{x}_t - \boldsymbol{x}_{t+1} \rangle$.
- Let's go back to use the notation of MWU.
  - $\boldsymbol{w}_1(i) = 1$ (initialization).
  - $\boldsymbol{w}_{t+1}(i) = \boldsymbol{w}_t(i) \cdot e^{-\ell_t(i)/c}$.
- So, $\boldsymbol{x}_t = \frac{\boldsymbol{w}_t(i)}{\sum_j \boldsymbol{w}_t(j)}$.
- Then,

$$\begin{aligned} \boldsymbol{x}_{t+1}(i) &= \frac{\boldsymbol{w}_{t+1}(i)}{\sum_j \boldsymbol{w}_{t+1}(j)} = \frac{\boldsymbol{w}_t(i) e^{-\ell_t(i)/c}}{\sum_j \boldsymbol{w}_{t+1}(j)} \geq \frac{\boldsymbol{w}_t(i) e^{-\ell_t(i)/c}}{\sum_j \boldsymbol{w}_t(j)} \\ &\geq \boldsymbol{x}_t(i) \cdot e^{-1/c} \geq (1 - 1/c)\boldsymbol{x}_t(i). \end{aligned}$$

assume $0 \leq \ell_t(i) \leq 1$

Regret of FTRL + Negative-Entropy Regularization

- At each step,
  $f_t(\mathbf{x}_t) - f_t(\mathbf{x}_{t+1}) = \langle \ell_t, \mathbf{x}_t - \mathbf{x}_{t+1} \rangle \leq \sum_i \ell_t(i) \cdot \frac{1}{c}\mathbf{x}_t(i) \leq \frac{1}{c}$.
- Let's go back to use the notation of MWU.
  - $\mathbf{w}_1(i) = 1$ (initialization).
  - $\mathbf{w}_{t+1}(i) = \mathbf{w}_t(i) \cdot e^{-\ell_t(i)/c}$.
- So, $\mathbf{x}_t = \frac{\mathbf{w}_t(i)}{\sum_j \mathbf{w}_t(j)}$.
- Then,

$$\begin{aligned} \mathbf{x}_{t+1}(i) &= \frac{\mathbf{w}_{t+1}(i)}{\sum_j \mathbf{w}_{t+1}(j)} = \frac{\mathbf{w}_t(i)e^{-\ell_t(i)/c}}{\sum_j \mathbf{w}_{t+1}(j)} \geq \frac{\mathbf{w}_t(i)e^{-\ell_t(i)/c}}{\sum_j \mathbf{w}_t(j)} \\ &\geq \mathbf{x}_t(i) \cdot e^{-1/c} \geq (1 - 1/c)\mathbf{x}_t(i). \end{aligned}$$

Regret of FTRL + Negative-Entropy Regularization

- By Theorem 3, for any $\boldsymbol{x}$,

$$\text{regret}_T(\boldsymbol{x}) \le \sum_{t=1}^{T} (f_t(\boldsymbol{x}_t) - f_t(\boldsymbol{x}_{t+1})) + R(\boldsymbol{x}) - R(\boldsymbol{x}_1) \le \frac{T}{c} + c \ln n.$$

Regret of FTRL + Negative-Entropy Regularization

- By Theorem 3, for any $\boldsymbol{x}$,

$$\mathrm{regret}_T(\boldsymbol{x}) \leq \sum_{t=1}^{T} (f_t(\boldsymbol{x}_t) - f_t(\boldsymbol{x}_{t+1})) + R(\boldsymbol{x}) - R(\boldsymbol{x}_1) \leq \frac{T}{c} + c \ln n.$$

$\because$ max entropy for uniform distribution

Regret of FTRL + Negative-Entropy Regularization

- By Theorem 3, for any $\boldsymbol{x}$,

$$\text{regret}_T(\boldsymbol{x}) \leq \sum_{t=1}^{T} (f_t(\boldsymbol{x}_t) - f_t(\boldsymbol{x}_{t+1})) + R(\boldsymbol{x}) - R(\boldsymbol{x}_1) \leq \frac{T}{c} + c \ln n.$$

Again, we have $\text{regret}_T \leq 2\sqrt{T \ln n}$ by choosing $c = \sqrt{\frac{T}{\ln n}}$.

Regret of FTRL + Negative-Entropy Regularization

- By Theorem 3, for any $\boldsymbol{x}$,

$$\text{regret}_T(\boldsymbol{x}) \leq \sum_{t=1}^{T} (f_t(\boldsymbol{x}_t) - f_t(\boldsymbol{x}_{t+1})) + R(\boldsymbol{x}) - R(\boldsymbol{x}_1) \leq \frac{T}{c} + c \ln n.$$

Again, we have $\text{regret}_T \leq 2\sqrt{T \ln n}$ by choosing $c = \sqrt{\frac{T}{\ln n}}$.

  - Note the slight difference b/w regret and regret*.

Outline

1. Follow The Regularized Leader (FTRL)
   - MWU Revisited
   - FTRL with 2-norm regularizer

L2 Regularization

- Let's try to apply the FTRL to the case that the regularizer is of L2 norm!
- Consider also linear cost functions but $\mathcal{K} = \mathbb{R}^n$ first.
- What kind of problem we might encounter?

L2 Regularization

- Let's try to apply the FTRL to the case that the regularizer is of L2 norm!
- Consider also linear cost functions but $\mathcal{K} = \mathbb{R}^n$ first.
- What kind of problem we might encounter?
- The offline optimum could be $-\infty$.
- FTL will also tend to find a solution of "big" size, too.

L2 Regularization

- Let's try to apply the FTRL to the case that the regularizer is of L2 norm!
- Consider also linear cost functions but $\mathcal{K} = \mathbb{R}^n$ first.
- What kind of problem we might encounter?
- The offline optimum could be $-\infty$.
- FTL will also tend to find a solution of "big" size, too.
- To fight this tendency, it makes sense to use a regularizer which penalizes the size of a solution.

$$R(\mathbf{x}) := c||\mathbf{x}||^2.$$

The regularizer of 2-norm tells us...

- $\boldsymbol{x}_1 = \mathbf{0}$.
- $\boldsymbol{x}_{t+1} = \arg\min_{\boldsymbol{x} \in \mathbb{R}^n} c||\boldsymbol{x}||^2 + \sum_{k=1}^{t} \langle \boldsymbol{\ell}_k, \boldsymbol{x} \rangle$.
- Compute the gradient:

$$2c\boldsymbol{x} + \sum_{k=1}^{t} \boldsymbol{\ell}_k = 0$$

$$\Rightarrow \quad \boldsymbol{x} = -\frac{1}{2c} \sum_{k=1}^{t} \boldsymbol{\ell}_k.$$

Hence, $\boldsymbol{x}_1 = \mathbf{0}, \boldsymbol{x}_{t+1} = \boldsymbol{x}_t - \frac{1}{2c}\boldsymbol{\ell}_t$.

The regularizer of 2-norm tells us...

- $\boldsymbol{x}_1 = \mathbf{0}$.
- $\boldsymbol{x}_{t+1} = \arg\min_{\boldsymbol{x}\in\mathbb{R}^n} c||\boldsymbol{x}||^2 + \sum_{k=1}^{t}\langle \boldsymbol{\ell}_k, \boldsymbol{x}\rangle$.
  convex
- Compute the gradient:

$$2c\boldsymbol{x} + \sum_{k=1}^{t} \boldsymbol{\ell}_k = 0$$

$$\Rightarrow \quad \boldsymbol{x} = -\frac{1}{2c}\sum_{k=1}^{t} \boldsymbol{\ell}_k.$$

Hence, $\boldsymbol{x}_1 = \mathbf{0}, \boldsymbol{x}_{t+1} = \boldsymbol{x}_t - \frac{1}{2c}\boldsymbol{\ell}_t$.

The regularizer of 2-norm tells us...

- $\boldsymbol{x}_1 = \mathbf{0}$.
- $\boldsymbol{x}_{t+1} = \arg\min_{\boldsymbol{x} \in \mathbb{R}^n} c||\boldsymbol{x}||^2 + \sum_{k=1}^{t} \langle \boldsymbol{\ell}_k, \boldsymbol{x} \rangle$.

- Compute the gradient:

$$2c\boldsymbol{x} + \sum_{k=1}^{t} \boldsymbol{\ell}_k = 0$$

$$\Rightarrow \quad \boldsymbol{x} = -\frac{1}{2c} \sum_{k=1}^{t} \boldsymbol{\ell}_k.$$

Hence, $\boldsymbol{x}_1 = \mathbf{0}, \boldsymbol{x}_{t+1} = \boldsymbol{x}_t - \frac{1}{2c}\boldsymbol{\ell}_t$.

The regularizer of 2-norm tells us...

- $\boldsymbol{x}_1 = \mathbf{0}$.
- $\boldsymbol{x}_{t+1} = \arg\min_{\boldsymbol{x}\in\mathbb{R}^n} c||\boldsymbol{x}||^2 + \sum_{k=1}^{t}\langle \boldsymbol{\ell}_k, \boldsymbol{x}\rangle$.

- Compute the gradient:

$$2c\boldsymbol{x} + \sum_{k=1}^{t} \boldsymbol{\ell}_k = 0$$

$$\Rightarrow \quad \boldsymbol{x} = -\frac{1}{2c}\sum_{k=1}^{t} \boldsymbol{\ell}_k.$$

Hence, $\boldsymbol{x}_1 = \mathbf{0}, \boldsymbol{x}_{t+1} = \boldsymbol{x}_t - \frac{1}{2c}\boldsymbol{\ell}_t$.
$\rightarrow$ penalize the experts that performed badly in the past!

The regret of FTRL with 2-norm regularization

- First, we have

$$f_t(\boldsymbol{x}_t) - f_t(\boldsymbol{x}_{t+1}) = \langle \boldsymbol{\ell}_t, \boldsymbol{x}_t - \boldsymbol{x}_{t+1} \rangle = \left\langle \boldsymbol{\ell}_t, \frac{1}{2c}\boldsymbol{\ell}_t \right\rangle = \frac{1}{2c}||\boldsymbol{\ell}_t||^2.$$

- So, with respect to a solution $\boldsymbol{x}$,

$$\begin{aligned} \text{regret}_T(\boldsymbol{x}) &\leq R(\boldsymbol{x}) - R(\boldsymbol{x}_1) + \sum_{t=1}^{T} f_t(\boldsymbol{x}_t) - f_t(\boldsymbol{x}_{t+1}) \\ &= c||\boldsymbol{x}||^2 + \frac{1}{2c}\sum_{t=1}^{T}||\boldsymbol{\ell}_t||^2. \end{aligned}$$

- Suppose that $||\boldsymbol{\ell}_t|| \leq L$ for each t and $||\boldsymbol{x}|| \leq D$. Then by optimizing $c = \sqrt{\frac{T}{2D^2L^2}}$, we have

$$\text{regret}_T(\boldsymbol{x}) \leq DL\sqrt{2T}.$$

Dealing with constraints

- Let's deal with the constraint that $\mathcal{K}$ is an arbitrary convex set instead of $\mathbb{R}^n$.
- Using the same regularizer, we have our FTRL which gives

$$\boldsymbol{x}_1 = \arg\min_{\boldsymbol{x}\in\mathcal{K}} c||\boldsymbol{x}||^2,$$

$$\boldsymbol{x}_{t+1} = \arg\min_{\boldsymbol{x}\in\mathcal{K}} c||\boldsymbol{x}||^2 + \sum_{k=1}^{t} \langle \boldsymbol{\ell}_t, \boldsymbol{x} \rangle.$$

Dealing with constraints

- Let's deal with the constraint that $\mathcal{K}$ is an arbitrary convex set instead of $\mathbb{R}^n$.
- Using the same regularizer, we have our FTRL which gives

$$\boldsymbol{x}_1 = \arg\min_{\boldsymbol{x}\in\mathcal{K}} c||\boldsymbol{x}||^2,$$

$$\boldsymbol{x}_{t+1} = \arg\min_{\boldsymbol{x}\in\mathcal{K}} c||\boldsymbol{x}||^2 + \sum_{k=1}^{t} \langle \boldsymbol{\ell}_t, \boldsymbol{x} \rangle.$$

- **The idea:** First solve the unconstrained optimization and then project the solution on K.

Unconstrained optimization + projection

$$\mathbf{y}_{t+1} = \arg\min_{\mathbf{y}\in\mathbb{R}^n} c||\mathbf{y}||^2 + \sum_{k=1}^{t} \langle \ell_t, \mathbf{y} \rangle.$$

$$\mathbf{x}'_{t+1} = \Pi_{\mathcal{K}}(\mathbf{y}_{t+1}) = \arg\min_{\mathbf{x}\in\mathcal{K}} ||\mathbf{x} - \mathbf{y}_{t+1}||.$$

Unconstrained optimization + projection

$$\boldsymbol{y}_{t+1} = \arg\min_{\boldsymbol{y}\in\mathbb{R}^n} c||\boldsymbol{y}||^2 + \sum_{k=1}^{t} \langle \ell_t, \boldsymbol{y} \rangle.$$

$$\boldsymbol{x}'_{t+1} = \Pi_{\mathcal{K}}(\boldsymbol{y}_{t+1}) = \arg\min_{\boldsymbol{x}\in\mathcal{K}} ||\boldsymbol{x} - \boldsymbol{y}_{t+1}||.$$

- **Claim:** $\boldsymbol{x}'_{t+1} = \boldsymbol{x}_{t+1}$.

Proof of the claim: $x'_{t+1} = x_{t+1}$

- First, we already have that $\boldsymbol{y}_{t+1} = -\frac{1}{2c}\sum_{k=1}^{t} \ell_t$.
- Then,

$$\begin{aligned} \boldsymbol{x}'_{t+1} &= \arg\min_{\boldsymbol{x}\in\mathcal{K}} ||\boldsymbol{x} - \boldsymbol{y}_{t+1}|| = \arg\min_{\boldsymbol{x}\in\mathcal{K}} ||\boldsymbol{x} - \boldsymbol{y}_{t+1}||^2 \\ &= \arg\min_{\boldsymbol{x}\in\mathcal{K}} ||\boldsymbol{x}||^2 - 2\langle \boldsymbol{x}, \boldsymbol{y}_{t+1}\rangle + ||\boldsymbol{y}_{t+1}||^2 \end{aligned}$$

Proof of the claim: $\boldsymbol{x}'_{t+1} = \boldsymbol{x}_{t+1}$

- First, we already have that $\boldsymbol{y}_{t+1} = -\frac{1}{2c}\sum_{k=1}^{t} \ell_t$.
- Then,

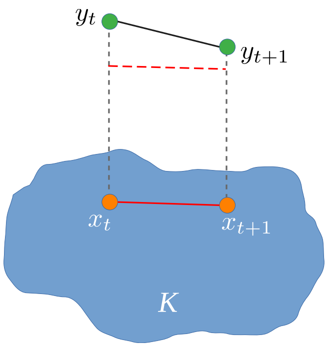
$$
\begin{aligned}
\boldsymbol{x}'_{t+1} &= \arg\min_{\boldsymbol{x}\in\mathcal{K}} ||\boldsymbol{x} - \boldsymbol{y}_{t+1}|| = \arg\min_{\boldsymbol{x}\in\mathcal{K}} ||\boldsymbol{x} - \boldsymbol{y}_{t+1}||^2 \\
&= \arg\min_{\boldsymbol{x}\in\mathcal{K}} ||\boldsymbol{x}||^2 - 2\langle \boldsymbol{x}, \boldsymbol{y}_{t+1}\rangle + ||\boldsymbol{y}_{t+1}||^2 \\
&= \arg\min_{\boldsymbol{x}\in\mathcal{K}} ||\boldsymbol{x}||^2 - 2\langle \boldsymbol{x}, \boldsymbol{y}_{t+1}\rangle \\
&= \arg\min_{\boldsymbol{x}\in\mathcal{K}} ||\boldsymbol{x}||^2 - 2\left\langle \boldsymbol{x}, -\frac{1}{2c}\sum_{k=1}^{t} \ell_t \right\rangle \\
&= \arg\min_{\boldsymbol{x}\in\mathcal{K}} c||\boldsymbol{x}||^2 + \left\langle \boldsymbol{x}, \sum_{k=1}^{t} \ell_t \right\rangle \\
&= \boldsymbol{x}_{t+1}.
\end{aligned}
$$

To bound the regret

$$f_t(\boldsymbol{x}_t) - f_t(\boldsymbol{x}_{t+1}) = \langle \ell_t, \boldsymbol{x}_t - \boldsymbol{x}_{t+1} \rangle \quad \leq \quad ||\ell_t|| \cdot ||\boldsymbol{x}_t - \boldsymbol{x}_{t+1}||$$

To bound the regret

$$
\begin{aligned}
f_t(\boldsymbol{x}_t) - f_t(\boldsymbol{x}_{t+1}) = \langle \ell_t, \boldsymbol{x}_t - \boldsymbol{x}_{t+1} \rangle &\leq ||\ell_t|| \cdot ||\boldsymbol{x}_t - \boldsymbol{x}_{t+1}|| \\
&\leq ||\ell_t|| \cdot ||\boldsymbol{y}_t - \boldsymbol{y}_{t+1}||.
\end{aligned}
$$



To bound the regret

$$
\begin{aligned}
f_t(\boldsymbol{x}_t) - f_t(\boldsymbol{x}_{t+1}) = \langle \ell_t, \boldsymbol{x}_t - \boldsymbol{x}_{t+1} \rangle &\leq ||\ell_t|| \cdot ||\boldsymbol{x}_t - \boldsymbol{x}_{t+1}|| \\
&\leq ||\ell_t|| \cdot ||\boldsymbol{y}_t - \boldsymbol{y}_{t+1}|| \\
&\leq \frac{1}{2c} ||\ell_t||^2.
\end{aligned}
$$

So, assume $\max_{\boldsymbol{x} \in \mathcal{K}} ||\boldsymbol{x}|| \leq D$ and $||\ell_t|| \leq L$ for all t, we have

$$
\begin{aligned}
\text{regret}_T &\leq c||\boldsymbol{x}^*||^2 - c||\boldsymbol{x}_1||^2 + \frac{1}{2c} \sum_{t=1}^{T} ||\ell_t||^2 \\
&\leq cD^2 + \frac{1}{2c} T L^2
\end{aligned}
$$

To bound the regret

$$
\begin{aligned}
f_t(\boldsymbol{x}_t) - f_t(\boldsymbol{x}_{t+1}) = \langle \boldsymbol{\ell}_t, \boldsymbol{x}_t - \boldsymbol{x}_{t+1} \rangle \quad &\leq \quad ||\boldsymbol{\ell}_t|| \cdot ||\boldsymbol{x}_t - \boldsymbol{x}_{t+1}|| \\
&\leq \quad ||\boldsymbol{\ell}_t|| \cdot ||\boldsymbol{y}_t - \boldsymbol{y}_{t+1}|| \\
&\leq \quad \frac{1}{2c} ||\boldsymbol{\ell}_t||^2.
\end{aligned}
$$

So, assume $\max_{\boldsymbol{x} \in \mathcal{K}} ||\boldsymbol{x}|| \leq D$ and $||\boldsymbol{\ell}_t|| \leq L$ for all t, we have

$$
\begin{aligned}
\text{regret}_T \quad &\leq \quad c||\boldsymbol{x}^*||^2 - c||\boldsymbol{x}_1||^2 + \frac{1}{2c} \sum_{t=1}^{T} ||\boldsymbol{\ell}_t||^2 \\
&\leq \quad cD^2 + \frac{1}{2c} TL^2 \leq DL\sqrt{2T}.
\end{aligned}
$$

Discussions