CSIE5410 Optimization algorithms

Lecture 12: Optimistic methods & accelerated gradient method

Yen-Huan Li (yenhuan.li@csie.ntu.edu.tw)

10.01.2019

Department of Computer Science and Information Engineering
National Taiwan University

Abstract

We have seen that the value of a convex-concave game can be computed via no-regret algorithms, and that a sublinear regret can be achieved via mirror descent or follow-the-leader-type methods.

In this lecture, we introduce no-regret optimistic methods, and exploit the results to re-derive Nesterov's accelerated gradient method.

Recommended reading

- H. H. Bauschke and Y. Lucet. 2012. What is a Fenchel conjugate?
- C.-K. Chiang *et al.* 2012. Online optimization with gradual variations.
- A. Rakhlin and K. Sridharan. 2013. Optimization, learning, and games with predictable sequences.
- J.-K. Wang and J. Abernethy. 2018. Acceleration through optimistic no-regret dynamics.

Table of contents

Optimistic FTL-type methods

Learning with a predictable sequence (1/2)

Consider the online optimization problem: Let $L_0 = 0$. For $t = 1, \ldots, T$, the following happen in order.

1. Learner announces $x_t \in \mathcal{X}$.
2. Reality announces $f_t : \mathcal{X} \to \mathbb{R}$.
3. Learner updates $L_t \leftarrow L_{t-1} + f_t(x_t)$.

Definition. The regret is defined as

$$R_T := \sum_{t=1}^{T} f_t(x_t) - \min_x \{ f_{1:T}(x) \mid x \in \mathcal{X} \}.$$

Learning with a predictable sequence (2/2)

Recall that a zero regret can be achieved, if one always does one-step look-ahead.

Lemma (FTL-BTL lemma). Consider the be-the-leader (BTL) approach:

$$x_t \in \arg\min_x \{ (f_{1:t-1} + f_t)(x) \mid x \in \mathcal{X} \}, \quad \forall t \in \mathbb{N}.$$

Then $R_T \leq 0$.

Question. What if we replace f_t with an estimate $\hat{f}_t$ of it?

Question If $\hat{f}_t \approx f_t$ for every t, then will the regret be small?

Optimistic follow-the-leader

Optimistic follow-the-leader. Let $(\hat{f}_t)_{t\in\mathbb{N}}$ be a sequence of functions. Compute

$$x_t \in \arg\min_x \left\{ \left(f_{1:t-1} + \hat{f}_t \right)(x) \,\middle|\, x \in \mathcal{X} \right\}, \quad \forall t \in \mathbb{N},$$

where we define $f_{1:0} := 0$.

Lemma (Optimistic FTL-BTL lemma). Denote by $\tilde{x}_t$ the t-th iterate of the BTL approach. Define $\delta_t := f_t - \hat{f}_t$. Then, the optimistic FTL achieves

$$R_T \leq \sum_{t=1}^{T} \left(\delta_t(x_t) - \delta_t(\tilde{x}_t) \right).$$

Remark. Therefore, the regret is small if $f_t \approx \hat{f}_t$ or $x_t \approx \tilde{x}_t$.

Proof of the proposition (1/2)

Proof. We write

$$
\begin{aligned}
R_T &= \sum_{t=1}^{T} f_t(x_t) - f_{1:T}(\tilde{x}_T) \\
&= \sum_{t=1}^{T-1} f_t(x_t) - f_{1:T-1}(\tilde{x}_T) + f_T(x_T) - f_T(\tilde{x}_T) \\
&= \sum_{t=1}^{T-1} f_t(x_t) - f_{1:T-1}(\tilde{x}_T) + f_T(x_T) - \hat{f}_T(\tilde{x}_T) - \delta_T(\tilde{x}_T) \\
&\leq \sum_{t=1}^{T-1} f_t(x_t) - f_{1:T-1}(x_T) + f_T(x_T) - \hat{f}_T(x_T) - \delta_T(\tilde{x}_T) \\
&\leq R_{T-1} + \delta_T(x_T) - \delta_T(\tilde{x}_T).
\end{aligned}
$$

Proof of the proposition (2/2)

Proof continued. Therefore, we obtain

$$R_T \leq R_1 + \sum_{t=2}^{T} \left[\delta_t(x_t) - \delta_t(\tilde{x}_t)\right].$$

We then bound R_1 as

$$\begin{aligned} R_1 &= f_1(x_1) - f_1(\tilde{x}_1) \\ &= \delta_1(x_1) - \left[f_1(\tilde{x}_1) - \hat{f}_1(x_1)\right] \\ &\leq \delta_1(x_1) - \left[f_1(\tilde{x}_1) - \hat{f}_1(\tilde{x}_1)\right] \\ &= \delta_1(x_1) - \delta_1(\tilde{x}_1). \end{aligned}$$

Optimistic FTL with strongly convex losses

Consider the online convex optimization problem defined by a sequence of convex functions $(f_t)_{t\in\mathbb{N}}$. Let $(\hat{f}_t)_{t\in\mathbb{N}}$ be a sequence of convex functions. Consider the optimistic FTL approach:

$$x_t \in \arg\min_x \left\{ (f_{1:t-1} + \hat{f}_t)(x) \;\middle|\; x \in \mathcal{X} \right\}, \quad \forall t \in \mathbb{N}.$$

Theorem. Suppose that every f_t is μ-strongly convex with respect to a norm $\|\cdot\|$. Then the optimistic FTL achieves

$$R_T \leq \frac{1}{\mu} \sum_{t=1}^{T} \frac{\|\nabla f_t(x_t) - \nabla \hat{f}_t(\tilde{x}_t)\|_*^2}{t}.$$

Proof of the theorem (1/3)

Proof. We start with the optimistic BTL-FTL lemma:

$$
\begin{aligned}
R_T &\leq \sum_{t=1}^{T} \left[\left(f_t - \hat{f}_t \right)(x_t) - \left(f_t - \hat{f}_t \right)(\tilde{x}_t) \right] \\
&= \sum_{t=1}^{T} \left\{ [f_t(x_t) - f_t(\tilde{x}_t)] - \left[\hat{f}_t(x_t) - \hat{f}_t(\tilde{x}_t) \right] \right\} \\
&\leq \sum_{t=1}^{T} \langle \nabla f_t(x_t), x_t - \tilde{x}_t \rangle - \langle \nabla \hat{f}_t(\tilde{x}_t), x_t - \tilde{x}_t \rangle \\
&\leq \sum_{t=1}^{T} \| \nabla f_t(x_t) - \nabla \hat{f}_t(\tilde{x}_t) \|_* \| x_t - \tilde{x}_t \|.
\end{aligned}
$$

Proof of the theorem (2/3)

Proof continued. By the optimality conditions, we have

$$\langle \nabla f_{1:t-1}(x_t) + \nabla \hat{f}_t(x_t), \tilde{x}_t - x_t \rangle \geq 0,$$
$$\langle \nabla f_{1:t-1}(\tilde{x}_t) + \nabla f_t(\tilde{x}_t), x_t - \tilde{x}_t \rangle \geq 0.$$

Summing up the two inequalities, we obtain

$$\langle \nabla f_{1:t}(x_t) - \nabla f_{1:t}(\tilde{x}_t), x_t - \tilde{x}_t \rangle \leq \langle \nabla f_t(x_t) - \nabla \hat{f}_t(x_t), x_t - \tilde{x}_t \rangle .$$

By the monotonicity of the subgradient, we have

$$\langle \nabla \hat{f}_t(x_t) - \nabla \hat{f}_t(\tilde{x}_t), x_t - \tilde{x}_t \rangle \geq 0.$$

Summing up the two inequalities, we obtain

$$\langle \nabla f_{1:t}(x_t) - \nabla f_{1:t}(\tilde{x}_t), x_t - \tilde{x}_t \rangle \leq \langle \nabla f_t(x_t) - \nabla \hat{f}_t(\tilde{x}_t), x_t - \tilde{x}_t \rangle .$$

Proof of the theorem (3/3)

Proof continued. Notice that $f_{1:t}$ is (μt)-strongly convex. By Hölder's inequality, we write

$$\mu t \|x_t - \tilde{x}_t\|^2 \leq \|\nabla f_t(x_t) - \nabla \hat{f}_t(\tilde{x}_t)\|_* \|x_t - \tilde{x}_t\|.$$

That is,

$$\|x_t - \tilde{x}_t\| \leq \frac{1}{\mu t} \|\nabla f_t(x_t) - \nabla \hat{f}_t(\tilde{x}_t)\|_*.$$

Then, we obtain

$$R_T \leq \frac{1}{\mu} \sum_{t=1}^{T} \frac{\|\nabla f_t(x_t) - \nabla \hat{f}_t(\tilde{x}_t)\|_*^2}{t}.$$

Optimistic FTRL

Consider the optimistic FTRL approach:

$$x_1 \in \arg\min_x \left\{ \mu h(x) + \hat{f}_1(x) \;\middle|\; x \in \mathcal{X} \right\},$$
$$x_t \in \arg\min_x \left\{ (f_{1:t-1} + \hat{f}_t)(x) + \mu h(x) \;\middle|\; x \in \mathcal{X} \right\}, \quad \forall t \geq 2,$$

where h is 1-strongly convex with respect to a norm $\|\cdot\|$.

Corollary. The optimistic FTRL achieves

$$R_T \leq \mu R^2 + \frac{1}{\mu} \sum_{t=1}^{T} \|\nabla f_t(x_t) - \nabla \hat{f}_t(\tilde{x}_t)\|_*^2,$$

where

$$R := \sqrt{\max_x \{\, h(x) \mid x \in \mathcal{X} \,\} - h(x_1)}.$$

Proof of the corollary

Proof. Consider the online convex optimization problem defined by the sequence of functions $(f_t)_{t\geq 0}$, with $f_0 := \mu h$. Set $\hat{f}_0 = \mu h$. Consider the optimistic FTL, whose iterates correspond to the iterates of the optimistic FTRL for all $t \in \mathbb{N}$. Then, we write

$$\mu h(x_0) + \sum_{t=1}^{T} f_t(x_t) - \mu h(x) - \sum_{t=1}^{T} f_t(x) \leq \sum_{t=1}^{T} \left[\delta_t(x_t) - \delta_t(\tilde{x}_t)\right],$$

where $\delta_t := f_t - \hat{f}_t$. Notice that $x_0 = x_1$. Then, we obtain

$$R_T \leq \mu R^2 + \sum_{t=1}^{T} \left[\delta_t(x_t) - \delta_t(\tilde{x}_t)\right].$$

The rest of the proof is similar to that of the strongly convex loss case.

Optimistic linearized FTL

Suppose that f_t are convex functions subdifferentiable on $\mathcal{X}$.

Optimistic linearized FTRL. Let $(\hat{g}_t)_{t\in\mathbb{N}}$ be a sequence of vectors in $\mathbb{R}^p$. Compute

$$x_t \in \operatorname*{arg\,min}_x \left\{ \sum_{\tau=1}^{t-1} \langle \nabla f_\tau(x_\tau), x\rangle + \langle \hat{g}_t, x\rangle + \mu h(x) \;\middle|\; x \in \mathcal{X} \right\}, \quad \forall t \in \mathbb{N},$$

where h is 1-strongly convex with respect to a norm $\|\cdot\|$.

Corollary. The optimistic linearized FTRL achieves

$$R_T \le \mu R^2 + \frac{1}{\mu} \sum_{t=1}^{T} \|\nabla f_t(x_t) - \hat{g}_t\|_*^2.$$

A. Rakhlin and K. Sridharan. 2013. Online learning with predictable sequences.

Optimistic lazy OMD

Recall that the linearized FTRL is equivalent to the lazy OMD. Analogously, the optimistic linearized FTRL is equivalent to the optimistic lazy OMD.

Optimistic lazy OMD Let y_1 be such that $\nabla h(y_1) = 0$. Set

$$
\begin{aligned}
x_1 &\in \arg\min_x \{\, \eta \langle \hat{g}_1, x \rangle + D_h(x, y_1) \mid x \in \mathcal{X} \,\}, \\
y_{t+1} &\leftarrow (\nabla h)^{-1} \left(\nabla h(y_t) - \eta \nabla f_t(x_t) \right), \quad \forall t \in \mathbb{N}, \\
x_{t+1} &\leftarrow \arg\min_x \{\, \eta \langle \hat{g}_{t+1}, x \rangle + D_h(x, y_{t+1}) \mid x \in \mathcal{X} \,\}, \quad \forall t \in \mathbb{N}.
\end{aligned}
$$

Proposition. The optimistic OMD is equivalent to the optimistic linearized FTRL, with $\eta = \mu^{-1}$.

Proof of the equivalence

Proof. The optimality condition for x_{t+1} says that

$$\langle \eta \hat{g}_{t+1} + \nabla h(x_{t+1}) - \nabla h(y_{t+1}), x - x_{t+1} \rangle \geq 0, \quad \forall x \in \mathcal{X}.$$

We have seen in the last lecture that

$$\nabla h(y_{t+1}) = -\eta \left(\nabla f_1(x_1) + \cdots + \nabla f_t(x_t) \right),$$

showing that x_{t+1} is also given by

$$x_{t+1} \in \arg\min_x \left\{ \eta \sum_{\tau=1}^{t} \langle \nabla f_\tau(x_\tau), x \rangle + \eta \langle \hat{g}_{t+1}, x \rangle + h(x) \;\middle|\; x \in \mathcal{X} \right\}.$$

Equivalent formulation of optimistic lazy OMD

Notice that the optimality condition shows that y_{t+1} can be written by

$$y_{t+1} \in \arg\min_{y} \{ \eta \langle \nabla f_t(x_t), y \rangle + D_h(y, y_t) \mid y \in \mathbb{R}^p \}.$$

Equivalent formulation of optimistic lazy OMD. Let y_0 be such that $\nabla h(y_0) = 0$.

$$x_t \leftarrow \arg\min_{x} \{ \eta \langle \hat{g}_t, x \rangle + D_h(x, y_{t-1}) \mid x \in \mathcal{X} \},$$

$$y_t \leftarrow \arg\min_{y} \{ \eta \langle \nabla f_t(x_t), y \rangle + D_h(y, y_{t-1}) \mid y \in \mathbb{R}^p \}.$$

Optimistic OMD

Optimistic ~~lazy~~ OMD

Equivalent formulation of optimistic lazy OMD. Let y_0 be a minimizer of h on $\mathbb{R}^p$.

$$x_t \leftarrow \arg\min_x \{ \eta \langle \hat{g}_t, x \rangle + D_h(x, y_{t-1}) \mid x \in \mathcal{X} \},$$

$$y_t \leftarrow \arg\min_y \{ \eta \langle \nabla f_t(x_t), y \rangle + D_h(y, y_{t-1}) \mid y \in \mathbb{R}^p \}.$$

Optimistic OMD. Let y_0 be a minimizer of h on $\mathcal{X}$.

$$x_t \leftarrow \arg\min_x \{ \eta \langle \hat{g}_t, x \rangle + D_h(x, y_{t-1}) \mid x \in \mathcal{X} \},$$

$$y_t \leftarrow \arg\min_y \{ \eta \langle \nabla f_t(x_t), y \rangle + D_h(y, y_{t-1}) \mid y \in \mathcal{X} \}.$$

C.-K. Chiang *et al.* 2012. Online optimization with gradual vairaitons.

A. Rakhlin and K. Sridharan. 2013. Online learning with predictable sequences.

A. Rakhlin and K. Sridharan. 2013. Optimization, learning, and games with predictable sequences.

Interpretations

Optimistic OMD. Let y_0 be a minimizer of h on $\mathcal{X}$.

$$x_t \leftarrow \arg\min_x \{ \eta \langle \hat{g}_t, x \rangle + D_h(x, y_{t-1}) \mid x \in \mathcal{X} \},$$

$$y_t \leftarrow \arg\min_y \{ \eta \langle \nabla f_t(x_t), y \rangle + D_h(y, y_{t-1}) \mid y \in \mathcal{X} \}.$$

- In the off-line setting, the optimistic OMD is Nemirovski's *mirror-prox algorithm*.
- In the online setting, the optimistic OMD is a slightly modified version of optimistic lazy OMD.

A. Nemirovski. 2004. Prox-method with rate of convergence $O(1/t)$ for variational inequalities with Lipschitz continuous monotone operators and smooth convex-concave saddle point problems.

T. Yang *et al.* 2014. Regret bounded by gradual variation for online convex optimization.

Regret bound

Theorem. The optimistic OMD achieves

$$R_T \leq \frac{1}{\eta} R^2 + \sum_{t=1}^{T} \|\nabla f_t(x_t) - \hat{g}_t\|_* \|y_t - x_t\| - \frac{1}{2\eta} \sum_{t=1}^{T} \left(\|y_t - x_t\|^2 + \|y_{t-1} - x_t\|^2 \right),$$

where $R := \sqrt{\max_x \{ D_h(x, y_0) \mid x \in \mathcal{X} \}}$.

Remark. By the same proof, one can prove the same bound for optimistic linearized FTRL.

A. Rakhlin and K. Sridharan. 2013. Optimization, learning, and games with predictable sequences.

V. Syrgkanis. 2015. Fast convergence of regularized learning in games.

Proof of the regret bound (1/2)

Proof. Let $g_t = \nabla f_t(x_t)$. By convexity of f_t, we write

$$\begin{aligned}
&\sum_{t=1}^{T} f_t(x_t) - f_{1:T}(x) \\
&\quad \leq \sum_{t=1}^{T} \langle \nabla f_t(x_t), x_t - x \rangle \\
&\quad = \sum_{t=1}^{T} \langle g_t - \hat{g}_t, x_t - y_t \rangle + \langle \hat{g}_t, x_t - y_t \rangle + \langle g_t, y_t - x \rangle \,.
\end{aligned}$$

By the Bregman proximal inequality, we have

$$\begin{aligned}
\eta \langle \hat{g}_t, x_t - y_t \rangle &\leq D_h(y_t, y_{t-1}) - D_h(y_t, x_t) - D_h(x_t, y_{t-1}), \\
\eta \langle g_t, y_t - x \rangle &\leq D_h(x, y_{t-1}) - D_h(x, y_t) - D_h(y_t, y_{t-1}).
\end{aligned}$$

Proof of the regret bound (2/2)

Proof continued. Then we obtain

$$
\begin{aligned}
&\langle \nabla f_t(x_t), x_t - x \rangle \\
&\quad \leq \|g_t - \hat{g}_t\|_* \|x_t - y_t\| + \frac{1}{\eta} \left[D_h(x, y_{t-1}) - D_h(x, y_t) \right] - \\
&\qquad \frac{1}{\eta} \left[D_h(y_t, x_t) + D_h(x_t, y_{t-1}) \right] \\
&\quad \leq \|g_t - \hat{g}_t\|_* \|x_t - y_t\| + \frac{1}{\eta} \left[D_h(x, y_{t-1}) - D_h(x, y_t) \right] - \\
&\qquad \frac{1}{2\eta} \left[\|y_t - x_t\|^2 + \|x_t - y_{t-1}\|^2 \right]
\end{aligned}
$$

Summing over $t = 1, \ldots, T$, the theorem follows.

Comparison with our previous results

Corollary. The optimistic OMD achieves

$$R_T \leq \frac{R^2}{\eta} + \frac{\eta}{2} \sum_{t=1}^{T} \|\nabla f_t(x_t) - \hat{g}_t\|_*^2.$$

Remark. This is basically the same as our bounds for the optimistic linearized FTRL.

Proof. Notice that for any $a, b \in \mathbb{R}$, we have that $2ab \leq a^2 + b^2$. Therefore, we obtain

$$\|\nabla f_t(x_t) - \hat{g}_t\|_* \|y_t - x_t\| \leq \frac{\eta}{2} \|\nabla f_t(x_t) - \hat{g}_t\|_*^2 + \frac{1}{2\eta} \|y_t - x_t\|^2.$$

Optimized regret bound

Notice that by optimizing over η, the smallest possible regret bound is

$$R_T = O\left(R\sqrt{\sum_{t=1}^{T} \|\nabla f_t(x_t) - \hat{g}_t\|_*^2}\right).$$

However, the optimal η cannot be known before the T-th round.

Corollary. There is an implementable sequence of step sizes $(\eta_t)_{t\in\mathbb{N}}$, such that the regret bound above is achieved.

A. Rakhlin and K. Sridharan. 2013. Optimization, learning, and games with predictable sequences.

Convex optimization as solving a convex-concave game

Fenchel conjugate

Definition. Let $f : \mathbb{R}^p \to [-\infty, +\infty]$. Its *Fenchel conjugate* is given by

$$f^*(y) := \sup_x \{ \langle y, x \rangle - f(x) \mid x \in \mathbb{R}^p \}.$$

Theorem (Fenchel-Young inequality). Let $f : \mathbb{R}^p \to]-\infty, +\infty]$ be proper. Then, it holds that

$$\langle y, x \rangle \leq f^*(y) + f(x), \quad \forall x, y \in \mathbb{R}^p.$$

Example. Therefore, $ab \leq (1/2)a^2 + (1/2)b^2$ for all $a, b \in \mathbb{R}$.

H. H. Bauschke and P. L. Combettes. 2017. *Convex Analysis and Monotone Operator Theory in Hilbert Spaces*.

Properties of the Fenchel conjugate (1/3)

Theorem. The Fenchel conjugate is closed and convex.

Theorem. It holds that $f^{**} \leq f$.

Proof. By the Fenchel-Young inequality, we write

$$f^{**}(x) = \max_y \{ \langle y, x \rangle - f^*(y) \mid y \in \mathbb{R}^p \} \leq f(x).$$

Theorem (Fenchel-Moreau). Let $f : \mathbb{R}^p \to]-\infty, +\infty]$ be proper. Then, $f = f^{**}$, if and only if f is closed and convex.

H. H. Bauschke and P. L. Combettes. 2017. *Convex Analysis and Monotone Operator Theory in Hilbert Spaces*.

Properties of the Fenchel conjugate (2/3)

Theorem. If f is proper closed convex, then $y \in \partial f(x)$ if and only if $x \in \partial f^*(y)$.

Proof. If $y \in \partial f(x)$, then $f^*(y) = \langle y, x \rangle - f(x)$. We write

$$f^*(v) \geq \langle v, x \rangle - f(x) = f^*(y) + \langle v - y, x \rangle .$$

The other direction follows since $f = f^{**}$.

Corollary. Let f be proper closed convex. If ∇f is one-to-one, then $(\nabla f)^{-1} = \nabla f^*$.

H. H. Bauschke and P. L. Combettes. 2017. *Convex Analysis and Monotone Operator Theory in Hilbert Spaces*.

Properties of the Fenchel conjugate (3/3)

Theorem. If f is proper closed convex and L-smooth, then f^* is $(1/L)$-strongly convex.

Proof. Let $y \in \operatorname{dom} \partial f^*$ and $x \in \partial f^*(y)$. For any $y' \in \operatorname{dom} \partial f^*$, we write

$$
\begin{aligned}
& f^*(y') \\
& \quad = \sup_{x'} \left\{ \langle y', x' \rangle - f(x') \;\middle|\; x' \in \mathbb{R}^p \right\} \\
& \quad \geq \sup_{x'} \left\{ \langle y', x' \rangle - f(x) - \langle \nabla f(x), x' - x \rangle - \frac{L}{2} \|x' - x\|_2^2 \;\middle|\; x \in \mathbb{R}^p \right\}
\end{aligned}
$$

R. Goebel and R. T. Rockafellar. 2007. Local strong convexity and local Lipschitz continuity of the gradient of convex functions.

Proof continued.

Proof continued. Notice that $y = \nabla f(x)$ and $f^*(y) = \langle y, x \rangle - f(x)$. Then we obtain

$$
\begin{aligned}
f^*(y') &\geq \langle y, x \rangle - f(x) + \sup_{x'} \left\{ \langle y' - y, x' \rangle - \frac{L}{2} \|x' - x\|_2^2 \;\middle|\; x \in \mathbb{R}^p \right\} \\
&= f^*(y) + \langle y' - y, x \rangle + \\
&\quad \sup_{x'} \left\{ \langle y' - y, x' - x \rangle - \frac{L}{2} \|x' - x\|_2^2 \;\middle|\; x \in \mathbb{R}^p \right\} \\
&\geq f^*(y) + \langle y' - y, x \rangle + \frac{1}{2L} \|y' - y\|_2^2.
\end{aligned}
$$

It remains to recall that $x \in \partial f(y)$.

R. Goebel and R. T. Rockafellar. 2007. Local strong convexity and local Lipschitz continuity of the gradient of convex functions.

Convex optimization as solving a convex-concave game

Consider the convex optimization problem

$$f^\star = \min_x \{\, f(x) \mid x \in \mathcal{X} \,\},$$

for some proper closed convex function f on $\mathbb{R}^p$ and closed convex set $\mathcal{X} \subseteq \mathbb{R}^p$.

By the Fenchel-Moreau theorem, we can equivalently write

$$\begin{aligned} f^\star &= \min_x \{\, f^{**}(x) \mid x \in \mathcal{X} \,\} \\ &= \min_x \max_y \{\, \langle y, x \rangle - f^*(y) \mid x \in \mathcal{X}, y \in \mathbb{R}^p \,\}. \end{aligned}$$

The objective function is convex in x, and concave in y.

Solving the convex-concave game (1/4)

Consider the convex-concave game

$$v = \min_x \max_y \{\, g(x,y) \mid x \in \mathcal{X}, y \in \mathcal{Y} \,\},$$

where $\mathcal{Y} = \mathbb{R}^p$, and $g(x,y) := \langle y, x \rangle - f^*(y)$.

Definition. A pair $(\hat{x}, \hat{y}) \in \mathcal{X} \times \mathcal{Y}$ is called an *ε-equilibrium*, if and only if

$$v - \varepsilon \leq g(x, \hat{y}) \leq v \leq g(\hat{x}, y) \leq v + \varepsilon, \quad \forall (x,y) \in \mathcal{X} \times \mathcal{Y}.$$

Proposition. If $(\hat{x}, \hat{y})$ is an ε-equilibrium, then $f(\hat{x}) \leq f^\star + \varepsilon$.

Solving the convex-concave game (2/4)

Let $(\alpha_t)_{t\in\mathbb{N}}$ be a sequence of positive real numbers.

x-Learner's perspective. Let the initial loss be zero. For every $t \in \mathbb{N}$, the following happen in order.

- x-Learner announces $x_t \in \mathcal{X}$.
- Reality announces $\xi_t : x \mapsto g(x, y_t) = \langle y_t, x\rangle - f^*(y_t)$.
- x-Learner's loss is increased by $\alpha_t \xi_t(x_t)$.

y-Learner's perspective. Let the initial loss be zero. For every $t \in \mathbb{N}$, the following happen in order.

- y-Learner announces $y_t \in \mathcal{Y}$.
- Reality announces $\eta_t : y \mapsto -g(x_t, y) = -\langle y, x_t\rangle + f^*(y)$.
- y-Learner's loss is increased by $\alpha_t \eta_t(y_t)$.

Solving the convex-concave game (3/4)

Let $(\alpha_t)_{t\in\mathbb{N}}$ be a sequence of positive real numbers.

x-Learner's weighted regret.

$$R_T^x := \sum_{t=1}^{T} \alpha_t \xi_t(x_t) - \min_x \left\{ \sum_{t=1}^{T} \alpha_t \xi_t(x) \;\middle|\; x \in \mathcal{X} \right\}.$$

y-Learner's weighted regret.

$$R_T^y := \sum_{t=1}^{T} \alpha_t \eta_t(y_t) - \min_y \left\{ \sum_{t=1}^{T} \alpha_t \eta_t(y) \;\middle|\; y \in \mathcal{Y} \right\}.$$

Solving the convex-concave game (4/4)

Proposition. Let $A_t = \alpha_1 + \cdots + \alpha_t$. Define

$$\hat{x}_T := \frac{1}{A_T} \sum_{t=1}^{T} \alpha_t x_t, \quad \hat{y}_T := \frac{1}{A_T} \sum_{t=1}^{T} \alpha_t y_t.$$

Then $(\hat{x}_T, \hat{y}_T)$ is an ε-equilibrium with

$$\varepsilon = \frac{1}{A_T} \left(R_T^x + R_T^y \right).$$

Proof. Similar to our previous proof with $\alpha_t = 1$ for all t.

J. Abernethy *et al.* 2018. Faster rates for convex-concave games.

An accelerated gradient method (1/4)

Question. What no-regret algorithms should x-Learner and y-Learner choose?

Assumption. The objective function f in the original optimization problem is L-smooth.

y-Learner's choice. Notice that for y-Learner, the losses $\eta_t : y \mapsto -\langle y, x_t \rangle + f^*(y)$ are $(1/L)$-strongly convex. Therefore, y-Learner chooses the *optimistic FTL algorithm* with $\hat{\eta}_t = \eta_{t-1}$.

x-Learner's choice x-Learner chooses the *standard online projected gradient descent*.

J.-K. Wang and J. Abernethy. 2018. Acceleration through optimistic no-regret dynamics.

An accelerated gradient method (2/4)

Resulting algorithm. For every $t = 1, \dots, T$, compute

$$y_t \in \arg\min_y \left\{ \alpha_t \eta_{t-1}(y) + \sum_{\tau=1}^{t-1} \alpha_\tau \eta_\tau(y) \;\middle|\; y \in \mathbb{R}^p \right\},$$

$$x_t \in \arg\min_x \left\{ \gamma_t \left\langle \nabla(\alpha_t \xi_t)(x_t), x - x_{t-1} \right\rangle + \frac{1}{2} \|x - x_{t-1}\|_2^2 \;\middle|\; x \in \mathcal{X} \right\},$$

given a step-size sequence $(\gamma_t)_{t \in \mathbb{N}}$. Output the weighted averages $\hat{x}_T$ and $\hat{y}_T$.

An accelerated gradient method (3/4)

Proposition. Define

$$\tilde{x}_t := \frac{1}{A_t}\left(\alpha_t x_{t-1} + \sum_{\tau=1}^{t-1} \alpha_\tau x_\tau\right).$$

Then, we can equivalently write

$$y_t = \nabla f(\tilde{x}_t).$$

Proof. We write

$$y_t \in \arg\min_y \left\{ -\langle \tilde{x}_t, y\rangle + f^*(y) \mid y \in \mathbb{R}^p \right\},$$

showing that $\tilde{x}_t \in \partial f^*(y_t)$.

An accelerated gradient method (4/4)

Resulting algorithm. For every $t = 1, \ldots, T$, compute

$$x_t \in \arg\min_x \left\{ \gamma \langle \alpha_t \nabla f(\tilde{x}_t), x - x_{t-1} \rangle + \frac{1}{2}\|x - x_{t-1}\|_2^2 \;\middle|\; x \in \mathcal{X} \right\},$$

given a step size $\gamma > 0$.

Theorem. Choose $\alpha_t = t$ and $\gamma = 1/(4L)$. Then

$$f(\hat{x}_T) - f^\star \leq \frac{8LR^2}{T^2},$$

where $R^2 := \max_{x,y} \left\{ (1/2)\|x - y\|_2^2 \;\middle|\; x, y \in \mathcal{X} \right\}$.

J.-K. Wang and J. Abernethy. 2018. Acceleration through optimistic no-regret dynamics.

Importance of choosing appropriate no-regret algorithms

Proposition. It holds that

$$R_T^y \leq L \sum_{t=1}^{T} \frac{\alpha_t^2}{A_t} \|x_{t-1} - x_t\|_2^2,$$

$$R_T^x \leq \frac{D}{\gamma} - \frac{1}{2\gamma} \sum_{t=1}^{T} \|x_{t-1} - x_t\|_2^2.$$

Proof. For R_T^y, the proof relies on the optimistic FTL-BTL lemma and smoothness of the objective function. For R_T^x, the proof is standard.

Remark. Then we choose α_t and γ such that R_T^y is compensated.

J.-K. Wang and J. Abernethy. 2018. Acceleration through optimistic no-regret dynamics.

Conclusions

Summary

- Optimistic methods try to mimic one-step look-ahead. The resulting regret can be much smaller than the worst case result, if the data is *predictable*.
- Optimistic OMD is almost the same as the optimistic lienarized FTRL.
- Fenchel duality allows us to formulate a convex optimization problem as a convex-concave game.
- Choosing appropriate no-regret algorithms for x-$\textsc{Learner}$ and y-$\textsc{Learner}$, we can design efficient optimization algorithms.

Next semester

- Prediction, Learning, and Games.