CSC 665: Online convex optimization

Chicheng Zhang

November 25, 2019

1 Background

1.1 Norms

Definition 1. *A function $\|\cdot\| : \mathbb{R}^d \to \mathbb{R}_+$ (that maps x to $\|x\|$) is called a norm, if the following holds:*

1. *(Homogeneity) $\forall a \in \mathbb{R}$, $\|ax\| = |a|\|x\|$.*
2. *(Triangle inequality) $\forall x, y \in \mathbb{R}^d$, $\|x+y\| \le \|x\| + \|y\|$.*
3. *(Point separation) If $\|v\| = 0$, then $v = \vec{0}$. In other words, all nonzero vectors have nonzero norms.*

Definition 2. *For a norm $\|\cdot\|$, define its dual norm as follows:*

$$\|z\|_\star = \sup_{x : \|x\| \le 1} \langle x, z \rangle .$$

(It can be checked that $\|\cdot\|_\star$ also satisfies the requirements of a norm.)

Example 1.

1. *$\|\cdot\|_2$ has dual norm $\|\cdot\|_2$.*
2. *In general, for $p, q \in [1, \infty]$ being conjugate exponents, that is $\frac{1}{p} + \frac{1}{q} = 1$, $\|\cdot\|_p$ has dual norm $\|\cdot\|_q$.*
3. *Given a positive definite matrix A, define $\|x\|_A = \sqrt{x^\top A x}$. It has dual norm $\|\cdot\|_{A^{-1}}$.*

Fact 1 ("Cauchy-Schwarz" for general norms)**.** *For any norm $\|\|$ and its dual norm $\|\|_\star$, and any two points $x, z \in \mathbb{R}^d$,*

$$\langle x, z \rangle \le \|x\| \|z\|_\star .$$

The fact simply follows from the definition of dual norm.

One might wonder, $\|\cdot\|$ has dual norm $\|\cdot\|_\star$, but what is the dual norm of $\|\cdot\|_\star$? It turns out that under mild assumptions, the dual of $\|\cdot\|_\star$ is $\|\cdot\|$.

1.2 Convexity

Definition 3. *Define convex sets and convex functions as follows:*

1. *A set $\mathcal{C} \subset \mathbb{R}^d$ is convex, if for any u, v in $\mathcal{C}$ and any $\alpha \in [0,1]$, $\alpha u + (1-\alpha) v \in \mathcal{C}$.*
2. *A function $f : \mathcal{C} \to \mathbb{R}$ is convex, if for any u, v in $\mathcal{C}$, and any $\alpha \in [0,1]$, $f(\alpha u + (1-\alpha) v) \le \alpha f(u) + (1-\alpha) f(v)$.*

Fact 2 (Local minimum vs. global minimum)**.** *Suppose f is a convex function. If x is a local minimum of f, in that there exists a radius $r > 0$ such that for all y such that $\|y - x\| \le r$, $f(x) \le f(y)$, then x is also a global minimum f.*

Definition 4 (Subgradient)**.** *Given a convex function $f : \mathcal{C} \to \mathbb{R}$ and a point $v \in \mathcal{C}$, define $\partial f(v)$ as the set of $g \in \mathbb{R}^d$'s such that:*

$$\forall u \in \mathcal{C}, \quad f(u) \geq f(v) + \langle g, u - v \rangle .$$

Therefore, for convex f, if $x^\star$ global minimum of f, then $0 \in \partial f(x^\star)$. It can be easily checked that the converse is also true.

Fact 3. *For any convex $f : \mathcal{C} \to \mathbb{R}$ and a point $v \in \mathcal{C}$, $\partial f(v) \neq \emptyset$, i.e. subgradient always exists. If f is differentiable at v, then $\partial f(v) = \big\{\nabla f(v)\big\}$.*

Example 2. *For function $f(x) = |x|$,*

$$\partial f(x) = \begin{cases} +1 & x > 0, \\ [-1, +1] & x = 0, \\ -1 & x < 0. \end{cases}$$

Definition 5 (Bregman divergence)**.** *For a differentiable convex function f, define its induced Bregman divergence on points u and v as:*

$$D_f(u, v) = f(u) - f(v) - \big\langle \nabla f(v), u - v \big\rangle .$$

In words, $D_f(u, v)$ is the gap between f and its first order approximation (using v) at location u. By convexity of f, $D_f(u, v)$ is always nonnegative. Interestingly, $D_f(u, v)$ may not agree with $D_f(v, u)$, as can be seen in the second example below.

Example 3.

1. *If $f(x) = \frac{\lambda}{2}\|x\|^2$, then $D_f(u, v) = \frac{\lambda}{2}\|u - v\|_2^2$.*
2. *If $f(x) = \sum_{i=1}^d x_i \ln x_i$, then $D_f(u, v) = \sum_{i=1}^d (u_i \ln \frac{u_i}{v_i} - u_i + v_i)$. This is the* unnormalized relative entropy *between u and v; if both u and v are in Δ^{d-1}, then $D_f(u, v)$ is the* relative entropy *between these two probability vectors.*

Fact 4 (Building convex functions from simple ones)**.** *Suppose $f_1, \ldots, f_n$ is a collection of convex functions.*

1. *If $w_1, \ldots, w_n \geq 0$, then $\sum_{i=1}^n w_i f_i(x)$ is convex.*
2. *Let $f(x) = \max(f_1(x), \ldots, f_n(x))$. Then f is convex. Moreover, given an x, $\partial f(x)$ contains elements of $\partial f_i(x)$, where $i \in \arg\max_{i=1}^n f_i(x)$.*

Definition 6. *f is L-Lipschitz with respect to norm $\|\cdot\|$ if for any u, v, $f(u) - f(v) \leq L\|u - v\|$.*

Fact 5. *For any convex $f : \mathcal{C} \to \mathbb{R}$,*

$$f \text{ is } L-\text{Lipschitz} \Leftrightarrow \forall v, \forall g \in \partial f(v), \|g\|_\star \leq L.$$

Therefore, for differentiable functions, to check Lipschitzness, it suffices to check that the gradients at all locations have uniformly-bounded norms.

1.3 Strong convexity

Definition 7 (Strong convexity)**.** *A function $f : \mathcal{C} \to \mathbb{R}$ is λ-strongly convex with respect to norm $\|\cdot\|$, if for any two points $u, v \in \mathcal{C}$, and $\alpha \in [0, 1]$,*

$$f(\alpha u + (1 - \alpha)v) \leq \alpha f(u) + (1 - \alpha) f(v) - \frac{\lambda}{2}\alpha(1 - \alpha)\|u - v\|^2.$$

Strong convexity requires that the gap between interpolated function values and the function value of the interpolated input to have a quadratic lower bound. Clearly, if f is λ-strongly convex, then f is λ'-strongly convex for $\lambda' < \lambda$. Moreover, a function f is 0-strongly convex iff f is convex.

Fact 6. *The following are equivalent:*

1. *f is λ-strongly convex.*
2. *For any v in $\mathcal{C}$, and $g \in \partial f(v)$,*
$$f(u) \geq f(v) + \langle g, u - v \rangle + \frac{\lambda}{2}\|u - v\|^2, \forall u \in \mathcal{C}\,.$$
3. *For any v in $\mathcal{C}$, there exists a vector g such that:*
$$f(u) \geq f(v) + \langle g, u - v \rangle + \frac{\lambda}{2}\|u - v\|^2, \forall u \in \mathcal{C}\,.$$

Properties 2 or 3 are sometimes easier to check than the original strong convexity definition. Specifically, if f is differentiable, then strong convexity is equivalent to a quadratic lower bound on Bregman divergence: $D_f(u, v) \geq \frac{\lambda}{2}\|u - v\|^2$.

Example 4.
1. *If $f(x) = \frac{\lambda}{2}\|x\|^2$, then $D_f(u, v) = \frac{\lambda}{2}\|u - v\|_2^2$. Therefore f is λ-strongly convex with respect to $\|\cdot\|_2$.*
2. *If $f(x) = \sum_{i=1}^d x_i \ln x_i, x \in \left\{x \in \mathbb{R}^d : x_i > 0 \forall i, \sum_{i=1}^d x_i \leq B_1\right\}$, then it can be checked by second-order Taylor's Theorem that $D_f(u, v) \geq \frac{1}{2B_1}\|u - v\|_1^2$, in other words, f is $\frac{1}{B_1}$-strongly convex with respect to $\|\cdot\|_1$.*

Strongly convex function has unique global minimum, as given by the following fact:

Fact 7. *If f is λ-strongly convex, and $x^\star$ is a global minimum of f, then $f(x) - f(x^\star) \geq \frac{\lambda}{2}\|x - x^\star\|^2$. Therefore, if $f(x) \leq f(x^\star)$, then $x = x^\star$.*

1.4 Smoothness

For twice-differentiable f, strong convexity with respect to $\|\cdot\|_2$ reduces to the following simple criterion.

Fact 8. *Suppose f is twice differentiable. f is λ-strongly convex with respect to $\|\cdot\|_2$ iff for any x, $\nabla^2 f(x) \succeq \lambda I$.*

Definition 8 (Smoothness)**.** *A differentiable function f is called β-smooth with respect to norm $\|\cdot\|$, if for any u, v, $\|\nabla f(u) - \nabla f(v)\|_\star \leq \beta\|u - v\|$. In other words, ∇f is β-Lipschitz with respect to $\|\cdot\|$.*

Fact 9. *The following are equivalent:*

1. *f is β-smooth with respect to norm $\|\cdot\|$.*
2. *For any u, v, $f(u) \leq f(v) + \big\langle \nabla f(v), u - v \big\rangle + \frac{\beta}{2}\|u - v\|^2$.*
3. *For any u, v, $f(u) \geq f(v) + \big\langle \nabla f(v), u - v \big\rangle + \frac{1}{2\beta}\|u - v\|^2$.*

It can be seen that, smoothness is opposite to strong convexity: it asks for a function f, $D_f(u, v) \leq \frac{\beta}{2}\|u - v\|^2$ for any u, v. Therefore, if f is both λ-strongly convex and β-smooth, then $\lambda \leq \beta$.

Again for twice-differentiable function f and ℓ_2 norm, we have a simpler way to check smoothness:

Fact 10. *Suppose f is twice differentiable. f is β-smooth with respect to $\|\cdot\|_2$ iff for any x, $\nabla^2 f(x) \preceq \beta I$.*

1.5 Legendre-Fenchel duality

Main idea: given convex function f, use all its tangents to charaterize it.

Fix a slope s, find a tangent of f with slope s. One charaterization of the tangent is that, go over all x's, look at the gaps between $f(x)$ and $\langle s, x\rangle$, and find the location with the smallest gap. This smallest gap is the offset b, such that $\langle s, x\rangle + b$ is the tangent of f with slope s.

As discussed above, the offeset can be written as:

$$b(s) = \min_x \big(f(x) - \langle s, x\rangle\big) .$$

We define the Legendre-Fenchel conjugate of f as $-b(s)$, denoted as $f^\star(s)$.

Definition 9. *The Legendre-Fenchel conjugate (dual) of f, $f^\star$, is defined as*

$$f^\star(s) = \max_x \big(\langle s, x\rangle - f(x)\big) .$$

As $f^\star$ is the pointwise maximum of a collection of convex functions, $f^\star$ is convex. Can we give a characterization of the subgradient of $f^\star$? Using a generalization of Fact 4, and the fact that $s \mapsto \langle s, x\rangle - f(x)$ has subgradient x, we can see that

$$\operatorname*{argmax}_x \big(\langle s, x\rangle - f(x)\big) \in \partial f^\star(s).$$

Let us look at the dual of $f^\star$, that is $f^{\star\star}(x) = \max_s \big(\langle x, s\rangle - f^\star(s)\big)$. Note that it has the following nice geometric interpretation: recall that for each s, $\langle x, s\rangle - f^\star(s)$ is the tangent of f of slope s; we get a collection of lines below f. Taking an upper envelope of these lines, we recover the original function f.

Fact 11. *Suppose f is closed (in that $\big\{(x,t) : f(x) \le t\big\}$ is a closed set) and convex, then $f^{\star\star} = f$. In words, the dual of the dual is the original function.*

The following simple fact is by the definition of Legendre-Fenchel conjugate function:

Fact 12 (Fenchel-Young's Inequality)**.** *For any pairs of x and s,*

$$f(x) + f^\star(s) \ge \langle x, s\rangle .$$

Example 5.

1. *For conjugate exponents $p, q \in (1, \infty)$ such that $\frac{1}{p} + \frac{1}{q} = 1$, if $f(x) = \frac{x^p}{p}$, then $f^\star(s) = \frac{s^q}{q}$. This is the classical Young's inequality.*
2. *For any norm $\|\cdot\|$, if $f(x) = \frac{\lambda}{2}\|x\|^2$, then $f^\star(s) = \frac{1}{2\lambda}\|s\|_\star^2$.*
3. *If $f(x) = \begin{cases} \sum_{i=1}^d x_i \ln x_i, & x \in \Delta^{d-1} \\ +\infty, & x \notin \Delta^{d-1} \end{cases}$, then $f^\star(s) = \ln \sum_{s=1}^d e^{s_i}$.*
4. *If $f(x) = \begin{cases} \sum_{i=1}^d x_i \ln x_i, & x \succ 0 \\ +\infty, & x \not\succ 0 \end{cases}$, then $f^\star(s) = \sum_{i=1}^d e^{s_i - 1}$.*

If $f \ge g$, then by the definition of conjugate function, $f^\star \le g^\star$.

It can be shown that for a strongly convex f, $f^\star$ is differentiable. Specifically,

$$\nabla f^\star(s) = \operatorname*{argmax}_x \big(\langle s, x\rangle - f(x)\big) ,$$

as f is strongly convex, the right hand side has unique element and the equality is thus well-defined.

Fact 13. *f is λ-strongly convex with respect to $\|\cdot\|$ iff $f^\star$ is $\frac{1}{\lambda}$-smooth with respect to $\|\cdot\|_\star$.*

Proof. We only show the "only if" here. Our goal is to show that for u, v,

$$\|x_u - x_v\|_\star \leq \frac{1}{\lambda}\|u - v\|,$$

where

$$x_u = \nabla f^\star(u) = \operatorname*{argmin}_x h_u(x), \text{ where } h_u(x) = \big(f(x) - \langle u, x\rangle\big),$$
$$x_v = \nabla f^\star(v) = \operatorname*{argmin}_x h_v(x), \text{ where } h_v(x) = \big(f(x) - \langle v, x\rangle\big).$$

Note that h_u and h_v are close to each other when u and v are close: but close functions may not necessarily imply that their optimal points are close to each other; for example, $f(x) = 0.01x$ has minimum at $-\infty$, and $f(x) = -0.01x$ has minimum at $+\infty$; luckily, for strongly convex functions that differ by a small linear function, we show that their minimum points are close.

By the strong convexity of $h_u(x)$ (resp. $h_v(x)$) and the optimality of x_u (resp. x_v),

$$h_u(x_v) \geq h_u(x_u) + \frac{\lambda}{2}\|x_u - x_v\|^2,$$

$$h_v(x_u) \geq h_v(x_v) + \frac{\lambda}{2}\|x_u - x_v\|^2.$$

Summing the two inequalities up,

$$\langle u - v, x_u - x_v\rangle \geq \lambda\|x_u - x_v\|^2.$$

By the generalized Cauchy-Schwarz, we have

$$\lambda\|x_u - x_v\|^2 \leq \|u - v\|\|x_u - x_v\|,$$

implying

$$\|x_u - x_v\|_\star \leq \frac{1}{\lambda}\|u - v\|.$$

□

The above fact shows that, if f is more "curved", then $f^\star$ is more "flat", and vice versa.

2 Online convex optimization

Setup:

Algorithm 1 Online convex optimization (OCO)

Require: Convex decision set $\mathcal{C}$.
  for timesteps $t = 1, 2, \dots, T$: **do**
    Learner chooses $x_t \in \mathcal{C}$,
    Learner receives a convex loss f_t.
  end for
  Goal: minimize cumulative loss $\sum_{t=1}^T f_t(x_t)$.

Definition 10. *Suppose for every f_t, $f_t(x) = \langle g_t, x\rangle$ for some vector g_t, then the OCO problem is called an online linear optimization (OLO) problem.*

2.1 Follow the regularized leader (FTRL) for OLO

Given a λ-strongly convex regularization function R, set

$$\begin{aligned} x_t &= \operatorname*{argmin}_x \sum_{s=1}^{t-1} \langle g_s, x\rangle + R(x) \\ &= \operatorname*{argmax}_x \langle -G_{t-1}, x\rangle - R(x) \\ &= \nabla R^\star(-G_{t-1}), \end{aligned}$$

where $G_t = \sum_{s=1}^t g_s$ is the cumulative gradients. the mapping $\nabla R^\star$ is called the *mirror map*, that "transports" the cumulative negative gradient to a point in the decision space.

Example 6. *We give a few instantiations of FTRL:*

1. *Hedge as FTRL: let $g_t = \ell_t$ for every t, and let $R(x) = \begin{cases} \frac{1}{\eta}\sum_{i=1}^d x_i \ln x_i, & x \in \Delta^{d-1} \\ +\infty, & x \notin \Delta^{d-1} \end{cases}$, then it can be checked that*
$$x_{t,i} = \exp\left(-\eta \sum_{s=1}^{t-1} \ell_{s,i}\right).$$

2. *Online gradient descent: let $R(x) = \frac{1}{2\eta}\|x\|_2^2$, then $R^\star(G) = \frac{\eta}{2}\|G\|_2^2$, and $\nabla R^\star(G) = \eta G$. Therefore, $x_t = -\eta G_{t-1} = -\sum_{s=1}^{t-1} \eta g_s$. This is the cumulative sum of negative gradients, times a stepsize of η.*

3. *Online gradient descent with lazy projections: let $R(x) = \begin{cases} \frac{1}{2\eta}\|x\|^2, & x \in \mathcal{C}, \\ +\infty, & x \notin \mathcal{C} \end{cases}$, then it can be shown that,*
$$x_t = \operatorname*{argmin}_{x\in\mathcal{C}} \|x - (-\eta G_{t-1})\|_2,$$
which is the ℓ_2-projection of the point returned by online gradient descent to the convex set $\mathcal{C}$.

In this theoreom below, we will show that FTRL has a small regret given an appropriately-tuned step size η.

Theorem 1. *If R is λ-strongly convex with respect to $\|\cdot\|$, then FTRL has the following regret:*

$$\mathrm{Reg}(T, x) = \sum_{t=1}^T \langle g_t, x_t - x\rangle \le R(x) - \min_{x'} R(x') + \frac{1}{\lambda}\sum_{t=1}^T \|g_t\|_\star^2.$$

Proof. Recall that $f_t(x) = \langle g_t, x\rangle$. We break the proof into two steps:

1. Consider a 'look-ahead' prediction strategy named the "be-the-regularized leader" (BTRL), that is, at time t, x_{t+1}'s are selected as the decision point. We will show that BTRL has a small regret.

2. Note that BTRL cannot be implemented as a real algorithm: x_{t+1} relies on information on g_t, which is unavailable at the beginning of round t. Nevertheless, we will show that x_t, the decision point selected by FTRL, is close to x_{t+1}, therefore the regret of FTRL can be bounded in terms of that of BTRL.

Step 1: Analysis of BTRL. Denote by $f_0(x) = R(x)$. Consider a modification of the original OCO game: there is an extra round of online convex optimization at the beginning, namely round 0. We will show that BTRL has nonpositive regret on this.

Lemma 1 (Be the leader)**.** *For any $x^\star$,*

$$\sum_{t=0}^{T} f_t(x_{t+1}) \leq \sum_{t=0}^{T} f_t(x^\star).$$

Proof. This is best illustrated by iteratively relaxing the right hand side; as $x_{T+1} = \mathrm{argmin}_x \sum_{t=0}^{T} f_t(x)$, we have that

$$\sum_{t=0}^{T} f_t(x_{T+1}) \leq \sum_{t=0}^{T} f_t(x^\star).$$

Now let us focus on all but the last term in the left hand side, that is, $\sum_{t=0}^{T-1} f_t(x_{t+1})$. As as $x_T = \mathrm{argmin}_x \sum_{t=0}^{T-1} f_t(x)$, we have that

$$\left(\sum_{t=0}^{T-1} f_t(x_T)\right) + f_T(x_{T+1}) \leq \sum_{t=0}^{T} f_t(x_{T+1}) \leq \sum_{t=0}^{T} f_t(x^\star).$$

By iteratively using the fact that $x_\tau = \mathrm{argmin}_x \sum_{t=0}^{\tau-1} f_t(x)$, we have that

$$\left(\sum_{t=0}^{\tau-1} f_t(x_\tau)\right) + f_\tau(x_{\tau+1}) + \ldots + f_T(x_{T+1}) \leq \sum_{t=0}^{T} f_t(x^\star).$$

The lemma is a direct consequence of the above inequality in the case of $\tau = 1$. $\square$

Lemma 1 immediately implies that:

$$\sum_{t=1}^{T} \langle g_t, x_{t+1} - x^\star \rangle \leq R(x^\star) - R(x_1). \tag{1}$$

Step 2: relating BTRL to FTRL. Our next task will be to upper bound $\sum_{t=1}^{T} \langle g_t, x_t - x_{t+1} \rangle$, the difference of the cumulative losses of FTRL and BTRL.

Lemma 2 (Stability)**.**

$$\sum_{t=1}^{T} \langle g_t, x_t - x_{t+1} \rangle \leq \frac{1}{\lambda} \sum_{t=1}^{T} \|g_t\|_\star^2. \tag{2}$$

Proof. We will show that for every t, $\langle g_t, x_t - x_{t+1} \rangle \leq \frac{1}{\lambda}\|g_t\|_\star^2$. To show this, by generalized Cauchy-Schwarz, it suffices to show that

$$\|x_t - x_{t+1}\| \leq \frac{1}{\lambda}\|g_t\|_\star.$$

By definition of $x_t = \nabla R^\star(-G_{t-1})$ and $x_{t+1} = \nabla R^\star(-G_t)$, we see that

$$\|x_t - x_{t+1}\| = \|\nabla R^\star(-G_{t-1}) - \nabla R^\star(-G_t)\|.$$

Recall that R is λ-strongly convex, by Fact 13, $R^\star$ is $\frac{1}{\lambda}$-smooth. Therefore the right hand side is indeed at most $\frac{1}{\lambda}\| - G_{t-1} - (-G_t)\| = \frac{1}{\lambda}\|g_t\|_\star$. $\square$

The theorem is proved by summing Equations (1) and (2) together. $\square$

2.2 FTRL for general OCO

It turns out that a low-regret algorithm for OLO immediately yields an algorithm for OCO. To see this, suppose that at every iteration t, f_t is a general convex function. Now, suppose that $g_t \in \partial f_t(x_t)$ is a subgradient of f_t at location x_t. We have that for any $x^\star$,

$$f_t(x_t) - f_t(x^\star) \le \langle g_t, x_t - x^\star \rangle .$$

Therefore, if we let $\tilde{f}_t(x) = \langle g_t, x \rangle$, and run FTRL on $\tilde{f}_t$'s, we get that

$$\sum_{t=1}^{T} \langle g_t, x_t - x^\star \rangle \le R(T)$$

for some regret function $R(T)$. This implies that

$$\mathrm{Reg}(T, x^\star) = \sum_{t=1}^{T} f_t(x_t) - f_t(x^\star) \le \sum_{t=1}^{T} \langle g_t, x_t - x^\star \rangle \le R(T).$$

2.3 Instantiations of FTRL: theoretical guarantees

1. Online gradient descent (OGD): $R(x) = \frac{1}{2\eta}\|x\|_2^2$, which is $\frac{1}{\eta}$-strongly convex wrt $\|\cdot\|_2$. FTRL with R has regret

   $$\mathrm{Reg}(T, x) \le \frac{\|x\|_2^2}{2\eta} + \eta \sum_{t=1}^{T} \|g_t\|_2^2,$$

   for all benchmark $x \in \mathbb{R}^d$.

   Suppose we would like to guarantee $\mathrm{Reg}(T, \mathcal{C})$ with $\mathcal{C} \subset \left\{x : \|x\| \le B_2\right\}$. If in addition, it is known apriori that $\|g_t\| \le R_2$, then

   $$\mathrm{Reg}(T, \mathcal{C}) \le \frac{B_2^2}{2\eta} + \eta T R_2^2.$$

   We can setting $\eta = \frac{B_2}{R_2\sqrt{2T}}$ that minimize the regret bound, which gives $B_2 R_2 \sqrt{2T}$.

2. OGD with lazy projections:

   $$R(x) = \begin{cases} \frac{1}{2\eta}\|x\|_2^2 & x \in \mathcal{C} \\ +\infty & x \notin \mathcal{C} \end{cases},$$

   which is also $\frac{1}{\eta}$-strongly convex wrt $\|\cdot\|_2$. Note that FTRL in this case gives $x_t \in \mathcal{C}$ at every round. FTRL with R has regret:

   $$\mathrm{Reg}(T, x) \le \frac{\|x\|_2^2}{2\eta} + \eta \sum_{t=1}^{T} \|g_t\|_2^2,$$

   for all benchmark $x \in \mathcal{C}$. Again, setting $\eta = \frac{B_2}{R_2\sqrt{2T}}$ guarantees $\mathrm{Reg}(T, \mathcal{C}) \le B_2 R_2 \sqrt{2T}$.

3. p-norm algorithms ($p \ge 2$): It is known that $R(x) = \frac{1}{2\eta}\|x\|_p^2$ is $\frac{1}{\eta}$-strongly convex wrt $\|\cdot\|_p$. FTRL with R has regret:

   $$\mathrm{Reg}(T, x) \le \frac{\|x\|_p^2}{2\eta} + \eta \sum_{t=1}^{T} \|g_t\|_q^2.$$

   If $\mathcal{C} \subset \left\{x : \|x\|_p \le B_p\right\}$, and for all t, $\|g_t\|_q \le R_q$, setting $\eta = \frac{B_p}{R_q\sqrt{2T}}$ implies that

   $$\mathrm{Reg}(T, \mathcal{C}) \le B_p R_q \sqrt{2T}.$$

4. Exponentiated gradient (Hedge): consider the negative entropy regularizer

$$R(x) = \begin{cases} \frac{1}{\eta}\sum_{i=1}^d x_i \ln x_i, & x \in \Delta^{d-1}, \\ +\infty, & \text{otherwise.} \end{cases}$$

Recall that by the calibration exercise, $R(x)$ is 1-strongly convex with respect to $\|\cdot\|_1$. Therefore, FTRL with R has regret:

$$\mathrm{Reg}(T, x) \le \frac{\sum_{i=1}^d x_i \ln x_i - \min_{x' \in \Delta^{d-1}} \sum_{i=1}^d x'_i \ln x'_i}{\eta} + \eta \sum_{t=1}^T \|g_t\|_\infty^2.$$

It can be seen that $\sum_{i=1}^d x_i \ln x_i \le 0$, on the other hand, $\min_{x' \in \Delta^{d-1}} \sum_{i=1}^d x'_i \ln x'_i = -\max_{x' \in \Delta^{d-1}} H(x)$, where $H(x)$ is the entropy of probability vector x. Therefore, it is $-\ln d$. This implies that the first term is at most $\frac{\ln d}{\eta}$. Now suppose we know that all t is such that $\|g_t\|_\infty \le R_\infty$, we have

$$\mathrm{Reg}(T, x) \le \frac{\ln d}{\eta} + \eta T R_\infty^2.$$

Setting $\eta = \frac{\sqrt{\ln d}}{R_\infty \sqrt{T}}$ gives that

$$\mathrm{Reg}(T, \Delta^{d-1}) \le 2R_\infty \sqrt{T \ln d}.$$

(The above regularizer can also be used to deal with a scaled version of probability simplex: $\left\{x : \forall i, x_i > 0, \sum_{i=1}^d x_i = B_1\right.$ for general $B_1 > 0$; we will skip the discussion for simplicity.)

2.4 Applications of FTRL to online linear classification

Algorithm 2 Online linear classification (with FTRL)

Require: Regularizer R, stepsize η.

for timesteps $t = 1, 2, \ldots, T$: **do**

Learner chooses $w_t = \nabla(\frac{R}{\eta})^\star(-\sum_{s=1}^{t-1} g_s) \in \mathbb{R}^d$,

Learner receives an example (x_t, y_t).

Learner suffers from zero-one loss $M_t = \mathbf{1}(\langle w_t, y_t x_t\rangle \le 0)$.

Induced loss $f_t(w) = \mathbf{1}(\langle w_t, y_t x_t\rangle \le 0)(1 - \langle w, y_t x_t\rangle)$.

Let $g_t = \nabla f_t(w)|_{w=w_t} = \begin{cases} 0 & M_t = 0 \\ -y_t x_t & M_t = 1 \end{cases} \in \partial f_t(w_t)$.

end for

Goal: minimize cumulative zero-one loss $\sum_{t=1}^T M_t$.

Theorem 2. *Suppose R is 1-strongly convex defined on $\mathcal{C}$ with with respect to $\|\cdot\|$, and for all x_t, $\|x_t\|_\star \le R$. Moreover, for all $w, w' \in \mathcal{C}$, $R(w) - R(w') \le \Delta$. Then, for any $w \in \mathcal{C}$,*

$$\sum_{t=1}^T M_t \le \frac{1}{1 - \eta R^2}(L_T(w) + \frac{\Delta}{\eta}),$$

where $L_T(w) = \sum_{t=1}^T (1 - \langle w, y_t x_t\rangle)_+$ is the cumulative hinge loss of w. Specifically, if there exists $w \in \mathcal{C}$ such that the data is separable by a margin of 1: $\forall t, \langle w, y_t x_t\rangle \ge 1$, then setting $\eta = \frac{1}{2R^2}$ implies that

$$\sum_{t=1}^T M_t \le 2R^2 \Delta,$$

in other words, the algorithm has a finite mistake bound.

Proof. As R is 1-strongly convex wrt $\|\cdot\|$, $\frac{R}{\eta}$ is $\frac{1}{\eta}$-strongly convex wrt $\|\cdot\|$. By the guarantees of OCO with respect to $\{f_t(\cdot)\}$'s, we have that for all w in $\mathcal{C}$,

$$\sum_{t=1}^{T} f_t(w_t) - \sum_{t=1}^{T} f_t(w) \le \frac{\Delta}{\eta} + \sum_{t=1}^{T} \eta \|g_t\|^2.$$

We have the following observations:

1. $g_t = 0$ if $M_t = 0$; therefore, the second term on the right hand side is at most $\eta R^2 (\sum_{t=1}^{T} M_t)$.
2. Moreover, $f_t(w_t) = \mathbf{1}(\langle w_t, y_t x_t\rangle \le 0)(1 - \langle w_t, y_t x_t\rangle)$. Observe that $f_t(w_t) \ge 0$. Moreover, if $M_t = 1$, then $f_t(w_t) \ge 1$. Therefore, $\sum_{t=1}^{T} M_t \le \sum_{t=1}^{T} f_t(w_t)$.
3. $f_t(w) \le \mathbf{1}(\langle w_t, y_t x_t\rangle \le 0)(1 - \langle w, y_t x_t\rangle)_+ \le (1 - \langle w, y_t x_t\rangle)_+$, which is the instantaneous hinge loss of w.

Combining the above insights, we get

$$\sum_{t=1}^{T} M_t \cdot (1 - \eta R^2) \le L_t(w) + \frac{\Delta}{\eta},$$

that is,

$$\sum_{t=1}^{T} M_t \le \frac{1}{1 - \eta R^2}(L_t(w) + \frac{\Delta}{\eta}).$$

The second claim of the theorem follows simply from algebra and the fact that $L_t(w) = 0$. □

Instantiations: Perceptron and Winnow. We consider two settings of R's:

1. Let $R(w) = \frac{1}{2}\|w\|^2$, and $\mathcal{C} = \{w : \|w\|_2 \le B\}$. Then it can be checked that Δ can be set as B^2. Suppose all examples lies in $\{x : \|x\|_2 \le R\}$. FTRL with R has a mistake bound of

   $$\sum_{t=1}^{T} M_t \le \min_{w \in \mathcal{C}} \frac{1}{1 - \eta R^2}(L_T(w) + \eta B^2).$$

   If the data is linearly separable by margin 1 by classifier w in $\mathcal{C}$, then setting $\eta = \frac{1}{2R^2}$ gives that

   $$\sum_{t=1}^{T} M_t \le 2R^2 B^2.$$

   This is a variant of the well-known Percetron convergence theorem by Novikoff.

2. Let $R(w) = \begin{cases} \sum_{i=1}^{d} w_i \ln w_i, & w \in \Delta^{d-1}, \\ +\infty, & \text{otherwise}. \end{cases}$. As seen before Δ can be set as $\ln d$. Suppose all examples lies in $\{x : \|x\|_\infty \le R\}$. FTRL with R has a mistake bound of

   $$\sum_{t=1}^{T} M_t \le \min_{w \in \mathcal{C}} \frac{1}{1 - \eta R^2}(L_T(w) + \eta \ln d).$$

   If the data is linearly separable by margin 1 by classifier w in Δ^{d-1}, then setting $\eta = \frac{1}{2R^2}$ gives that

   $$\sum_{t=1}^{T} M_t \le 2R^2 \ln d.$$

2.5 FTRL with adaptive regularization

As we have seen before, the choice of regularizer is crucial to obtain good online prediction performance. However, if we are faced with a stream of data, it is difficult to know which regularizer to choose ahead of the time. In this section, we will look at FTRL with adaptive regularization, which is a systematic way to achieve online performance guarantees that adapts to the geometry of the data on the fly.

Our starting point is to consider the following algorithm:

$$x_t = \nabla R_{t-1}^\star(-G_{t-1}),$$

recall that $G_{t-1} = \sum_{s=1}^{t-1} g_s$ is the sum of the gradients up to time $t-1$. We called the above algorithm FTRL-AR. Specifically, we will be looking at a sequence of monotonically increasing regularizers $\{R_t\}$'s, where R_t's are generated on the fly, and can thus carry over information on the past g_t's.

Theorem 3. *Suppose FTRL-AR uses R_t that are 1-strongly convex with respect to $\|\cdot\|_t$. Then it has the following upper bound on its cumulative loss guarantee:*

$$\sum_{t=1}^T \langle g_t, x_t \rangle \leq R_0^\star(0) - R_T^\star(-G_T) + \sum_{t=1}^T \|g_t\|_{\star,t-1}^2.$$

Consequently,

$$\mathrm{Reg}(T, x^\star) = \sum_{t=1}^T \langle g_t, x_t - x^\star \rangle \leq R_T(x^\star) + R_0^\star(0) + \sum_{t=1}^T \|g_t\|_{\star,t-1}^2.$$

Note that the above theorem supercedes Theorem 1, as it is a direct consequence of the above theorem by taking $R_t \equiv R_0$ and observing that $R_0^\star(0) = -\min_{x'} R_0(x')$.

Proof. It suffices to show that

$$\langle g_t, x_t \rangle \leq R_{t-1}^\star(-G_{t-1}) - R_t^\star(-G_t) + \|g_t\|_{\star,t-1}^2,$$

as the theorem concludes by summing this inequality up over all t's.

To show the above inequality, it suffices for us to show that

$$R_t^\star(-G_t) - R_{t-1}^\star(-G_{t-1}) + \langle g_t, x_t \rangle \leq \|g_t\|_{\star,t-1}^2.$$

The above inequality is true by the following observations: first, as $R_t \geq R_{t-1}$, $R_t^\star \leq R_{t-1}^\star$; second, $x_t = \nabla R_{t-1}^\star(-G_{t-1})$, therefore, the left hand side of the inequality is at most

$$R_{t-1}^\star(-G_t) - R_{t-1}^\star(-G_{t-1}) - \left\langle \nabla R_{t-1}^\star(-G_{t-1}), -g_t \right\rangle = D_{R_{t-1}^\star}(-G_t, -G_{t-1}).$$

third, as R_{t-1} is 1-strongly convex wrt $\|\cdot\|_{t-1}$, $R_{t-1}^\star$ is 1-smooth wrt $\|\cdot\|_{\star,t-1}$, implying that the right hand side is at most $\frac{1}{2}\| -G_t - (-G_{t-1})\|_{\star,t-1}^2 = \frac{1}{2}\|g_t\|_{\star,t-1}^2$. □

Using the above meta-theorem, we can instantiate with different adpative regularizers and get online learning algorithms with different degrees of adaptivity.

Online gradient descent with adaptive step-sizes. One instantiation of the above result is to let

$$R_t(x) = \frac{\sqrt{t+1}}{\eta_0}\|x\|_2^2.$$

This implies that

$$\mathrm{Reg}(T, x^\star) \leq \frac{\sqrt{T+1}}{\eta_0}\|x^\star\|_2^2 + \sum_{t=1}^T \eta_0 \cdot \frac{\|g_t\|^2}{\sqrt{t}}.$$

Note that if $\eta =$, then this algorithm automatically achieves a $O(\sqrt{T})$ regret for all timesteps T.

There is a variant of the above ℓ_2 regularization scheme with another setting of the regularization strength:

$$R_t(x) = \frac{\sqrt{\sigma + \sum_{s=1}^t \|g_s\|^2}}{2\eta_0} \|x\|^2.$$

for some $\sigma > 0$.

Adaptive subgradient methods (Adagrad). More generally we can allow adaptive Mahalanobis norm-based regularization. Specifically, we can let

$$R_t(x) = \frac{1}{2}\|x\|_{A_t}^2,$$

for some adaptively generated A_t.

Specifically, one can let

$$A_t = \frac{1}{\eta}(\sigma I + \mathrm{diag}(\sum_{s=1}^t g_s g_s^\top))^{\frac{1}{2}}$$

be an "diagonal" adaptive regularizer.

Alternatively, one can let

$$A_t = \frac{1}{\eta}(\sigma I + \sum_{s=1}^t g_s g_s^\top)^{\frac{1}{2}}$$

be an "nondiagnoal" adaptive regularizer.

3 OCO for strongly convex functions

Motivating example: SVM optimization:

$$\min_w \sum_{t=1}^T \left(\frac{\lambda}{2}\|w\|_2^2 + (1 - \langle w, y_t x_t \rangle)_+ \right).$$

Here $f_t(w) = \frac{\lambda}{2}\|w\|_2^2 + (1 - \langle w, y_t x_t \rangle)_+$. If one can get a low regret $R(T)$, then one can use online-to-batch conversion to get a f that has excess expected regularized loss $\frac{R(T)}{T}$.

One can show that if all f_t's are λ-strongly convex, one can design a better OCO algorithm with regret bound much better than $O(\sqrt{T})$, that is, $O(\ln T)$.

How to achieve this? We will use the adaptive regularization method developed in the last section. Recall that AR-FTRL has the following regret guarantee:

$$\sum_{t=1}^T \langle g_t, x_t - x^\star \rangle \le R_0^\star(0) + R_T(x^\star) + \sum_{t=1}^T \|g_t\|_{\star,t-1}^2.$$

How can the above regret relate to $\mathrm{Reg}(T, x^\star) = \sum_{t=1}^T f_t(x_t) - f_t(x^\star)$? Now because f_t is λ-strongly convex, we have a tighter bound on it. Specifically,

$$\mathrm{Reg}(T, x^\star) \le \sum_{t=1}^T \langle g_t, x_t - x^\star \rangle - \sum_{t=1}^T \frac{\lambda}{2}\|x_t - x^\star\|^2.$$

This motivates us to define $R_t(x) = \frac{\lambda}{2}\|x\|^2 + \sum_{s=1}^t \frac{\lambda}{2}\|x_s - x\|^2$ so that $R_T(x^\star)$ cancels out the negative terms induced by linear approximation. Observe that R_t is 1-strongly convex with respect to $\|\cdot\|_{\lambda(t+1)I}$. We therefore get:

$$\mathrm{Reg}(T, x^\star) \le \frac{\lambda}{2}\|x\|^2 + \sum_{t=1}^T \frac{\|g_t\|^2}{\lambda t}.$$

4 OCO for exp-concave functions

Motivating example 1: sequential investing. There are d stocks, with different growth rates every day.

$W_1 \leftarrow 1$.

For $t = 1, 2, \ldots, T$:

1. Given the current wealth W_t, allocate $p_t \in \Delta^{d-1}$ (spend $p_{t,i}$ fraction of current wealth to stock i)
2. Receive loss $f_t(p_t) = -\ln\big(\langle c_t, p_t \rangle\big)$, where $c_t \in \mathbb{R}_+^d$, and $c_{t,i}$ is the ratio of the stock i at the .
3. Sell all stocks, get new wealth W_{t+1}. Observe that

$$W_{t+1} = W_t \left(\sum_{t=1}^{T} p_{t,i} c_{t,i} \right),$$

i.e. $\ln(W_{t+1}) = \ln(W_t) - f_t(p_t)$. Therefore, maximizing W_{T+1} amounts to minimizing the cumulative loss $\sum_{t=1}^T f_t(p_t)$.

Goal: compete with the best constant rebalanced portfolio in hindsight (abbrev. CRP; that is, at the beginning of every day, allocate a constant fraction $q \in \Delta^{d-1}$ to all stocks.) Concretely,

$$\mathrm{Reg}(T, q) = \sum_{t=1}^{T} f_t(p_t) - \sum_{t=1}^{T} f_t(q).$$

Motivating example 2: online least squares regression. For $t = 1, 2, \ldots, T$:

1. Output a linear predictor $w_t \in \mathbb{R}^d$.
2. Receive example $(x_t, y_t) \in \mathbb{R}^d \times \mathbb{R}$.
3. Suffer loss $f_t(w_t)$, where $f_t(w) = \frac{1}{2}(\langle w, x_t \rangle - y_t)^2$.

$$\mathrm{Reg}(T, w^\star) = \sum_{t=1}^{T} f_t(w_t) - \sum_{t=1}^{T} f_t(w^\star).$$

The common characteristic of the above two OCO problems are that the f_t's are structured: they are compositions of a univariate "strongly convex" function and a linear function. It turns out that they both belong to the family called *exp-concave* functions.

Definition 11. *f is called α-exp-concave, if $\exp\big(-\alpha f(x)\big)$ is a concave function.*

Clearly, $f(x) = -\ln\big(\langle c, x \rangle\big)$ is 1-exp-concave.

Lemma 3. *f is α-exp-concave, iff for every x,*

$$\nabla^2 f(x) \succeq \alpha \nabla f(x) \cdot \nabla f(x)^\top.$$

Proof. $h = \exp\big(-\alpha f(x)\big)$ is concave iff for every x, the hessian of h is negative semidefinite. Observe that

$$\nabla^2 h(x) = \alpha^2 \nabla f(x) \nabla f(x)^\top \exp\big(-\alpha f(x)\big) - \alpha \nabla^2 f(x) \exp\big(-\alpha f(x)\big) \preceq 0.$$

$\square$

It can be readily seen that for $\alpha < \gamma$, if f is γ-exp-concave, then f is α-exp-concave.

Lemma 4. *Suppose h is λ-strongly convex and has gradient at most G. Then for any sw, $h(\langle w, x\rangle)$ is $\frac{\lambda}{G^2}$-exp-concave.*

For online least-square regression with domain $\{w : \|w\|_2 \le B\}$ and all $x \in \{x : \|x\|_2 \le R\}$ and $y \in [-Y, Y]$, one can take $h(z) = \frac{1}{2}(z-y)^2$, which is 1-strongly convex, and has gradient norm at most $RB + Y$. Therefore, $\frac{1}{2}(\langle w, x\rangle - y)^2$ is $\frac{1}{(RB+Y)^2}$-exp-concave.

For exp-concave functions, one can have a more refined lower bound than linear approximation.

Lemma 5. *If f is α-exp-concave and G-Lipschitz, then for any two points $u, v \in \{x : \|x\|_2 \le B\}$, we have*

$$f(u) \ge f(v) + \langle \nabla f(v), u - v\rangle + \frac{\tilde{\alpha}}{2}(u-v)^\top \nabla f(v) \nabla f(v)^\top (u-v),$$

where $\tilde{\alpha} = \min(\frac{1}{8BR}, \frac{1}{2\alpha})$.

Algorithm with logarithmic regret: adaptive regularization. We will be using Lemma 5 and the insights similar to OCO for strongly-convex optimization to develop an algorithm with a $O(\log T)$ regret.

Recall that AR-FTRL has the following regret guarantee:

$$\sum_{t=1}^{T} \langle g_t, x_t - x^\star\rangle \le R_0^\star(0) + R_T(x^\star) + \sum_{t=1}^{T} \|g_t\|_{\star, t-1}^2.$$

In addition, by Lemma 5, we have that

$$\sum_{t=1}^{T} f_t(x_t) - f_t(x^\star) \le \sum_{t=1}^{T} \langle g_t, x_t - x^\star\rangle - \sum_{t=1}^{T} \frac{\tilde{\alpha}}{2}(x^\star - x_t)^\top \nabla f(x_t) \nabla f(x_t)^\top (x^\star - x_t)$$

This motivates us to set $R_T(x) = \frac{\sigma}{2}\|x\|_2^2 + \sum_{t=1}^{T} \frac{\tilde{\alpha}}{2}(x - x_t)^\top \nabla f(x_t) \nabla f(x_t)^\top (x - x_t)$. Observe that for every t, $R_t(x)$ is σ-strongly convex with respect to $\|\cdot\|_t = \|\cdot\|_{A_t}$, where $A_t = \sigma I + \sum_{t=1}^{T} \nabla f(x_t) \nabla f(x_t)^\top$.

This gives that

$$\mathrm{Reg}(T, x^\star) \le \frac{\sigma}{2}\|x^\star\|_2^2 + \sum_{t=1}^{T} \|g_t\|_{A_{t-1}^{-1}}^2.$$