CSC 665: Online convex optimization

Chicheng Zhang

December 5, 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. *For any u, v and any $\alpha \in [0, 1]$, the* convex combination *between u and v with coefficient α is defined as $\alpha u + (1 - \alpha)v$.*
2. *A set $\mathcal{C} \subset \mathbb{R}^d$ is* convex*, if for u and v in $\mathcal{C}$, and any coefficient $\alpha \in [0, 1]$, their convex combination with coefficient α is in $\mathcal{C}$.*
3. *A function $f : \mathcal{C} \to \mathbb{R}$ is* convex*, if (1) its domain $\mathcal{C}$ is convex, (2) for any u, v in $\mathcal{C}$, and any $\alpha \in [0, 1]$, $f(\alpha u + (1 - \alpha)v) \le \alpha f(u) + (1 - \alpha)f(v)$.*

If we have a convex function f on a convex domain $\mathcal{C}$, we define its extension to $\mathbb{R}^d$ as

$$\bar{f}(x) = \begin{cases} f(x) & x \in \mathcal{C} \\ +\infty & x \notin \mathcal{C} \end{cases}. \tag{1}$$

Sometimes we will use $f : \mathcal{C} \to \mathbb{R}$ and $\bar{f} : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ interchangably.

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:*

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

Therefore, for convex f, if $0 \in \partial f(x^\star)$, then $x^\star$ is the global minimum of f. However for $f : \mathcal{C} \to \mathbb{R}$, a global minimum of f in $\mathcal{C}$ may not necessarily have zero subgradient: for example, suppose $\mathcal{C} = [-1, +1]$ and $f(x) = x$, then the global minimum $x^\star = -1$, but f has subgradient 1 on $x^\star$. Nevertheless, we have the following first order optimality condition.

Fact 3 (First order optimality condition)**.** *For a convex set $\mathcal{C}$ and $f : \mathcal{C} \to \mathbb{R}$. Suppose $x^\star \in \mathcal{C}$ is the global minimum of f, then we have that there exists $g \in \partial f(x^\star)$:*

$$\forall x \in \mathcal{C}, \quad \langle g, x - x^\star \rangle \ge 0. \tag{2}$$

The proof of this fact is not trivial and can be found at [1, Proposition 4.7.2]. We make the following two remarks:

1. The "exists $g \in \partial f(x^\star)$" cannot be replaced with "for any $g \in \partial f(x^\star)$": for example, if $f(x) = |x|$ over $\mathcal{C} = [-1, +1]$, $x^\star = 0$, but we can only take $g = 0 \in \partial f(0)$ such that Equation (2) is true.

2. If f is differentiable, then the above fact is not hard to show: indeed, we only need to check that $\forall x \in \mathcal{C}, \quad \left\langle \nabla f(x^\star), x - x^\star \right\rangle \ge 0$. If this were not true, i.e. $\left\langle \nabla f(x^\star), x - x^\star \right\rangle < 0$, then it can be seen that
$$f(x^\star + \alpha(x - x^\star)) = f(x^\star) + \alpha \cdot \left\langle \nabla f(x^\star), x - x^\star \right\rangle + o(\alpha),$$
and is smaller than $f(x^\star)$ when α is small enough; contradiction.

Fact 4. *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 $\partial f(v) = \left\{\nabla f(v)\right\}$.*

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:*

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

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 5 (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 6. *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.

We have the following simple additivity property on strong convexity simply by definition:

Lemma 1. *If f_1 and f_2 are λ_1- and λ_2- strongly convex with respect to $\|\cdot\|$ respectively, then $f_1 + f_2$ is $\lambda_1 + \lambda_2$-strongly convex. Specifically, a λ-strongly convex function plus a convex function is still λ-strongly convex.*

Fact 7. *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, using the equivalence between items 1 and 2, 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 $f(x) = \sum_{i=1}^d x_i \ln x_i, x \in \left\{x \in \mathbb{R}^d : x_i > 0, \forall i, \text{ and } \sum_{i=1}^d x_i \le B_1\right\}$, then it can be checked by second-order Taylor's Theorem that $D_f(u,v) \ge \frac{1}{2B_1}\|u-v\|_1^2$, in other words, f is $\frac{1}{B_1}$-strongly convex with respect to $\|\cdot\|_1$.

Strongly convex functions have unique global minima, as given by the following fact:

Fact 8. *If $f : \mathcal{C} \to \mathbb{R}$ is λ-strongly convex, and $x^\star$ is a global minimum of f in $\mathcal{C}$, then $f(x) - f(x^\star) \ge \frac{\lambda}{2}\|x - x^\star\|^2$. Consequently, if $x \in \mathcal{C}$ is such that $f(x) \le f(x^\star)$, then $x = x^\star$.*

Proof. Note that for all $g \in \partial f(x^\star)$, we have that for all $x \in \mathcal{C}$,

$$f(x) - f(x^\star) \ge \langle g, x - x^\star \rangle + \frac{\lambda}{2}\|x - x^\star\|^2.$$

Now, by first order optimality condition (Fact 3), we also have that there exists $g_0 \in \partial f(x^\star)$, such that for all $x \in \mathcal{C}$,

$$\langle g_0, x - x^\star \rangle \ge 0.$$

Combining the above two inequalities, we immediately conclude that

$$f(x) - f(x^\star) \ge \frac{\lambda}{2}\|x - x^\star\|^2.$$

The second statement directly follows from the point separation property of norms. □

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

Fact 9. *Suppose f is twice differentiable. f is λ-strongly convex with respect to $\|\cdot\|_2$ iff for any x, $\nabla^2 f(x) \succeq \lambda I$.*

1.4 Smoothness

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

Fact 10. *The following are equivalent:*

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

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

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

Fact 11. *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 $f : \mathcal{C} \to \mathbb{R}$, use all its tangents to charaterize it.

Fix a slope s, we would like 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:

$$b(s) = \min_{x\in\mathcal{C}} \big(f(x) - \langle s, x\rangle\big) .$$

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

Definition 9. *Given convex function $f : \mathcal{C} \to \mathbb{R}$, its Legendre-Fenchel conjugate (dual), $f^\star$, is defined as*

$$f^\star(s) = \max_{x\in\mathcal{C}} \big(\langle s, x\rangle - f(x)\big) .$$

Remark. Alternatively, if we extend f to domain $\mathbb{R}^d$ using the definition of $\bar{f}$ in Equation 1, and taking the Legendre-Fenchel dual, we get the same $f^\star$. Namely,

$$\max_{x\in\mathcal{C}} \big(\langle s, x\rangle - f(x)\big) = \max_{x\in\mathbb{R}^d} \big(\langle s, x\rangle - \bar{f}(x)\big) .$$

This can be easily seen by noting that if $x \notin \mathcal{C}$, then it must not achieve the maximum on the function of x on the right hand side, as $\langle s, x\rangle - \bar{f}(x) = -\infty$.

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 5, and the facts that $h_x(s) = \langle s, x\rangle - f(x)$ has subgradient x, and $f^\star(s) = \max_x h_x(s)$, we can see that

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

Let us look at the dual of $f^\star$, that is $f^{\star\star}(x) = \max_s \big(\langle x, s\rangle - f^\star(s)\big)$. This equation has a nice geometric interpretation. Recall that for each s, $\langle x, s\rangle - f^\star(s)$ is the tangent of f of slope s; therefore, by varying s in $\mathbb{R}$, we get a collection of lines below f. $f^{\star\star}$ is an upper envelope of these lines. Curiously, under mild assumptions, $f^{\star\star}$ is exactly the original function f.

Fact 12. *Suppose f is closed (in that $\big\{(x, t) \in \mathbb{R}^{d+1} : f(x) \le t\big\}$ 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 13 (Fenchel-Young's Inequality)**.** *For any pairs of x and s in $\mathbb{R}^d$,*

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

Example 5.

1. *Suppose $f(x) = \begin{cases} 0 & \|x\| \le 1 \\ +\infty & \|x\| > 1 \end{cases}$. Its conjugate function $f^\star(s) = \max_{x:\|x\|\le 1} \langle x, s\rangle = \|s\|_\star$.*
2. *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.*
3. *For any norm $\|\cdot\|$, if $f(x) = \frac{\lambda}{2}\|x\|^2$, then $f^\star(s) = \frac{1}{2\lambda}\|s\|_\star^2$.*
4. *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 $f^\star(s) = \ln \sum_{s=1}^d e^{s_i}$.*

5. *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 $f \geq g$, then by the definition of conjugate function, $f^\star \leq g^\star$.

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

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

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

Fact 14. *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. The proof of the "if" statement can be found at [9, Theorem 3]. Our goal is to show that for u, v,

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

where

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

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), and Fact 8, we have:

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

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 [5, 17]: see Framework 1.

Equivalent goal: minimize regret against *the best fixed point in hindsight*:

$$\mathrm{Reg}(T, \mathcal{C}) = \max_{w^\star \in \mathcal{C}} \mathrm{Reg}(T, w^\star) = \sum_{t=1}^T f_t(w_t) - \min_{w^\star \in \mathcal{C}} \sum_{t=1}^T f_t(w^\star),$$

where

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

Algorithm 1 Online convex optimization (OCO)

Require: Convex decision set $\mathcal{C}$.

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

Learner chooses $w_t \in \mathcal{C}$,

Learner receives a convex loss f_t.

end for

Goal: minimize cumulative loss $\sum_{t=1}^T f_t(w_t)$.

Definition 10. *Suppose for every f_t, $f_t(w) = \langle g_t, w\rangle$ 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 Φ, set

$$\begin{aligned} w_t &= \operatorname*{argmin}_w \sum_{s=1}^{t-1} \langle g_s, w\rangle + \Phi(w) \\ &= \operatorname*{argmax}_w \langle -G_{t-1}, w\rangle - \Phi(w) \\ &= \nabla\Phi^\star(-G_{t-1}), \end{aligned}$$

where $G_t = \sum_{s=1}^t g_s$ is the cumulative gradients. the mapping $\nabla\Phi^\star$ is called the *mirror map* or *link function*, 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 $\Phi(w) = \begin{cases} \frac{1}{\eta}\sum_{i=1}^d w_i \ln w_i, & w \in \Delta^{d-1} \\ +\infty, & w \notin \Delta^{d-1} \end{cases}$, then it can be checked that*
$$w_{t,i} = \exp\left(-\eta \sum_{s=1}^{t-1} \ell_{s,i}\right).$$

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

3. *Online gradient descent with lazy projections: let $\Phi(w) = \begin{cases} \frac{1}{2\eta}\|w\|^2, & w \in \mathcal{C}, \\ +\infty, & w \notin \mathcal{C} \end{cases}$, then it can be shown that,*
$$w_t = \operatorname*{argmin}_{w\in\mathcal{C}} \|w - (-\eta G_{t-1})\|_2,$$
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 agains benchmark $w^\star$:*
$$\mathrm{Reg}(T, w^\star) = \sum_{t=1}^T \langle g_t, w_t - w\rangle \le \Phi(w^\star) - \min_{w'} \Phi(w') + \frac{1}{\lambda}\sum_{t=1}^T \|g_t\|_\star^2.$$

Proof. Recall that $f_t(w) = \langle g_t, w \rangle$. 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, w_{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: w_{t+1} relies on information on g_t, which is unavailable at the beginning of round t. Nevertheless, we will show that w_t, the decision point selected by FTRL, is close to w_{t+1}, therefore the regret of FTRL can be bounded in terms of that of BTRL.

Step 1: Analysis of BTRL. Denote by $f_0(w) = \Phi(w)$. Consider a modification of the original OCO game: there is an extra round of online convex optimization at the beginning, namely round 0. Therefore, algorithmically, BTRL is equivalent to Be-the-leader (BTL) on $\{f_0, f_1, \ldots, f_T\}$. We will show that BTL has nonpositive regret on this modified OCO game, and relate this regret guarantee to that of the original OCO game.

Lemma 2 (Be the leader)**.** *For any $w^\star$,*

$$\sum_{t=0}^{T} f_t(w_{t+1}) \le \sum_{t=0}^{T} f_t(w^\star).$$

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

$$\sum_{t=0}^{T} f_t(w_{T+1}) \le \sum_{t=0}^{T} f_t(w^\star).$$

Now let us focus on all but the last term in the left hand side, that is, $\sum_{t=0}^{T-1} f_t(w_{t+1})$. As as $w_T = \operatorname{argmin}_w \sum_{t=0}^{T-1} f_t(w)$, we have that

$$\left(\sum_{t=0}^{T-1} f_t(w_T)\right) + f_T(w_{T+1}) \le \sum_{t=0}^{T} f_t(w_{T+1}) \le \sum_{t=0}^{T} f_t(w^\star).$$

By iteratively using the fact that $w_\tau = \operatorname{argmin}_w \sum_{t=0}^{\tau-1} f_t(w)$, we have that

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

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

Lemma 2 immediately implies that:

$$\sum_{t=1}^{T} \langle g_t, w_{t+1} - w^\star \rangle \le \Phi(w^\star) - \Phi(w_1). \tag{3}$$

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

Lemma 3 (Stability)**.**

$$\sum_{t=1}^{T} \langle g_t, w_t - w_{t+1} \rangle \le \frac{1}{\lambda} \sum_{t=1}^{T} \|g_t\|_\star^2. \tag{4}$$

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

$$\|w_t - w_{t+1}\| \le \frac{1}{\lambda}\|g_t\|_\star.$$

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

$$\|w_t - w_{t+1}\| = \|\nabla\Phi^\star(-G_{t-1}) - \nabla\Phi^\star(-G_t)\|.$$

Recall that Φ is λ-strongly convex, by Fact 14, $\Phi^\star$ is $\frac{1}{\lambda}$-smooth. Therefore the right hand side is indeed at most $\frac{1}{\lambda}\| - G_{t-1} - (-G_t)\| = \frac{1}{\lambda}\|g_t\|_\star$. □

The theorem is proved by summing Equations (3) and (4) together. □

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 $g_t \in \partial f_t(w_t)$ is a subgradient of f_t at location x_t. We have that for any $w^\star$,

$$f_t(w_t) - f_t(w^\star) \le \langle g_t, w_t - w^\star\rangle .$$

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

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

for some regret function $\mathrm{R}(T)$. This implies that

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

2.3 Instantiations of FTRL: theoretical guarantees

1. Online gradient descent (OGD) [17]: $\Phi(w) = \frac{1}{2\eta}\|w\|_2^2$, which is $\frac{1}{\eta}$-strongly convex wrt $\|\cdot\|_2$. FTRL with Φ has regret

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

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

   Suppose we would like to guarantee $\mathrm{Reg}(T, \mathcal{C})$ with $\mathcal{C} \subset \big\{w : \|w\| \le B_2\big\}$. 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:

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

which is also $\frac{1}{\eta}$-strongly convex wrt $\|\cdot\|_2$. Note that FTRL in this case ensures $w_t \in \mathcal{C}$ at every round. This is useful in error or safety critical settings (for exmaple, taking actions in $\mathcal{C}$ prevents self-driving cars from falling off cliffs). FTRL with Φ has regret:

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

for all benchmark $w^\star \in \mathcal{C}$. 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 ($p \in (1, 2]$) [6, 4]: It is known that $\Phi(w) = \frac{1}{2\eta}\|w\|_p^2$ is $\frac{p-1}{\eta}$-strongly convex wrt $\|\cdot\|_p$ (see [15, Lemma 17] or [cesa2006prediction , Corollary 2.1]). FTRL with R has regret:

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

If $\mathcal{C} \subset \left\{w : \|w\|_p \le B_p\right\}$, and for all t, $\|g_t\|_q \le R_q$, setting $\eta = \frac{B_p}{R_q\sqrt{2(p-1)T}}$ implies that

$$\mathrm{Reg}(T, \mathcal{C}) \le B_p R_q \sqrt{\frac{2T}{p-1}}.$$

4. Exponentiated gradient (Hedge) [3, 10]: consider the negative entropy regularizer

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

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

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

It can be seen that $\sum_{i=1}^{d} w_i^\star \ln w_i^\star \le 0$, on the other hand, $\min_{w' \in \Delta^{d-1}} \sum_{i=1}^{d} w_i' \ln w_i' = -\max_{w' \in \Delta^{d-1}} H(w)$, where $H(w)$ is the entropy of probability vector w. 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

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

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:

$$\left\{ w : \forall i, w_i > 0, \sum_{i=1}^{d} w_i = B_1 \right\},$$

for general $B_1 > 0$; we skip the discussion for brevity.)

Algorithm 2 Online linear classification (with FTRL)

Require: Regularizer R, stepsize η.

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

Learner chooses $w_t = \operatorname{argmin}_w \left(\frac{1}{\eta}\Phi(w) + \sum_{s=1}^{t-1} \langle g_s, w \rangle \right) = \nabla(\frac{1}{\eta}\Phi)^\star(-\sum_{s=1}^{t-1} g_s) \in \mathbb{R}^d$,

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 $\sum_{t=1}^T M_t$.

2.4 Applications of FTRL to online linear classification

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, suppose for all w, $\Phi(w) \ge \Phi_{\min}$. Then, for any $w^\star \in \mathcal{C}$,*

$$\sum_{t=1}^T M_t \le \frac{1}{1 - \eta R^2} \left(L_T(w^\star) + \frac{\Phi(w^\star) - \Phi_{\min}}{\eta} \right),$$

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 $w^\star \in \mathcal{C}$ such that the data is separable by a margin of 1: $\forall t, \langle w^\star, y_t x_t \rangle \ge 1$, then setting $\eta = \frac{1}{2R^2}$ implies that

$$\sum_{t=1}^T M_t \le 2R^2 \cdot (\Phi(w^\star) - \Phi_{\min}),$$

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

Proof. As R is 1-strongly convex wrt $\|\cdot\|$, $\frac{\Phi}{\eta}$ 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 $w^\star$,

$$\sum_{t=1}^T f_t(w_t) - \sum_{t=1}^T f_t(w^\star) \le \frac{\Phi(w^\star) - \min_{w'} \Phi(w')}{\eta} + \sum_{t=1}^T \eta \|g_t\|^2 \le \frac{\Phi(w^\star) - \Phi_{\min}}{\eta} + \sum_{t=1}^T \eta \|g_t\|^2,$$

where the second inequality uses the uniform lower bound of Φ.

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

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

that is,

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

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

Instantiations. We consider two settings of Φ:

1. Let $\Phi(w) = \frac{1}{2}\|w\|^2$. This gives the well-known Perceptron algorithm [14]:

$$w_t = \operatorname*{argmin}_w \left(\frac{1}{2\eta}\|w\|_2^2 + \sum_{s=1}^{t-1} \langle g_s, w \rangle \right) = -\eta \cdot \sum_{s=1}^{t-1} g_s.$$

   Suppose all examples lies in $\{x : \|x\|_2 \le R_2\}$. By Theorem 2, Perceptron has a mistake bound of

$$\sum_{t=1}^{T} M_t \le \frac{1}{1-\eta R_2^2}\left(L_T(w^\star) + \eta \|w^\star\|^2 \right),$$

   for any $w^\star \in \mathbb{R}^d$.

   Now, if the data is linearly separable by margin 1 by classifier w such that $\|w\|_2 \le B_2$, then setting $\eta = \frac{1}{2R_2^2}$ gives that

$$\sum_{t=1}^{T} M_t \le 2R_2^2 B_2^2.$$

   This is a variant of the well-known Percetron convergence theorem by Novikoff [14].

2. Let $\Phi(w) = \begin{cases} \sum_{i=1}^d w_i \ln w_i, & w \in \Delta^{d-1}, \\ +\infty, & \text{otherwise.} \end{cases}$. This gives the Winnow [11] algorithm:

$$w_{t,i} = \exp\left\{ -\eta \sum_{s=1}^{t-1} g_{s,i} \right\}, \forall i \in \{1, \ldots, d\}.$$

   Suppose all examples lies in $\{x : \|x\|_\infty \le R_\infty\}$. Also, as discussed before, we can set $\Phi_{\min} = -\ln d$ and $\Phi(w) - \Phi_{\min} \le \ln d$ for all $w^\star \in \Delta^{d-1}$. Therefore, FTRL with Φ has a mistake bound of

$$\sum_{t=1}^{T} M_t \le \frac{1}{1-\eta R_\infty^2}\left(L_T(w^\star) + \eta \ln d \right).$$

   for all $w^\star \in \Delta^{d-1}$.

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

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

   This mistake bound is in general incomparable with the Perceptron mistake bound (see our discussions on ℓ_2-ℓ_2 vs. ℓ_1-ℓ_∞ margin bounds before.)

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:

$$w_t = \operatorname*{argmin}_w \left(\Phi_{t-1}(w) + \sum_{s=1}^{t-1} \langle g_s, w \rangle \right) = \nabla \Phi_{t-1}^\star(-G_{t-1}),$$

where $\{\Phi_t\}_{t=0}^T$ is a sequence of regularizers, and 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 with adaptive regularization, abbreviated as FTRL-AR. Specifically, we will be looking at sequences of $\{\Phi_t\}$'s such that they are generated on the fly, and can thus carry information on the past g_t's.

Theorem 3 (Modified from Lemma 1 of [13])**.** *Suppose FTRL-AR uses Φ_t's that are 1-strongly convex with respect to time-varying norm $\|\cdot\|_t$. Then it has the following upper bound on its cumulative loss guarantee:*

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

Consequently,

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

Note that the above theorem supercedes Theorem 1, as it is a direct consequence of the above theorem by taking $R_t \equiv R_0$ for all t, and observing that $R_0^\star(0) = -\min_{w'} R_0(w')$.

Proof. It suffices to show that

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

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

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

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, $w_t = \nabla R_{t-1}^\star(-G_{t-1})$, therefore, the left hand side of the inequality is at most

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

recall that $D_f(\cdot,\cdot)$ is the Bregman divergence induced by f. 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 $\frac{1}{2}\| -G_t - (-G_{t-1})\|_{\star,t-1}^2 = \frac{1}{2}\|g_t\|_{\star,t-1}^2$. $\square$

Using the above meta-theorem, we can instantiate with different adpative regularizers and get online learning algorithms with different degrees of adaptivity. Below, we focus on a specific family of regularizer; that is, the squared Mahalanobis norm regularizer:

$$\Phi_t(w) = \frac{1}{2}\|w\|_{A_t}^2,$$

where $A_t \succeq A_{t-1}$ for all $t \geq 1$. Observe that $\Phi_t(w)$ is 1-strongly convex with norm $\|w\|_t = \|w\|_{A_t}$. Meanwhile, $\|g\|_{t,\star} = \|g\|_{A_t^{-1}}$. FTRL-AR selects the following point at round t:

$$w_t = \nabla \Phi_{t-1}^\star(-G_{t-1}) = -A_{t-1}^{-1} G_{t-1}.$$

Therefore, we have the following simple corollary:

Corollary 1. *Suppose FTRL-AR is executed with $\Phi_t(w) = \frac{1}{2}\|w\|_{A_t}^2$ for a sequence of monotonically increasing positive definite matrices $\{A_t\}$. Then,*

$$\mathrm{Reg}(T, w^\star) \leq \frac{1}{2}\|w\|_{A_T}^2 + \sum_{t=1}^{T} \|g_t\|_{A_{t-1}^{-1}}^2.$$

We discuss several nice consequences of the corollary below.

Online gradient descent with adaptive step-sizes [17]. One instantiation of Corollary 1 is to let $A_t = \frac{\sqrt{t+1}}{\eta_0} I_d$, which implies that,

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

Suppose the benchmark set $\mathcal{C}$ is defined as $\left\{w : \|w^\star\| \leq B_2\right\}$. If one knows that $\|g_t\|_2 \leq R_2$, then setting $\eta_0 = \frac{R_2}{B_2}$ gives

$$\mathrm{Reg}(T, w^\star) \leq O\left(R_2 B_2 \sqrt{T}\right), \quad \forall w \in \mathcal{C}\,.$$

Even we don't have any prior knowledge on the norm of the g_t's, setting $\eta_0 = 1$ gives

$$\mathrm{Reg}(T, w^\star) \leq O\left((R_2^2 + B_2^2)\sqrt{T}\right), \quad \forall w \in \mathcal{C}\,.$$

Regularization that depends on historical gradient lengths. We let $\sigma > 0$, and $A_t = \frac{\sqrt{\sigma + \sum_{s=1}^t \|g_s\|^2}}{\eta_0} I_d$.

Corollary 1 implies that, with this setting of Φ_t,

$$\mathrm{Reg}(T, w^\star) \leq \frac{\sqrt{\sigma + \sum_{s=1}^t \|g_s\|^2}}{2\eta_0}\|w^\star\|^2 + \sum_{s=1}^{t} \frac{\eta_0 \|g_s\|^2}{\sqrt{\sigma + \sum_{s=1}^{t-1} \|g_s\|^2}}$$

If $\sigma \geq \max_{t=1}^T \|g_t\|_2^2$ [1], it can be shown that second term on the right hand side is at most

$$2\sum_{s=1}^{t} \frac{\eta_0 \|g_s\|^2}{\sqrt{\sigma + \sum_{s=1}^{t} \|g_s\|^2}} \quad \leq \quad 4\sqrt{\sigma + \sum_{s=1}^{t} \|g_s\|^2}.$$

where the inequality is from Lemma 7.

Therefore, the regret is at most:

$$\mathrm{Reg}(T, w^\star) = O\left(\sqrt{\sigma + \sum_{s=1}^{t} \|g_s\|^2}\left(\frac{\|w^\star\|^2}{\eta_0} + \eta_0\right)\right).$$

If $\eta_0 = \|w^\star\|$, and σ is a constant factor away from $\max_{t=1}^T \|g_t\|_2^2$, then the regret guarantee is $O(\|w^\star\|\sqrt{\sum_{s=1}^t \|g_s\|^2})$, which can be much better than $R_2^2 B_2^2$.

[1] It turns out that a sibling of FTRL, namely Online Mirror Descent, can get rid of this extra σ while achieving the same guarantee. We refer the reader to [12, Lecture 5].

Adaptive subgradient methods (Adagrad) [2]. More generally, we allow adaptive regularization matrix A_t being a general diagnoal matrix, or even a matrix with nonzero diagonal entries.

Specifically, one can let

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

be a diagonal adaptive regularizer. Here the $\mathrm{diag}(M)$ takes a full $d \times d$ matrix and set all its off-diagnoal entries to be zero. The induced FTRL algorithm is called *AdaGrad with diagnonal matrices*. This is one of the most widely used gradient-based optimization algorithm in modern machine learning.

Specifically, we can look at the point it selects at iteration t, w_t:

$$w_{t,i} = -\eta \cdot \frac{\sum_{s=1}^{t-1} g_{s,i}}{\sqrt{\sigma + \sum_{s=1}^{t-1} g_{s,i}^2}}, \quad \forall i \in \{1, \ldots, d\}\,.$$

Intuitively, the algorithm is performing online gradient descent on every coordinate separately: for coordinate i, if we have seen a big cumulative gradient along the direction of e_i, then we decrease the learning rate on that direction, as we have already learned a lot there.

Corollary 1 gives that,

$$\mathrm{Reg}(T, w^\star) \le O\left(\frac{\|w^\star\|_{A_T}^2}{2} + \eta \sum_{t=1}^{T}\sum_{i=1}^{d} \frac{g_{t,i}^2}{\sqrt{\sigma + \sum_{s=1}^{t-1} g_{s,i}^2}}\right).$$

If $\sigma \ge \max_{t=1}^{T} \max_{i=1}^{d} g_{t,i}^2$, then using Lemma 7 and a similar reasoning as the last subsection (such that we can replace the $t-1$ by t in the denominator with only a constant factor overhead), we can show that the second term is at most $\sum_{i=1}^{d} \sqrt{\sum_{t=1}^{T} g_{t,i}^2}$.

Thus, note that $\|w\|_M^2 \le \|w\|_\infty^2 \sum_{i=1}^{d} M_{ii}$ for diagnoal M, we get that the above is at most

$$\mathrm{Reg}(T, w^\star) \le O\left(\left(\frac{\|w^\star\|_\infty^2}{\eta} + \eta\right) \cdot \left(\sum_{i=1}^{d} \sqrt{\sum_{t=1}^{T} g_{t,i}^2}\right)\right).$$

If $\eta = \|w^\star\|_\infty$, then AdaGrad gives a regret bound of $O\left(\|w^\star\|_\infty \cdot \left(\sum_{i=1}^{d} \sqrt{\sum_{t=1}^{T} g_{t,i}^2}\right)\right)$, which is a new regret guarantee incomparable with the ones obtained by (variants of) online gradient descent discussed above.

Example 7. *Let us compare the regret bound of AdaGrad with that obtained by online gradient descent with optimal tuning of step size, that is:*

$$\mathrm{Reg}(T, w^\star) = O\left(\|w^\star\|_2 \sqrt{\sum_{t=1}^{T} \|g_t\|_2^2}\right) = O\left(\|w^\star\|_2 \sqrt{\sum_{t=1}^{T}\sum_{i=1}^{d} g_{t,i}^2}\right).$$

Suppose the $g_{t,i}$'s are such that only $g_{t,1}$'s are nonzero. Then the second fact on the regret bounds agree with each other. Therefore, in terms of the final regret bound, AdaGrad is better, as $\|w^\star\|_\infty \le \|w^\star\|_2$.

Alternatively, one can let

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

be a nondiagnoal adaptive regularizer. The induced FTRL algorithm is called *AdaGrad with full matrices*. We can still apply Corollary 1 to obtain a regret guarantee, but the interpretation is slightly more involved, and we refer the reader to [7, Section 5.6] for details.

3 OCO for strongly convex functions

Motivating example: we would like a fast optimizer for regularized loss minimization, e.g. soft-margin SVM or logistic regression:

$$\min_w F(w), \quad \text{where } F(w) = \mathbb{E}_{(x,y)\sim D}\left(\frac{\lambda}{2}\|w\|_2^2 + (1 - \langle w, yx\rangle)_+\right),$$

or $F(w) = \mathbb{E}_{(x,y)\sim D}\left(\frac{\lambda}{2}\|w\|_2^2 + \ln\Big(1 + \exp\left(-\langle w, yx\rangle\right)\Big)\right)$. Throughout the rest of the section, let us consider soft-margin SVM for concreteness.

Here, letting $f(w,(x,y)) = \frac{\lambda}{2}\|w\|_2^2 + (1 - \langle w, yx\rangle)_+$, we can write $F(w) = \mathbb{E}_{(x,y)\sim D} f(w,(x,y))$.

If one can develop a fast OCO algorithm with $\left\{f_t(w) \triangleq f(w,(x_t,y_t))\right\}_{t=1}^T$, with a small regret guarantee $\mathrm{R}(T)$, as we have seen before, one can use online-to-batch conversion, and run the OCO algorithm on f_t's induced by iid $(x_t, y_t) \sim D$ to get a $\bar{w}_T$ that has excess expected regularized loss $\frac{\mathrm{R}(T)}{T}$, in other words,

$$\mathbb{E}\, F(\bar{w}_T) - \min_w F(w) \le \frac{\mathrm{R}(T)}{T}.$$

One baseline is the FTRL algorithm with squared norm regularizer $\Phi(w) = \frac{1}{2\eta}\|w\|^2$, with surrogate convex function $\tilde{f}_t(w) = \langle g_t, w\rangle$, where $g_t \in \partial f(w_t)$. This achieves a regret of $O(\sqrt{T})$; moreover, each step, the algorithm simply calculates $w_t = -\eta\sum_{s=1}^t g_{t-1}$, which can be maintained efficiently on the fly. It can be checked that to guarantee an expected excess loss of ϵ, the computational complexity is $O(\frac{d}{\epsilon^2})$.

In fact, we can do better! In this section, we show that by utilizing the structure that 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, $O(\frac{\ln T}{\lambda})$.

How to achieve this? We will use the adaptive regularization method developed in the last section.

Theorem 4. *Suppose all f_t's are λ-strongly convex and L-Lipschitz wrt $\|\cdot\|_2$. Then FTRL-AR with adaptive regularizer $R_t(w) = \frac{\lambda}{2}\|w\|^2 + \sum_{s=1}^t \frac{\lambda}{2}\|w_s - w\|^2$ has regret*

$$\mathrm{Reg}(T, w^\star) = O\left(\frac{L^2 \ln T}{\lambda}\right).$$

Proof. Recall that FTRL-AR has the following regret guarantee:

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

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

$$f_t(w_t) - f_t(w^\star) \le \langle g_t, w_t - w^\star\rangle - \frac{\lambda}{2}\|w_t - w^\star\|^2.$$

This implies that,

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

This motivates the definition of R_T, so that $R_T(w^\star)$ cancels out the negative terms induced by linear approximation. Observe that R_t is 1-strongly convex with respect to $\|\cdot\|_t = \|\cdot\|_{\lambda(t+1)I}$. We therefore get:

$$\mathrm{Reg}(T, w^\star) \le \frac{\lambda}{2}\|w^\star\|^2 + \sum_{t=1}^{T} \frac{\|g_t\|^2}{\lambda t} \le \frac{\lambda}{2}\|w^\star\|^2 + \frac{L^2}{\lambda}(1+\ln T) = O\left(\frac{L^2 \ln T}{\lambda}\right).$$

where the penultimate inequality uses the L-Lipschitzness of f and Fact 6, and the last inequality uses the simple fact that $\sum_{t=1}^{T} \frac{1}{t} \le 1 + \ln T$. □

What is the induced FTRL-AR algorithm? It can be shown that

$$\begin{aligned} w_t &= \operatorname*{argmin}_w \left(\sum_{s=1}^{t-1} \langle g_s, w \rangle + \frac{\lambda}{2}\|w\|^2 + \sum_{s=1}^{t-1} \frac{\lambda}{2}\|w_s - w\|^2 \right) \\ &= \frac{1}{t}\left(\sum_{s=1}^{t-1} w_s - \frac{1}{\lambda}\sum_{s=1}^{t-1} g_s \right), \end{aligned}$$

which can be obtained on the fly with $O(d)$ time per round by maintaining $\sum_{s=1}^{t-1} w_s$ and $\sum_{s=1}^{t-1} g_s$ online. Therefore, to obtain an excess loss guarantee of ϵ, one can let run FTRL-AR with the specified regularizer with $T = O\left(\frac{1}{\lambda\epsilon}\ln\frac{1}{\lambda\epsilon}\right)$, which has a total running time of $\tilde{O}(\frac{d}{\lambda\epsilon})$ (where $\tilde{O}$ ignores logarithmic factors).

Below are materials not covered in the class and can be safely skipped!

A brief introduction to online gradient descent.* Here we introduce an alternative method for online strongly convex optimization called *online gradient descent (OGD)*. OGD and FTRL are closely related; it can be shown that FTRL with certain regularizer is algorithmically equivalent to OGD with a fixed stepsize, under natural assumptions. In a nutshell, OGD performs the following steps of update at every round:

$$w_{t+1} \triangleq w_t - \eta_t g_t, \text{ where } g_t \in \partial f_t(w_t),$$

here, $\eta_t > 0$ is the time-varying step size parameter. Intuitively, if all $f_t \equiv f$'s are the same, then OGD selects w_t's that walks in the negative gradient direction of f, making the function value smaller.

The following theroem is now well-known; see [8, Theorem 1] or [16, Section 14.4.4 and 14.5.3].

Theorem 5. *If all f_t's are λ-strongly convex and L-Lipschitz wrt $\|\cdot\|_2$, then OGD with $\eta_t = \frac{1}{\lambda t}$ has the following regret guarantee:*

$$\mathrm{Reg}(T, w^\star) = O(\frac{L^2 \ln T}{\lambda}).$$

Proof. We have the following key claim:

$$\begin{aligned} \langle g_t, w_t - w^\star \rangle &= \frac{\|w_t - w^\star\|^2 - \|w_{t+1} - w^\star\|^2}{2\eta_t} + \eta_t \|g_t\|^2 & (5) \\ &\le \frac{\lambda t \|w_t - w^\star\|^2}{2} - \frac{\lambda t \|w_{t+1} - w^\star\|^2}{2} + \frac{1}{\lambda t} L^2. & (6) \end{aligned}$$

This inequality has a nice intuitive explanation: if we have a larger instaneous regret $\langle g_t, w_t - w^\star \rangle$ at round t, then w_{t+1} will be brought closer to $w^\star$, as the right hand side is larger. Hence, we can think of $\|w_t - w^\star\|^2$ as some kind of "potential function" as in the analysis of Consistency and Halving in online classification.

Equation (5) is true, because of the following simple calculation:

$$\|w_{t+1} - w^\star\|^2 = \|w_t - w^\star - \eta_t g_t\|^2 = \|w_t - w^\star\|^2 - 2\langle \eta_t g_t, w_t - w^\star \rangle + \|\eta_t g_t\|^2.$$

Now continuing Equation 6, we would like to relate it to the instanteous regret $f_t(w_t) - f_t(w^\star)$. Recall that by strong convexity, we have

$$f_t(w_t) - f_t(w^\star) \le \langle g_t, w_t - w^\star \rangle - \frac{\lambda}{2}\|w_t - w^\star\|^2,$$

This implies that

$$f_t(w_t) - f_t(w^\star) \le \frac{\lambda(t-1)\|w_t - w^\star\|^2}{2} - \frac{\lambda t \|w_{t+1} - w^\star\|^2}{2} + \frac{1}{\lambda t} L^2.$$

Observe that the first two terms on the right hand side is *telescoping*: summing over all $t = 1, 2, \ldots, T$, we get

$$\mathrm{Reg}(T, w^\star) = \sum_{t=1}^T f_t(w_t) - f_t(w^\star) \le \frac{\lambda \cdot 0 \|w_1 - w^\star\|^2}{2} - \frac{\lambda \cdot T \|w_{T+1} - w^\star\|^2}{2} + \left(\sum_{t=1}^T \frac{1}{\lambda t} \right) L^2 \le \frac{L^2(1 + \ln T)}{\lambda}.$$

□

4 OCO for exp-concave functions*

Motivating example 1: sequential investing. There are d types of assets (e.g. stocks, bonds, commodities), with different growth rates at every timestep.

Suppose we start with unit wealth, $W_1 \leftarrow 1$.

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

1. Given the current wealth W_t, allocate $w_t \in \Delta^{d-1}$ (spend $w_{t,i}$ fraction of current wealth, i.e, $W_t \cdot w_{t,i}$, to asset i)
2. Receive where $c_t \in \mathbb{R}^d_+$, and $c_{t,i}$ is the growth of the asset i at timestep t (defined as the ratio between the prices of stock i at timestep $t+1$ and t).
3. Cash out all assets, get new wealth W_{t+1}. Observe that
$$W_{t+1} = W_t \left(\sum_{t=1}^{T} w_{t,i} c_{t,i} \right),$$
i.e. $\ln(W_{t+1}) = \ln(W_t) - f_t(w_t)$, where loss
$$f_t(w_t) = -\ln\big(\langle c_t, w_t \rangle\big).$$
Consequently, $\ln(W_{T+1}) = -\sum_{t=1}^{T} f_t(w_t)$, and therefore, maximizing W_{T+1} amounts to minimizing the cumulative loss.

Goal: compete with the best constant rebalanced portfolio in hindsight (abbrev. CRP; that is, at the beginning of every day, allocate a constant fraction $w^\star \in \Delta^{d-1}$ to all assets. This has the advantage that the portion of each asset is well-controlled, and therefore can effective control risk.) Concretely,

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

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

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

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

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

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

The lemma is proved in light of the fact that $\exp\big(-\alpha f(x)\big) > 0$. $\square$

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

The following lemma shows that a univariate strongly convex function, composed with a linear function, is exp-concave.

Lemma 5. *Suppose h is λ-strongly convex and has gradient at most G. Then for any vector w, $h(\langle w, x\rangle)$ is $\frac{\lambda}{G^2}$-exp-concave.*

Proof. Because $\nabla h(\langle w, x\rangle) = h'(\langle w, x\rangle) w$, we have the following two simple observations:

$$\nabla^2 h(\langle w, x\rangle) = h''(\langle w, x\rangle) w w^\top \succeq \lambda w w^\top.$$

$$\nabla h(\langle w, x\rangle) \cdot \nabla h(\langle w, x\rangle)^\top \preceq h'(\langle w, x\rangle)^2 w w^\top \leq G^2 w w^\top.$$

The lemma is shown in light of Lemma 4. $\square$

For online least-square regression with domain $\big\{w : \|w\|_2 \leq B\big\}$ and all $x \in \big\{x : \|x\|_2 \leq R\big\}$ 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, $f(w) = \frac{1}{2}(\langle w, x\rangle - y)^2$ is $\frac{1}{(RB+Y)^2}$-exp-concave.

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

Lemma 6. *If f is α-exp-concave and L-Lipschitz, then for any two points $u, v \in \big\{w : \|w\|_2 \leq B\big\}$, we have*

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

where $\tilde{\alpha} = \min(\frac{1}{16BL}, \frac{\alpha}{2})$.

Proof. Let $\gamma \leq \alpha$ be a number to be decided. By the concavity of $\exp\Big(-\gamma \tilde{f}(w)\Big)$, we have that for any two points u, v with norm at most B,

$$\exp\big(-\gamma f(u)\big) \leq \exp\big(-\gamma f(v)\big) - \gamma \exp\big(-\gamma f(v)\big) \big\langle \nabla f(v), u - v \big\rangle .$$

In other words,

$$\exp\big(-\gamma (f(u) - f(v))\big) \leq 1 - \gamma \big\langle \nabla f(v), u - v \big\rangle .$$

Therefore,

$$-\gamma (f(u) - f(v)) \leq \ln\Big(1 - \gamma \big\langle \nabla f(v), u - v \big\rangle\Big).$$

Let us take $\gamma = \min(\alpha, \frac{1}{8BL})$. This ensures that $\gamma \leq \alpha$, and $\Big|\gamma \big\langle \nabla f(v), u - v \big\rangle\Big| \leq \frac{1}{8BL} \cdot 2BL \leq \frac{1}{4}$. Using the simple fact that $\ln(1 - z) \leq -z - \frac{1}{4} z^2$ for all $z \in [-\frac{1}{4}, \frac{1}{4}]$, we have that

$$-\gamma (f(u) - f(v)) \leq -\gamma \big\langle \nabla f(v), u - v \big\rangle - \frac{1}{4} \gamma^2 (\big\langle \nabla f(v), u - v \big\rangle)^2.$$

The lemma is proved by dividing both sides by $-\gamma$ and recognizing that $\tilde{\alpha} = \frac{\gamma}{2}$. $\square$

Algorithm with logarithmic regret: adaptive regularization. We will be using Lemma 6 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:

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

In addition, by Lemma 6, we have that

$$\sum_{t=1}^{T} f_t(w_t) - f_t(w^\star) \le \sum_{t=1}^{T} \langle g_t, w_t - w^\star \rangle - \sum_{t=1}^{T} \frac{\tilde{\alpha}}{2}(w^\star - w_t)^\top \nabla f(w_t) \nabla f(w_t)^\top (w^\star - w_t). \tag{7}$$

Using the same idea behind logarithmic regret for strongly convex functions (Theorem 4), we choose a specialized regularizer that takes advantage of the negative terms relating OCO with $\{f_t\}$ and OLO with $\big\{w \mapsto \langle g_t, w\rangle\big\}$, that induces a logarithmic regret algorithm.

Theorem 6. *Suppose f_t's are all α-exp-concave, and are L-Lipschitz wrt $\|\cdot\|_2$. Denote by $\mathcal{C} = \big\{w : \|w\|_2 \le B\big\}$ be the feasible set. Then, FTRL-AR with adaptive regularizer $\{\Phi_t\}$, where $\Phi_t(w) = \frac{\alpha L^2}{2}\|w\|_2^2 + \sum_{s=1}^{t} \frac{\tilde{\alpha}}{2}(w - w_s)^\top \nabla f_s(w_s) \nabla f_s(w_s)^\top (w - w_s)$ has regret*

$$\mathrm{Reg}(T, w^\star) \le O\left(d \cdot (\frac{1}{\alpha} + LB) \ln T\right),$$

against all $w^\star \in \mathcal{C}$. Recall that $\tilde{\alpha} = \min(\frac{1}{16BL}, \frac{\alpha}{2})$ is defined in Lemma 6.

Proof. Observe that for every t, $\Phi_t(x)$ is 1-strongly convex with respect to $\|\cdot\|_t = \|\cdot\|_{A_t}$, where $A_t = \tilde{\alpha}\left(L^2 I + \sum_{s=1}^{t} \nabla f_s(w_s) \nabla f_s(w_s)^\top\right)$.

With this choice of $\{\Phi_t\}$, theorem 3 implies that

$$\sum_{t=1}^{T} \langle g_t, x_t - w^\star \rangle \le \frac{\tilde{\alpha} L^2}{2}\|w^\star\|^2 + \sum_{t=1}^{T} \frac{\tilde{\alpha}}{2}(w^\star - w_t)^\top \nabla f_t(w_t) \nabla f_t(w_t)^\top (w - w_t) + \sum_{t=1}^{T} \|g_t\|_{\star,t-1}^2.$$

In combination with Equation (7), note that the quadratic terms on w cancel out completely. We have:

$$\mathrm{Reg}(T, w^\star) \le \frac{\tilde{\alpha} L^2}{2}\|w^\star\|_2^2 + \sum_{t=1}^{T} \|g_t\|_{A_{t-1}^{-1}}^2. \tag{8}$$

As $A_{t-1} \succeq \tilde{\alpha} L^2 I$, $A_t = A_{t-1} + \tilde{\alpha} g_t g_t^\top \preceq 2A_{t-1}$. This implies that we can simplify the right hand side of Equation (8) as follows:

$$\begin{aligned}
(8) \quad &\le \quad \frac{\tilde{\alpha} L^2 B^2}{2} + 2\sum_{t=1}^{T} \|g_t\|_{A_t^{-1}}^2 \\
&= \quad O\left(\frac{1}{\tilde{\alpha}} \sum_{t=1}^{T} \mathrm{tr}\Big((g_t g_t^\top) \cdot A_t^{-1}\Big)\right) \\
&= \quad O\left(\frac{1}{\tilde{\alpha}} \big(\ln\det(A_T) - \ln\det(A_0)\big)\right) \\
&= \quad O\left(\frac{1}{\tilde{\alpha}} \left(\sum_{i=1}^{d} \ln \frac{\lambda_i}{\tilde{\alpha} L^2}\right)\right) = O\left(\frac{d}{\tilde{\alpha}} \ln T\right),
\end{aligned}$$

where the first equality is based on the observation that $\tilde{\alpha} \leq \frac{1}{LB}$; the second equality is from Lemma 10; in the penultimate equality, we denote by $\{\lambda_i\}_{i=1}^d$ the d eigenvalues of A_T, and the last inequality uses the AM-GM inequality and the fact that $\sum_{i=1}^d \lambda_i = \mathrm{tr}(A_T) \leq \tilde{\alpha}(T+1)L^2$. The theorem follows from expanding the definition of $\tilde{\alpha}$. $\square$

References

[1] Dimitri Bertsekas and Angelia Nedic. Convex analysis and optimization (conservative). 2003.

[2] John Duchi, Elad Hazan, and Yoram Singer. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research*, 12(Jul):2121–2159, 2011.

[3] Yoav Freund and Robert E Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of computer and system sciences*, 55(1):119–139, 1997.

[4] Claudio Gentile. The robustness of the p-norm algorithms. *Machine Learning*, 53(3):265–299, 2003.

[5] Geoffrey J Gordon. Regret bounds for prediction problems. In *COLT 99*, 1999.

[6] Adam J Grove, Nick Littlestone, and Dale Schuurmans. General convergence results for linear discriminant updates. *Machine Learning*, 43(3):173–210, 2001.

[7] Hazan. Introduction to online convex optimization. *Foundations and Trends® in Optimization*, 2(3-4):157–325, 2016.

[8] Elad Hazan, Amit Agarwal, and Satyen Kale. Logarithmic regret algorithms for online convex optimization. *Machine Learning*, 69(2-3):169–192, 2007.

[9] Sham M Kakade, Shai Shalev-Shwartz, and Ambuj Tewari. Regularization techniques for learning with matrices. *Journal of Machine Learning Research*, 13:1865–1890, 2012.

[10] Jyrki Kivinen and Manfred K Warmuth. Exponentiated gradient versus gradient descent for linear predictors. *information and computation*, 132(1):1–63, 1997.

[11] Nick Littlestone. Learning quickly when irrelevant attributes abound: A new linear-threshold algorithm. *Machine learning*, 2(4):285–318, 1988.

[12] Francesco Orabona. Introduction to online learning. `https://parameterfree.com/lecture-notes-on-online-learning/`, 2019.

[13] Francesco Orabona, Koby Crammer, and Nicolo Cesa-Bianchi. A generalized online mirror descent with applications to classification and regression. *Machine Learning*, 99(3):411–435, 2015.

[14] Frank Rosenblatt. The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological review*, 65(6):386, 1958.

[15] Shai Shalev-Shwartz. Online learning: Theory, algorithms, and applications. 2007.

[16] Shai Shalev-Shwartz and Shai Ben-David. *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014.

[17] Martin Zinkevich. Online convex programming and generalized infinitesimal gradient ascent. In *Proceedings of the 20th International Conference on Machine Learning (ICML-03)*, pages 928–936, 2003.

A Auxiliary lemmas

Lemma 7. *Suppose $a_1, \ldots, a_T \geq 0$ is a sequence of positive numbers, with $a_1 > 0$. Let $A_t = \sum_{s=1}^t a_s$. Then*

$$\sum_{t=1}^T \frac{a_t}{\sqrt{A_t}} \leq 2\sqrt{A_T}.$$

Proof. We note that each term on the left hand side is

$$\frac{a_t}{\sqrt{A_t}} = \frac{A_t - A_{t-1}}{\sqrt{A_t}} \leq 2 \cdot \frac{A_t - A_{t-1}}{\sqrt{A_t} + \sqrt{A_{t-1}}} \leq 2(\sqrt{A_t} - \sqrt{A_{t-1}}).$$

The lemma is concluded by summing over all t's from 1 to T. □

We have the following matrix generalization of the above lemma.

Lemma 8. *Suppose $M_1, \ldots, M_T \succeq 0$ is a sequence of positive semidefinite matrices, with $M_1 \succ 0$. Let $N_t = \sum_{s=1}^t M_s$. Then*

$$\sum_{t=1}^T \operatorname{tr}\left(N_t^{-\frac{1}{2}} M_t\right) \leq 2 \operatorname{tr}\left(N_T^{\frac{1}{2}}\right).$$

Proof. We claim that

$$\operatorname{tr}\left(N_t^{-\frac{1}{2}} M_t\right) = \operatorname{tr}\left(N_t^{-\frac{1}{2}} (N_t - N_{t-1})\right) \leq 2 \operatorname{tr}\left(N_t^{\frac{1}{2}}\right) - 2 \operatorname{tr}\left(N_{t-1}^{\frac{1}{2}}\right).$$

Indeed, by the concavity of function $f(N) = 2 \operatorname{tr}\left(N^{\frac{1}{2}}\right)$, and the fact that $\nabla f(N) = N^{-\frac{1}{2}}$, we have that $f(N) - f(N') \leq \left\langle \nabla f(N'), N - N' \right\rangle$, which implies the last equality above by taking $N = N_{t-1}$ and $N' = N_t$. □

Similarly as above, we also have the following lemma regarding the sum of a sequence of numbers divided by their cumulative sums.

Lemma 9. *Suppose $a_1, \ldots, a_T \geq 0$ is a sequence of positive numbers, with $a_1 > 0$. Let $A_t = \sum_{s=1}^t a_s$. Then*

$$\sum_{t=1}^T \frac{a_t}{A_t} \leq \ln \frac{A_T}{A_1}.$$

Proof. Each term on the left hand side can be upper bounded as:

$$\frac{a_t}{A_t} \leq -\ln\left(1 - \frac{a_t}{A_t}\right) = \ln \frac{A_t}{A_{t-1}}.$$

The lemma is concluded by summing over all t's from 1 to T. □

We also have its matrix generalization:

Lemma 10. *Suppose $M_1, \ldots, M_T \succeq 0$ is a sequence of positive semidefinite matrices. Given $N_0 \succ 0$, let $N_t = N_0 + \sum_{s=1}^t M_s$. Then*

$$\sum_{t=1}^T \operatorname{tr}\left(N_t^{-1} M_t\right) \leq \ln\det(N_T) - \ln\det(N_0).$$

Proof. We claim that

$$\operatorname{tr}\left(N_t^{-1} M_t\right) \leq \ln\det(N_t) - \ln\det(N_{t-1}).$$

Ineed, by the concavity of function $f(N) = \ln\det(N)$, and the fact that $\nabla f(N) = N^{-1}$, we have that $f(N) - f(N') \leq \left\langle \nabla f(N'), N - N' \right\rangle$, which implies the last equality above by taking $N = N_{t-1}$ and $N' = N_t$. $\square$