Online Methods in Machine Learning: Recitation 1.

As seen in class, many offline machine learning problems can be written as:

$$\min_{w \in C} \frac{1}{n} \sum_{t=1}^{n} \ell(w, (x_t, y_t)) + \lambda R(w), \tag{1}$$

where C is a subset of $\mathbb{R}^d$, R is a penalty function, and ℓ is the loss function that measures how well our predictor, w, explains the relationship between x and y. Example: Support Vector Machines with $R(w) = \frac{\lambda}{2}||w||_2^2$ and $\ell(w, (x_t, y_t)) = \max(0, 1 - y_t \langle w, x_t \rangle)$ where $\langle \frac{w}{||w||_2}, x_t \rangle$ is the signed distance to the hyperplane $\{x \mid \langle w, x \rangle = 0\}$ used to classify the data. A couple of remarks:

1. In the context of classification, ℓ is typically a **convex surrogate** loss function in the sense that (i) $w \to \ell(w, (x, y))$ is convex for any example (x, y) and (ii) $\ell(w, (x, y))$ is larger than the true prediction error given by $1_{y\langle w,x \rangle \leq 0}$ for any example (x, y). In general, we introduce this surrogate loss function because it would be difficult to solve (1) computationally. On the other hand, convexity enables the development of general-purpose methods (e.g. Gradient Descent) that are fast, easy to implement, and simple to analyze. Moreover, the techniques and the proofs generalize very well to the online setting, where the data (x_t, y_t) arrive sequentially and we have to make predictions online. This theme will be explored later in the class.

2. If ℓ is a surrogate loss for a classification problem and we solve (1) (e.g with Gradient Descent), we cannot guarantee that our predictor will perform well with respect to our original classification problem

$$\min_{w \in C} \sum_{t=1}^{n} 1_{y\langle w,x \rangle \leq 0} + \lambda R(w), \tag{2}$$

hence it is crucial to choose a surrogate loss function that reflects accurately the classification problem at hand. Of course, our predictor gets better as $|\ell(w, (x, y)) - 1_{y\langle w,x \rangle \leq 0}|$ gets smaller for all possible examples (x, y).

3. R is typically a **strongly** convex function. This guarantees that our predictor is stable, in the sense that it does not vary much when only a few examples are modified. Stability is crucial in Machine Learning to derive good generalization properties.

As a result,

$$w \to \frac{1}{n} \sum_{t=1}^{n} \ell(w, (x_t, y_t)) + \lambda R(w)$$

is a convex function and our goal is to minimize this function on C.

Basics of Convex analysis.

Consider C a subset of $\mathbb{R}^d$ and $f : \mathbb{R}^d \to \mathbb{R}$ a real-valued function. $\langle \cdot, \cdot \rangle$ denotes the usual scalar product in $\mathbb{R}^d$ and $|| \cdot ||_2$ is the norm associated with this scalar product, i.e. $||w||_2 = \sqrt{\langle w, w \rangle}$. Most of the proofs are omitted because they are not of direct interest for the class (but you can find them in [1] if you are interested).

Definition 1. *C is said to be convex if $\alpha w + (1-\alpha)z \in C$ for any $\alpha \in [0,1]$ and for any two points w and z in C. Assuming C is convex, f is convex on C if*

$$f(\alpha w + (1-\alpha)z) \leq \alpha f(w) + (1-\alpha)f(z) \tag{3}$$

for any α in $[0,1]$ and for any two points w and z in C.

In other words, C is convex if every line segment joining any two points $w \in C$ and $z \in C$ is contained in C, i.e. $[w,z] \subset C$. Typical examples of convex sets include $\mathbb{R}^d$, any set defined by linear inequalities or equalities, balls $\{w \in \mathbb{R}^d \mid ||w-z|| \leq r\}$ for any norm $||\cdot||$, $z \in C$, and $r > 0$. From now on we assume that C is convex. f is convex on C if every line segment joining $(w, f(w))$ and $(z, f(z))$ is above the graph of f on $[w,z]$ for any two points w and z in C.
The next property is crucial as it enables the development of optimization methods based on local information about f (such as Gradient Descent).

Lemma 1. *Suppose that f is convex on C. Any local minimum of f on C is also a global minimum of f on C. If f is continuously differentiable, w^* is a global minimum of f if and only if:*

$$\langle \nabla f(w^*), w - w^* \rangle \geq 0 \quad \forall w \in C. \tag{4}$$

The optimality condition can be interpreted as follows. The first-order approximation of f cannot decrease when we start at w^* and move along any valid direction $w - w^*$. Observe that (4) simplifies to $\nabla f(w^*) = 0$ if $C = \mathbb{R}^d$ (and $f'(w^*) = 0$ if $d = 1$). When $C = \mathbb{R}^d$ and f is convex, Lemma 1 shows that finding a global minimum is equivalent to solving the equation $\nabla f(w) = 0$. The rational behind Gradient Descent is best understood with a first-order Taylor series expansion. Assuming w is not a minimum of f:

$$f(w - \eta \nabla f(w)) - f(w) \sim -\eta ||\nabla f(w)||^2 < 0,$$

for η small enough. When f is not convex, f may have several local minima and maxima, all of which satisfy $\nabla f(w) = 0$ and methods based on local information, including Gradient Descent, may fail. A standard example when $d = 1$ is given by $f(w) = w^3$ and $C = [-1, 1]$ (we have $f'(0) = 0$ but 0 is not a local extremum).
Next, we give equivalent characterizations of convexity.

Lemma 2. *(a) If f is continuously differentiable, f is convex on C if and only if:*

$$f(z) \geq f(w) + \langle z - w, \nabla f(w) \rangle \quad \forall (z, w) \in C,$$

i.e. f lies above each of its tangent lines.

(b) If f is continuously differentiable, f is convex on C if and only if:

$$\langle z - w, \nabla f(z) - \nabla f(w) \rangle \geq 0 \quad \forall (z, w) \in C.$$

When $d = 1$, this is equivalent to imposing that f' be non-decreasing.

(c) If f is twice continuously differentiable and C is open, f is convex on C if and only if:

$$\nabla^2 f(w) \text{ is positive-semidefinite} \quad \forall w \in C,$$

where $\nabla^2 f(w)$ is the Jacobian matrix of f at w. This is equivalent to imposing that the eigenvalues of $\nabla^2 f(w)$ be non-negative for any $w \in C$. When $d = 1$, this condition is equivalent to $f'' \geq 0$.

We have used (a) a number of times in Lectures 3 and 4. Once we know that a function is convex, all of these properties prove useful when we want to show that algorithms based on gradient methods work. If we have a closed-form expression for f, it is usually easier to show (c) but it might be even easier to apply the next result.

Lemma 3. *Consider a collection of convex functions* $(f_k)_{1\leq k\leq K}$*.*

1. $w \to \sum_{k=1}^{K} \alpha_k \cdot f_k(w)$ *is convex for any choice of non-negative coefficients* $(\alpha_k)_{1\leq k\leq K}$*.*

2. $w \to \max_{k=1,\cdots,K} f_k(w)$ *is convex.*

Typical examples of convex functions include norm functions $w \to ||w||$, linear functions, quadratic functions $w \to \langle w, Qw\rangle$ with Q symmetric positive-semidefinite, the hinge loss $w \to \max(0, 1 - y_t\langle w, x_t\rangle)$ (using Lemma 3 since 0 and $w \to 1 - y_t\langle w, x_t\rangle$ are linear functions). In Lecture 3, we have seen that a stronger notion of convexity makes Gradient Descent methods converge faster for an appropriate choice of stepsize.

Definition 2. *For* $\sigma > 0$*,* f *is* σ*-strongly convex on* C *if:*

$$f(\alpha \cdot w + (1-\alpha)\cdot z) \leq \alpha \cdot f(w) + (1-\alpha)\cdot f(z) - \sigma\frac{\alpha(1-\alpha)}{2}||z - w||_2^2, \tag{5}$$

for any α *in* $[0,1]$ *and for any two points* w *and* z *in* C*.*

Just like for convexity, there are several equivalent characterizations of strong convexity.

Lemma 4. *(a) If* f *is continuously differentiable,* f *is* σ*-strongly convex on* C *if and only if:*

$$f(z) \geq f(w) + \langle z - w, \nabla f(w)\rangle + \frac{\sigma}{2}||z-w||_2^2 \quad \forall (z,w) \in C.$$

(b) If f *is continuously differentiable,* f *is* σ*-strongly on* C *if and only if:*

$$\langle z - w, \nabla f(z) - \nabla f(w)\rangle \geq \sigma||z-w||_2^2 \quad \forall (z,w) \in C.$$

(c) If f *is twice continuously differentiable,* f *is* σ*-strongly on* C *if and only if:*

$$\nabla^2 f(w) - \sigma I_d \text{ is semidefinite positive} \qquad \forall w \in C,$$

where $\nabla^2 f(w)$ *is the Jacobian matrix of* f *at* w *and* I_d *is the identity matrix. When* $d = 1$*, this is equivalent to* $f'' \geq \sigma$*.*

(d) f *is* σ*-strongly on* C *if and only if the function:*

$$w \to f(w) - \frac{\sigma}{2}||w||_2^2$$

is convex.

Observe that these properties are strictly stronger than their counterparts of Lemma 2. In particular properties (a) and (b) of Lemma 2 are satisfied with some (quadratic) margin. Property (d) tells us that a σ-strongly convex function is nothing more than the sum of a convex function and of $w \to \frac{\sigma}{2}||w||_2^2$. Property (b) shows that a strongly convex function has a unique global minimum. Additionally, property (b) provides intuition as to why Gradient Descent methods converge faster for strongly-convex functions. Indeed, using Cauchy-Schwarz:

$$\begin{aligned}\sigma||z-w||_2^2 &\leq \langle z - w, \nabla f(z) - \nabla f(w)\rangle \\ &\leq ||z-w||_2||\nabla f(z) - \nabla f(w)||_2,\end{aligned}$$

which implies:

$$||\nabla f(z) - \nabla f(w)||_2 \geq \sigma ||z - w||_2 \quad \forall z, w \in C. \tag{6}$$

Taking w^* as the minimum of f on C, this tells us that $||\nabla f(w)||_2 \geq ||w - w^*||_2$, i.e. the magnitude of a step in Gradient Descent is at least proportional to the distance to the optimal solution.
In Lecture 3, we have introduced another notion, called β-smoothness.

Definition 3. *Suppose that f is differentiable. For $\beta > 0$, f is β-smooth on C if its gradient maps are β-Lipschitz, i.e:*

$$||\nabla f(z) - \nabla f(w)||_2 \leq \beta ||z - w||_2 \quad \forall (z, w) \in C. \tag{7}$$

The notion of smoothness is dual to that of strong convexity (compare (6) and (7)).

Lemma 5. *Suppose that f is convex on C.*

(a) f is β-smooth on C if and only if:

$$f(z) \leq f(w) + \langle z - w, \nabla f(w) \rangle + \frac{\beta}{2} ||z - w||_2^2 \quad \forall (z, w) \in C.$$

(b) If f is β-smooth on C, then:

$$\langle z - w, \nabla f(z) - \nabla f(w) \rangle \geq \frac{1}{\beta} ||\nabla f(z) - \nabla f(w)||_2^2 \quad \forall (z, w) \in C.$$

Proof. Let us start with (a). Suppose that f is β-smooth on C. Consider $(z, w) \in C$ and define $g : t \in \mathbb{R} \to f(w + t(z - w))$. We have:

$$\begin{aligned}
f(z) - f(w) &= g(1) - g(0) \\
&= \int_0^1 g'(t) \mathrm{d}t \\
&= \int_0^1 \langle z - w, \nabla f(w + t(z - w)) \rangle \mathrm{d}t \\
&= \langle z - w, \nabla f(w) \rangle + \langle z - w, \int_0^1 [\nabla f(w + t(z - w)) - \nabla f(w)] \mathrm{d}t \rangle \\
&\leq \langle z - w, \nabla f(w) \rangle + ||z - w||_2 || \int_0^1 [\nabla f(w + t(z - w)) - \nabla f(w)] \mathrm{d}t ||_2 \\
&\leq \langle z - w, \nabla f(w) \rangle + ||z - w||_2 \int_0^1 ||\nabla f(w + t(z - w)) - \nabla f(w)||_2 \mathrm{d}t \\
&\leq \langle z - w, \nabla f(w) \rangle + ||z - w||_2 \int_0^1 \beta t ||z - w||_2 \mathrm{d}t \\
&\leq \langle z - w, \nabla f(w) \rangle + \frac{\beta}{2} ||z - w||_2^2,
\end{aligned}$$

where we use Cauchy-Schwarz and the definition of β-smoothness. Conversely, suppose that

$$f(z) \leq f(w) + \langle z - w, \nabla f(w) \rangle + \frac{\beta}{2} ||z - w||_2^2 \quad \forall (z, w) \in C.$$

Consider $(z,w) \in C$ and define $x = z - \frac{1}{\beta}(\nabla f(z) - \nabla f(w))$. We have:

$$\begin{aligned}
f(z) - f(w) &= f(x) - f(w) - [f(x) - f(z)] \\
&\geq \langle \nabla f(w), x - w \rangle - [\langle \nabla f(z), x - z \rangle + \frac{\beta}{2}||x - z||_2^2] \\
&\geq \langle \nabla f(w), x - w \rangle - [\langle \nabla f(z), x - z \rangle + \frac{\beta}{2}||x - z||_2^2] \\
&\geq \langle \nabla f(w), z - w \rangle + \langle \nabla f(w) - \nabla f(z), x - z \rangle - \frac{\beta}{2}||x - z||_2^2 \\
&\geq \langle \nabla f(w), z - w \rangle + \frac{1}{\beta}||\nabla f(w) - \nabla f(z)||_2^2 - \frac{\beta}{2}\frac{1}{\beta^2}||\nabla f(z) - \nabla f(w)||_2^2 \\
&\geq \langle \nabla f(w), z - w \rangle + \frac{1}{2\beta}||\nabla f(w) - \nabla f(z)||_2^2,
\end{aligned}$$

where we use the convexity of f, property (a) for the points x and z, and the definition of x. Using property (a) once again but for the points z and w, we get:

$$\begin{aligned}
\langle z - w, \nabla f(w) \rangle + \frac{\beta}{2}||z - w||_2^2 &\geq f(z) - f(w) \\
&\geq \langle z - w, \nabla f(w) \rangle + \frac{1}{2\beta}||\nabla f(w) - \nabla f(z)||_2^2,
\end{aligned}$$

which implies:

$$||\nabla f(z) - \nabla f(w)||_2 \leq \beta||z - w||_2.$$

We move on to prove (b). We sum the inequalities

$$f(z) - f(w) \geq \langle \nabla f(w), z - w \rangle + \frac{1}{2\beta}||\nabla f(w) - \nabla f(z)||_2^2$$

and

$$f(w) - f(z) \geq \langle \nabla f(z), w - z \rangle + \frac{1}{2\beta}||\nabla f(z) - \nabla f(w)||_2^2$$

obtained in the course of proving (a) to derive:

$$0 \geq -\langle z - w, \nabla f(z) - \nabla f(w) \rangle + \frac{1}{2\beta}||\nabla f(z) - \nabla f(w)||_2^2.$$

□

Using property (a), a β-smooth convex function can be sandwiched between its first order approximation at any point and a quadratic function:

$$f(w) + \langle z - w, \nabla f(w) \rangle \leq f(z) \leq f(w) + \langle z - w, \nabla f(w) \rangle + \frac{\beta}{2}||z - w||_2^2 \quad \forall (z,w) \in C.$$

It turns out that smoothness also makes Gradient Descent methods converge faster for an appropriate choice of stepsize. We point out that smoothness and strong convexity are not mutually exclusive, for instance $w \rightarrow \frac{1}{2}||w||_2^2$ is both 1-smooth and 1-strongly convex. In fact, Gradient Descent converges much faster for convex functions that are both smooth and convex. Typically we need $O(\log(\frac{1}{\epsilon}))$ iterations to derive an ϵ-optimal solution, as opposed to $O(\frac{1}{\epsilon})$ iterations for functions that are either smooth or strongly convex but not both.
We end with a technical point that was left out in Lecture 3.

Lemma 6. *Suppose that f is continuously differentiable on an open set S. If f is K-Lipschitz, i.e.:*

$$|f(w) - f(z)| \leq K||w - z||_2 \quad \forall (w, z) \in S$$

then:

$$||\nabla f(w)||_2 \leq K \quad \forall w \in S.$$

Proof. Consider $w \in S$. Since S is open, there exists $t > 0$ such that $w + t\nabla f(w) \in S$. We have:

$$|f(w + t\nabla f(w)) - f(w)| \leq Kt||\nabla f(w)||_2.$$

Using the mean value theorem, there exists $\tau \in [0, t]$ such that:

$$f(w + t\nabla f(w)) - f(w) = t\langle \nabla f(w), \nabla f(w + \tau \nabla f(w))\rangle.$$

This yields:

$$|\langle \nabla f(w), \nabla f(w + \tau \nabla f(w))\rangle| \leq K||\nabla f(w)||_2.$$

Taking the limit $t \to 0$, we get $\tau \to 0$ and:

$$||\nabla f(w)||_2^2 \leq K||\nabla f(w)||_2,$$

by continuity of ∇f. □

Non-differentiable convex functions. A convex function may not be differentiable everywhere. This typically occurs when f is defined as a maximum of convex functions even if the latter are differentiable (e.g. $w \in \mathbb{R} \to |w| = \max(w, -w)$ which is not differentiable at 0). In this case, as seen in Lecture 3, we need to invoke the notion of subdifferentiability to properly define Gradient Descent.

Definition 4. *Suppose that f is convex. u is a subgradient of f at $w \in C$ if:*

$$f(z) \geq f(w) + \langle z - w, u\rangle \quad \forall z \in C. \tag{8}$$

We define $\partial f(w)$ as the set of subgradients of f at w. It can be shown that $\partial f(w)$ is never empty and that it is always a convex set.

Compare (8) and property (a) of Lemma 2. When f is differentiable at w, $\nabla f(w)$ is the only subgradient of f at w and this is actually an equivalence. The following results provide a convenient way to compute subgradients when f is expressed as a maximum or as a sum of convex functions.

Lemma 7. *Consider a collection of convex functions $(f_k)_{1\leq k\leq K}$ and $f : w \to \max\limits_{k=1,\cdots,K} f_k(w)$. Take $w \in C$ and any $k_w \in \operatorname*{argmax}\limits_{k=1,\cdots,K} f_k(w)$. Any subgradient of f_{k_w} at w is also a subgradient of f at w. In particular:*

$$\nabla f_{k_w}(w) \in \partial f(w)$$

if f_{k_w} is differentiable at w.

Proof. Take $z \in C$ and u a subgradient of f_{k_w} at w. We have:

$$\begin{aligned} f(z) - f(w) &\geq f_{k_w}(z) - f_{k_w}(w) \\ &\geq \langle u, z - w\rangle. \end{aligned}$$

□

Example: the hinge loss $w \to \max(0, 1 - y_t\langle w, x_t\rangle)$. The hinge loss is differentiable everywhere but at any point w such that $1 = y_t\langle w, x_t\rangle$, in which case, using Lemma 7, $-y_t x_t$ is a subgradient of f at w. Similarly, for $f : w \in \mathbb{R} \to |w|$, $\partial f(0) = [-1, 1]$.

Lemma 8. *Consider a collection of convex functions $(f_k)_{1\leq k\leq K}$ and non-negative coefficients $(\alpha_k)_{1\leq k\leq K}$. Define $f : w \to \sum_{k=1}^K \alpha_k f_k(w)$. Take $w \in C$ and, for any $k \in \{1, \cdots, K\}$, u_k, a subgradient of f_k at w. Then:*

$$\sum_{k=1}^{K} \alpha_k u_k$$

is a subgradient of f at w.

Proof. Take $z \in C$. We have:

$$\begin{aligned} f(z) &= \sum_{k=1}^{K} \alpha_k f_k(z) \\ &\geq \sum_{k=1}^{K} \alpha_k (f_k(w) + \langle u_k, z - w\rangle) \\ &\geq \sum_{k=1}^{K} \alpha_k f_k(w) + \langle \sum_{k=1}^{K} \alpha_k u_k, z - w\rangle \\ &\geq f(w) + \langle \sum_{k=1}^{K} \alpha_k u_k, z - w\rangle. \end{aligned}$$

□

For a non-differentiable convex function f, Gradient Descent can be adapted by using any subgradient as a substitute for the gradient. The proof of convergence of Gradient Descent (i.e. Lemma 1 in Lecture 3) remains the same except that we use (8) as opposed to property (a) of Lemma 2.

Gradient Descent. We want to adapt Gradient Descent when C is any closed convex subset of $\mathbb{R}^d$, e.g. $C = \{w \mid ||w||_2 \leq 1\}$. Consider f a differentiable convex function. Gradient Descent is modified as follows:

$$w_{t+1} = \text{Proj}(w_t - \eta\nabla f(w_t)), \tag{9}$$

where Proj is the Euclidean projection onto C defined as follows:

$$\text{Proj}(w) \in \operatorname*{argmin}_{z\in C} ||z - w||_2.$$

We can show that $\text{Proj}(w)$ is always uniquely defined. We point out that carrying out the update (9) can be computationally difficult depending on the exact definition of C. However, a closed-form expression may be readily available for some sets, e.g. when $C = \{w \mid ||w||_2 \leq 1\}$: $\text{Proj}(w) = \frac{w}{||w||_2}$ if $w \neq 0$ and 0 otherwise. The following result will be useful to show that Projected Gradient Descent converges.

Lemma 9. *Proj is 1-Lipschitz.*

Borrowing the notations from Lecture 3, the result of Lemma 1 of Lecture 3 can still be shown to hold. To prove the result, we only need to adapt (8) of Lecture 3 as follows:

$$\begin{aligned} ||w_{t+1} - w^*||_2^2 &= ||\text{Proj}(w_t - \eta\nabla f(w_t)) - w^*||_2^2 \\ &= ||\text{Proj}(w_t - \eta\nabla f(w_t)) - \text{Proj}(w^*)||_2^2 \\ &\leq ||w_t - \eta\nabla f(w_t) - w^*||_2^2, \end{aligned}$$

where we use the fact that $w^* \in C$ for the second equality and Lemma 9 for the last inequality.

References

[1] Dimitri P Bertsekas. *Nonlinear programming.* Athena scientific, 1999.