Mathematical Tools in Machine Learning

Fadoua Balabdaoui

Seminar für Statistik, ETH

28 novembre 2019

Lecture 8 (Week 11)

Convex Learning Problems (Chapter 12)

Stochastic Gradient Descent (Chapter 14)

Lecture 8

Convexity, Lipschitzness and smoothness

- **Convexity** of the loss function, when it holds, makes learning **efficient**. Examples of convex learning problems include :
  - Linear regression with the quadratic loss $\ell_{sq}(h_w, (x, y)) = (h_w(x) - y)^2$ with $h_w(x) = \langle w, x \rangle$
  - Logistic regression with the loss $\ell(h_w, (x, y)) = \log(1 + \exp(-y\langle w, x \rangle))$

Classification with the ℓ_{0-1} is an example of a **non-convex** learning problem.

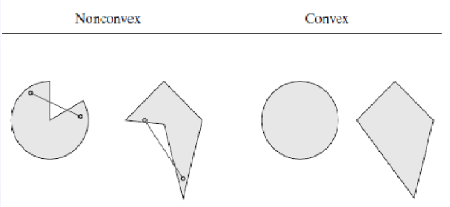
Definition (convex set). A set C in a vector space is convex if for **any** two vectors $\mathbf{u}, \mathbf{v} \in C$, the line segment between u and v is contained in C : for any $\alpha \in [0, 1]$, $\alpha\mathbf{u} + (1 - \alpha)\mathbf{v} \in C$.

Convexity

Definition (convex function). Let C be a convex set. A function $f : C \mapsto \mathbb{R}$ is **convex** if $\forall\ \mathbf{u}, \mathbf{v} \in C$ and $\forall\ \alpha \in [0,1]$,
$f(\alpha\mathbf{u} + (1-\alpha)\mathbf{v}) \leq \alpha f(\mathbf{u}) + (1-\alpha) f(\mathbf{v})$.

- The following characterization can be shown : f is convex on C iff

epigraph$(f) = \{(\mathbf{x}, \beta) \in C \times \mathbb{R} : f(\mathbf{x}) \leq \beta\}$ is a convex set of $C \times \mathbb{R}$



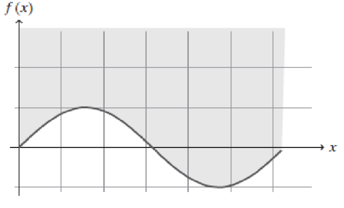


FIGURE – Left : examples for convex and non-convex 2-dimensional sets. Right : example of a non-convex function

Convexity

Property 1. An important consequence of convexity of some function f is that a local minimizer of f is necessarily a **global minimizer** of f.

Proof. Let $\mathbf{u}$ be a local minimum of f defined on C. Then, there exists $r > 0$ such that for all $\mathbf{v} \in B(\mathbf{u}, r)$ (the Euclidean ball of radius r and centered at $\mathbf{u}$) we have that

$$f(\mathbf{u}) \leq f(\mathbf{v}).$$

Let $\mathbf{w} \in C$ (not necessarily in $B(\mathbf{u}, r)$). Then, we can find some small $\alpha > 0$ such that $\mathbf{u} + \alpha(\mathbf{w} - \mathbf{u}) \in B(\mathbf{u}, r)$. Therefore,

$$f(\mathbf{u}) \leq f\big(\mathbf{u} + \alpha(\mathbf{w} - \mathbf{u})\big) = f\big((1-\alpha)\mathbf{u} + \alpha\mathbf{w}\big).$$

If f is convex, the latter implies that $f(\mathbf{u}) \leq (1-\alpha)f(\mathbf{u}) + \alpha f(\mathbf{w})$, which is equivalent to

$f(\mathbf{u}) \leq f(\mathbf{w}) \iff \mathbf{u}$ is a global minimizer, since $\mathbf{w}$ was arbitrarily chosen. $\square$

Convexity

Property 2. Suppose that f is convex on a convex set $C \subset \mathbb{R}^d$ and is differentiable at $\mathbf{w} \in C$, that is

$$\nabla f(\mathbf{w}) = \left(\frac{\partial f(\mathbf{w})}{\partial w_1}, \ldots, \frac{\partial f(\mathbf{w})}{\partial w_d} \right)^T \quad \text{exists.}$$

Then, the function f stays **above** the tangent at $\mathbf{w}$, that is

$$\forall \, \mathbf{u} \quad f(\mathbf{u}) \geq f(\mathbf{w}) + \nabla f(\mathbf{w})^T (\mathbf{u} - \mathbf{w})$$

Lemma. Let $f : \mathbb{R} \to \mathbb{R}$ be a twice differential function. Then, the following assertions are equivalent :

1. f is convex.
2. f' is nondecreasing.
3. $f'' \geq 0$.

Convexity

- **Examples.** The functions $f(x) = x^2$ and $f(x) = \log(1 + \exp(x))$ are convex on $\mathbb{R}$ since their respective derivatives $f'(x) = 2x$ and $f'(x) = \exp(x)/(1 + \exp(x))$ are nondecreasing.

Result. Let $g : \mathbb{R} \to \mathbb{R}$ be convex. Then, the function $f(\mathbf{w}) = g(\langle \mathbf{w}, \mathbf{x} \rangle + y)$ for some fixed $\mathbf{x} \in \mathbb{R}^d$ and $y \in \mathbb{R}$ is **convex**.

Proof. For $\mathbf{w}_1, \mathbf{w}_2 \in \mathbb{R}^d$ and $\alpha \in [0, 1]$, we have that

$$
\begin{aligned}
f(\alpha \mathbf{w}_1 + (1-\alpha)\mathbf{w}_2) &= g(\alpha \langle \mathbf{w}_1, \mathbf{x} \rangle + (1-\alpha)\langle \mathbf{w}_2, \mathbf{x} \rangle + y) \\
&= g\Big(\alpha(\langle \mathbf{w}_1, \mathbf{x} \rangle + y) + (1-\alpha)(\langle \mathbf{w}_2, \mathbf{x} \rangle + y)\Big) \\
&\le \alpha g(\langle \mathbf{w}_1, \mathbf{x} \rangle + y) + (1-\alpha) g(\langle \mathbf{w}_2, \mathbf{x} \rangle + y) \\
&= \alpha f(\mathbf{w}_1) + (1-\alpha) f(\mathbf{w}_2). \qquad \square
\end{aligned}
$$

Convexity

- **Examples.** The previous result implies that
  - $f(\mathbf{w}) = (\langle \mathbf{w}, \mathbf{x} \rangle - y)^2$ is convex on $\mathbb{R}^d$ as the composition of $g(t) = t^2$ and the linear function $\mathbf{w} \mapsto \langle \mathbf{w}, \mathbf{x} \rangle - y$
  - $f(\mathbf{w}) = \log\left(1 + \exp(-y\langle \mathbf{w}, \mathbf{x} \rangle\right)$ is convex on $\mathbb{R}^d$ (with $y \in \{-1, 1\}$) as the composition of the convex function $g(t) = \log(1 + \exp(t))$ or $g(t) = \log(1 + \exp(-t))$ and the linear function $\mathbf{w} \mapsto \langle \mathbf{w}, \mathbf{x} \rangle$.

Result. For $i \in \{1, \ldots, r\}$, let $f_i : \mathbb{R}^d \mapsto \mathbb{R}$ be a convex function. Then, the functions

- $g(x) = \max_{1 \leq i \leq r} f_i(\mathbf{x})$,
- $g(x) = \sum_{i=1}^r w_i f_i(\mathbf{x})$, for $w_i \geq 0, i = 1, \ldots, r$

are also **convex**.

Still on Convexity... and Lipschitzness

Proof. We prove only the claim for the first function. We have that

$$\begin{aligned} g(\alpha \mathbf{x}_1 + (1-\alpha)\mathbf{x}_2) &= \max_{1\le i\le r} f_i(\alpha \mathbf{x}_1 + (1-\alpha)\mathbf{x}_2) \\ &\le \max_{1\le i\le r} [\alpha f_i(\mathbf{x}_1) + (1-\alpha) f_i(\mathbf{x}_2)] \\ &\le \alpha \max_{1\le i\le r} f_i(\mathbf{x}_1) + (1-\alpha) \max_{1\le i\le r} f_i(\mathbf{x}_2) \\ &= \alpha g(\mathbf{x}_1) + (1-\alpha) g(\mathbf{x}_2). \end{aligned}$$

Definition (Lipschitzness). Let $C \subset \mathbb{R}^d$. A function $f : \mathbb{R}^d \to \mathbb{R}^k$ is ρ-Lipschitz over C if $\forall\ \mathbf{w}_1, \mathbf{w}_2 \in C$ $\|f(\mathbf{w}_2) - f(\mathbf{w}_1)\| \le \rho \|\mathbf{w}_2 - \mathbf{w}_1\|$.

- **Remark.** Lipschitz functions cannot change too fast. If f is a differentiable real function, then ρ-Lipschitzness of f implies that $\sup_t |f'(t)| \le \rho$ since $\lim_{x\to t} |(f(x) - f(t))/(x-t)| \le \rho$.

Lipschitzness

- **Remark (continued).** The converse is true. Suppose that f satisfies $\sup_t |f'(t)| \leq \rho$. For any x and y we have that

$$\begin{aligned} |f(x) - f(y)| &= |f'(u^*)||x-y|, \text{ for some } u^* = \lambda^* x + (1-\lambda^*)y \\ &\leq \rho|x-y| \text{ if } \sup_t |f'(t)| \leq \rho. \end{aligned}$$

Examples :

- $f(x) = |x|$ is 1-Lipschitz over $\mathbb{R}$ using the well-known inequality $\Big||x| - |y|\Big| \leq |x-y|$.
- $f(x) = \log(1+\exp(x))$ is also 1-Lipschitz since for all $x \in \mathbb{R}$ $|f'(x)| = f'(x) = \exp(x)/(1+\exp(x)) \leq 1$.
- $f(x) = x^2$ is not ρ-Lipschitz on $\mathbb{R}$ for any $\rho > 0$ since with $(x_1, x_2) = (0, 1+\rho)$ we can check that $|f(x_2) - f(x_1)| > \rho|x_2 - x_1|$.

Liptschitzness

Examples (continued) :

- However, $f(x) = x^2$ is ρ-Lipschitz on $C_\rho = [-\rho/2, \rho/2]$ on which $|f'(x)| = 2|x| \leq \rho$.
- Consider $f(\mathbf{w}) = \langle \mathbf{v}, \mathbf{w} \rangle + b$ defined on $\mathbb{R}^d$ to $\mathbb{R}$ for some fixed $\mathbf{v} \in \mathbb{R}^d$. Then, $|f(\mathbf{w}_2) - f(\mathbf{w}_1)| = |\langle \mathbf{v}, \mathbf{w}_2 - \mathbf{w}_1 \rangle| \leq \|\mathbf{v}\| \; \|\mathbf{w}_2 - \mathbf{w}_1\|$ by the Cauchy-Schwartz inequality, so that f is $\|\mathbf{v}\|$-Lipschitz.

Result. Let $f(\mathbf{x}) = g_1(g_2(\mathbf{x}))$, where g_1 is ρ_1-Lipschitz and g_2 is ρ_2-Lipschitz. Then, f is $(\rho_1\rho_2)$-Lipschitz. In particular, if $g_2(\mathbf{x}) = \langle \mathbf{v}, \mathbf{x} \rangle + b$ for some $\mathbf{v} \in \mathbb{R}^d, b \in \mathbb{R}$, then f is $(\rho_1\|\mathbf{v}\|)$-Lipschitz.

Proof. Write $|f(\mathbf{w}_2) - f(\mathbf{w}_1)| = |g_1(g_2(\mathbf{w_2})) - g_1(g_2(\mathbf{w_1}))| \leq \rho_1|g_2(\mathbf{w_2}) - g_2(\mathbf{w_1})| \leq \rho_1\rho_2 \; \|\mathbf{w}_2 - \mathbf{w}_1\|$. $\square$

Smoothness

- Recall that if $f : \mathbb{R}^d \to \mathbb{R}$ is differentiable at some $\mathbf{w} \in \mathbb{R}$, then its **gradient** at $\mathbf{w}$ is given by

$$\nabla f(\mathbf{w}) = \left(\frac{\partial f}{\partial w_1}, \dots, \frac{\partial f}{\partial w_d} \right)^T .$$

- **Definition.** A differentiable function $f : \mathbb{R}^d \to \mathbb{R}$ is β-smooth if its gradient is β-Lipschitz :

$$\|\nabla f(\mathbf{v}) - \nabla f(\mathbf{w})\| \leq \beta \|\mathbf{v} - \mathbf{w}\|.$$

- **Result.** If $f : \mathbb{R} \to \mathbb{R}$ is β-smooth, then for all $\mathbf{v}, \mathbf{w}$

$$f(\mathbf{v}) \leq f(\mathbf{w}) + \langle \nabla f(\mathbf{w}), \mathbf{v} - \mathbf{w} \rangle + \frac{\beta}{2} \|\mathbf{v} - \mathbf{w}\|^2.$$

Proof. Define the function h on $[0, 1]$ by

$$h(t) = f\Big(t\mathbf{v} + (1 - t)\mathbf{w}\Big) = f\Big(\mathbf{w} + t(\mathbf{v} - \mathbf{w})\Big).$$

Smoothness

Proof (continued). The function h is differentiable on $(0,1)$ (as a composition of two differentiable functions) with derivative at $t \in (0,1)$ $h'(t) = \langle \nabla f(\mathbf{w} + t(\mathbf{v} - \mathbf{w})), \mathbf{v} - \mathbf{w} \rangle$. Hence,

$$\begin{aligned} f(\mathbf{v}) - f(\mathbf{w}) = h(1) - h(0) &= \int_0^1 h'(t)dt \\ &= \int_0^1 \langle \nabla f(\mathbf{w} + t(\mathbf{v} - \mathbf{w})), \mathbf{v} - \mathbf{w} \rangle dt. \end{aligned}$$

It follows that

$$\begin{aligned} & f(\mathbf{v}) - f(\mathbf{w}) - \langle \nabla f(\mathbf{w}), \mathbf{v} - \mathbf{w} \rangle \\ & \int_0^1 \Big(\langle \nabla f(\mathbf{w} + t(\mathbf{v} - \mathbf{w})) - \nabla f(\mathbf{w}), \mathbf{v} - \mathbf{w} \rangle \Big) dt \\ & \leq \int_0^1 \|\langle \nabla f(\mathbf{w} + t(\mathbf{v} - \mathbf{w})) - \nabla f(\mathbf{w})\| \|\mathbf{v} - \mathbf{w}\| dt \end{aligned}$$

by the Cauchy-Schwartz inequality.

Smoothness

Proof (continued). Now, by the β-smoothness of f we have that

$$\int_0^1 \|\langle \nabla f(\mathbf{w} + t(\mathbf{v} - \mathbf{w})) - \nabla f(\mathbf{w})\| dt \quad \leq \quad \beta \int_0^1 t\|\mathbf{w} - \mathbf{v}\| dt = \frac{\beta}{2}\|\mathbf{w} - \mathbf{v}\|$$

yielding $f(\mathbf{v}) \leq f(\mathbf{w}) + \langle \nabla f(\mathbf{w}), \mathbf{v} - \mathbf{w}\rangle + \beta\|\mathbf{w} - \mathbf{v}\|^2/2$ $(\star)$. □

- Note that if f is both **convex** and **β-smooth** on $\mathbb{R}^d$, then for all $\mathbf{v}, \mathbf{w}$

$$f(\mathbf{w}) + \langle \nabla f(\mathbf{w}), \mathbf{v} - \mathbf{w}\rangle \leq f(\mathbf{v}) \leq f(\mathbf{w}) + \langle \nabla f(\mathbf{w}), \mathbf{v} - \mathbf{w}\rangle + \frac{\beta}{2}\|\mathbf{w} - \mathbf{v}\|^2.$$

- Consider the case $\mathbf{v} = \mathbf{w} - \nabla f(\mathbf{w})/\beta$. Then, $\mathbf{v} - \mathbf{w} = -\nabla f(\mathbf{w})/\beta$, and

$$\frac{1}{2\beta}\|\nabla f(\mathbf{w})\|^2 \leq f(\mathbf{w}) - f(\mathbf{v}), \quad \text{using the inequality in } (\star)$$

Smoothness

- If $f \geq 0$ on $\mathbb{R}^d$, then β-smoothness of f implies that

$$\|\nabla f(\mathbf{w})\|^2 \leq 2\beta f(\mathbf{w}), \quad \forall\ \mathbf{w} \in \mathbb{R}^d,$$

(we say that the function f is **self-bounded**).

Result. Let $g : \mathbb{R} \to \mathbb{R}$ be a β-smooth function and consider the function $f(\mathbf{w}) = g(\langle \mathbf{w}, \mathbf{x} \rangle + b)$ for some $\mathbf{x} \in \mathbb{R}^d$ and $b \in \mathbb{R}$. Then, f is $(\beta\|\mathbf{x}\|^2)$-smooth.

Proof. By taking the derivative of the composition, we have that

$$\begin{aligned}
\|\nabla f(\mathbf{w}) - \nabla f(\mathbf{v})\| &= \left\|\Big(g'\big(\langle \mathbf{w}, \mathbf{x} \rangle + b\big) - g'\big(\langle \mathbf{v}, \mathbf{x} \rangle + b\big)\Big)\mathbf{x}\right\| \\
&= \|\mathbf{x}\| \left|g'\big(\langle \mathbf{w}, \mathbf{x} \rangle + b\big) - g'\big(\langle \mathbf{v}, \mathbf{x} \rangle + b\big)\right| \\
&\leq \|\mathbf{x}\|\ \beta\big|\langle \mathbf{w} - \mathbf{v}, \mathbf{x} \rangle\big| \leq \beta\|\mathbf{x}\|^2\|\mathbf{w} - \mathbf{v}\|
\end{aligned}$$

by the Cauchy-Schwartz inequality. $\square$

Smoothness

- **Examples.**
  - The function $x \mapsto x^2$ is 2-smooth and hence $f(\mathbf{w}) = \left(\langle \mathbf{w}, \mathbf{x} \rangle - y\right)^2$ for some $\mathbf{x} \in \mathbb{R}^d, y \in \mathbb{R}$ is $(2\|\mathbf{x}\|^2)$-smooth.
  - Consider the function $g(x) = \log(1 + \exp(-yx))$, for some fixed $y \in \{-1, 1\}$. Then,

$$\begin{aligned} |g''(x)| &= \frac{\exp(-xy)}{\left(1 + \exp(-xy)\right)^2} = \frac{1}{\left(1 + \exp(-xy)\right)\left(1 + \exp(xy)\right)} \\ &\leq \frac{1}{4}. \end{aligned}$$

    Hence, $|g'(x) - g'(y)| \leq |x - y|/4$ (g' is $(1/4)$-Lipschitz) and g is $1/4$-smooth. Thus, the function

$$f(\mathbf{w}) = \log\left(1 + \exp(-y\langle \mathbf{w}, \mathbf{x} \rangle\right)$$

    is $(\|\mathbf{x}\|^2/4)$-smooth.

Convex Learning Problems

- Recall that a **learning problem** needs a hypothesis class $\mathcal{H}$, a domain $\mathcal{Z}(= \mathcal{X} \times \mathcal{Y})$, and a loss function $\ell : \mathcal{H} \times \mathcal{Z} \to [0, \infty)$.

- Up to now, the elements in $\mathcal{H}$ were functions $h : \mathcal{X} \mapsto \mathcal{Y}$. Here, we will assume that each hypothesis function h can be identified with a real d-dimensional vector : $\mathbf{w} \in \mathbb{R}^d$.

Definition (Convex Learning Problem). A learning problem, $(\mathcal{H}, \mathcal{Z}, \ell)$ is called **convex** if $\mathcal{H}$ is a **convex set** and for all $z \in \mathcal{Z}$, the function

$$f(\mathbf{w}) = \ell(\mathbf{w}, z)$$

is **convex**, for any fixed $z \in \mathcal{Z}$.

- **Example.** Consider a regression problem, where the hypothesis class $\mathcal{H}$ can be **identified** with $\mathbb{R}^d$ since $h(\mathbf{x}) = \langle \mathbf{w}, \mathbf{x} \rangle$ for some $\mathbf{w} \in \mathbb{R}^d$, and the quadratic loss function

$$\ell(\mathbf{w}, (\mathbf{x}, y)) = \left(\langle \mathbf{w}, \mathbf{x} \rangle - y \right)^2.$$

Convex Learning Problems

Lemma. If ℓ is a convex loss function and $\mathcal{H}$ is convex, then the $\mathrm{ERM}_{\mathcal{H}}$ problem (of minimizing the empirical loss over $\mathcal{H}$), is a **convex optimization problem** (the problem of minimizing a convex function over a convex set).

Proof. Let $S = \{(\mathbf{x}_1, y_1), \ldots, \mathbf{x}_m, y_m)\}$ be some training set. Then, when searching for the $\mathrm{ERM}_{\mathcal{H}}$ rule we aim at minimizing the function

$$\mathbf{w} \mapsto L_S(\mathbf{w}) = \frac{1}{m} \sum_{i=1}^{m} \ell(\mathbf{w}, (\mathbf{x}_i, y_i))$$

which is a convex function (by a previous result with weights equal to $w_i = 1/m, i = 1, \ldots, m$). $\square$

Learnability of Convex Learning Problems. A counterexample

- Question : Is convexity enough for a problem to be learnable ?
- Answer : **no**. It can be shown that even linear regression for $d = 1$ with
  - $\mathcal{H} = \mathbb{R}$
  - $\ell(w, (x, y)) = (wx - y)^2, w \in \mathbb{R}, (x, y) \in \mathbb{R}^2$

is not agnostic PAC learnable : For any size $m \geq 1$ and any leaning algorithm $A : S \to \mathbb{R}$ we can find $\epsilon_0 \in (0, 1)$ and $\delta_0 \in (0, 1)$ and a distribution $\mathcal{D}$ such that for

$$\mathbb{P}_{S \sim \mathcal{D}^m}\left(L_{\mathcal{D}}(A(S)) > \min_{w \in \mathbb{R}} L_{\mathcal{D}}(w) + \epsilon_0\right) \geq \delta_0.$$

Convex-Lipschitz/Smooth-Bounded Learning Problems

Definition (convex-Lipschitz-bounded Learning Problem). A learning problem, $(\mathcal{H}, \mathcal{Z}, \ell)$, is called **convex-Lipschitz-bounded**, with parameters ρ, B if the following holds

- The class $\mathcal{H}$ is a convex set and for all $\mathbf{w} \in \mathcal{H}$ we have $\|\mathbf{w}\| \leq B$,
- for all $z \in \mathcal{Z}$, $\mathbf{w} \mapsto \ell(\mathbf{w}, z)$ is convex and ρ-Lipschitz.

- **Example.** Consider the setting :
  - $\mathcal{X} = \{\mathbf{x} \in \mathbb{R}^d : \|\mathbf{x}\| \leq \rho\}$ and $\mathcal{Y} = \mathbb{R}$,
  - $\mathcal{H} = \{\mathbf{w} \in \mathbb{R}^d : \|\mathbf{w}\| \leq B\}$,
  - $\ell(\mathbf{w}, (\mathbf{x}, y)) = |\langle \mathbf{w}, \mathbf{x} \rangle - y|$.

Since the functions $t \mapsto |t|$ and $\mathbf{w} \mapsto \langle \mathbf{w}, \mathbf{x} \rangle - y$ are 1- and ρ-Lipschitz, it follows that $\mathbf{w} \mapsto \ell(\mathbf{w}, (\mathbf{x}, y))$ is ρ-Lipschitz.

Convex-Lipschitz/Smooth-Bounded Learning Problems

Definition (convex-smooth-bounded Learning Problem. A learning problem, $(\mathcal{H}, \mathcal{Z}, \ell)$, is called **convex-smooth-bounded**, with parameters β, B if the following holds

- The class $\mathcal{H}$ is a convex set and for all $\mathbf{w} \in \mathcal{H}$ we have $\|\mathbf{w}\| \leq B$.
- for all $z \in \mathcal{Z}$, $\mathbf{w} \mapsto \ell(\mathbf{w}, z)$ is convex and β-smooth.

- **Example.** Consider the setting :
  - $\mathcal{X} = \{\mathbf{x} \in \mathbb{R}^d : \|\mathbf{x}\| \leq \sqrt{\beta/2}\}$ and $\mathcal{Y} = \mathbb{R}$,
  - $\mathcal{H} = \{\mathbf{w} \in \mathbb{R}^d : \|\mathbf{w}\| \leq B\}$,
  - $\ell(\mathbf{w}, (\mathbf{x}, y)) = \big(\langle \mathbf{w}, \mathbf{x} \rangle - y\big)^2$.

We have seen that $\mathbf{w} \mapsto \big(\langle \mathbf{w}, \mathbf{x} \rangle - y\big)^2$ is $(2\|\mathbf{x}\|^2)$-smooth, and $2\|\mathbf{x}\|^2 \leq \beta$. Then, it follows that the loss function β-smooth.

Surrogate loss functions

- Consider the classification problem with halfspaces with domain $\mathcal{Z} = \mathbb{R}^d \times \{-1, 1\}$ and loss function

$$\ell_{0-1}(\mathbf{w}, (\mathbf{x}, y)) = \mathbb{1}_{[y \neq \text{sign}(\langle \mathbf{w}, \mathbf{x} \rangle)]}$$

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

- The function $\mathbf{w} \mapsto \ell_{0-1}(\mathbf{w}, (\mathbf{x}, y))$ is **not convex**. It can be shown that finding the ERM rule in the non-separable case (the case where we cannot find $\mathbf{w}^*$ such that $y_i = \text{sign}(\langle \mathbf{w}^*, \mathbf{x}_i \rangle)$) is **NP-hard**.

- To make the minimization problem **easier**, one solution is to **upper bound** the non-convex function (to be minimized) by a **convex surrogate** function. For example, consider

$$\ell^{\text{hinge}}(\mathbf{w}, (\mathbf{x}, y)) = \max\left(0, 1 - y\langle \mathbf{w}, \mathbf{x} \rangle\right).$$

Surrogate loss functions

- For all $\mathbf{w}$ and $(\mathbf{x}, y)$ we have that

$$\ell_{0-1}(\mathbf{w}, (\mathbf{x}, y)) \leq \ell^{\text{hinge}}(\mathbf{w}, (\mathbf{x}, y))$$

Indeed, $y \neq \text{sign}(\langle \mathbf{w}, \mathbf{x} \rangle) \iff y\langle \mathbf{w}, \mathbf{x} \rangle \leq 0$, and hence,

$$\ell_{0-1}(\mathbf{w}, (\mathbf{x}, y)) = 1 \Longrightarrow \ell^{\text{hinge}}(\mathbf{w}, (\mathbf{x}, y)) = 1 - y\langle \mathbf{w}, \mathbf{x} \rangle \geq 1.$$

- Also, $\mathbf{w} \mapsto \ell^{\text{hinge}}(\mathbf{w}, (\mathbf{x}, y))$ is **convex** by convexity of the maximum of convex functions.

- Let A be a learning algorithm which can learn $\mathbf{w}$ using the hinge loss. We
aim to achieve

$$L_{\mathcal{D}}^{\text{hinge}}(A(S)) \leq \min_{\mathbf{w} \in \mathcal{H}} L_{\mathcal{D}}^{\text{hinge}}(\mathbf{w}) + \epsilon$$

for some small estimation error ϵ.

Surrogate loss functions

- Here, $L_{\mathcal{D}}^{\text{hinge}}(\mathbf{w}) = \mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}}[\max\left(0, 1 - y\langle\mathbf{w},\mathbf{x}\rangle\right)]$ for any $\mathbf{w} \in \mathbb{R}^d$.
- Thus, we have

$$\begin{aligned}
L_{\mathcal{D}}^{0-1}(A(S)) &\leq L_{\mathcal{D}}^{\text{hinge}}(A(S)) \\
&\leq \min_{\mathbf{w}\in\mathcal{H}} L_{\mathcal{D}}^{\text{hinge}}(\mathbf{w}) + \epsilon \\
&= \underbrace{\min_{\mathbf{w}\in\mathcal{H}} L_{\mathcal{D}}^{0-1}(\mathbf{w})}_{\text{approximation error}} + \underbrace{\left(\min_{\mathbf{w}\in\mathcal{H}} L_{\mathcal{D}}^{\text{hinge}}(\mathbf{w}) - \min_{\mathbf{w}\in\mathcal{H}} L_{\mathcal{D}}^{0-1}(\mathbf{w})\right)}_{\text{optimization error}} \\
&\quad + \underbrace{\epsilon}_{\text{estimation error}}
\end{aligned}$$

- The optimization error depends on the unknown distribution $\mathcal{D}$ (and also on our choice for the surrogate function).

Lecture 8

What is the goal ?

- We consider again the setting where
  - $\mathcal{H}$ can be identified with some convex subset of vectors $\mathbf{w} \in \mathbb{R}^d$,
  - the loss function $\mathbf{w} \mapsto \ell(\mathbf{w}, z)$ is convex for any $z \in \mathcal{Z}$.
- Here, we will study the properties of a new learning method : **Stochastic gradient descent** (SGD) .
- We start with the simpler version called **gradient descent** and analyze its convergence.
- We will show how the SGD can be employed in learning problems.

Gradient descent

- **Idea :** If f is a differentiable function on $\mathbb{R}^d$ with gradient $\nabla f(\mathbf{w})$, then $\nabla f(\mathbf{w})$ points in the direction of the greatest rate of increase of f around $\mathbf{w}$.
- If f admits a minimum at $\mathbf{w}^*$, then we"hunt" for this minimizer by iteratively updating the operation $\mathbf{w}^{(t+1)} = \mathbf{w}^{(t)} - \eta \nabla f(\mathbf{w}^{(t)})$.
- Starting from $\mathbf{w}^{(1)} = \mathbf{0}$, it can be shown that under some conditions, the output $\bar{\mathbf{w}} = 1/T \sum_{t=1}^{T} \mathbf{w}^{(t)}$ converges to $\mathbf{w}^*$ for a large enough T.
- Suppose that f is convex. Then, the starting point is to write that

$$
\begin{aligned}
f(\bar{\mathbf{w}}) - f(\mathbf{w}^*) = f\left(\frac{1}{T}\sum_{t=1}^{T}\mathbf{w}^{(t)}\right) - f(\mathbf{w}^*) \quad &\leq \quad \frac{1}{T}\sum_{t=1}^{T} f(\mathbf{w}^{(t)}) - f(\mathbf{w}^*) \\
&= \quad \frac{1}{T}\sum_{t=1}^{T}\Big(f(\mathbf{w}^{(t)}) - f(\mathbf{w}^*)\Big)
\end{aligned}
$$

Gradient descent : Analysis

- Using convexity, we have that $f(\mathbf{w}^{(t)}) \leq f(\mathbf{w}^*) + \langle \mathbf{w}^{(t)} - \mathbf{w}^*, \nabla f(\mathbf{w}^{(t)}) \rangle$ and hence

$$f(\bar{\mathbf{w}}) - f(\mathbf{w}^*) \leq \frac{1}{T} \sum_{t=1}^{T} \langle \mathbf{w}^{(t)} - \mathbf{w}^*, \nabla f(\mathbf{w}^{(t)}) \rangle,$$

and the goal now is to **upper bound** the term on the right side :

Lemma. Let $\mathbf{v}_1, \ldots, \mathbf{v}_T$ be an arbitrary sequence of vectors. Any algorithm with an **initialization $\mathbf{w}^{(1)} = \mathbf{0}$** and an **update rule** of the form

$$\mathbf{w}^{(t+1)} = \mathbf{w}^{(t)} - \eta \mathbf{v}_t$$

for some $\eta > 0$ satisfies

$$\sum_{t=1}^{T} \langle \mathbf{w}^{(t)} - \mathbf{w}^*, \mathbf{v}_t \rangle \leq \frac{\|\mathbf{w}^*\|^2}{2\eta} + \frac{\eta}{2} \sum_{t=1}^{T} \|\mathbf{v}_t\|^2.$$

Gradient descent : Analysis

Lemma (continued). In particular, for $\forall\ B > 0, \rho > 0$, if we have $\|\mathbf{v}_t\| \leq \rho$ and if $\eta = \sqrt{B^2/(\rho^2 T)}$, then for any $\mathbf{w}^* : \|\mathbf{w}^*\| \leq B$ we have

$$\frac{1}{T}\sum_{t=1}^{T}\langle \mathbf{w}^{(t)} - \mathbf{w}^*, \nabla f(\mathbf{w}^{(t)})\rangle \leq \frac{B\rho}{\sqrt{T}}.$$

Proof. We can write that

$$\begin{aligned}
\langle \mathbf{w}^{(t)} - \mathbf{w}^*, \mathbf{v}_t\rangle &= \frac{1}{\eta}\langle \mathbf{w}^{(t)} - \mathbf{w}^*, \eta\mathbf{v}_t\rangle \\
&= \frac{1}{2\eta}\left(-\|\mathbf{w}^{(t)} - \mathbf{w}^* - \eta\mathbf{v}_t\|^2 + \|\mathbf{w}^{(t)} - \mathbf{w}^*\|^2 + \eta^2\|\mathbf{v}_t\|^2\right) \\
&= \frac{1}{2\eta}\left(-\|\mathbf{w}^{(t+1)} - \mathbf{w}^*\|^2 + \|\mathbf{w}^{(t)} - \mathbf{w}^*\|^2\right) + \frac{\eta}{2}\|\mathbf{v}_t\|^2
\end{aligned}$$

by definition of $\mathbf{w}^{(t+1)}$.

Gradient descent : Analysis

Proof (continued). By summing over t, it follows that

$$\begin{aligned}\sum_{t=1}^{T}\langle \mathbf{w}^{(t)} - \mathbf{w}^*, \mathbf{v}_t\rangle &= \frac{1}{2\eta}\left(-\|\mathbf{w}^{(T+1)} - \mathbf{w}^*\|^2 + \|\mathbf{w}^{(1)} - \mathbf{w}^*\|^2\right) + \frac{\eta}{2}\sum_{t=1}^{T}\|\mathbf{v}_t\|^2 \\ &\leq \frac{1}{2\eta}\|\mathbf{w}^{(1)} - \mathbf{w}^*\|^2 + \frac{\eta}{2}\sum_{t=1}^{T}\|\mathbf{v}_t\|^2 \\ &= \frac{1}{2\eta}\|\mathbf{w}^*\|^2 + \frac{\eta}{2}\sum_{t=1}^{T}\|\mathbf{v}_t\|^2, \quad \text{since } \mathbf{w}^{(1)} = \mathbf{0}.\end{aligned}$$

If $\|\mathbf{w}^*\| \leq B$, $\|\mathbf{v}_t\| \leq \rho$ and $\eta = \sqrt{B^2/(\rho^2 T)}$, then we can further bound the right term $/T$ by

$$\frac{1}{2T}\frac{\rho\sqrt{T}}{B}B^2 + \frac{B}{2\rho\sqrt{T}}\rho^2 = \frac{B\rho}{\sqrt{T}}. \qquad \square$$

Gradient descent : Analysis

Corollary. Let f be a convex, ρ-Lipschitz function and differentiable, and let $\mathbf{w}^* \in \operatorname{argmin}_{\mathbf{w}:\|\mathbf{w}\|\le B} f(\mathbf{w})$. If the GD algorithm is run for T steps with $\eta = \sqrt{B^2/(\rho^2 T)}$, then

$$f(\bar{\mathbf{w}}) - f(\mathbf{w}^*) \le \frac{B\rho}{\sqrt{T}}.$$

Thus, to have $f(\bar{\mathbf{w}}) - f(\mathbf{w}^*) \le \epsilon$ for some $\epsilon > 0$, it suffices to take $T \ge B^2\rho^2/\epsilon^2$.

Proof. Since f is ρ-Lipschitz and differentiable, we have that $\|\nabla f(\mathbf{w}^{(t)})\| \le \rho$. Take

$$\mathbf{v}_t = \nabla f(\mathbf{w}^{(t)})$$

and apply the previous Lemma. □

Subgradients

- We can generalize the GD algorithm to convex **non-differentiable** functions, using **subgradients**.
- Recall that if f is a convex differentiable function, then for all $\mathbf{u}$

$$f(\mathbf{u}) \geq f(\mathbf{w}) + \langle \mathbf{u} - \mathbf{w}, \nabla f(\mathbf{w}) \rangle.$$

This property can be strengthened through the following result :

Lemma. Let S be an open convex set. A function $f : S \to \mathbb{R}$ is convex iff $\forall\ \mathbf{w} \in S\ \exists\ \mathbf{v} : f(\mathbf{u}) \geq f(\mathbf{w}) + \langle \mathbf{u} - \mathbf{w}, \mathbf{v} \rangle$ for all $\mathbf{u} \in S$. $(\star)$

Definition (subgradients). A vector $\mathbf{v}$ that satisfies $(\star)$ is called a subgradient of f at $\mathbf{w}$. The set of all subgradients of f at $\mathbf{w}$ is called the differential set and is denoted by $\partial f(\mathbf{w})$.

Subgradients : calculation and examples

- **Result.** If f is differentiable at $\mathbf{w}$, then $\partial f(\mathbf{w}) = \{\nabla f(\mathbf{w})\}$.
- **Example.** Consider $f(x) = |x|$. This function is differentiable on $(-\infty, 0) \cup (0, \infty)$ and hence $\partial f(x) = \{-1\}$ if $x < 0$ and $\partial f(x) = \{1\}$ if $x > 0$. For $x = 0$, note that

$$f(t) \geq f(0) + a(t-0) \iff |t| \geq at \iff a \leq 1 \text{ or } a \geq -1.$$

Hence,

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

- **Result.** Let $g_1, \ldots, g_r$ be r convex differentiable functions and $g = \max_{1 \leq i \leq r} g_i$. For a given $\mathbf{w}$, let $j \in \{1, \ldots, r\}$ such that $g(\mathbf{w}) = g_j(\mathbf{w})$. Then,

$$\nabla g_j(\mathbf{w}) \in \partial g(\mathbf{w}).$$

Subgradients : calculation and examples

- **Proof.** Convexity of g_j implies that for all $\mathbf{u}$

$$g_j(\mathbf{u}) \geq g_j(\mathbf{w}) + \langle \mathbf{u} - \mathbf{w}, \nabla g_j(\mathbf{w}) \rangle.$$

  Since $g(\mathbf{w}) = g_j(\mathbf{w})$ and $g(\mathbf{u}) \geq g_j(\mathbf{u})$, it follows that

$$g(\mathbf{u}) \geq g(\mathbf{w}) + \langle \mathbf{u} - \mathbf{w}, \nabla g_j(\mathbf{w}) \rangle.$$

  As this is true for all $\mathbf{u}$, this means that $\nabla g_j(\mathbf{w}) \in \partial g(\mathbf{w})$. $\square$

- **Example.** Consider the hinge loss function $f(\mathbf{w}) = \max(0, 1 - y\langle \mathbf{w}, \mathbf{x} \rangle)$ for some vector $\mathbf{x} \in \mathbb{R}^d$ and $y \in \{-1, 1\}$. Then, for a given $\mathbf{w} \in \mathbb{R}^d$, the vector

$$\mathbf{v} = \begin{cases} \mathbf{0}, & \text{if } 1 - y\langle \mathbf{w}, \mathbf{x} \rangle \leq 0 \\ -y\mathbf{x}, & \text{if } 1 - y\langle \mathbf{w}, \mathbf{x} \rangle > 0 \end{cases}$$

  is a subgradient of f at $\mathbf{w}$.

Subgradients of Lipschitz functions

- Recall that a function : $A \to \mathbb{R}$ is ρ-Lipschitz if for all $\mathbf{u}, \mathbf{v} \in A$, we have $|f(\mathbf{v}) - f(\mathbf{u})| \leq \rho\|\mathbf{v} - \mathbf{u}\|$.

Lemma. Let A be a convex open set and let $f : A \to \mathbb{R}$ be a convex function. Then,

$$f \text{ is } \rho\text{-Lipschitz over } A \iff \forall\, \mathbf{w} \in A, \mathbf{v} \in \partial f(\mathbf{w}) \text{ we have that } \|\mathbf{v}\| \leq \rho.$$

Proof. Suppose that any $\mathbf{v} \in \partial f(\mathbf{w})$ satisfies $\|\mathbf{v}\| \leq \rho$. By definition of $\partial f(\mathbf{w})$, we have that

$$f(\mathbf{w}) - f(\mathbf{u}) \leq \langle \mathbf{v}, \mathbf{w} - \mathbf{u} \rangle.$$

By the Cauchy-Schwartz inequality applied to the right term, the latter inequality implies that

$$f(\mathbf{w}) - f(\mathbf{u}) \leq \|\mathbf{v}\|\ \|\mathbf{w} - \mathbf{u}\| \leq \rho\|\mathbf{w} - \mathbf{u}\|$$

Subgradients of Lipschitz functions

Proof (continued). A similar argument can be applied to show that $f(\mathbf{u}) - f(\mathbf{w}) \leq \rho \|\mathbf{u} - \mathbf{w}\|$. Hence, f is ρ-Lipschitz.

Suppose now that f is ρ-Lipschitz, and let $\mathbf{w} \in A$ and $\mathbf{v} \in \partial f(\mathbf{w})$. If $\mathbf{v} = \mathbf{0}$, then we are done. Suppose now that $\mathbf{v} \neq \mathbf{0}$. Since A is open, we can find a small $\epsilon > 0$ such that $\mathbf{u} = \mathbf{w} + \epsilon \mathbf{v} / \|\mathbf{v}\| \in A$. Then,

$$\langle \mathbf{u} - \mathbf{w}, \mathbf{v} \rangle = \epsilon \|\mathbf{v}\|, \text{ and } \|\mathbf{u} - \mathbf{w}\| = \epsilon.$$

From the definition of the subgradient and ρ-Lipschitzness, we have that

$$\rho\epsilon = \rho \|\mathbf{u} - \mathbf{w}\| \geq f(\mathbf{u}) - f(\mathbf{w}) \geq \langle \mathbf{u} - \mathbf{w}, \mathbf{v} \rangle = \epsilon \|\mathbf{v}\|$$

implying that $\|v\| \leq \rho$. $\square$

Subgradient descent

- In case f is non-differentiable but convex and ρ-Lipschitz, we can construct a subgradient descent algorithm, where $\mathbf{v}_t \in \partial f(\mathbf{w}^{(t)})$:
  - Start with $\mathbf{w}^{(1)} = \mathbf{0}$.
  - For $t = 1, \ldots, T$, take $w^{(t+1)} = w^{(t)} - \eta \mathbf{v}_t$ with $\mathbf{v}_t \in \partial f(w^{(t)})$.
  - Output $\bar{\mathbf{w}} = \sum_{t=1}^{T} \mathbf{w}^{(t)} / T$.
- If we again take $\eta = \sqrt{B^2/(\rho^2 T)}$, then $f(\bar{\mathbf{w}}) - f(\mathbf{w}^*) \leq \frac{B\rho}{\sqrt{T}}$ under the assumption that the minimizer $\mathbf{w}^*$ of f satisfies $\|\mathbf{w}^*\| \leq B$.
- **Justification** : the following two ingredients can be again used in the proof (as for GD)
  - any $\mathbf{v}_t \in \partial f(w^{(t)})$ satisfies $\|\mathbf{v}_t\| \leq \rho$ (by the ρ-Lipschitzness of f).
  - $f(\mathbf{w}^{(t)}) - f(\mathbf{w}^*) \leq \langle \mathbf{w}^{(t)} - \mathbf{w}^*, \mathbf{v}_t \rangle$ (by the properties of subgradients).

Stochastic gradient descent

- **Idea :** The function f we want to minimize is unknown. Thus, the gradient or sub-gradient at any vector $\mathbf{w}$ is also unknown. What should we do ?
- At some iteration t, we can replace the unknown gradient or subgradient by a random vector $\mathbf{v}_t$ such that

$$\mathbb{E}[\mathbf{v}_t|\mathbf{w}^{(t)}] \in \partial f(\mathbf{w}^{(t)})$$

- This random step yields the stochastic gradient descent algorithm (SDG) : for some $\eta > 0$ and $T > 0$ an integer
  - Start with $\mathbf{w}^{(1)} = \mathbf{0}$
  - For $t = 1, \ldots, T$
    - generate $\mathbf{v}_t$ from a distribution such that $\mathbb{E}[\mathbf{v}_t|\mathbf{w}^{(t)}] \in \partial f(\mathbf{w}^{(t)})$
    - update $\mathbf{w}^{(t+1)} = \mathbf{w}^{(t)} - \eta \mathbf{v}_t$
  - Output $\bar{\mathbf{w}} = \sum_{t=1}^{T} \mathbf{w}^{(t)}/T$