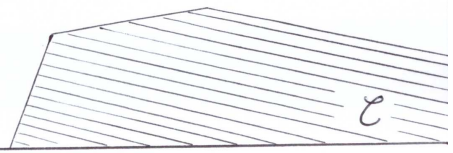
SL = GRADIENT DESCENT ALGORITHMS

Gradient descent methods are a family of algorithms designed to optimize $\min_{x \in \mathcal{C}} f(x)$, where both f and $\mathcal{C}$ are convex. We present in this chapter some popular gradient techniques, with perhaps the most important one: Stochastic Gradient Descent (SGD). We then adapt these algorithms to learning problems, where the objective is to minimize the empirical risk: $\min_{f \in \mathcal{F}} \hat{R}_n(f)$ (or some regularized version of it). We first recall some important definitions and properties of functions, such as convexity, Lipschitzness, and subdifferentiability.

I. PRELIMINARIES

I.1. Convexity.

Convexity of sets and functions were introduced in the chapter SL: SUPPORT VECTOR MACHINE. A set $\mathcal{C}$ is convex if $\forall x, y \in \mathcal{C}$, $0 \leq \lambda \leq 1$, $\lambda x + (1-\lambda) y \in \mathcal{C}$ [$\mathcal{C}$ contains line segments between any two points]. A function $f: \mathbb{R}^d \to \mathbb{R}$ is convex if its domain is a convex set, and $f(\lambda x + (1-\lambda) y) \leq \lambda f(x) + (1-\lambda) f(y)$, $\forall x, y \in \text{dom } f$, $\forall \lambda \in [0,1]$.



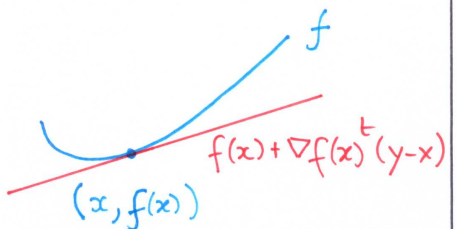
When f is differentiable / smooth, it is easy to check whether it is convex or not.

→ First order condition (f differentiable)

"f lies above its tangent line"

$\forall x, y \in \text{dom } f$,

$f(y) \geq f(x) + \nabla f(x)^t (y - x)$



→ Second order condition (f twice differentiable)

f is convex (strictly) $\Leftrightarrow$ $\nabla^2 f(x) \succcurlyeq 0$ ($\succ$) $\quad \forall x \in \text{dom } f$.

Ex: (i) Quadratic functions $f(x) = \frac{1}{2} x^t P x + q^t x + r$

$\nabla f(x) = Px + q$

$\nabla^2 f(x) = P \Rightarrow f$ is convex $\Leftrightarrow P \succcurlyeq 0$

(ii) Least Squares objective $f(\beta) = \| y - X\beta \|_2^2$

$\nabla f(\beta) = -2 X^t (y - X\beta)$

$\nabla^2 f(\beta) = 2 X^t X \succcurlyeq 0$ (convex $\forall X$).

Theorem. Let $f_i : \mathbb{R}^d \to \mathbb{R}$ be convex functions $i = 1, \dots, n$.

Then

(i) $f(x) := \max_{1 \leq i \leq n} f_i(x)$ (pointwise maximum) is also convex

(ii) $f(x) := \sum_{i=1}^n \alpha_i f_i(x)$, $\alpha_i \geq 0$ (weighted sum) is cnvx.

I.2. Lipschitzness

Let $\mathcal{C} \subset \mathbb{R}^d$ (not necessarily convex). A function $f : \mathbb{R}^d \to \mathbb{R}^k$ is L-Lipschitz on $\mathcal{C}$ if $\forall x, y \in \mathcal{C}$,

$$\| f(x) - f(y) \|_2 \leq L \| x - y \|_2$$

Euclidean norms in $\mathbb{R}^k$ / $\mathbb{R}^d$ respectively; but the notion of Lipschitzness can be generalized to arbitrary norms. We drop the subscript 2 in what follows.

- **Consequences**: An L-Lipschitz function cannot move too fast. For example, if $f : \mathbb{R} \to \mathbb{R}$ is differentiable $\forall x \in \operatorname{dom} f$, then $\forall h > 0$, $\frac{|f(x+h) - f(x)|}{h} \leq L$

$$\frac{|f(x+h) - f(x)|}{h} \xrightarrow{h \to 0} f'(x)$$

$\Rightarrow$ The derivative ($\equiv$ rate of change of f) is bounded by the Lipschitz constant (we may reach a similar conclusion using the mean value theorem : $\exists u \in [x, y]$, such that $f(x) - f(y) = f'(u)(x - y)$).

Ex: (i) $f(x) = |x|$ is 1-Lipschitz on $\mathbb{R}$ since $\forall x, y$,

$$|x| - |y| = |x - y + y| - |y| \leq |x - y| + |y| - |y| = |x - y|$$

Likewise $|y| - |x| \leq |x - y| \Rightarrow ||x| - |y|| \leq |x - y|$.

(ii) $f(x) = \log(1 + e^x)$ is 1-Lipschitz on $\mathbb{R}$ since

$$|f'(x)| = \left| \frac{e^x}{1 + e^x} \right| \leq 1$$

(iii) $f(x)$ is not L-Lipschitz, for any L. Indeed, taking $x = 0$ and $y = 1 + L$, we have

$$f(y) - f(x) = (1 + L)^2 > L(1 + L) = L|y - x|$$

However, $f(x)$ is L-Lipschitz on $\mathcal{C} = \{x \mid |x| \leq \frac{L}{2}\}$ since $\forall x, y \in \mathcal{C}$,

$$|f(x) - f(y)| = |x^2 - y^2| = |x - y| \underbrace{|x + y|}_{\leq 2 \times \frac{L}{2}} \leq L|x - y|$$

I.3. Subgradients.

A convex and differentiable function f is such that the graph of f lies above its tangent line at any point of its domain. The slope of the tangent line being the gradient of f evaluated at that point. When dealing with non-differentiable functions, we introduce the notion of subdifferential $\partial f(x)$ of f at $x \in \operatorname{dom} f$:

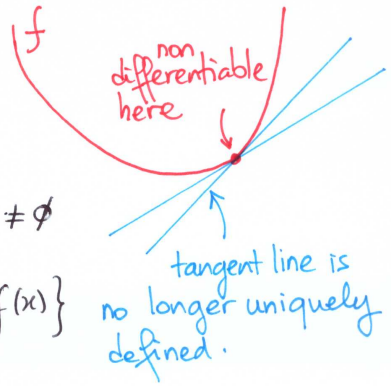
$$\partial f(x) = \{u \in \mathbb{R}^d \mid f(y) \geq f(x) + \langle u, y - x \rangle \ \forall y \in \operatorname{dom} f\}$$

$u \in \partial f(x)$ is called the SUBGRADIENT of f at x.

In fact, we have :

f is convex $\Leftrightarrow \forall x \in \operatorname{dom} f \quad \partial f(x) \neq \emptyset$

f is cnvx & differentiable at $x \Leftrightarrow \partial f(x) = \{\nabla f(x)\}$



Ex: $\partial \|x\|_1 = \left\{ u \in \mathbb{R}^d \;\middle|\; \begin{array}{ll} u_j = \text{sign } x_j & \text{if } x_j \neq 0 \\ u_j \in [-1,1] & \text{if } x_j = 0 \end{array} \right\}$

(see chapter SL: RIDGE REGRESSION & LASSO)

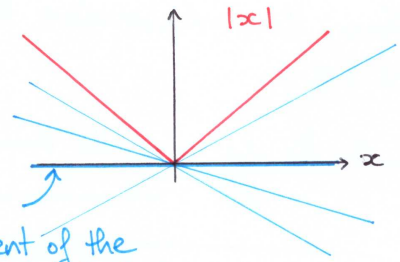
Theorem: The minimizers of a convex function $f: \mathbb{R}^d \to \mathbb{R}$ are characterized by

$$x_* \in \operatorname{argmin}_{x \in \mathbb{R}^d} f(x) \iff 0 \in \partial f(x_*)$$

↑ Follows directly from $\forall x \in \text{dom } f$

$f(x) \geq f(x^*) + \langle 0, x - x^* \rangle$,

so that $0 \in \partial f(x^*)$ indeed ■



slope 0 is an element of the subgradient of $|\cdot|$ at 0.

Theorem: Let $\mathcal{C}$ be a convex open set, $f: \mathcal{C} \to \mathbb{R}$ a convex function.

Then

f is L-Lipschitz on $\mathcal{C} \iff \forall x \in \mathcal{C}, \; \forall u \in \partial f(x), \; \|u\| \leq L$

↳ For differentiable L-Lipschitz functions, the gradient is bounded by the Lipschitz constant. This theorem extends the result to non-differentiable L-Lipschitz functions, stating that the norm of the subgradient is necessarily less than L.

proof: ⇐ Suppose that $\forall x, \forall u \in \partial f(x), \; \|u\| \leq L$.

Since $u \in \partial f(x)$, we have $f(y) \geq f(x) + \langle u, y - x \rangle$.

Thus $f(x) - f(y) \leq \langle u, x - y \rangle$ ↘ Cauchy-Schwartz

$\leq \|u\| \, \|x - y\|$

$\leq L \|x - y\|$

Likewise, we can show that $f(y) - f(x) \leq L \|x - y\|$, which concludes the first part of the theorem.

⇒ Suppose that f is L-Lipschitz.

Take $x \in \mathcal{C}$ and $u \in \partial f(x)$.

Since $\mathcal{C}$ is open, $\exists \varepsilon > 0$ s.t. $y := x + \varepsilon \frac{u}{\|u\|} \in \mathcal{C}$.

Thus $\rightarrow \langle y - x, u \rangle = \varepsilon \langle \frac{u}{\|u\|}, u \rangle = \varepsilon \|u\|$

$\rightarrow \|x - y\| = \varepsilon$.

Since $u \in \partial f(x)$, we have $f(y) \geq f(x) + \langle u, y - x \rangle$,

or $f(y) - f(x) \geq \langle u, y - x \rangle = \varepsilon \|u\|$. (*)

Moreover, $f(y) - f(x) \leq L \|y - x\| = L \varepsilon$ (**)

↑ f is L-Lipschitz

Putting (*) & (**) together, we obtain $\|u\| \leq L$ indeed ■

II. GRADIENT DESCENT ALGORITHMS.

II.1. Standard Approach

We consider the unconstrained optimization problem

$$\text{minimize}_{x \in \mathbb{R}^d} \; f(x).$$

where $f: \mathbb{R}^d \to \mathbb{R}$ is a convex function. If in addition f is differentiable, a necessary & sufficient condition for

a point x^* to be optimal is that the gradient of f at x^* vanishes : $\nabla f(x^*) = 0$ ($\equiv$ the advantage of working with convex functions: obtain global properties from local properties: a local minimum is a global minimum).

The optimization task is usually performed iteratively : an algorithm computes a sequence of points $x_0, x_1, \dots$ such that (hopefully) : $f(x_k) \to f(x^*)$ as $k \to +\infty$.

Descent algorithms produce a sequence $\{x_k\}_{k \geq 0}$ s.t.

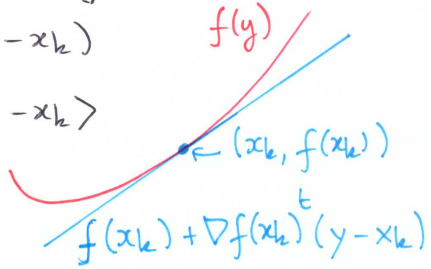
$$x_{k+1} = x_k + \eta_k \, \Delta x_k \quad \& \quad f(x_{k+1}) < f(x_k)$$

(η_k: step size (>0); Δx_k: search direction)

For a cnvx & differentiable function f,

$$f(y) \geq f(x_k) + \nabla f(x_k)^t (y - x_k)$$

$$f(y) - f(x_k) \geqslant \langle \nabla f(x_k), y - x_k \rangle$$



To get $f(y) < f(x_k)$, the inner product must be negative, so that $y - x_k$ must go in the direction of the negative gradient.

The search direction is then $\Delta x_k = -\nabla f(x_k)$, so that $x_{k+1} = x_k - \eta_k \nabla f(x_k)$.

If we loose differentiability, we loose the uniqueness of the tangent line, and $y - x_k$ must be in the direction of one of the subgradients. We obtain the following procedure:

Algorithm I : Gradient Descent Algorithm

1. Init: $x_1 \in \mathbb{R}^d$ and a positive sequence $\{\eta_j\}_{j \geq 1}$
2. For $j = 1, \dots, k-1$ do
3. $\quad x_{j+1} = x_j - \eta_j g_j$, $g_j \in \partial f(x_j)$
4. End For
5. Return either $\bar{x} = \frac{1}{k} \sum_{j=1}^{k} x_j$ or $x^\circ \in \operatorname{argmin}_{x \in \{x_1, \dots, x_k\}} f(x)$

Goal: $\min_{x \in \mathbb{R}^d} f(x)$; $f: \mathbb{R}^d \to \mathbb{R}$

Theorem: Let f = convex + L-Lipschitz function on $\mathbb{R}^d$, such that x^* exists.

Assume that $\| x_1 - x^* \|_2 \leqslant R$.

Then, if $\eta_j = \eta = \frac{R}{L\sqrt{k}}$, $\forall j \geqslant 1$, we have

$$\to f\left(\frac{1}{k} \sum_{j=1}^{k} x_j \right) - f(x^*) \leqslant \frac{LR}{\sqrt{k}}$$

$$\to \min_{1 \leqslant j \leqslant k} f(x_j) - f(x^*) \leqslant \frac{LR}{\sqrt{k}}$$

Alternatively, we may consider $f: \mathbb{R}^d \to \mathbb{R}$, and look for $x \in \mathcal{C} \subset \mathbb{R}^d$ s.t $\mathcal{C} = \{ x : \| x \| \leqslant R/2 \}$ ($\equiv \operatorname{diam}(\mathcal{C}) \leqslant R$)

proof = First, note that

$$\min_{1 \leqslant j \leqslant k} f(x_j) - f(x^*) \leqslant \frac{1}{k} \sum_{j=1}^{k} f(x_j) - f(x^*)$$

$$f\left(\frac{1}{k} \sum_{j=1}^{k} x_j \right) - f(x^*)$$

Jensen

⇒ We turn our attention to the term $f(x_j) - f(x^*)$.

$$f(x_j) - f(x^*) \leq g_j^t (x_j - x^*) \quad \text{(def of subgradient)}$$

$$= \frac{1}{\eta} (x_j - x_{j+1})^t (x_j - x^*) \quad \text{(since } \eta g_j = x_j - x_{j+1})$$

Since $\forall u, v \in \mathbb{R}^d$, $2u^t v = \|u\|^2 + \|v\|^2 - \|u-v\|^2$

$$= \frac{1}{2\eta} \left\{ \|x_j - x_{j+1}\|^2 + \|x_j - x^*\|^2 - \|x_{j+1} - x^*\|^2 \right\}$$

$$= \frac{\eta}{2} \|g_j\|^2 + \frac{1}{2\eta} \left\{ d_j^2 - d_{j+1}^2 \right\},$$

where $d_j := \|x_j - x^*\|$.

f is L-Lipschitz $\Rightarrow \|g_j\|^2 \leq L^2$ (Theorem p. 5), so that

$$f(x_j) - f(x^*) \leq \frac{\eta L^2}{2} + \frac{1}{2\eta} (d_j^2 - d_{j+1}^2) \qquad (1)$$

summing up from 1 to k: (telescoping sum)

$$\frac{1}{k} \sum_{j=1}^{k} f(x_j) - f(x^*) \leq \frac{\eta L^2}{2} + \frac{1}{2\eta k} (d_1^2 - d_{k+1}^2)$$

$$\leq \frac{\eta L^2}{2} + \frac{1}{2\eta k} d_1^2$$

$$\leq \frac{\eta L^2}{2} + \frac{R^2}{2\eta k}.$$

Optimizing with respect to η ($\eta = \frac{R}{L\sqrt{k}}$) gives the result. ∎

Remark = The learning rate depends on the maximum number of iterations, k. We would prefer to obtain theoretical guaranties with a learning rate independent of k. To this end, note that

$$\left(\sum_{j=1}^{k} \eta_j \right) \left(\min_{1 \leq j \leq k} f(x_j) - f(x^*) \right) \leq \sum_{j=1}^{k} \eta_j (f(x_j) - f(x^*))$$

Thus, making use of relation (1) on page 9, we see that

$$\eta_j (f(x_j) - f(x^*)) \leq \frac{\eta_j^2 L^2}{2} + \frac{1}{2} (d_j^2 - d_{j+1}^2)$$

summing up:

$$\sum_{j=1}^{k} \eta_j (f(x_j) - f(x^*)) \leq \frac{L^2}{2} \sum_{j=1}^{k} \eta_j^2 + \frac{1}{2} \sum_{j=1}^{k} (d_j^2 - d_{j+1}^2)$$

$$\leq \frac{L^2}{2} \sum_j \eta_j^2 + \frac{R^2}{2}$$

dividing by $\sum \eta_j$

$$\min_{1 \leq j \leq k} f(x_j) - f(x^*) \leq \frac{L^2}{2} \frac{\sum_j \eta_j^2}{\sum_j \eta_j} + \frac{R^2}{2} \frac{1}{\sum_j \eta_j}$$

For this term to go to 0 with k, we need $\sum \eta_j \to +\infty$ & $\frac{\sum \eta_j^2}{\sum \eta_j} \to 0$

One possible candidate is to take $\eta_j = \frac{c}{\sqrt{j}}$, for some $c > 0$. Indeed,

$$\to \sum_{j=1}^{k} \eta_j^2 \leq c^2 \sum_{j=1}^{k} \frac{1}{j} \leq c^2 \left(1 + \int_1^k \frac{dx}{x}\right) \leq c^2 c_1 \log k$$

$$\to \sum_{j=1}^{k} \eta_j \geq \sum_{j=1}^{k} \eta_k = \sum_{j=1}^{k} \frac{c}{\sqrt{k}} = c\sqrt{k}.$$

We obtain:

$$\min_{1\leq j\leq k} f(x_j) - f(x^*) \leq \frac{L^2}{2}\frac{cc_1 \log k}{\sqrt{k}} + \frac{R^2}{2}\frac{1}{c\sqrt{k}}$$

Optimizing with respect to c: $c = \frac{R}{L\sqrt{c_1 \log k}}$,

we get a rate $O\left(LR\sqrt{\frac{\log k}{k}}\right)$.

↖ We get an extra $\sqrt{\log k}$ factor.

However, noticing that

$$\min_{1\leq j\leq k} f(x_j) - f(x^*) \leq \min_{\frac{k}{2}\leq j\leq k} f(x_j) - f(x^*)$$

$$\leq \frac{1}{\sum_{j=k/2}^{k} \eta_j}\left(\sum_{j=k/2}^{k} f(x_j) - f(x^*)\right)$$

$$\leq c_2 \frac{LR}{\sqrt{k}},$$

since with $\eta_j = \frac{c}{\sqrt{j}}$, we still have $\sum_{j=\frac{k}{2}}^{k} \eta_j \geq c\sqrt{k}$,

but $\sum_{j=k/2}^{k} \eta_j^2 \leq c^2 \sum_{j=k/2}^{k} \frac{1}{j} \leq c' \int_{k/2}^{k} \frac{1}{x}\,dx \leq c''$,

for $c', c'' > 0$ constants. However, we need to ensure that $\|x_{k/2} - x^*\| \leq R$ which is not always true (unless we are restricted to $\mathcal{C} = \{x \mid \|x\| \leq R/2\}$). ■

II.2. Projected Gradient Descent.

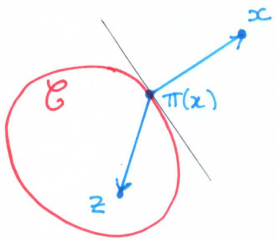
When dealing with constrained optimization problem of the form $\underset{x\in\mathcal{C}}{\text{minimize}}\ f(x)$, where $\begin{cases} f, \mathcal{C} = \text{convex} \\ f: \mathbb{R}^d \to \mathbb{R} \end{cases}$, we need to make sure that at each iteration $x_j \in \mathcal{C}$. The standard form of gradient descent does not ensure this. The projected gradient descent algorithm will incorporate this condition by adding an extra step at each iteration, and projecting the current estimate onto the convex set $\mathcal{C}$. We have the following result.

Theorem. Let $\mathcal{C}$ be a closed convex set $\subset \mathbb{R}^d$.
$\forall x \in \mathbb{R}^d$, let $\pi(x) \in \mathcal{C}$ be such that

$$\|x - \pi(x)\| = \min_{z\in\mathcal{C}} \|x - z\|.$$

Then $\pi(x)$ is unique, and

$$\langle \pi(x) - x, \pi(x) - z \rangle \leq 0 \quad \forall z \in \mathcal{C}$$



proof • Uniqueness.

If $\pi_1, \pi_2 \in \mathcal{C}$ both satisfy $\langle \pi_1 - x, \pi_1 - z \rangle \leq 0$
$\langle \pi_2 - x, \pi_2 - z \rangle \leq 0 \quad \forall z,$

Then taking $z = \pi_2$ in the first equality
& $z = \pi_1$ in the second,

$$\left.\begin{aligned} \langle \pi_1 - x, \pi_1 - \pi_2 \rangle &\leq 0 \\ \langle x - \pi_2, \pi_1 - \pi_2 \rangle &\leq 0 \end{aligned}\right\} \to \text{adding them} \quad \|\pi_1 - \pi_2\|^2 \leq 0$$

$$\Rightarrow \pi_1 = \pi_2.$$

- Proof that $\langle \pi - x, \pi - z \rangle \leq 0$.

By definition, $\|x - \pi\|^2 \leq \|x - z\|^2 \quad \forall z \in \mathcal{C}$.

Fix $u \in \mathcal{C}$ and put $z := (1-\varepsilon)\pi + \varepsilon u \; ; \; \varepsilon \in (0,1]$.

$\uparrow$ $\in \mathcal{C}$ since $\mathcal{C}$ is convex, and both u, π belong to $\mathcal{C}$.

$$\Rightarrow \|x-\pi\|^2 \leq \|x - z\|^2 = \underbrace{\|x - \pi - \varepsilon(u-\pi)\|^2}_{\|}$$

$$\|x-\pi\|^2 - 2\varepsilon\langle x-\pi, u-\pi\rangle + \varepsilon^2\|u-\pi\|^2$$

$$\Updownarrow$$

$$\langle x-\pi, u-\pi\rangle \leq \varepsilon \|u-\pi\|^2$$

Valid $\forall \varepsilon \in (0,1]$. Taking the limit $\varepsilon \to 0$ yields the desired result. ■

Algorithm 2 : Projected Gradient Descent Algorithm

1. Init: $x_1 \in \mathcal{C}$ and a positive sequence $\{\eta_j\}_{j \geq 1}$.
2. For $j = 1, \dots, k-1$ do
3. $\quad z_{j+1} = x_j - \eta_j g_j \; , \quad g_j \in \partial f(x_j)$
4. $\quad x_{j+1} = \pi(z_{j+1})$
5. End For
6. Return either $\bar{x} = \frac{1}{k}\sum_{j=1}^{k} x_j$ or $x^\circ \in \operatorname{argmin}_{x \in \{x_1, \dots, x_k\}} f(x)$

↑ Note that the projection is in terms of the Euclidean norm $\|\cdot\|_2$.

Goal: minimize $f(x)$; $f: \mathbb{R}^d \to \mathbb{R}$, $x \in \mathcal{C}$ ↖ close & convex

— Theorem. Let $\mathcal{C}$ = closed, convex, non-empty subset of $\mathbb{R}^d$, such that $\operatorname{diam}(\mathcal{C}) \leq R$.
Let f be a convex, L-Lipschitz function on $\mathcal{C}$, such that $x_* \in \operatorname{argmin}_{x \in \mathcal{C}} f(x)$ exists. ↖ but f is defined $\forall x \in \mathbb{R}^d$

- Then, with $\eta_j = \frac{R}{L\sqrt{k}}$, we have

$$f(\bar{x}) - f(x^*) \leq \frac{LR}{\sqrt{k}} \quad \& \quad f(x^\circ) - f(x^*) \leq \frac{LR}{\sqrt{k}}.$$

- In addition, if $\eta_j = \frac{R}{L\sqrt{j}}$, there exist $c > 0$ s.t.

$$f(\bar{x}) - f(x^*) \leq \frac{cLR}{\sqrt{k}} \quad \& \quad f(x^\circ) - f(x^*) \leq \frac{cLR}{\sqrt{k}}$$

proof : We proceed as before ; and evaluate the difference

$$f(x_j) - f(x^*) \leq g_j^t (x_j - x^*)$$

$$= \frac{1}{\eta}(x_j - z_{j+1})^t (x_j - x^*)$$

(With $\eta = \eta_j$)

$$= \frac{1}{2\eta}\left\{\|x_j - z_{j+1}\|^2 + \|x_j - x^*\|^2 - \|z_{j+1} - x^*\|^2\right\}.$$

We turn our attention to the term $\|z_{j+1} - x^*\|^2$:

$$\|z_{j+1} - x^*\|^2 = \|z_{j+1} - x_{j+1}\|^2 + \|x_{j+1} - x^*\|^2 + 2\langle z_{j+1} - x_{j+1}, x_{j+1} - x^*\rangle$$

$\pm x_{j+1}$ & expanding

$$\underbrace{\pi(z_{j+1}) \quad \pi(z_{j+1})}_{\geq 0}$$

$$\geq \|x_{j+1} - x^*\|^2.$$

Moreover, $\|x_j - z_{j+1}\|^2 = \eta^2 \|g_j\|^2 \leq \eta^2 L^2$. We obtain:

$$\frac{1}{k}\sum_{j=1}^{k} f(x_j) - f(x^*) \leq \frac{1}{k}\sum_{j=1}^{k} \frac{1}{2\eta}\left(\eta^2 L^2 + \|x_j - x^*\|^2 - \|x_{j+1} - x^*\|^2\right)$$

$$\leq \frac{\eta L^2}{2} + \frac{1}{2\eta k}\underbrace{\|x_1 - x^*\|^2}_{\leq R^2}$$

Minimizing the RHS yields $\eta = \frac{R}{L\sqrt{k}}$, and the desired upper bound.

The case $\eta_j = \frac{R}{L\sqrt{j}}$ is treated as in the proof of the performance of the standard gradient descent, since we have $\|x_{k/2} - x^*\|^2 \leq R^2$. ■

II.3. Stochastic Gradient Descent.

We turn our attention to the optimization of random functions

$$x \mapsto \ell(x, Z),$$

where x = optimization parameter
Z = RV with distribution $\mathbb{P}_Z$.

We assume that $x \mapsto \ell(x, Z)$ is convex, $\mathbb{P}_Z$-a.s.
(so that in a practical setting, for each observation z_i we make, $x \mapsto \ell(x, z_i)$ is convex).
In particular, $x \mapsto \mathbb{E}\{\ell(x, Z)\}$ is also convex in x.
Our goal is to solve the optimization problem

$$\min_{x \in \mathcal{C}} \mathbb{E}\{\ell(x, Z)\}$$

($\mathcal{C}$: convex set & close) ← put $f(x) := \mathbb{E}\{\ell(x, Z)\}$

Remark: In the learning context, Z correspond to pairs (X, Y), ℓ = loss function, and $\mathbb{E}\{\ell(x, Z)\}$ represents the risk of a predictor parametrized by the variable x.

Ex: square loss + linear classifier:

$$\mathbb{E}\{\ell(x, Z)\} = \mathbb{E}(Y - \beta_0 - \beta^t X)^2$$
$$= R(f); \text{ for } f \in \mathcal{F} = \{x \mid x \mapsto \beta_0 + \beta^t x\}$$

→ Usually, the distribution $\mathbb{P}_Z$ is unknown, and we cannot compute explicitely the objective function $\mathbb{E}\{\ell(x, Z)\}$. Instead, we are given the observations $Z_1, Z_2, \ldots$ and the associated values of $\ell(x, Z_i)$. This means that a gradient descent method cannot be directly applied here, as the direction of the steepest descent cannot be calculated. Stochastic Gradient Descent circumvents this problem by taking a step at each iteration in a random direction, as long as the expected value of this direction corresponds to the negative gradient.

→ Consider two distinct scenarios:

(i) At each iteration of the algorithm, consider a new pair $z_i = (X_i, Y_i)$. Each pair is used at most once. We do **ONE PASS** on the data.

(ii) External Randomization: the SGD algorithm can be applied in the context of ERM, where the objective function $f(x)$ is of the form $\frac{1}{n}\sum_{i=1}^{n} f_i(x)$.

Once observations $\{(x_1, y_1), \dots, (x_n, y_n)\}$ are collected, the functions $f_i(x) = \ell(y_i, g(x_i))$, $g \in \mathcal{F}$, are deterministic. We artificially introduce randomness as follows: let I be a RV uniformly distributed on $\{1, \dots, n\}$. We then have the representation $f(x) = \mathbb{E}_I\{f_I(x)\}$, which falls into the context of stochastic (convex) optimization, with $\mathcal{Z} = I$, and $\ell(x, I) = f_I(x)$.

↑ the sample size.

↳ In the notation of ERM, $g_n \in \operatorname{augmin}_{g \in \mathcal{F}} \hat{R}_n(g)$,

where $\hat{R}_n(g) = \frac{1}{n} \sum_{i=1}^{n} \ell(y_i, g(x_i))$

$= \mathbb{E}_I\{\ell(y_I, g(x_I))\}$

When performing k passes of the algorithm, we may take k as large as possible, regardless of the value of n, by doing **MULTIPLE PASSES** on the data (the performance of the minimizer is however limited by the number of observations).

Summary:

- In (i), observations are received on-line, each $(X_i, Y_i) \sim \mathbb{P}_{X,Y}$, and used only one.
  Objective: minimize the true risk $\mathbb{E}\{\ell(Y, f(X))\}$
- In (ii), we have received n observations. We pick randomly k observations from $\{(X_1, Y_1), \dots, (X_n, Y_n)\}$, with replacement.
  Objective: minimize the empirical risk $\frac{1}{n} \sum_{i=1}^{n} \ell(Y_i, f(X_i))$.

- In the SGD algorithm, each time an observation $\mathcal{Z} = z_j$ is made, a subgradient $\tilde{g}_j$ of $\ell(x, z_j)$ at the current estimate x_j is taken. For any x, we have

↑ And used to update the parameters

$$\ell(x, z_j) - \ell(x_j, z_j) \geq \langle \tilde{g}_j, x - x_j \rangle$$

↳ Taking $\mathbb{E}(\cdot)$ with respect to the distribution of $\mathcal{Z}$, conditionally on x_t, we obtain

$$\mathbb{E}\{\ell(x, \mathcal{Z})\} - \mathbb{E}\{\ell(x_j, \mathcal{Z}) | x_j\}$$

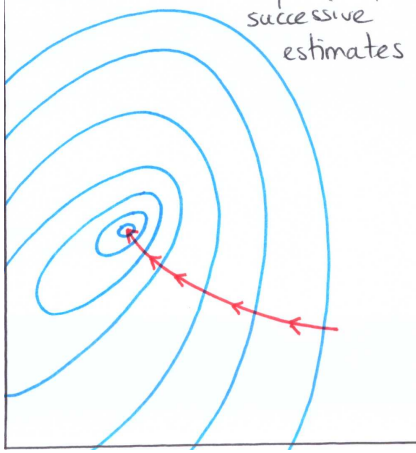
Always hold in scenario (ii) → $\geq \langle \mathbb{E}(\tilde{g}_j | x_j), x - x_j \rangle$,

so that $\mathbb{E}(\tilde{g}_j | x_j) =: g_j$ is a subgradient of $\mathbb{E}\{\ell(x_j, \mathcal{Z}) | x_j\}$

↳ "on average", $\tilde{g}_j$ points towards the right direction

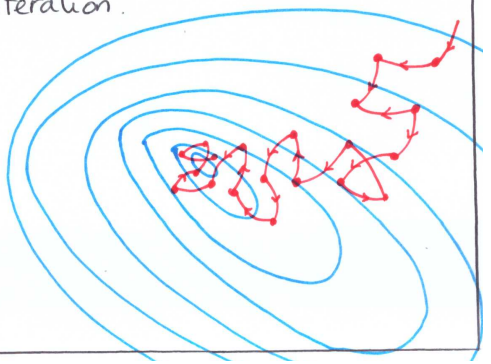
• Gradient Descent

Updates point toward the direction of the largest decrease ⇒ smooth paths of successive estimates



→ Stochastic Gradient Descent

Updates point "on average" towards the direction of the largest decrease. ⇒ Rough path. "Not a descent" algorithm as such: no guarantee to decrease the objective at each iteration.



Algorithm 3: Stochastic Gradient Descent (SGD)

1. Init: $x_1 \in \mathcal{C}$
2. positive sequence $\{\eta_j\}_{j \geq 1}$
3. independent RVs $Z_1, \dots, Z_k$ with distribution $\mathbb{P}_Z$
4. For $j = 1, \dots, k-1$ do
5. $\quad y_{j+1} = x_j - \eta_j \tilde{g}_j$, $\quad \tilde{g}_j \in \partial \ell(x_j, Z_j)$ — a random sequence
6. $\quad x_{j+1} = \pi(y_{j+1})$
7. End For
8. Return $\bar{x}_k = \frac{1}{k} \sum_{j=1}^{k} x_j$

Goal: $\min_{x \in \mathcal{C}} \mathbb{E}\{\ell(x, Z)\}$;
$\ell(\cdot, Z) : \mathbb{R}^d \to \mathbb{R}$

Typically, the objective function $\mathbb{E}\{\ell(x, Z)\}$ is unknown, and x^* cannot be computed.

Theorem: Let. $\mathcal{C}$ = closed convex subset of $\mathbb{R}^d$, diam $\mathcal{C} \leq R$.

- $f(x) = \mathbb{E}\{\ell(x, Z)\}$ attains its minimum on $\mathcal{C}$ at x^*; $f : \mathbb{R}^d \to \mathbb{R}$
- $\ell(x, Z)$ is convex on x, $\mathbb{P}_Z$-a.s.
- $\mathbb{E}\|\tilde{g}\|^2 \leq L^2 \quad \forall \tilde{g} \in \partial \ell(x, Z)$, $\forall x$.

Then, with $\eta_j = \eta = \frac{R}{L\sqrt{k}}$, $\mathbb{E}\{f(\bar{x}_k)\} - f(x^*) \leq \frac{LR}{\sqrt{k}}$

proof = We have $f(x_j) - f(x^*) \leq g_j^t (x_j - x^*)$ — by definition of subgradient of f

$= \mathbb{E}\{\tilde{g}_j^t (x_j - x^*) \mid x_j\}$ — since $\mathbb{E}(\tilde{g}_j \mid x_j) = g_j$

$= \frac{1}{2\eta} \mathbb{E}\{(y_{j+1} - x_j)^t (x_j - x^*) \mid x_j\}$ — $\eta_j = \eta$

$$\Rightarrow f(x_j) - f(x^*) = \frac{1}{2\eta} \Big\{ \mathbb{E}\big[\|x_j - y_{j+1}\|^2 + \|x_j - x^*\|^2 - \|y_{j+1} - x^*\|^2 \mid x_j \big] \Big\}$$

$$\leq \frac{1}{2\eta} \Big\{ \mathbb{E}\big(\eta^2 \|\tilde{g}_j\|^2 \mid x_j\big) + \mathbb{E}\big(\|x_j - x^*\|^2 \mid x_j\big) - \mathbb{E}\big(\|x_{j+1} - x^*\|^2 \mid x_j\big) \Big\}$$

argue as on page 14 for this term

Taking $\mathbb{E}(\dots)$, summing over j, using Jensen's inequality, and taking $\eta = \frac{R}{L\sqrt{k}}$ yields the upper bound. ∎

* **Remark:** On-line algorithms that compute the minimum (or root) of a function $x \mapsto \mathbb{E}\{\ell(x, Z)\}$ date back to Robbins & Monro in the 50's. Such algorithms are known as Stochastic Approximation Algorithms in the statistics litterature. Under mild assumptions on ℓ and on the learning rate η_j ($\eta_j \to 0$, $\sum \eta_j = +\infty$, $\sum \eta_j^2 < \infty$), almost sure convergence and convergence in mean square towards the minimum (or root) can be established, using tools from martingale theory. The two most famous procedures are due to Robbins & Monro (1951) and Kiefer & Wolfowitz (1952). Some basic convergence results of the Robbins & Monro algorithm are presented in the Appendix of this chapter.

III - APPLICATION TO LEARNING PROBLEMS.

In the context of empirical risk minimization, the goal is to minimize the empirical risk $\hat{R}_n(f_\beta) = \frac{1}{n}\sum_{i=1}^n \ell(y_i, f_\beta(x_i))$ with respect to a vector of parameters β. The key difference between GD & SGD algorithms applied in this context is that at each iteration, either we update according to:

$$\beta_{j+1} \leftarrow \beta_j - \gamma_j \left\{ \frac{1}{n}\sum_{i=1}^n \nabla_\beta \ell(y_i, f_\beta(x_i)) \right\} \quad \text{(GD)}$$

↖ slow: one update requires a pass through n observations (smooth trajectories of the estimates $\{\beta_j\}$)

or

$$\beta_{j+1} \leftarrow \beta_j - \gamma_j \left\{ \nabla_\beta \ell(y_i, f_\beta(x_i)) \right\} \quad \text{(SGD)}$$

gradient, or subgradient ↘

↖ much faster: one update requires only one observation. (⇒ rough trajectories) (May not settle at the minimum, even if we reach it: wiggles around). On the other hand, the noisy updates can allow the algorithm to avoid local minima.

↗ Theoretical guarantees for particular choices of $\{\gamma_j\}$

MINI-BATCH SGD is a compromise between GD (all training data used at once) and the SGD (only one observation is used at a time) that splits the learning sample into small batches, used to update the paramaters at each iteration:

$$\beta_{j+1} \leftarrow \beta_j - \gamma_j \left\{ \frac{1}{B}\sum_{i} \nabla_\beta \ell(y_i, f_\beta(x_i)) \right\}$$

↖ sum over all observations belonging to the mini-batch, of size B.

↑ Updates are more frequent than for GD; local minima may be avoided; smoothees up the trajectories of the SGD updates.

(popular in SGD in deep neural nets)

• Example: We implement soft-SVM using SGD. Recall that the optimization problem is

$$\min_{\beta \in \mathbb{R}^d} \frac{1}{n}\sum_{i=1}^n (1 - y_i \langle \beta, x_i \rangle)_+ + \frac{\lambda}{2}\|\beta\|^2$$

↖ hinge loss $\varphi(z) = \max(0, 1+z)$

Consider for simplicity the homogeneous case; by setting the bias term β_0 to 0. Note that a bias term can be introduced by adding one more feature equal to 1 into the input vectors x_i (in which case we regularize the bias term as well).

The hinge loss is not differentiable at $z=-1$, and we need to consider subgradients. Note that the function $\ell(\beta) := (1 - y\langle \beta, x \rangle)_+$ is $\|x\|$-Lipschitz, and that its subgradients are all of the form αx, with $|\alpha| \leq 1$. Specifically, the vector

$$v = \begin{cases} 0 & \text{if } y\langle \beta, x \rangle \geq 1 \\ -yx & \text{if } y\langle \beta, x \rangle < 1 \end{cases} \quad \text{belongs to } \partial\ell(\beta).$$

More generally, if $h(\beta) = \max_{1 \leq i \leq m} h_i(\beta)$ where h_i = convex + different, then $\nabla h_j(\beta) \in \partial h(\beta)$, where $j = \operatorname{argmax}_{1 \leq i \leq m} h_i(\beta)$.

Indeed, h_j convex $\Rightarrow \forall u \quad h_j(u) \geq \underbrace{h_j(\beta)}_{= h(\beta)} + \langle \nabla h_j(\beta), u - \beta \rangle$

$\Rightarrow h(u) \geq h_j(u) \geq h(\beta) + \langle \nabla h_j(\beta), u - \beta \rangle \Rightarrow \nabla h_j(\beta) \in \partial h(\beta)$.

SGD: . Current estimate: β_j.

. Select at random an observation (x_i, y_i) from $\mathcal{L}_n$.

Then $(v_j + \lambda \beta_j)$, where $v_j = \begin{cases} -y_i x_i & \text{if } y_i \langle \beta_j, x_i \rangle < 1, \\ 0 & \text{otherwise} \end{cases}$

is a subgradient of $(1 - y_i \langle \beta_j, x_i \rangle)_+ + \frac{\lambda}{2}\|\beta_j\|^2$ at β_j.

$\Rightarrow \beta_{j+1} \leftarrow \beta_j - \gamma_j (v_j + \lambda \beta_j)$ (with no projection step).

In the case of soft-SVM, and more generally, when a quadratic regularizer is added to the convex empirical risk, we can achieve faster rates of convergence than the ones stated on page 19, since the objective function belongs to the family of **λ-STRONGLY CONVEX** functions.

A function f is said to be λ-strongly convex if its curvature remains bounded away from 0. Specifically, if $f(x) - \frac{\lambda}{2}\|x\|^2$ is convex $\forall x$. There are equivalent definitions of λ-strongly convex functions. Another characterization is

$$\forall x, y \quad \forall u \in \partial f(x), \quad \langle u, x-y \rangle \geq f(x) - f(y) + \frac{\lambda}{2}\|x-y\|^2.$$

↖ If $v \in \partial\left(f(x) - \frac{\lambda}{2}\|x\|^2\right)$, then $v = u + \lambda x$ for some $u \in \partial f(x)$. Then

$$\underbrace{f(y) - \frac{\lambda}{2}\|y\|^2}_{x \mapsto f(x) - \frac{\lambda}{2}\|x\|^2 \text{ is convex}} \geq \underbrace{f(x) - \frac{\lambda}{2}\|x\|^2}_{} + \langle u + \lambda x, y - x \rangle$$

$$\Updownarrow$$

$$\langle u, x-y \rangle \geq f(x) - f(y) + \frac{\lambda}{2}\underbrace{\left(\|y\|^2 - \|x\|^2 - 2\langle x, y-x \rangle\right)}_{= \|x-y\|^2}$$

Algorithm 4: SGD for λ-strongly convex functions.

1. **Init:** $x_1 \in \mathcal{C}$ / indpt RVs $z_1, \dots, z_k \sim \mathbb{P}_z$.
2. For $j = 1, \dots, k-1$ do
3. $\quad y_{j+1} = x_j - \eta_j \tilde{g}_j$, $\eta_j = \frac{1}{\lambda j}$, $\tilde{g}_j \in \partial \ell(x_j, z_j)$
4. $\quad x_{j+1} = \pi(y_{j+1})$
5. End For
6. **Return** $\bar{x}_k = \frac{1}{k}\sum_{j=1}^{k} x_j$

Goal: $\min_{x \in \mathcal{C}} \mathbb{E}\{\ell(x,z)\}$, where $\mathbb{E}\{\ell(x,z)\}$ is λ-str. cnv. in x

Theorem: Let

- $\mathcal{C}$ = closed convex subset of $\mathbb{R}^d$
- $f(x) = \mathbb{E}\{\ell(x,z)\}$ attains its minimum on $\mathcal{C}$ at x^*
- f is λ-strongly convex in x.
- $\mathbb{E}\|\tilde{g}\|^2 \leq L^2 \quad \forall \tilde{g} \in \partial \ell(x,z)$, $\forall x$.

Then, with $\eta_j := \frac{1}{\lambda j}$, $\mathbb{E}\{f(\bar{x}_k)\} - f(x^*) \leq \frac{L^2}{2\lambda k}(1 + \log k)$

The diameter of $\mathcal{C}$ does not appear in the bound.

The "log k" can be removed, see Rakhlin, Shamir, Sridharan (2012). ⇒ Faster convergence in $\frac{1}{k}$ instead of $\frac{1}{\sqrt{k}}$.

proof = Put $g_j := \mathbb{E}(\tilde{g}_j \mid x_j) \in \partial f(x_j)$. Since f is λ-strongly convex, we can write

$$(0) \quad \langle g_j, x_j - x^* \rangle \geq f(x_j) - f(x^*) + \frac{\lambda}{2}\|x_j - x^*\|^2.$$

On the other hand, we have

$$(1) \quad \langle g_j, x_j - x^* \rangle \leq \frac{1}{2\eta_j}\left\{\mathbb{E}\left(\|x_j - x^*\|^2 - \|x_{j+1} - x^*\|^2\right)\right\} + \frac{\eta_j}{2}L^2$$

↖ Indeed, since $x_{j+1} = \pi(y_{j+1})$, we have that $\|y_{j+1} - x^*\|^2 \geq \|x_{j+1} - x^*\|^2$.

$$\Rightarrow \|x_j - x^*\|^2 - \|x_{j+1} - x^*\|^2 \geq \underbrace{\|x_j - x^*\|^2 - \|y_{j+1} - x^*\|^2}_{\| x_j - \eta_j \tilde{g}_j - x^* \|^2}$$

$$= -\eta_j^2\|\tilde{g}_j\|^2 + 2\langle x_j - x^*, \eta_j \tilde{g}_j \rangle$$

Rearranging the terms, taking expectations, and using $\mathbb{E}\|\tilde{g}\|^2 \leq L^2$ gives (1).

Putting (0) & (1) together, and summing over j yields:

$$\sum_{j=1}^{k} \mathbb{E}\{f(x_j)\} - f(x^*)$$

$$\leq \mathbb{E}\left\{ \sum_{j=1}^{k} \frac{\|x_j - x^*\|^2 - \|x_{j+1} - x^*\|^2}{2\eta_j} - \frac{\lambda}{2}\|x_j - x^*\|^2 \right\} + \frac{L^2}{2} \underbrace{\sum_{j=1}^{k} \eta_j}_{\leq \frac{1}{\lambda}(1+\log k)}$$

With $\eta_j = \frac{1}{\lambda_j}$, we get

$$\frac{\lambda}{2} \sum_{j=1}^{k} \{(j-1)\|x_j - x^*\|^2 - j\|x_{j+1} - x^*\|^2\}$$

$$= -\frac{\lambda}{2} k \|x_{k+1} - x^*\|^2 \leq 0$$

$$\Rightarrow \sum_{j=1}^{k} \mathbb{E}\{f(x_j)\} - f(x^*) \leq \frac{L^2}{2\lambda}(1+\log k).$$ ∎

→ Back to soft-SVM. Recall that we wish to optimize

$$\frac{1}{n}\sum_{i=1}^{n}(1 - y_i \langle \beta, x_i \rangle)_+ + \frac{\lambda}{2}\|\beta\|^2.$$

Given the current estimate β_j, and with $\mathcal{C} = \mathbb{R}^d$, we obtain

$$\beta_{j+1} \leftarrow \beta_j - \frac{1}{\lambda_j}(v_j + \lambda \beta_j), \text{ where}$$

$$v_j = \begin{cases} 0 & \text{if } y_i \langle \beta_j, x_i \rangle \geq 1 \\ -y_i x_i & \text{if } \text{—}''\text{—} < 1 \end{cases}$$

associated with the randomly chosen observation (x_i, y_i).

The RHS is equal to

$$\left(1 - \frac{1}{j}\right)\beta_j - \frac{1}{\lambda j} v_j = \frac{j-1}{j}\left\{ \frac{j-2}{j-1}\beta_{j-1} - \frac{1}{\lambda(j-1)} v_{j-1} \right\} - \frac{1}{\lambda j} v_j$$

$$= \ldots = -\frac{1}{\lambda j} \sum_{i=1}^{j} v_i.$$

$\frac{j-1}{j}$

Starting with $\beta_1 \equiv 0$.

We obtain the following simple algorithm:

Goal: $\min_{\beta \in \mathbb{R}^d} \frac{1}{n}\sum_{i=1}^{n}(1 - y_i \langle \beta, x_i \rangle)_+ + \frac{\lambda}{2}\|\beta\|^2$.

Init: $s_1 = 0$

For $j = 1, \ldots, k-1$ do

- $\beta_j = \frac{1}{\lambda_j} s_j$
- Select (x_i, y_i) at random from $\mathcal{L}_n$.
- If $y_i \langle \beta_j, x_i \rangle < 1$

  Put $s_{j+1} \leftarrow s_j + y_i x_i$

  Else

  Put $s_{j+1} \leftarrow s_j$

End For

Return $\bar{\beta} = \frac{1}{k}\sum_{j=1}^{k} \beta_j$.

"Pegasos" Algorithm

See Shalev-Shwartz, Singer, Srebro, Cotter (2011)

SGD for Soft-SVM.

- Example 2 = K-means.

The K-means objective is the minimization of the cost

$$C_0(M, \pi) = \frac{1}{n}\sum_{k=1}^{K}\sum_{i=1}^{n} \pi_{ik} \|x_i - m_k\|_2^2,$$

where $\pi_{ik} = \begin{cases} 1 & \text{if } x_i \in k\text{-th cluster} \\ 0 & \text{otherwise} \end{cases}$

$M = \{m_1, \ldots, m_K\}$ = set of K representatives, summarizing the data $\mathcal{L}_n = \{x_1, \ldots, x_n\}$, $x_i \in \mathbb{R}^d$.

For more details, we refer to the chapter UL: CLUSTERING, page 4. We make the K-means algorithm presented p.7 there

online, by considering a SGD:

At each iteration, pick randomly an observation x_i and update the nearest centroid using the gradient of $\sum_{i=1}^{K} \pi_{ik} \| x_i - m_k^{(j)} \|_2^2$ with respect to the current estimate $m_k^{(j)}$.

$$\nabla_{m_k^{(j)}}(\dots) = -2\pi_{ik}(x_i - m_k^{(j)}).$$

↑ All zero, except for the closest centroid of x_i.

Update: $m_k^{(j+1)} \leftarrow m_k^{(j)} + \eta_j (x_i - m_k^{(j)})$.

↑ learning rate

SGD for K-MEANS.

See MacQueen (1967).

- SGD is a general algorithm that may be applied to a wide variety of optimization problems. For instance, a stochastic version of the coordinate descent algorithm was succesfully applied to the lasso criterion (see Shalev-Shwartz & Tewari (2011)). In the next chapter, we discuss two important classes of model where SGD are particularly useful: the perceptron & neural networks.

References:

- S. Shalev-Shwartz & S. Ben-David (2014). Understanding Machine Learning: From Theory to Algorithms. Cambridge.
- S. Shalev-Shwartz, Y. Singer, N. Srebro, A. Cotter (2011). Pegasos: Primal Estimated Sub-Gradient Solver for SVM. Mathematical Programming, vol 127, no 1, p. 3-30.
- S. Shalev-Shwartz & A. Tewari (2011). Stochastic Methods For ℓ_1-Regularized Loss Minimization. Journal of Machine Learning Research. Vol 12, p. 1865-1892.
- P. Rigollet. Lecture notes on "Mathematics of Machine Learning".

APPENDIX = STOCHASTIC APPROXIMATION ALGORITHMS

I - THE ROBBINS - MONRO ALGORITHM

I.1. Introductive remarks

- Goal: Find the root of a real-valued function $f(\cdot)$.

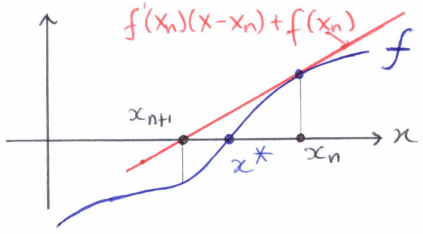
If the function f were known, f assumed differentiable, $f: \mathbb{R} \to \mathbb{R}$, Newton's procedure could be used: the sequence $\boxed{x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)}}$ converges to the root x^* of f.

(f' = derivative of f.)

↳ Idea: approximate f by its tangent line: given a current approximation x_n, the tangent line at x_n is $f'(x_n)(x - x_n) + f(x_n)$.

↑ the x-intercept of this line is used as the next approximation:

$$f'(x_n)(x_{n+1} - x_n) + f(x_n) = 0$$



If f is not differentiable, a less efficient algorithm would be $\boxed{x_{n+1} = x_n - \gamma f(x_n)}$ (*)

↑ step size; $\gamma > 0$

Will converge provided the initial estimate x_0 is not too far away from the root.

→ Assume now that f is not directly observable, but that only noisy measurements of $f(x)$ are available $\forall x$.

We observe: $f(x) + U$ = noisy measurement

↑ noise term, with zero mean $\mathbb{E}U = 0$

One can use the procedure (*), with $f(x_n)$ replaced by a good estimate of its value, obtained by averaging many observations: $f(x_n) \approx \frac{1}{m} \sum_{i=1}^{m} (f(x_n) + U_i)$

call this term $F(x_i, U_i)$, so that $\frac{1}{m} \sum_{i=1}^{n} F(x_i, U_i) \xrightarrow{a.s.} f(x_i)$, under the assumption that $\mathbb{E}U = 0$.

$$\Rightarrow \boxed{x_{n+1} = x_n - \frac{\gamma}{m} \sum_{i=1}^{m} F(x_i, U_i)}$$

↑ Robbins & Monro argued that taking an excessive number of observations at each iteration is inefficient, since x_n is used only as an intermediate step in the approximation of the root. They suggested to consider instead:

$$\boxed{x_{n+1} = x_n - \gamma_n F(x_n, U_{n+1})}$$

↑ The sequence $\gamma_n > 0$ converges to zero as $n \to \infty$, and is such that $\sum \gamma_n = \infty$. The decreasing step size actually provides an averaging of the observations, required to proved that the resulting sequence x_n converges to the root x^* of f.

Example:
Let $U_1, \dots, U_{n+1}$ = sequence of iid RVs with mean $\mathbb{E}U$.

$$\bar{X}_{n+1} := \frac{1}{n+1} \sum_{i=1}^{n+1} U_i$$

$$= \frac{n}{n+1} \frac{1}{n} \sum_{i=1}^{n} U_i + \frac{1}{n+1} U_{n+1}$$

$$\bar{X}_{n+1} = \left(\frac{n}{n+1}\right)\bar{X}_n + \frac{1}{n+1}U_{n+1}$$

$$= \bar{X}_n - \frac{1}{n+1}(\bar{X}_n - U_{n+1}) = \bar{X}_n - \gamma_n F(\bar{X}_n, U_{n+1}),$$

with $\cdot\ \gamma_n = \frac{1}{n+1}$ (note that $\sum \gamma_n = +\infty$; while $\sum \gamma_n^2 < +\infty$)

$\cdot\ F(x,U) = x - U$

$= (x - \mathbb{E}U) + (\mathbb{E}U - U)$ — $\forall x$, you do not observe $x - \mathbb{E}U$, but only $x - U$.

$= f(x) + \underbrace{(\mathbb{E}U - U)}_{\text{noise term}}$

$\cdot\ f(x) = x - \mathbb{E}U$

And indeed, $\bar{X}_{n+1} \xrightarrow{a.s.} \mathbb{E}U$ (SLLN) ; corresponding to the root of f.

⇒ We are interested in the convergence of the sequence of estimates $X_{n+1} = X_n - \gamma_n F(X_n, U_{n+1})$ (sequence of RVs).

Put $f(x) = \mathbb{E}[F(x,U)]$. We show that under some conditions on F and f, the sequence converges (in some sense) to the unique root x^* of f.

I.2. The Robbins-Monro algorithm.

We first establish convergence under restrictive assumptions on f and F. The conditions are later weakened.

Theorem. Let $\cdot\ F : \mathbb{R}^2 \to \mathbb{R}$ measurable

$\cdot\ U$: a RV with distribution G.

Consider the sequence of RVs $X_0, X_1, \dots$, satisfying

↘ $X_0 \in \mathcal{L}^2$

↘ $X_{n+1} = X_n - \gamma_n F(X_n, U_{n+1})$, where $\{U_n\}$ = iid with distribution G.

Under the assumptions that

[H1] $\exists K > 0 \quad \forall x \in \mathbb{R}, \forall u \in \mathbb{R} \quad |F(x,u)| \leq K$

[H2] f = continuous, strictly increasing. Moreover, there exists a unique $x^* \in \mathbb{R}$ such that $f(x^*) = 0$

[H3] $\forall n\ \gamma_n > 0$; $\sum \gamma_n = +\infty$; $\sum \gamma_n^2 < +\infty$

Then $X_n \xrightarrow{a.s.} x^*$

proof = • Step 1: Put $\mathcal{F}_n = \sigma(X_0, X_1, \dots, X_n)$

$$S_n = \sum_{k=0}^{n-1} \gamma_k^2\, \mathbb{E}\{F^2(X_k, U_{k+1}) \mid \mathcal{F}_k\}$$

We show that the sequence $M_n = (X_n - x^*)^2 - S_n$ is an $\mathcal{F}_n$-supermartingale.

$$\mathbb{E}(M_{n+1} \mid \mathcal{F}_n) = \mathbb{E}\{(X_{n+1} - x^*)^2 \mid \mathcal{F}_n\} - \mathbb{E}\{S_{n+1} \mid \mathcal{F}_n\}$$

$$= \mathbb{E}\{(X_n - \gamma_n F(X_n, U_{n+1}) - x^*)^2 \mid \mathcal{F}_n\} - \mathbb{E}\{S_{n+1} \mid \mathcal{F}_n\}$$

$$= (X_n - x^*)^2 - 2(X_n - x^*)\gamma_n \mathbb{E}\{F(X_n, U_{n+1}) \mid \mathcal{F}_n\} + \gamma_n^2 \mathbb{E}\{F^2(X_n, U_{n+1}) \mid \mathcal{F}_n\}$$

On the other hand,

$$\mathbb{E}\{S_{n+1} \mid \mathcal{F}_n\} = \mathbb{E}\left\{\sum_{k=0}^{n} \gamma_k^2\, \mathbb{E}\{F^2(X_k, U_{k+1}) \mid \mathcal{F}_k\} \,\middle|\, \mathcal{F}_n\right\}$$

$$= S_n + \gamma_n^2\, \mathbb{E}\{F^2(X_n, U_{n+1}) \mid \mathcal{F}_n\}$$

Thus

$$\mathbb{E}\{M_{n+1} \mid \mathcal{F}_n\} = (X_n - x^*)^2 - 2(X_n - x^*)\gamma_n \mathbb{E}\{F(X_n, U_{n+1}) \mid \mathcal{F}_n\} - S_n$$

$$= M_n - 2(X_n - x^*)\gamma_n \mathbb{E}\{F(X_n, U_{n+1}) \mid \mathcal{F}_n\}$$

In addition, $(X_n - x^*)\,\mathbb{E}\{F(X_n, U_{n+1}) \mid \mathcal{F}_n\}$

$= (X_n - x^*)\,\mathbb{E}\{F(X_n, U_{n+1}) - F(x^*, U_{n+1}) \mid \mathcal{F}_n\}$

$= (X_n - x^*)\,(f(X_n) - f(x^*))$

↑ same sign, since f is strictly positive.

Thus $\mathbb{E}(M_{n+1} \mid \mathcal{F}_n) \leq M_n$ i.e. $M_n = \overline{\text{SMG}}$.

- **Step 2**: The sequence $\{M_n\}$ converges almost surely and in mean to some $M_\infty \in \mathcal{L}^1$.

$M_n = (X_n - x^*)^2 - S_n = M_n^+ - M_n^-$,

where $M_n^+ = \max(0, M_n)$

$M_n^- = -\min(0, M_n)$

Since $(X_n - x^*)^2$ and S_n are positive, we conclude that $M_n^+ \leq (X_n - x^*)^2$ and $M_n^- \leq S_n$

Thus $\mathbb{E}\,M_n^- \leq \mathbb{E}\,S_n \leq K \sum_{k=0}^{+\infty} \gamma_k^2 < +\infty$ under [H3]

Doob's first martingale convergence theorem implies that $M_n \xrightarrow{a.s} M_\infty \in \mathcal{L}^1$

Moreover, $M_n = \overline{\text{SMG}} \Rightarrow \mathbb{E}\,M_{n+1} \leq \mathbb{E}\,M_n$, and

$\mathbb{E}\,M_n \geq -K \sum_{k \geq 0} \gamma_k^2$

$\Rightarrow$ sequence $\{\mathbb{E}\,M_n\}$ is non-increasing & bounded below $\Rightarrow$ it converges to $\mathbb{E}\,M_\infty$.

- **Step 3**: $\sum_{k \geq 0} \gamma_k\,\mathbb{E}\{f(X_k)(X_k - x^*)\} < +\infty$

Indeed,

$\mathbb{E}\{f(X_n)(X_n - x^*)\} = \mathbb{E}\left[\mathbb{E}\{(X_n - x^*)F(X_n, U_{n+1}) \mid \mathcal{F}_n\}\right]$

↑ definition of f

and $\gamma_n\,\mathbb{E}\{(X_n - x^*)F(X_n, U_{n+1}) \mid \mathcal{F}_{n+1}\}$

$= \frac{1}{2}(M_n - \mathbb{E}\{M_{n+1} \mid \mathcal{F}_n\})$

$\Rightarrow \gamma_n\,\mathbb{E}\{f(X_n)(X_n - x^*)\} = \frac{1}{2}(\mathbb{E}\,M_n - \mathbb{E}\,M_{n+1})$.

$\Rightarrow \sum_{k=0}^{n} \gamma_k\,\mathbb{E}\{f(X_k)(X_k - x^*)\} = \frac{1}{2}(\mathbb{E}\,M_0 - \mathbb{E}\,M_{n+1})$

$\longrightarrow \frac{1}{2}(\mathbb{E}\,M_0 - \mathbb{E}\,M_\infty)$.

from Step 2.

- **Step 4**: Conclusion.

First, note that since $M_n \xrightarrow{a.s} M_\infty$ and

$0 \leq S_n = \sum_{k=0}^{n-1} \gamma_k^2\,\mathbb{E}\{F^2(X_k, U_{k+1}) \mid \mathcal{F}_k\} \leq S_{n+1} \leq K^2 \underbrace{\sum_{k \geq 0} \gamma_k^2}_{< +\infty}$

$\Rightarrow S_n$ converges.

$\Rightarrow (X_n - x^*)^2$ converges almost surely to some RV L.

It remains to show that $L \equiv 0$.

Suppose $\exists\, \omega \in \Omega$ such that $L(\omega) > 0$.

$\Rightarrow$ there exists $m_1 > 0$ and n_0 such that $\forall n \geq n_0$,

$|X_n - x^*|(\omega) \geq m_1$

Since f is continuous, and x^* such that $f(x^*) = 0$, there exists $m_2 > 0$ such that $\forall n \geq n_0$,

$f(X_n)(X_n - x^*) \geq m_2$, which implies that

$\sum_{k \geq 0} \gamma_k\,\mathbb{E}\,f(X_k)(X_k - x^*) \geq m_2 \underbrace{\sum_{k \geq 0} \gamma_k}_{= +\infty \text{ [H3]}}$,

which contradicts what we established in Step 3.

We conclude that $L \equiv 0$ and that $X_n \xrightarrow{a.s} x^*$ ∎

Remarks: (i) Assumption [H1] can be replaced with the weaker assumption that

> [H1'] $\exists K>0 \quad \forall x \in \mathbb{R}^d \quad \mathbb{E}\{\|F(x,U)\|^2\} \leq K(1+\|x\|^2)$
>
> for $F: \mathbb{R}^d \times \mathbb{R} \to \mathbb{R}$

Almost sure convergence still holds, replacing [H1] with [H1']

↑ The idea is to show that the sequence

$$M_n = B_{n-1}\|X_n - x^*\|^2 - \sum_{k=0}^{n-1} B_k \gamma_k (K\gamma_k - 2f(X_k)(X_k - x^*))$$

with $B_n := \prod_{i=1}^{n-1} \frac{1}{1+K\gamma_i^2}$, is a supermartingale, banded in $\mathcal{L}^1$, and then proceed as before.

(ii) Assumption [H2] can be weakened either to

> [H2'] $\exists x^* \in \mathbb{R}^d$ & $c>0$ s.t. $\forall x \in \mathbb{R}^d$
> $f(x).(x-x^*) \geq c\|x-x^*\|^2$

or

> [H2''] $\exists x^* \in \mathbb{R}^d$ s.t. $f(x^*)=0$ and $\forall x \neq x^*$,
> $f(x).(x-x^*) > 0$.

Then it is possible to show that under
[H1'] + [H2'] + [H3], $X_n \xrightarrow{\mathcal{L}^2} x^*$, and
[H1'] + [H2''] + [H3], $X_n \xrightarrow{a.s.} x^*$.

Note that [H2''] is weaker than [H2'], and that these two assumptions imply that x^* is the unique root of the equation $f(x)=0$.

(iii) Assumption $\sum_{k\geq 0} \gamma_k^2 < \infty$ in [H3] can also be weakened considerably.

I.3. Applications.

(a): Quantile estimation.

Take . $F(x,u) = \mathbb{1}(u \leq x) - \alpha$, for $\alpha \in (0,1)$.
. G continuous (so that the quantile is well defined).

Then $f(x) = \mathbb{E}[F(x,U)]$
$= \mathbb{P}(U \leq x) - \alpha = G(x) - \alpha$.

The root of this function is precisely q_α: α-quantile of G; $G(q_\alpha) = \alpha$.

⇒ Use the recursion $X_{n+1} = X_n - \gamma_n(\mathbb{1}(U_{n+1} \leq X_n) - \alpha)$.

(b): Maximum likelihood estimation.

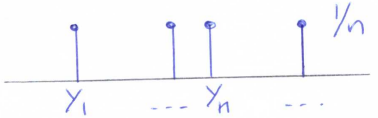
Let $Y_1, \dots, Y_n$ = iid observation with distribution P_θ, and density g_θ (twice differentiable)

The MLE is typically solution of an equation

$$\frac{1}{n}\sum_{i=1}^n F(\theta, Y_i) = 0,$$

where $F(\theta, y) := g'_\theta(y)/g_\theta(y)$.

Take * $\mathbb{P}_n = \frac{1}{n}\sum_{i=1}^n \delta_{Y_i}$ = empirical probability of obs $Y_1, \dots, Y_n$



* $f(\theta) = \mathbb{E}_{\mathbb{P}_n} F(\theta, U)$, where $U \sim \mathbb{P}_n$.

⇒ The sequence $\theta_{n+1} = \theta_n - \gamma_n F(\theta_n, U_{n+1}) \xrightarrow[n\to\infty]{a.s.}$ MLE $\hat{\theta}_{MLE}$

↑ drawn uniformly in $\{Y_1, \dots, Y_n\}$ at each iteration