Machine Learning 8: Convex Optimization for Machine Learning

Master 2 Computer Science

Aurélien Garivier

2018-2019

Table of contents

Convex functions in $\mathbb{R}^d$

Convex functions and subgradients

Convex function

Let $\mathcal{X} \subset \mathbb{R}^d$ be a convex set. The function $f : \mathcal{X} \to \mathbb{R}$ is *convex* if

$$\forall x, y \in X, \forall \lambda \in [0,1],\ f\big((1-\lambda)x + \lambda y\big) \leq (1-\lambda) f(x) + \lambda f(y)\ .$$

Subgradients

A vector $g \in \mathbb{R}^n$ is a *subgradient* of f at $x \in \mathcal{X}$ if for any $y \in \mathcal{X}$,

$$f(y) \geq f(x) + \langle g, y - x \rangle\ .$$

The set of subgradients of f at x is denoted $\partial f(x)$.

Proposition

- If $\partial f(x) \neq \emptyset$ for all $x \in \mathcal{X}$, then f is convex.
- If f is convex, then $\forall x \in \mathring{\mathcal{X}}, \partial f(x) \neq \emptyset$.
- If f is convex and differentiable at x, then $\partial f(x) = \big\{\nabla f(x)\big\}$.

Convex functions and optimization

Proposition

Let f be convex. Then

- x is local minimum of f iff $0 \in \partial f(x)$,
- and in that case, x is a *global* minimum of f;
- if $\mathcal{X}$ is closed and if f is differentiable on $\mathcal{X}$, then

$$x = \arg\min_{x \in \mathcal{X}} f(x) \quad \text{iff} \quad \forall y \in \mathcal{X}, \langle \nabla f(x), y - x \rangle \geq 0 \,.$$

Black-box optimization model

The set $\mathcal{X}$ is known, $f : \mathcal{X} \to \mathbb{R}$ is unknown but accessible thru:

- a zeroth-order oracle: given $x \in \mathcal{X}$, yields $f(x)$,
- and possibly a first-order oracle: given $x \in \mathcal{X}$, yields $g \in \partial f(x)$.

Gradient Descent

Gradient Descent algorithms

A memoryless algorithm for first-order black-box optimization:

Algorithm: Gradient Descent

Input: convex function f, step size γ_t, initial point x_0

for $t = 0 \dots T-1$ **do**
- Compute $g_t \in \partial f(x_t)$
- $x_{t+1} \leftarrow x_t - \gamma_t \, g_t$

return x_T or $\dfrac{x_0 + \cdots + x_{T-1}}{T}$

Questions:

- $x_T \underset{T\to\infty}{\to} x^* \overset{def}{=} \arg\min f$?
- $f(x_T) \underset{T\to\infty}{\to} f(x^*) = \min f$?
- under which conditions?
- what about $\dfrac{x_0 + \cdots + x_{T-1}}{T}$?
- at what speed?
- works in high dimension?
- do some properties help?
- can other algorithms do better?

Monotonicity of gradient

Property

Let f be a convex function on $\mathcal{X}$, and let $x, y \in \mathcal{X}$. For every $g_x \in \partial f(x)$ and every $g_y \in \partial f(y)$,

$$\langle g_x - g_y, x - y \rangle \geq 0 \; .$$

In fact, a differentiable mapping f is convex iff

$$\forall x, y \in \mathcal{X}, \langle \nabla f(x) - \nabla f(y), x - y \rangle \geq 0 \; .$$

In particular, $\langle g_x, x - x^* \rangle \geq 0$.

$\implies$ the negative gradient does not point the the wrong direction.

Under some assumptions (to come), this inequality can be strenghtened, making gradient descent more relevant.

Convergence of GD for Convex-Lipschitz functions

Lipschitz Assumption

For every $x \in \mathcal{X}$ and every $g \in \partial f(x)$, $\|g\| \le L$.

This implies $\big|f(y) - f(x)\big| \le \big|\langle g, y - x\rangle\big| \le L\|y - x\|$.

Theorem

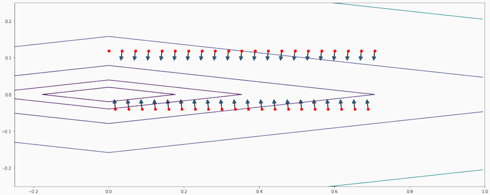
Under the Lipschitz assumption, GD with $\gamma_t \equiv \gamma = \frac{R}{L\sqrt{T}}$ satisfies

$$f\left(\frac{1}{T}\sum_{i=0}^{T-1} x_i\right) - f(x^*) \le \frac{RL}{\sqrt{T}} \, .$$

- Of course, can return $\arg\min_{1 \le i \le T} f(x_i)$ instead (not always better).
- It requires $T_\epsilon \approx \frac{R^2 L^2}{\epsilon^2}$ steps to ensure precision ϵ.
- Online version $\gamma_t = \frac{R}{L\sqrt{t}}$: bound in $3RL/\sqrt{T}$ (see Hazan).
- Works just as well for constrained optimization with $x_{t+1} \leftarrow \Pi_{\mathcal{X}}\big(x_t - \gamma_t \nabla f(x_t)\big)$ thanks to Pythagore projection theorem.

Intuition: what can happen?

The step must be large enough to reach the region of the minimum, but not too large too avoid skipping over it.



Let $\mathcal{X} = \mathcal{B}(0, R) \subset \mathbb{R}^2$ and

$$f(x^1, x^2) = \frac{R}{\sqrt{2}\gamma T}|x^1| + L|x^2|$$

which is L-Lipschitz for $\gamma \geq \frac{R}{\sqrt{2}LT}$. Then, if $x_0 = \left(\frac{R}{\sqrt{2}}, \frac{3L\gamma}{4}\right) \in \mathcal{X}$,

- $x_t^1 = \frac{R}{\sqrt{2}} - \frac{Rt}{\sqrt{2}T}$ and $\bar{x}_T^1 \gtrapprox \frac{R}{2\sqrt{2}}$;
- $x_{2s+1}^2 = \frac{3L\gamma}{4} - \gamma L = -\frac{L\gamma}{4}$, $x_{2s}^2 = \frac{3L\gamma}{4}$, and $\bar{x}_T^2 \gtrapprox \frac{L\gamma}{4}$.

Hence

$$f\left(\bar{x}_T^1, \bar{x}_T^2\right) \gtrapprox \frac{R}{\sqrt{2}\gamma T}\frac{R}{2\sqrt{2}} + L\frac{\gamma L}{4} = \frac{1}{4}\left(\frac{R^2}{T\gamma} + L^2\gamma\right) ,$$

which is minimal for $\gamma = \frac{R}{L\sqrt{T}}$ where $f\left(\bar{x}_T^1, \bar{x}_T^2\right) \approx \frac{RL}{2\sqrt{T}}$.

Proof

Cosinus theorem = generalized Pythagore theorem = Alkashi's theorem:

$$2\langle a, b\rangle = \|a\|^2 + \|b\|^2 - \|a - b\|^2 .$$

Hence for every $0 \le t < T$:

$$\begin{aligned}
f(x_t) - f(x^*) &\le \langle g_t, x_t - x^*\rangle \\
&= \frac{1}{\gamma}\langle x_t - x_{t+1}, x_t - x^*\rangle \\
&= \frac{1}{2\gamma}\Big(\|x_t - x^*\|^2 + \|x_t - x_{t+1}\|^2 - \|x_{t+1} - x^*\|^2\Big) \\
&= \frac{1}{2\gamma}\Big(\|x_t - x^*\|^2 - \|x_{t+1} - x^*\|^2\Big) + \frac{\gamma}{2}\|g_t\|^2 ,
\end{aligned}$$

and hence

$$\sum_{t=0}^{T-1} f(x_t) - f(x^*) \le \frac{1}{2\gamma}\Big(\|x_0 - x^*\|^2 - \|x_T - x^*\|^2\Big) + \frac{L^2\gamma T}{2} \le \frac{L\sqrt{T}\,R^2}{2R} + \frac{L^2\,RT}{2L\sqrt{T}}$$

and by convexity $\quad f\left(\frac{1}{T}\sum_{i=0}^{T-1} x_i\right) \le \frac{1}{T}\sum_{t=0}^{T-1} f(x_t).$

Smoothness

Smoothness

Definition

A continuously differentiable function f is β-smooth if the gradient ∇f is β-Lipschitz, that is if for all $x, y \in \mathcal{X}$,

$$\|\nabla f(y) - \nabla f(x)\| \leq \beta \|y - x\| \,.$$

Property

If f is β-smooth, then for any $x, y \in \mathcal{X}$:

$$\left| f(y) - f(x) - \langle \nabla f(x), y - x \rangle \right| \leq \frac{\beta}{2} \|y - x\|^2 \,.$$

- f is convex and β-smooth iff $x \mapsto \frac{\beta}{2}\|x\|^2 - f(x)$ is convex iff $\forall x, y \in \mathcal{X}$

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

- If f is twice differentiable, then f is α-strongly convex iff all the eigenvalues of the Hessian of f are at most equal to β.

Convergence of GD for smooth convex functions

Theorem

Let f be a convex and β-smooth function on $\mathbb{R}^d$. Then GD with $\gamma_t \equiv \gamma = \frac{1}{\beta}$ satisfies:

$$f(x_T) - f(x^*) \leq \frac{2\beta\|x_0 - x^*\|^2}{T+4} \; .$$

Thus it requires $T_\epsilon \approx \frac{2\beta R}{\epsilon}$ steps to ensure precision ϵ.

Majoration/minoration

Taking $\gamma = \frac{1}{\beta}$ is a "safe" choice ensuring progress:

$$x^+ \stackrel{def}{=} x - \frac{1}{\beta}\nabla f(x) = \arg\min_y f(x) + \left\langle \nabla f(x), y - x \right\rangle + \frac{\beta}{2}\|y - x\|^2$$

is such that $f(x^+) \leq f(x) - \frac{1}{2\beta}\left\|\nabla f(x)\right\|^2$. Indeed,

$$\begin{aligned} f(x^+) - f(x) &\leq \left\langle \nabla f(x), x^+ - x \right\rangle + \frac{\beta}{2}\|x^+ - x\|^2 \\ &= -\frac{1}{\beta}\left\|\nabla f(x)\right\|^2 + \frac{1}{2\beta}\left\|\nabla f(x)\right\|^2 = -\frac{1}{2\beta}\left\|\nabla f(x)\right\|^2 . \end{aligned}$$

$\implies$ *Descent* method.

Moreover, x^+ is "on the same side of x^* as x" (no overshooting): since $\left\|\nabla f(x)\right\| = \left\|\nabla f(x) - \nabla f(x^*)\right\| \leq \beta\|x - x^*\|$,
$\left\langle \nabla f(x), x - x^* \right\rangle \leq \left\|\nabla f(x)\right\|\|x - x^*\| \leq \beta\|x - x^*\|^2$ and thus

$$\left\langle x^* - x^+, x^* - x \right\rangle = \left\|x^* - x\right\|^2 + \left\langle \frac{1}{\beta}\nabla f(x), x^* - x \right\rangle \geq 0 .$$

Lemma: the gradient shoots in the right direction

Lemma

For every $x \in \mathcal{X}$,

$$\left\langle \nabla f(x), x - x^* \right\rangle \geq \frac{1}{\beta} \left\| \nabla f(x) \right\|^2 .$$

We already now that $f(x^*) \leq f\left(x - \frac{1}{\beta}\nabla f(x)\right) \leq f(x) - \frac{1}{2\beta}\left\|\nabla f(x)\right\|^2$. In addition, taking $z = x^* + \frac{1}{\beta}\nabla f(x)$:

$$\begin{aligned} f(x^*) &= f(z) + f(x^*) - f(z) \\ &\geq f(x) + \left\langle \nabla f(x), z - x \right\rangle - \frac{\beta}{2}\|z - x^*\|^2 \\ &= f(x) + \left\langle \nabla f(x), x^* - x \right\rangle + \left\langle \nabla f(x), z - x^* \right\rangle - \frac{1}{2\beta}\left\|\nabla f(x)\right\|^2 \\ &= f(x) + \left\langle \nabla f(x), x^* - x \right\rangle + \frac{1}{2\beta}\left\|\nabla f(x)\right\|^2 . \end{aligned}$$

Thus $f(x) + \left\langle \nabla f(x), x^* - x \right\rangle + \dfrac{1}{2\beta}\left\|\nabla f(x)\right\|^2 \leq f(x^*) \leq f(x) - \dfrac{1}{2\beta}\left\|\nabla f(x)\right\|^2$.

In fact, this lemma is a corrolary of the *co-coercivity of the gradient*: $\forall x, y \in \mathcal{X}$,

$$\left\langle \nabla f(x) - \nabla f(y), x - y \right\rangle \geq \frac{1}{\beta}\left\|\nabla f(x) - \nabla f(y)\right\|^2 ,$$

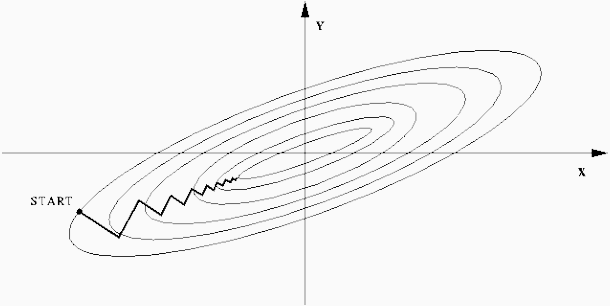
which holds iff the convex, differentiable function f is β-smooth.

Proof step 1: the iterates get closer to x^*

Applying the preceeding lemma to $x = x_t$, we get

$$\begin{aligned}
\|x_{t+1} - x^*\|^2 &= \left\|x_t - \frac{1}{\beta}\nabla f(x_t) - x^*\right\|^2 \\
&= \left\|x_t - x^*\right\|^2 - \frac{2}{\beta}\langle \nabla f(x_t), x_t - x^*\rangle + \frac{1}{\beta^2}\left\|\nabla f(x_t)\right\|^2 \\
&\leq \left\|x_t - x^*\right\|^2 - \frac{1}{\beta^2}\left\|\nabla f(x_t)\right\|^2 \\
&\leq \left\|x_t - x^*\right\|^2 \ .
\end{aligned}$$

$\rightarrow$ it's good, but it can be slow...



Proof step 2: the values of the iterates converge

We have seen that $f(x_{t+1}) - f(x_t) \leq -\frac{1}{2\beta}\left\|\nabla f(x_t)\right\|^2$. Hence, if $\delta_t = f(x_t) - f(x^*)$, then

$$\delta_0 = f(x_0) - f(x^*) \leq \frac{\beta}{2}\left\|x_0 - x^*\right\|^2$$

and $\delta_{t+1} \leq \delta_t - \frac{1}{2\beta}\left\|\nabla f(x_t)\right\|^2$. But

$$\delta_t \leq \left\langle \nabla f(x_t), x_t - x^* \right\rangle \leq \left\|\nabla f(x_t)\right\| \left\|x_t - x^*\right\| \ .$$

Therefore, since δ_t is decreasing with t, $\delta_{t+1} \leq \delta_t - \frac{\delta_t^2}{2\beta\|x_0 - x^*\|^2}$. Thinking to the coresponding ODE, one sets $u_t = 1/\delta_t$, which yields:

$$u_{t+1} \geq \frac{u_t}{1 - \frac{1}{2\beta\|x_0 - x^*\|^2 u_t}} \geq u_t\left(1 + \frac{1}{2\beta\|x_0 - x^*\|^2\, u_t}\right) = u_t + \frac{1}{2\beta\|x_0 - x^*\|^2}$$

Hence, $u_T \geq u_0 + \frac{T}{2\beta\|x_0 - x^*\|^2} \geq \frac{2}{\beta\|x_0 - x^*\|^2} + \frac{T}{2\beta\|x_0 - x^*\|^2} = \frac{T+4}{2\beta\|x_0 - x^*\|^2}$.

Strong convexity

Strong convexity

Definition

$f : \mathcal{X} \to \mathbb{R}$ is α-strongly convex if for all $x, y \in \mathcal{X}$, for any $g_x \in \partial f(x)$,

$$f(y) \geq f(x) + \langle g_x, y - x \rangle + \frac{\alpha}{2} \|y - x\|^2 .$$

- f is α-strongly convex iff $f(x) - \frac{\alpha}{2}\|x\|^2$ is convex.
- α measures the *curvature* of f.
- If f is twice differentiable, then f is α-strongly convex iff all the eigenvalues of the Hessian of f are larger than α.

Faster rates for Lipschitz functions through strong convexity

Theorem

Let f be a α-strongly convex and L-Lipschitz. Under the Lipschitz assumption, GD with $\gamma_t = \frac{1}{\alpha(t+1)}$ satisfies:

$$f\left(\frac{1}{T}\sum_{i=0}^{T-1} x_i\right) - f(x^*) \leq \frac{L^2 \log(T)}{\alpha T} .$$

Note : returning another weighted average with $\gamma_t = \frac{2}{\alpha(t+1)}$ yields:

$$f\left(\sum_{i=0}^{T-1} \frac{2(i+1)}{T(T+1)} x_i\right) - f(x^*) \leq \frac{2L^2}{\alpha(T+1)} .$$

Thus it requires $T_\epsilon \approx \frac{2L^2}{\alpha\epsilon}$ steps to ensure precision ϵ.

Proof

Cosinus theorem = generalized Pythagore theorem = Alkashi's theorem:

$$2\langle a, b\rangle = \|a\|^2 + \|b\|^2 - \|a - b\|^2 \,.$$

Hence for every $0 \le t < T$, by α-strong convexity:

$$\begin{aligned}
f(x_t) - f(x^*) &\le \langle g_t, x_t - x^*\rangle - \frac{\alpha}{2}\|x_t - x^*\|^2 \\
&= \frac{1}{\gamma_t}\langle x_t - x_{t+1}, x_t - x^*\rangle - \frac{\alpha}{2}\|x_t - x^*\|^2 \\
&= \frac{1}{2\gamma_t}\left(\|x_t - x^*\|^2 + \|x_t - x_{t+1}\|^2 - \|x_{t+1} - x^*\|^2\right) - \frac{\alpha}{2}\|x_t - x^*\|^2 \\
&= \frac{t\alpha}{2}\|x_t - x^*\|^2 - \frac{(t+1)\alpha}{2}\|x_{t+1} - x^*\|^2 + \frac{1}{2(t+1)\alpha}\|g_t\|^2
\end{aligned}$$

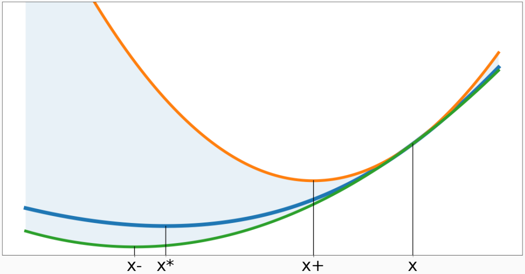
since $\gamma_t = \frac{1}{\alpha(t+1)}$, and hence

$$\sum_{t=0}^{T-1} f(x_t) - f(x^*) \le \frac{0 \times \alpha}{2}\|x_0 - x^*\|^2 - \frac{T\alpha}{2}\|x_T - x^*\|^2 + \frac{L^2}{2\alpha}\sum_{t=0}^{T-1}\frac{1}{t+1} \le \frac{L^2 \log(T)}{2\alpha}$$

and by convexity $f\left(\frac{1}{T}\sum_{i=0}^{T-1} x_i\right) \le \frac{1}{T}\sum_{t=0}^{T-1} f(x_t)$.

Outline

Smoothness and strong convexity: sandwiching f by squares



Let $x \in \mathcal{X}$.

For every $y \in \mathcal{X}$,

β-smoothness implies:

$$f(y) \leq f(x) + \langle \nabla f(x), y - x \rangle + \frac{\beta}{2}\|y - x\|^2$$

$$\overset{def}{=} \bar{f}(x) = \bar{f}(x^+) + \frac{\beta}{2}\left\|y - x^+\right\|^2 .$$

Moreoever, α-strong convexity implies, with $x^- = x - \frac{1}{\alpha}\nabla f(x)$,

$$f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle + \frac{\alpha}{2}\|y - x\|^2$$

$$\overset{def}{=} \underline{f}(x) = \underline{f}(x^-) + \frac{\alpha}{2}\left\|y - x^-\right\|^2 .$$

Convergence of GD for smooth and strongly convex functions

Theorem

Let f be a β-smooth and α-strongly convex function. Then GD with the choice $\gamma_t \equiv \gamma = \frac{1}{\beta}$ satisfies

$$f(x_T) - f(x^*) \leq e^{-\frac{T}{\kappa}} \left(f(x_0) - f(x^*) \right) ,$$

where $\kappa = \frac{\beta}{\alpha} \geq 1$ is the *condition number* of f.

Linear convergence: it requires $T_\epsilon = \kappa \log \left(\frac{\mathrm{osc}(f)}{\epsilon} \right)$ steps to ensure precision ϵ.

Proof: every step fills a constant part of the gap

In particular, with the choice $\gamma = \frac{1}{\beta}$,

$$f(x_{t+1}) = f(x_t^+) \leq \bar{f}(x_t^+) = f(x_t) - \frac{1}{2\beta}\left\|\nabla f(x_t)\right\|^2 ,$$

and

$$f(x^*) \geq \underline{f}(x^*) \geq \underline{f}(x_t^-) = f(x_t) - \frac{1}{2\alpha}\left\|\nabla f(x_t)\right\|^2 .$$

Hence, every step fills at least a part $\frac{\alpha}{\beta}$ of the gap:

$$f(x_t) - f(x_{t+1}) \geq \frac{\alpha}{\beta}\big(f(x_t) - f(x^*)\big) .$$

It follows that

$$\begin{aligned} f(x_T) - f(x^*) &\leq \left(1 - \frac{\alpha}{\beta}\right)\big(f(x_{T-1}) - f(x^*)\big) \\ &\leq \left(1 - \frac{\alpha}{\beta}\right)^T \big(f(x_0) - f(x^*)\big) \leq e^{-\frac{\alpha}{\beta}T}\big(f(x_0) - f(x^*)\big) . \end{aligned}$$

Using coercivity

Lemma

If f is α-strongly convex then for all $x, y \in \mathcal{X}$,

$$\left\langle \nabla f(x) - \nabla f(y), x - y \right\rangle \geq \alpha \|x - y\|^2 \,.$$

Proof: monotonicity of the gradient of the convex function $x \mapsto f(x) - \alpha\|x\|^2/2$.

Lemma

If f is α-strongly convex and β-smooth, then for all $x, y \in \mathcal{X}$,

$$\left\langle \nabla f(x) - \nabla f(y), x - y \right\rangle \geq \frac{\alpha\beta}{\alpha + \beta}\|x - y\|^2 + \frac{1}{\alpha + \beta}\left\|\nabla f(x) - \nabla f(y)\right\|^2 \,.$$

Proof: co-coercivity of the $(\beta - \alpha)$-smooth and convex function $x \mapsto f(x) - \alpha\|x\|^2/2$.

Stronger result by coercivity

Theorem

Let f be a β-smooth and α-strongly convex function. Then GD with the choice $\gamma_t \equiv \gamma = \frac{2}{\alpha+\beta}$ satisfies

$$\|x_T - x^*\|^2 \leq e^{-\frac{4T}{\kappa+1}} \|x_0 - x^*\|^2 ,$$

where $\kappa = \frac{\beta}{\alpha} \geq 1$ is the *condition number* of f.

Corollary: since by β-smoothness $f(x_T) - f(x^*) \leq \dfrac{\beta}{2}\|x_T - x^*\|^2$, this bound implies

$$f(x_T) - f(x^*) \leq \frac{\beta}{2} \exp\left(-\frac{4T}{\kappa+1}\right) \|x_0 - x^*\|^2 .$$

NB: *Bolder* jumps: $\gamma = \left(\dfrac{\alpha+\beta}{2}\right)^{-1} \geq \beta^{-1}$.

Proof

Using the coercivity inequality,

$$
\begin{aligned}
\|x_t - x^*\|^2 &= \left\|x_{t-1} - \gamma\nabla f(x_{t-1}) - x^*\right\|^2 \\
&= \left\|x_{t-1} - x^*\right\|^2 - 2\gamma\left\langle\nabla f(x_{t-1}), x_{t-1} - x^*\right\rangle + \gamma^2\left\|\nabla f(x_{t-1})\right\|^2 \\
&\leq \left(1 - 2\frac{\alpha\beta\gamma}{\alpha+\beta}\right)\|x_{t-1} - x^*\|^2 + \left(\underbrace{\gamma^2 - \frac{2\gamma}{\alpha+\beta}}_{=0}\right)\left\|\nabla f(x_{t-1})\right\|^2 \\
&= \left(1 - \frac{2}{\kappa+1}\right)^2\|x_{t-1} - x^*\|^2 \\
&\leq \exp\left(-\frac{4t}{\kappa+1}\right)\|x_0 - x^*\|^2 .
\end{aligned}
$$

Lower bounds lower bound for Lipschitz convex optimization

Lower bounds

General first-order black-box optimization algorithm = sequence of maps $(x_0, g_0, \ldots, x_t, g_t) \mapsto x_{t+1}$. We assume:

- $x_0 = 0$
- $x_{t+1} \in \mathrm{Span}\big(g_0, \ldots, g_t\big)$,

Theorem

For every $T \geq 1, L, R > 0$ there exists a convex and L-Lipschitz function f on $\mathbb{R}^{T+1}$ such that for any black-box procedure as above,

$$\min_{0 \leq t \leq T} f(x_t) - \min_{\|x\| \leq R} f(x) \geq \frac{RL}{2\big(1 + \sqrt{T+1}\big)} .$$

- Minimax lower bound: f and even d depend on T...
- ...but not limited to gradient descent algorithms.
- For a fixed dimension, exponential rates are always possible by other means (e.g. center of gravity method).
- $\implies$ the above GD algorithm is minimax rate-optimal!

Proof

Let $d = T + 1$, $\rho = \frac{L\sqrt{d}}{1+\sqrt{d}}$ and $\alpha = \frac{L}{R\left(1+\sqrt{d}\right)}$, and let

$$f(x) = \rho \max_{1 \leq i \leq d} x^i + \frac{\alpha}{2} \|x\|^2 .$$

Then

$$\partial f(x) = \alpha x + \rho \, \mathrm{Conv}\left(\left\{e_i : i \text{ s.t. } x_i = \max_{1 \leq j \leq d} x_j\right\}\right) .$$

If $\|x\| \leq R$, then $\forall g \in \partial f(x)$, $\|g\| \leq \alpha R + \rho$ which means that f is $\alpha R + \rho = L$-Lipschitz. For simplicity of notation, we assume that the first-order oracle returns $\alpha x + \rho e_i$ where i is the *first* coordinate such that $x_i = \max_{1 \leq j \leq d} x^j$.

- Thus $x_1 \in \mathrm{Span}(e_1)$, and by induction $x_t \in \mathrm{Span}(e_1, \ldots, e_t)$.
- Hence for every $j \in \{t+1, \ldots, d\}$, $x_t^j = 0$, and $f(x_t) \geq 0$ for all $t \leq T = d - 1$.
- f reaches its minimum at $x^* = \left(-\frac{\rho}{\alpha d}, \ldots, -\frac{\rho}{\alpha d}\right)$ since $0 \in \partial f(x^*)$, $\|x^*\|^2 = \frac{\rho^2}{\alpha^2 d} = R^2$ and

$$f(x^*) = -\frac{\rho^2}{\alpha d} + \frac{\alpha}{2}\frac{\rho^2}{\alpha^2 d} = -\frac{\rho^2}{2\alpha d} = -\frac{RL}{2\left(1+\sqrt{T+1}\right)} .$$

Other lower bounds

For α-**strongly convex and Lipschitz functions**, lower bound in $\dfrac{L^2}{\alpha T}$.
$\implies$ GD is order-optimal.

For β-**smooth convex functions**, the lower bound is in $\dfrac{\beta\|x_0 - x^*\|^2}{T^2}$.
$\implies$ room for improvement over GD with reaches $\dfrac{2\beta\|x_0 - x^*\|^2}{T+4}$.

For α-**strongly convex and** β-**smooth functions**, lower bound in $\|x_0 - x^*\|^2 e^{-\frac{T}{\sqrt{\kappa}}}$.
$\implies$ room for improvement over GD which reaches $\|x_0 - x^*\|^2 e^{-\frac{T}{\kappa}}$.

For proofs, see [Bubeck].

What more?

Need more?

- Constrained optimization
  - projected gradient descent

$$y_t = x_t - \gamma_t g_t, \qquad x_{t+1} = \Pi_{\mathcal{X}}(y_{t+1}) \,.$$

  - Frank-Wolfe

$$y_{t+1} = \arg\min_{y \in \mathcal{X}} \left\langle \nabla f(x_t), y \right\rangle, \qquad x_{t+1} = (1-\gamma_t)x_t + \gamma_t y_{t+1} \,.$$

- Nesterov acceleration

$$y_{t+1} = x_t - \frac{1}{\beta}\nabla f(x_t), \qquad x_{t+1} = \left(1 + \frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right) y_{t+1} - \frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1} y_t \,.$$

- Second-order methods, Newton and quasi-Newton

$$x_{t+1} = x_t - \left[\nabla^2 f(x)\right]^{-1} \nabla f(x) \,.$$

- Mirror descent: for a given convex potention Φ,

$$\nabla\Phi(y_{t+1}) = \nabla\Phi(x_t) - \gamma g_t, \qquad x_{t+1} \in \Pi_{\mathcal{X}}^{\Phi}(y_{t+1}) \,.$$

- Structured optimization, proximal methods
  - Example: $f(x) = L(x) + \lambda \|x\|_1$

Taken from `https://blogs.princeton.edu/imabandit`

Algorithm: Nesterov Accelerated Gradient Descent

Input: convex function f, initial point x_0

$d_0 \leftarrow 0,\ \lambda_0 \leftarrow 1$;
for $t = 0 \ldots T-1$ **do**
    $y_t \leftarrow x_t + d_t$;
    $x_{t+1} \leftarrow y_t - \frac{1}{\beta}\nabla f(y_t)$;
    $\lambda_{t+1} \leftarrow$ largest solution of $\lambda_{t+1}^2 - \lambda_{t+1} = \lambda_t^2$;
    $d_{t+1} \leftarrow \frac{\lambda_t - 1}{\lambda_{t+1}}\left(x_{t+1} - x_t\right)$;
return x_T

- d_t = momentum term ("heavy ball"), well-known practical trick to accelerate convergence, intensity $\frac{\lambda_t - 1}{\lambda_{t+1}} \lessapprox 1$.
- $\lambda_t \gtrapprox t/2 + 1$: let $\delta_t = \lambda_t - \lambda_{t-1} \geq 0$ and observe that $\lambda_t^2 - \lambda_{t-1}^2 = \delta_t(2\lambda_t - \delta_t) = \lambda_t$, from which one deduces that $1/2 \leq \delta_t = \frac{1}{2 - \delta_t/\lambda_t} \leq 1$, thus $1 + t/2 \leq \lambda_t \leq 1 + t$, hence $\delta_t \leq \frac{1}{2 - 1/(1+t/2)} \leq 1/2 + 1/(t+1)$ and $1 + t/2 \leq \lambda_t \leq t/2 + \log(t+1) + 1$.

Nesterov Acceleration

Theorem

Let f be a convex and β-smooth function on $\mathbb{R}^d$. Then Nesterov Accelerated Gradient descent algorithm satisfies:

$$f(x_T) - f(x^*) \leq \frac{2\beta\|x_0 - x^*\|^2}{T^2} .$$

- Thus it requires $T_\epsilon \approx \frac{2\beta R}{\sqrt{\epsilon}}$ steps to ensure precision ϵ.
- Nesterov acceleration also works for β-smooth, α-strongly convex functions and permits to reach the minimax rate $\|x_0 - x^*\|^2 e^{-\frac{T}{\sqrt{\kappa}}}$: see for example [Bubeck].

Proof

Let $\delta_t = f(x_t) - f(x^*)$. Denoting $g_t = -\beta^{-1}\nabla f(x_t + d_t)$, one has:

$$\begin{aligned}\delta_{t+1} - \delta_t &= f(x_{t+1}) - f(x_t + d_t) + f(x_t + d_t) - f(x_t)\\ &\le -\frac{2}{\beta}\left\|\nabla f(x_t + d_t)\right\|^2 + \left\langle \nabla f(x_t + d_t), d_t \right\rangle = -\frac{\beta}{2}\left(\|g_t\|^2 + 2\langle g_t, d_t\rangle\right) ,\end{aligned}$$

and $$\begin{aligned}\delta_{t+1} &= f(x_{t+1}) - f(x_t + d_t) + f(x_t + d_t) - f(x^*)\\ &\le -\frac{2}{\beta}\left\|\nabla f(x_t + d_t)\right\|^2 + \left\langle \nabla f(x_t + d_t), x_t + d_t - x^* \right\rangle = -\frac{\beta}{2}\left(\|g_t\|^2 + 2\langle g_t, x_t + d_t - x^*\rangle\right) .\end{aligned}$$

Hence, $$\begin{aligned}(\lambda_t - 1)\left(\delta_{t+1} - \delta_t\right) + \delta_{t+1} &\le -\frac{\beta}{2}\left(\lambda_t\|g_t\|^2 + 2\langle g_t, x_t + \lambda_t d_t - x^*\rangle\right)\\ &= -\frac{\beta}{2\lambda_t}\left(\left\|\lambda_t g_t + x_t + \lambda_t d_t - x^*\right\|^2 - \left\|x_t + \lambda_t d_t - x^*\right\|^2\right)\\ &= -\frac{\beta}{2\lambda_t}\left(\left\|x_{t+1} + \lambda_{t+1} d_{t+1} - x^*\right\|^2 - \left\|x_t + \lambda_t d_t - x^*\right\|^2\right) ,\end{aligned}$$

since the choice of the momentum intensity is precisely ensuring that $x_t + \lambda_t g_t + \lambda_t d_t = x_{t+1} + (\lambda_t - 1)(g_t + d_t) = x_{t+1} + (\lambda_t - 1)(x_{t+1} - x_t) = x_{t+1} + \lambda_{t+1} \underbrace{\frac{\lambda_t - 1}{\lambda_{t+1}}(x_{t+1} - x_t)}_{d_{t+1}}$.

It follows from the choice of λ_t that

$$\lambda_t^2\delta_{t+1} - \lambda_{t-1}^2\delta_t = \lambda_t^2\delta_{t+1} - (\lambda_t^2 - \lambda_t)\delta_t \le -\frac{\beta}{2}\left(\left\|x_{t+1} + \lambda_{t+1} d_{t+1} - x^*\right\|^2 - \left\|x_t + \lambda_t d_t - x^*\right\|^2\right)$$

and hence, since $\lambda_{-1} = 0$ and $\lambda_t \ge (t+1)/2$:

$$\left(\frac{T}{2}\right)^2\delta_T \le \lambda_{T-1}^2\delta_T \le \frac{\beta}{2}\left\|x_0 + \lambda_0 d_0 - x^*\right\|^2 = \frac{\beta\left\|x_0 - x^*\right\|^2}{2} .$$

Research article 4

Incremental Majorization-Minimization Optimization with Application to Large-Scale Machine Learning

by Julien Mairal

SIAM Journal on Optimization Vol. 25 Issue 2, 2015 Pages 829-855

SIAM J. OPTIM.
Vol. 25, No. 2, pp. 829–855

INCREMENTAL MAJORIZATION-MINIMIZATION OPTIMIZATION WITH APPLICATION TO LARGE-SCALE MACHINE LEARNING*

JULIEN MAIRAL†

Abstract. Majorization-minimization algorithms consist of successively minimizing a sequence of upper bounds of the objective function. These upper bounds are tight at the current estimate, and each iteration monotonically drives the objective function downhill. Such a simple principle is widely applicable and has been very popular in various scientific fields, especially in signal processing and statistics. We propose an incremental majorization-minimization scheme for minimizing a large sum of continuous functions, a problem of utmost importance in machine learning. We present convergence guarantees for nonconvex and convex optimization when the upper bounds approximate the objective up to a smooth error; we call such upper bounds "first-order surrogate functions." More precisely, we study asymptotic stationary point guarantees for nonconvex problems, and for convex ones, we provide convergence rates for the expected objective function value. We apply our scheme to composite optimization and obtain a new incremental proximal gradient algorithm with linear convergence rate for strongly convex functions. Our experiments show that our method is competitive with the state of the art for solving machine learning problems such as logistic regression when the number of training samples is large enough, and we demonstrate its usefulness for sparse estimation with nonconvex penalties.

Key words. nonconvex optimization, convex optimization, majorization-minimization

AMS subject classifications. 90C06, 90C26, 90C25

DOI. 10.1137/140957639

1. Introduction. The principle of successively minimizing upper bounds of the objective function is often called *majorization-minimization* [35] or *successive upper-bound minimization* [48]. Each upper bound is locally tight at the current estimate, and each minimization step decreases the value of the objective function. Even though this principle does not provide any theoretical guarantee about the quality of the returned solution, it has been very popular and widely used because of its simplicity. Various existing approaches can indeed be interpreted from the majorization-minimization point of view. This is the case of many gradient-based or proximal methods [3, 14, 28, 45, 54], expectation-maximization (EM) algorithms in statistics [20, 42], difference-of-convex (DC) programming [30], boosting [13, 19], some variational Bayes techniques used in machine learning [53], and the mean-shift algorithm for finding modes of a distribution [25]. Majorizing surrogates have also been used successfully in the signal processing literature about sparse estimation [11, 16, 26], linear inverse problems in image processing [1, 23], and matrix factorization [37, 40].

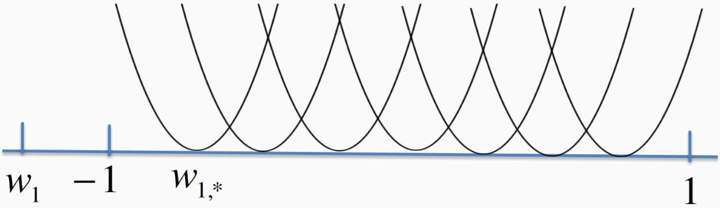
*Received by the editors February 18, 2014; accepted for publication (in revised form) January 27, 2015; published electronically April 14, 2015. This work was partially supported by the Gargantua project (Program Mastodons, CNRS), the Microsoft Research-Inria Joint Centre, and Agence Nationale de la Recherche (MACARON project ANR-14-CE23-0003-01 and LabEx PERSYVAL-Lab ANR-11-LABX-0025). A short version of this work was presented at the International Conference of Machine Learning (ICML) in 2013.
http://www.siam.org/journals/siopt/25-2/95763.html
†Inria, LEAR Team, Laboratoire Jean Kuntzmann, CNRS, Université Grenoble Alpes, 655 avenue de l'Europe, 38330 Montbonnot, France (julien.mairal@inria.fr).

Stochastic Gradient Descent

Motivation

Big data: an evaluation of f can be very expensive, and useless! (especially at the begining).

$$L(\theta) = \frac{1}{m} \sum_{i=1}^{m} \ell\big(y_i \langle \theta, x_i \rangle\big) \ .$$



Src: `https://arxiv.org/pdf/1606.04838.pdf`

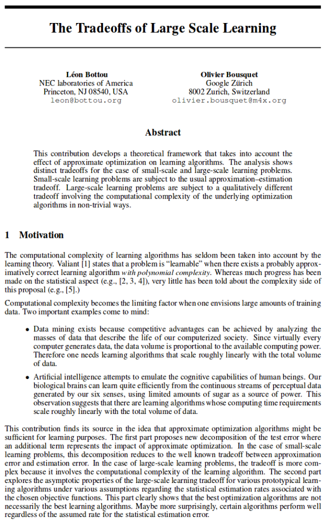
$\rightarrow$ often faster and cheaper for the required precision.

The Tradeoffs of Large Scale Learning

by Léon Bottou and Olivier Bousquet

Advances in Neural Information Processing Systems, NIPS Foundation (http://books.nips.cc) (2008), pp. 161-168

NeurIPS 2018 award: "test of time"



The Tradeoffs of Large Scale Learning

Léon Bottou
NEC laboratories of America
Princeton, NJ 08540, USA
leon@bottou.org

Olivier Bousquet
Google Zürich
8002 Zurich, Switzerland
olivier.bousquet@m4x.org

Abstract

This contribution develops a theoretical framework that takes into account the effect of approximate optimization on learning algorithms. The analysis shows distinct tradeoffs for the case of small-scale and large-scale learning problems. Small-scale learning problems are subject to the usual approximation–estimation tradeoff. Large-scale learning problems are subject to a qualitatively different tradeoff involving the computational complexity of the underlying optimization algorithms in non-trivial ways.

1 Motivation

The computational complexity of learning algorithms has seldom been taken into account by the learning theory. Valiant [1] states that a problem is "learnable" when there exists a probably approximatively correct learning algorithm *with polynomial complexity*. Whereas much progress has been made on the statistical aspect (e.g., [2, 3, 4]), very little has been told about the complexity side of this proposal (e.g., [5].)

Computational complexity becomes the limiting factor when one envisions large amounts of training data. Two important examples come to mind:

- Data mining exists because competitive advantages can be achieved by analyzing the masses of data that describe the life of our computerized society. Since virtually every computer generates data, the data volume is proportional to the available computing power. Therefore one needs learning algorithms that scale roughly linearly with the total volume of data.
- Artificial intelligence attempts to emulate the cognitive capabilities of human beings. Our biological brains can learn quite efficiently from the continuous streams of perceptual data generated by our six senses, using limited amounts of sugar as a source of power. This observation suggests that there are learning algorithms whose computing time requirements scale roughly linearly with the total volume of data.

This contribution finds its source in the idea that approximate optimization algorithms might be sufficient for learning purposes. The first part proposes new decomposition of the test error where an additional term represents the impact of approximate optimization. In the case of small-scale learning problems, this decomposition reduces to the well known tradeoff between approximation error and estimation error. In the case of large-scale learning problems, the tradeoff is more complex because it involves the computational complexity of the learning algorithm. The second part explores the asymptotic properties of the large-scale learning tradeoff for various prototypical learning algorithms under various assumptions regarding the statistical estimation rates associated with the chosen objective functions. This part clearly shows that the best optimization algorithms are not necessarily the best learning algorithms. Maybe more surprisingly, certain algorithms perform well regardless of the assumed rate for the statistical estimation error.

Stochastic Gradient Descent Algorithms

We consider a function to miminize $f(x) = \frac{1}{m}\sum_{i=1}^{m} f_i(x)$:

Algorithm: Stochastic Gradient Descent

Input: convex function f, step size γ_t, initial point x_0

for $t = 0 \dots T-1$ **do**
    Pick $I_t \sim \mathcal{U}(\{1, \dots, m\})$
    Compute $g_t \in \partial f_{I_t}(x_t)$
    $x_{t+1} \leftarrow x_t - \gamma_t g_t$
return x_T or $\frac{x_0 + \cdots + x_{T-1}}{T}$

- $x_T \underset{T\to\infty}{\to} x^* \stackrel{def}{=} \arg\min f$?
- $f(x_T) \underset{T\to\infty}{\to} f(x^*) = \min f$?
- under which conditions?
- what about $\frac{x_0 + \cdots + x_{T-1}}{T}$?
- at what speed?
- works in high dimension?
- do some properties help?
- can other algorithms do better?

Noisy Gradient Descent

Let $\mathcal{F}_t = \sigma(I_0, \dots, I_t)$, where $\mathcal{F}_{-1} = \{\Omega, \emptyset\}$. Note that x_t is $\mathcal{F}_{t-1}$-measurable, i.e. x_t depends only on $I_0, \dots, I_{t-1}$.

Lemma

For all $t \geq 0$,

$$\mathbb{E}\big[g_t | \mathcal{F}_{t-1}\big] \in \partial f(x_t) \; .$$

Proof: let $y \in \mathcal{X}$. Since $g_t \in \partial f_{I_t}(x_t)$, $f_{I_t}(y) \geq f_{I_t}(x_t) + \langle g_t, y - x_t \rangle$. Taking expectation conditionnal on F_{t-1} (i.e. integrating on I_t), and using that x_t is F_{t-1}-measurable, one obtains:

$$f(y) \geq f(x_t) + \mathbb{E}\big[\langle g_t, y - x_t | \mathcal{F}_{t-1} \rangle\big] = f(x_t) + \Big\langle \mathbb{E}\big[g_t | \mathcal{F}_{t-1}\big], y - x_t \Big\rangle \; .$$

More generally, SGD for the optimization of functions f that are accessible by a *noisy first-order example*, i.e. for which it is possible to obtain at every point an independent, unbiased estimate of the gradient. Two distinct objective functions:

$$L_S(\theta) = \frac{1}{m} \sum_{i=1}^{m} \ell_i\big(h_\theta(x_i), y_i\big) \qquad \text{and } L_D(\theta) = \mathbb{E}\Big[\ell\big(h_\theta(X), Y\big)\Big] \; .$$

Convergence for Lipschitz convex functions

Theorem

Assume that for all i, all $x \in \mathcal{X}$ and all $g \in \partial f_i(x)$, $\|g_t\| \leq L$. Then SGD with $\gamma_t \equiv \gamma = \frac{R}{L\sqrt{T}}$ satisfies

$$\mathbb{E}\left[f\left(\frac{1}{T}\sum_{i=0}^{T-1} x_i\right)\right] - f(x^*) \leq \frac{RL}{\sqrt{T}}\,.$$

- Exactly the same bound as for GD in the Lipschitz convex case.
- As before, it requires $T_\epsilon \approx \frac{R^2 L^2}{\epsilon^2}$ steps to ensure precision ϵ.
- Bound only in expectation.
- In contrast to the deterministic case, smoothness does not improve the speed of convergence in general.

Proof: exactly the same as for GD

Cosinus theorem = generalized Pythagore theorem = Alkashi's theorem:

$$2\langle a, b\rangle = \|a\|^2 + \|b\|^2 - \|a-b\|^2 \ .$$

Hence for every $0 \leq t < T$, since $\mathbb{E}\big[g_t | \mathcal{F}_{t-1}\big] \in \partial f(x_t)$,

$$\begin{aligned}
\mathbb{E}\big[f(x_t) - f(x^*) | \mathcal{F}_{t-1}\big] &\leq \Big\langle \mathbb{E}\big[g_t|\mathcal{F}_{t-1}\big], x_t - x^* \Big\rangle \\
&= \frac{1}{\gamma}\Big\langle \mathbb{E}\big[x_t - x_{t+1}|\mathcal{F}_{t-1}\big], x_t - x^* \Big\rangle \\
&= \mathbb{E}\left[\frac{1}{2\gamma}\Big(\|x_t - x^*\|^2 + \|x_t - x_{t+1}\|^2 - \|x_{t+1} - x^*\|^2\Big)\Big|\mathcal{F}_{t-1}\right] \\
&= \mathbb{E}\left[\frac{1}{2\gamma}\Big(\|x_t - x^*\|^2 - \|x_{t+1} - x^*\|^2\Big) + \frac{\gamma}{2}\|g_t\|^2\Big|\mathcal{F}_{t-1}\right] \ ,
\end{aligned}$$

and hence, taking expectation:

$$\begin{aligned}
\mathbb{E}\left[\sum_{t=0}^{T-1} f(x_t) - f(x^*)\right] &\leq \frac{1}{2\gamma}\Big(\|x_0 - x^*\|^2 - \cancel{\mathbb{E}\big[\|x_T - x^*\|^2\big]}\Big) + \frac{L^2\gamma T}{2} \\
&\leq \frac{L\sqrt{T}\,R^2}{2R} + \frac{L^2\,RT}{2L\sqrt{T}} \ .
\end{aligned}$$

A lot more to know

- Faster rate for the strongly convex case:
  same proof as before.
- No improvement in general by using smoothness only.
- Ruppert-Polyak averaging.
- Improvement for sums of smooth and strongly convex functions, etc.
- Analysis in expectation only is rather weak.
- Mini-batch SGD: the best of the two worlds.
- Beyond SGD methods: momentum, simulated annealing, etc.

Convergence in quadratic mean

Theorem

Let $(\mathcal{F}_t)_t$ be an increasing family of σ-fields. For every $t \geq 0$, let f_t be a convex, differentiable, β-smooth, square-integrable, $\mathcal{F}_t$-measurable function on $\mathcal{X}$. Further, assume that for every $x \in \mathcal{X}$ and every $t \geq 1$, $\mathbb{E}\big[\nabla f_t(x) | \mathcal{F}_{t-1}\big] = \nabla f(x)$, where f is an α-strongly convex function reaching its minimum at $x^* \in \mathcal{X}$. Also assume that for all $t \geq 0$, $\mathbb{E}\big[\|\nabla f_t(x^*)\|^2 \big| \mathcal{F}_{t-1}\big] \leq \sigma^2$. Then, denoting $\kappa = \frac{\beta}{\alpha}$, the SGD with $\gamma_t = \frac{1}{\alpha\left(t+1+2\kappa^2\right)}$ satisfies:

$$\mathbb{E}\Big[\|x_T - x^*\|^2\Big] \leq \frac{2\kappa^2\|x_0 - x^*\|^2 + \frac{2\sigma^2}{\alpha^2}\log\left(\frac{T}{2\kappa^2} + 1\right)}{T + 2\kappa^2} .$$

Proof 1/2: induction formula for the quadratic risk

We observe that

$$\mathbb{E}\Big[\big\|\nabla f_{I_t}(x_{t-1})\big\|^2 \,\big|\mathcal{F}_{t-1}\Big] \le 2\mathbb{E}\Big[\big\|\nabla f_{I_t}(x_{t-1}) - f_{I_t}(x^*)\big\|^2 \,\big|\mathcal{F}_{t-1}\Big] + 2\mathbb{E}\Big[\big\|\nabla f_{I_t}(x^*)\big\|^2 \,\big|\mathcal{F}_{t-1}\Big]$$
$$\le 2\beta^2\big\|x_{t-1} - x^*\big\|^2 + 2\sigma^2 \ .$$

Hence,

$$\begin{aligned}
\mathbb{E}\Big[\big\|x_t - x^*\big\|^2 \,\big|\mathcal{F}_{t-1}\Big] &= \big\|x_{t-1} - x^*\big\|^2 - 2\gamma_{t-1}\big\langle x_{t-1} - x^*, \nabla f(x_{t-1})\big\rangle + \gamma_{t-1}^2\mathbb{E}\Big[\big\|\nabla f_{I_t}(x_{t-1})\big\|^2 \,\big|\mathcal{F}_{t-1}\Big] \\
&\le \big\|x_{t-1} - x^*\big\|^2 - 2\gamma_{t-1}\alpha\big\|x_{t-1} - x^*\big|^2 + \gamma_{t-1}^2\mathbb{E}\Big[\big\|\nabla f_{I_t}(x_{t-1})\big\|^2 \,\big\|\mathcal{F}_{t-1}\Big] \\
&\le \big(1 - 2\alpha\gamma_{t-1} + 2\beta^2\gamma_{t-1}^2\big)\big\|x_{t-1} - x^*\big\|^2 + 2\sigma^2\gamma_{t-1}^2 \\
&\le \big(1 - \alpha\gamma_{t-1}\big)\big\|x_{t-1} - x^*\big\|^2 + 2\sigma^2\gamma_{t-1}^2
\end{aligned}$$

thanks to the fact that for all $t \ge 0$, $\alpha\gamma_t \ge 2\beta^2\gamma_t^2 \iff \gamma_t \le \alpha/(2\beta^2) = \gamma_{-1}$, and γ_t is decreasing in t. Hence, denoting $\delta_t = \mathbb{E}\big[\|x_t - x^*\|^2\big]$, by taking expectation we obtain that

$$\delta_t \le \big(1 - \alpha\gamma_{t-1}\big)\delta_{t-1} + 2\sigma^2\gamma_{t-1}^2 \ .$$

Note that unfolding the induction formula leads to an explicit upper-bound for δ_t:

$$\delta_t \le \prod_{k=0}^{t-1}\big(1 - \mu\gamma_k\big) + 2\sigma^2\sum_{k=0}^{t-1}\gamma_k^2\prod_{i=k+1}^{t-1}\big(1 - \mu\gamma_i\big) \ .$$

Proof 2/2: solving the induction

One may either use the closed form for δ_t, or (with the hint of the corresponding ODE) set $u_t = (t + 2\kappa^2)\delta_t$ and note that

$$
\begin{aligned}
u_t &= (t + 2\kappa^2)\delta_t \\
&\leq (t + 2\kappa^2)\left((1 - \alpha\gamma_{t-1}) \frac{u_{t-1}}{(t - 1 + 2\kappa^2)} + 2\sigma^2\gamma_{t-1}^2 \right) \\
&\leq (t + 2\kappa^2)\frac{t - 1 + 2\kappa^2}{t + 2\kappa^2} \frac{u_{t-1}}{(t - 1 + 2\kappa^2)} + \frac{2\sigma^2(t + 2\kappa^2)}{\alpha^2(t + 2\kappa^2)^2} \\
&= u_{t-1} + \frac{2\sigma^2}{\alpha^2} \frac{1}{(t + 2\kappa^2)} \\
&\leq u_0 + \frac{2\sigma^2}{\alpha^2} \sum_{s=1}^{t} \frac{1}{(s + 2\kappa^2)} \\
&\leq 2\kappa^2\delta_0 + \frac{2\sigma^2}{\alpha^2} \log \frac{t + 2\kappa^2}{2\kappa^2} \ .
\end{aligned}
$$

Hence for every t

$$
\delta_t \leq \frac{2\kappa^2\|x_0 - x^*\|^2 + \frac{2\sigma^2}{\alpha^2} \log \frac{t+2\kappa^2}{2\kappa^2}}{t + 2\kappa^2} \ .
$$

Remark: with some more technical work, the analysis works for all γ_t, possibly of the form $\gamma_t = t^{-\beta}$ for $\beta \leq 1$: see [Bach&Moulines '11].

Non-Asymptotic Analysis of Stochastic Approximation Algorithms for Machine Learning

by Francis Bach and Eric Moulines

Advances in Neural Information Processing Systems 24 (NIPS 2011)

`https://papers.nips.cc/paper/4316-non-asymptotic-analysis-of-stochastic-approximation-algorithms-for-machine-learning`

Non-Asymptotic Analysis of Stochastic Approximation Algorithms for Machine Learning

Francis Bach
INRIA - Sierra Project-team
Ecole Normale Supérieure, Paris, France
francis.bach@ens.fr

Eric Moulines
LTCI
Telecom ParisTech, Paris, France
eric.moulines@enst.fr

Abstract

We consider the minimization of a convex objective function defined on a Hilbert space, which is only available through unbiased estimates of its gradients. This problem includes standard machine learning algorithms such as kernel logistic regression and least-squares regression, and is commonly referred to as a stochastic approximation problem in the operations research community. We provide a non-asymptotic analysis of the convergence of two well-known algorithms, stochastic gradient descent (a.k.a. Robbins-Monro algorithm) as well as a simple modification where iterates are averaged (a.k.a. Polyak-Ruppert averaging). Our analysis suggests that a learning rate proportional to the inverse of the number of iterations, while leading to the optimal convergence rate in the strongly convex case, is not robust to the lack of strong convexity or the setting of the proportionality constant. This situation is remedied when using slower decays together with averaging, robustly leading to the optimal rate of convergence. We illustrate our theoretical results with simulations on synthetic and standard datasets.

1 Introduction

The minimization of an objective function which is only available through unbiased estimates of the function values or its gradients is a key methodological problem in many disciplines. Its analysis has been attacked mainly in three communities: stochastic approximation [1, 2, 3, 4, 5, 6], optimization [7, 8], and machine learning [9, 10, 11, 12, 13, 14, 15]. The main algorithms which have emerged are stochastic gradient descent (a.k.a. Robbins-Monro algorithm), as well as a simple modification where iterates are averaged (a.k.a. Polyak-Ruppert averaging).

Traditional results from stochastic approximation rely on strong convexity and asymptotic analysis, but have made clear that a learning rate proportional to the inverse of the number of iterations, while leading to the optimal convergence rate in the strongly convex case, is not robust to the wrong setting of the proportionality constant. On the other hand, using slower decays together with averaging robustly leads to optimal convergence behavior (both in terms of rates and constants) [4, 5].

The analysis from the convex optimization and machine learning literatures however has focused on differences between strongly convex and non-strongly convex objectives, with learning rates and roles of averaging being different in these two cases [11, 12, 13, 14, 15].

A key desirable behavior of an optimization method is to be adaptive to the hardness of the problem, and thus one would like a single algorithm to work in all situations, favorable ones such as strongly convex functions and unfavorable ones such as non-strongly convex functions. In this paper, we unify the two types of analysis and show that (1) a learning rate proportional to the inverse of the number of iterations is not suitable because it is not robust to the setting of the proportionality constant and the lack of strong convexity, (2) the use of averaging with slower decays allows (close to) optimal rates in *all* situations.

More precisely, we make the following contributions:

– We provide a direct non-asymptotic analysis of stochastic gradient descent in a machine learning context (observations of real random functions defined on a Hilbert space) that includes

1

Micro-bibliography: optimization and learning

- *Convex Optimization*, by **Stephen Boyd and Lieven Vandenberghe**, Cambridge University Press. Available online (with slides) on `http://web.stanford.edu/~boyd/cvxbook/`.
- *Convex Optimization: Algorithms and Complexity*, by **Sébastien Bubeck**, Foundations and Trends in Machine Learning, Vol. 8: No. 3-4, pp 231-357, 2015.
- *Introductory Lectures on Convex Optimization, A Basic Course*, by **Yurii Nesterov**, Springer, 2004
- *Introduction to Online Convex Optimization*, by **Elad Hazan**, Foundations and Trends in Optimization: Vol. 2: No. 3-4, pp 157-325, 2016.
- *Optimization Methods for Large-Scale Machine Learning*, by **Léon Bottou, Frank E. Curtis, and Jorge Nocedal**, SIAM Review, 2018, Vol. 60, No. 2 : pp. 223-311