Gradient descent

Olivier Fercoq

Télécom Paris

Application cases

Minimize a differentiable function f

$$\min_{x \in \mathbb{R}^n} f(x)$$

A very simple algorithm, and the basis for more evolved ones

Problems that fit in this framework

- Least squares $f(x) = \frac{1}{2}\|Zx - y\|^2$
- logistic regression $f(x) = \log(1 + \exp(-y_i z_i^\top x)) + \lambda\|x\|^2$
- Collaborative filtering $f(P, Q) = \sum_{(i,j)\in C}(R_{i,j} - \sum_{k=1}^{K} P_{i,k} Q_{k,j})^2$
- ...

The algorithm

- Goal: Minimize a differentiable function f

$$\min_{x \in \mathbb{R}^n} f(x)$$

- Algorithm: Let $(\gamma_k)_k$ with $\gamma_k > 0$ be a step size sequence. Fix $x_0 \in \mathbb{R}^n$ and for all $k \in \mathbb{N}$, do

$$x_{k+1} = x_k - \gamma_k \nabla f(x_k)$$

- Requirements:
  - Be able to compute ∇f
  - Set (γ_k) so that $f(x_k) \to \min_x f(x)$

Analysis without convexity

$$x_{k+1} = x_k - \gamma_k \nabla f(x_k)$$

Assumptions:

∇f is L-Lipschitz continuous: $\|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\|$ and

$\inf_x f(x) > -\infty$

Theorem:

i) $(f(x_k))$ *is decreasing and converges*

ii) $\lim_{k\to+\infty} \|\nabla f(x_k)\| = 0$

Proof:

By Taylor-Lagrange inequality

$$f(x_{k+1}) \leq f(x_k) + \langle \nabla f(x_k), x_{k+1} - x_k \rangle + \frac{L}{2}\|x_{k+1} - x_k\|^2 \leq f(x_k) - \left(\gamma_k - \frac{L}{2}\gamma_k^2\right)\|\nabla f(x_k)\|^2$$

We choose $\gamma_k = \gamma < \frac{2}{L}$ then $\|\nabla f(x_k)\| \leq \frac{1}{\gamma - \frac{L}{2}\gamma^2}\left(f(x_k) - f(x_{k+1})\right)$

Additional result with convexity

Suppose f is convex and ∇f is L-Lipschitz: $x_0 \in \mathbb{R}^n$ and $x_{k+1} = x_k - \frac{1}{L}\nabla f(x_k)$

$$\begin{aligned} f(x_{k+1}) &\leq f(x_k) + \langle \nabla f(x_k), x_{k+1} - x_k \rangle + \frac{L}{2}\|x_{k+1} - x_k\|^2 \\ &\leq f(x_k) + \langle \nabla f(x_k), x_* - x_k \rangle + \frac{L}{2}\|x_* - x_k\|^2 - \frac{L}{2}\|x_* - x_{k+1}\|^2 \\ &\leq f(x_*) + \frac{L}{2}\|x_* - x_k\|^2 - \frac{L}{2}\|x_* - x_{k+1}\|^2 \end{aligned}$$

Moreover, $f(x_{k+1}) \leq f(x_k)$. Hence,

$$\frac{K}{L}(f(x_K) - f(x_*)) \leq \frac{1}{L}\left(\sum_{k=0}^{K-1} f(x_{k+1}) - f(x_*)\right) \leq \frac{1}{2}\|x_* - x_0\|^2 - \frac{1}{2}\|x_* - x_K\|^2$$

$$f(x_K) - f(x_*) \leq \frac{L\|x_* - x_0\|^2}{2K}$$

Additional result with convexity

Suppose f is convex and ∇f is L-Lipschitz: $x_0 \in \mathbb{R}^n$ and $x_{k+1} = x_k - \frac{1}{L}\nabla f(x_k)$

Taylor-Lagrange inequality

$$
\begin{aligned}
f(x_{k+1}) &\leq f(x_k) + \langle \nabla f(x_k), x_{k+1} - x_k \rangle + \frac{L}{2}\|x_{k+1} - x_k\|^2 \\
&\leq f(x_k) + \langle \nabla f(x_k), x_* - x_k \rangle + \frac{L}{2}\|x_* - x_k\|^2 - \frac{L}{2}\|x_* - x_{k+1}\|^2 \\
&\leq f(x_*) + \frac{L}{2}\|x_* - x_k\|^2 - \frac{L}{2}\|x_* - x_{k+1}\|^2
\end{aligned}
$$

Moreover, $f(x_{k+1}) \leq f(x_k)$. Hence,

$$
\frac{K}{L}(f(x_K) - f(x_*)) \leq \frac{1}{L}\left(\sum_{k=0}^{K-1} f(x_{k+1}) - f(x_*)\right) \leq \frac{1}{2}\|x_* - x_0\|^2 - \frac{1}{2}\|x_* - x_K\|^2
$$

$$
f(x_K) - f(x_*) \leq \frac{L\|x_* - x_0\|^2}{2K}
$$

Additional result with convexity

Suppose f is convex and ∇f is L-Lipschitz: $x_0 \in \mathbb{R}^n$ and $x_{k+1} = x_k - \frac{1}{L}\nabla f(x_k)$

$\boxed{\text{RHS independent of } x_*}$

$$
\begin{aligned}
f(x_{k+1}) &\leq f(x_k) + \langle \nabla f(x_k), x_{k+1} - x_k \rangle + \frac{L}{2}\|x_{k+1} - x_k\|^2 \\
&\leq f(x_k) + \langle \nabla f(x_k), x_* - x_k \rangle + \frac{L}{2}\|x_* - x_k\|^2 - \frac{L}{2}\|x_* - x_{k+1}\|^2 \\
&\leq f(x_*) + \frac{L}{2}\|x_* - x_k\|^2 - \frac{L}{2}\|x_* - x_{k+1}\|^2
\end{aligned}
$$

Moreover, $f(x_{k+1}) \leq f(x_k)$. Hence,

$$
\frac{K}{L}(f(x_K) - f(x_*)) \leq \frac{1}{L}\left(\sum_{k=0}^{K-1} f(x_{k+1}) - f(x_*)\right) \leq \frac{1}{2}\|x_* - x_0\|^2 - \frac{1}{2}\|x_* - x_K\|^2
$$

$$
f(x_K) - f(x_*) \leq \frac{L\|x_* - x_0\|^2}{2K}
$$

Additional result with convexity

Suppose f is convex and ∇f is L-Lipschitz: $x_0 \in \mathbb{R}^n$ and $x_{k+1} = x_k - \frac{1}{L}\nabla f(x_k)$

$\boxed{f \text{ is convex}}$

$$
\begin{aligned}
f(x_{k+1}) &\leq f(x_k) + \langle \nabla f(x_k), x_{k+1} - x_k \rangle + \frac{L}{2}\|x_{k+1} - x_k\|^2 \\
&\leq f(x_k) + \langle \nabla f(x_k), x_* - x_k \rangle + \frac{L}{2}\|x_* - x_k\|^2 - \frac{L}{2}\|x_* - x_{k+1}\|^2 \\
&\leq f(x_*) + \frac{L}{2}\|x_* - x_k\|^2 - \frac{L}{2}\|x_* - x_{k+1}\|^2
\end{aligned}
$$

Moreover, $f(x_{k+1}) \leq f(x_k)$. Hence,

$$
\frac{K}{L}(f(x_K) - f(x_*)) \leq \frac{1}{L}\left(\sum_{k=0}^{K-1} f(x_{k+1}) - f(x_*)\right) \leq \frac{1}{2}\|x_* - x_0\|^2 - \frac{1}{2}\|x_* - x_K\|^2
$$

$$
f(x_K) - f(x_*) \leq \frac{L\|x_* - x_0\|^2}{2K}
$$

Summary of convergence results for gradient descent

No convexity	f is convex	f is μ-strongly convex

Subgradient method

What if f is not differentiable?

$g \in \partial f(x) \Leftrightarrow$ $f(y) \geq f(x) + \langle g, y-x \rangle \;\forall y.$

Algorithm

$$g_k \in \partial f(x_k)$$
$$x_{k+1} = x_k - \gamma_k g_k$$
$$\bar{x}_k^{\gamma} = \frac{1}{\sum_{l=0}^{k} \gamma_l} \sum_{j=0}^{k} \gamma_j x_j$$

Theorem

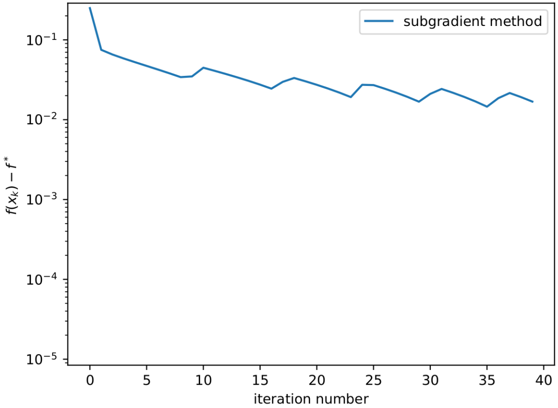
If f is convex and Lipschitz continuous and $\gamma_k = \dfrac{\gamma_0}{\sqrt{k+1}}$

$$f(\bar{x}_k^{\gamma}) - f(x^*) \in O\left(\frac{\ln(k)}{\sqrt{k}}\right)$$

Example on Lasso problem

$$\min_x \frac{1}{2}\|Ax - b\|_2^2 + \lambda\|x\|_1 = \min_x f(x)$$

$$x_{k+1} = x_k - \frac{\gamma_0}{\sqrt{k}} g_k, \qquad g_k \in \partial f(x_k)$$

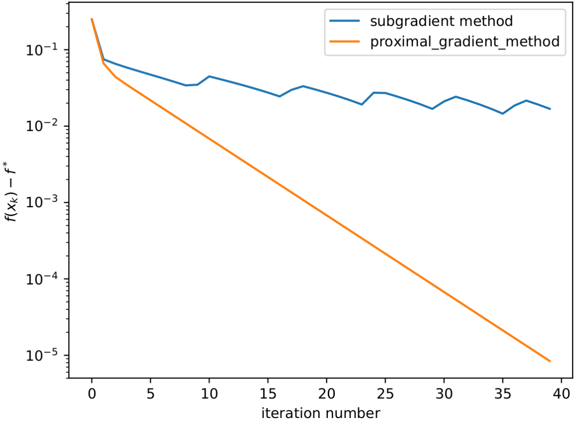


Example on Lasso problem

$$\min_x \frac{1}{2}\|Ax - b\|_2^2 + \lambda\|x\|_1 = \min_x f(x)$$

$$x_{k+1} = x_k - \frac{\gamma_0}{\sqrt{k}} g_k, \qquad g_k \in \partial f(x_k)$$

$$x_{k+1} = \mathrm{prox}_{\frac{1}{L}g}(x_k - \tfrac{1}{L}\nabla f(x_k))$$



Proximal gradient descent

- The subgradient method is a general but slow algorithm
- Idea: use structure of the problem to design a faster algorithm
- Composite objective: $\min_x f(x) + g(x)$
  f differentiable and ∇f is L-Lipschitz
  g not differentiable but $\mathrm{prox}_g(x) = \arg\min_y g(y) + \frac{1}{2}\|x - y\|^2$ easy to compute

Algorithm

$x_{k+1} = \mathrm{prox}_{\frac{1}{L}g}\left(x_k - \frac{1}{L}\nabla f(x_k)\right)$

Theorem

If f is convex, ∇f is L-Lipschitz and g is convex, then

$$f(x_k) + g(x_k) - f(x^*) - g(x^*) \leq \frac{L}{2k}\|x_0 - x^*\|^2$$

Moreover, if $f + g$ is strongly convex, we have linear convergence

Why does it work?

Set $T(x) = \text{prox}_{\frac{1}{L}g}\left(x - \frac{1}{L}\nabla f(x)\right) = T_2 \circ T_1(x)$

Proposition

$\| \text{prox}_{\gamma g}(x) - \text{prox}_{\gamma g}(y)\| \leq \|x - y\|$

Show that T is a contraction as soon as T_1 is a contraction.

Examples of proximal operators

- Let C be a convex set and $g(x) = \iota_C(x) = \begin{cases} 0 & \text{if } x \in C \\ +\infty & \text{if } \ x \notin C \end{cases}$

  $\mathrm{prox}_{\gamma \iota_C}(x) = \arg\min_y \gamma \iota_C(y) + \frac{1}{2}\|x - y\|^2 = \arg\min_{y \in C} \frac{1}{2}\|x - y\|^2 = \mathrm{Proj}_C(x)$

  Proximal gradient descent generalizes projected gradient descent
- $g(x) = |x|$

  $$\mathrm{prox}_{\gamma|\cdot|}(x) = \begin{cases} x + \gamma & \text{if } x < -\gamma \\ 0 & \text{if } -\gamma \le x \le \gamma \\ x - \gamma & \text{if } x > \gamma \end{cases}$$
- $g : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ such that $g(x) = \sum_{i=1}^n g_i(x_i)$
  (g is thus a separable function)

  For all i, the ith coordinate of $\mathrm{prox}_{\gamma g}(x)$ is $(\mathrm{prox}_{\gamma g}(x))_i = \mathrm{prox}_{\gamma g_i}(x_i)$