Line search: Gradient, Newton and proximal gradient methods

Olivier Fercoq

Télécom Paris

Introduction

- Gradient algorithm: $x_0 \in \mathbb{R}^n$ and $x_{k+1} = x_k - \frac{1}{L}\nabla f(x_k)$

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

- The Lipschitz constant L of ∇f is needed to run the algorithm
- What can we do
  - if L is not known?
  - if ∇f is only locally Lipschitz continuous?

Convergence proof of the gradient method (Exercices 6 and 7)

Suppose f is convex and ∇f is L-Lipschitz

$$\text{Gradient algorithm: } x_0 \in \mathbb{R}^n \text{ and } x_{k+1} = x_k - \frac{1}{L}\nabla f(x_k)$$

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

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

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

Convergence proof of the gradient method (Exercices 6 and 7)

Suppose f is convex and ∇f is L-Lipschitz

$$\text{Gradient algorithm: } x_0 \in \mathbb{R}^n \text{ and } x_{k+1} = x_k - \frac{1}{L}\nabla f(x_k)$$

Taylor-Lagrange inequality

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

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

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

Convergence proof of the gradient method (Exercices 6 and 7)

Suppose f is convex and ∇f is L-Lipschitz

$$\text{Gradient algorithm: } x_0 \in \mathbb{R}^n \text{ and } x_{k+1} = x_k - \frac{1}{L}\nabla f(x_k)$$

RHS independent of x_*

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

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

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

Convergence proof of the gradient method (Exercices 6 and 7)

Suppose f is convex and ∇f is L-Lipschitz

$$\text{Gradient algorithm: } x_0 \in \mathbb{R}^n \text{ and } x_{k+1} = x_k - \frac{1}{L}\nabla f(x_k)$$

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

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

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

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

Any point in the proof where you would like more details?

Unknown Lipschitz constant

Suppose f is convex and ∇f is L-Lipschitz.

$$\text{Gradient algorithm: } x_0 \in \mathbb{R}^n \text{ and } x_{k+1} = x_k - \frac{1}{L}\nabla f(x_k)$$

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

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

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

Unknown Lipschitz constant

Suppose f is convex and $f(x_{k+1}) \leq f(x_k) + \langle \nabla f(x_k), x_{k+1} - x_k \rangle + \frac{L_k}{2}\|x_{k+1} - x_k\|^2$.

$$\text{Gradient algorithm: } x_0 \in \mathbb{R}^n \text{ and } x_{k+1} = x_k - \frac{1}{L_k}\nabla f(x_k)$$

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

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

$$\Big(\sum_{j=0}^{K-1} \frac{1}{L_j}\Big)(f(x_K) - f(x_*)) \leq \sum_{k=0}^{K-1} \frac{1}{L_k}\big(f(x_{k+1}) - f(x_*)\big) \leq \frac{1}{2}\|x_* - x_0\|^2$$

How to choose L_k?

$$f(x_K) - f(x_*) \le \frac{1}{2\sum_{j=0}^{K-1} \frac{1}{L_j}} \|x_* - x_0\|^2$$

Do we have $$\sum_{j=0}^{K-1} \frac{1}{L_j} \in O(K) \text{ ?}$$

Let us define

$$x^+(\gamma) = x_k - \gamma \nabla f(x_k)$$

We set $b > 0, a \in (0,1)$ and we find the first integer l such that

$$f(x^+(ba^l)) \le f(x_k) + \langle \nabla f(x_k), x^+(ba^l) - x_k \rangle + \frac{1}{2ba^l}\|x_k - x^+(ba^l)\|^2$$

Proposition

If ∇f is L-Lipschitz and $L_k = \frac{1}{ba^l}$, then $L_k < \cancel{aL}.$ $\frac{L}{a}$

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 $\text{prox}_g(x) = \arg\min_y g(y) + \frac{1}{2}\|x - y\|^2$ easy to compute

Algorithm

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

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

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

- $g(x) = |x|$

  $$\mathrm{prox}_{\gamma|\cdot|}(x) = \begin{cases} x + \gamma & \text{if } x < -\gamma \\ 0 & \text{if } \ -\gamma \leq x \leq \gamma \\ x - \gamma & \text{if } x > \gamma \end{cases}$$

- $g : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ 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 $(\mathrm{prox}_{\gamma g}(x))_i = \mathrm{prox}_{\gamma g_i}(x_i)$

Fixed points of the proximal gradient operators are minimizers (Exercise 5)

Proposition

If $x = \operatorname{prox}_{\gamma g}\left(x - \gamma \nabla f(x)\right)$ *then* $x \in \arg\min_y f(y) + g(y)$.

Convergence proof of the proximal gradient method (Exercice 10)

Suppose f and g are convex, ∇f is L-Lipschitz

$$\text{Proximal gradient algorithm: } x_0 \in \mathbb{R}^n \text{ and } x_{k+1} = \mathrm{prox}_{\frac{1}{L}g}\left(x_k - \frac{1}{L}\nabla f(x_k)\right)$$

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

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

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

Convergence proof of the proximal gradient method (Exercice 10)

Suppose f and g are convex, ∇f is L-Lipschitz

$$\text{Proximal gradient algorithm: } x_0 \in \mathbb{R}^n \text{ and } x_{k+1} = \mathrm{prox}_{\frac{1}{L}g}\left(x_k - \frac{1}{L}\nabla f(x_k)\right)$$

$\boxed{\text{Taylor-Lagrange inequality}}$

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

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

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

Convergence proof of the proximal gradient method (Exercice 10)

Suppose f and g are convex, ∇f is L-Lipschitz

$$\text{Proximal gradient algorithm: } x_0 \in \mathbb{R}^n \text{ and } x_{k+1} = \operatorname{prox}_{\frac{1}{L}g}\left(x_k - \frac{1}{L}\nabla f(x_k)\right)$$

Property of the proximal operator

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

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

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

Convergence proof of the proximal gradient method (Exercice 10)

Suppose f and g are convex, ∇f is L-Lipschitz

$$\text{Proximal gradient algorithm: } x_0 \in \mathbb{R}^n \text{ and } x_{k+1} = \operatorname{prox}_{\frac{1}{L}g}\left(x_k - \frac{1}{L}\nabla f(x_k)\right)$$

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

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

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

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

Proof of the 3 point inequality

Lemma

Let $p = \operatorname{prox}_{\gamma g}(x) = \arg\min_y g(y) + \frac{1}{2\gamma}\|y - x\|^2$*. For all* $y \in \mathbb{R}^n$*,*

$$g(p) + \frac{1}{2\gamma}\|p - x\|^2 \leq g(y) + \frac{1}{2\gamma}\|y - x\|^2 - \frac{1}{2\gamma}\|p - y\|^2$$

Line search for proximal gradient

$$\min_x f(x) + g(x) \qquad f \text{ smooth, } g \text{ nonsmooth}$$

- Find L_k such that

$$x^+(L_k) = \operatorname{prox}_{\frac{1}{L_k}g}\left(x_k - \frac{1}{L_k}\nabla f(x_k)\right)$$

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

- Set $x_{k+1} = x^+(L_k)$.

Line search for Newton's method

$$\min_x f(x) \qquad f \text{ is } C^2$$

- Newton's method: $x_{k+1} = x_k - (\nabla^2 f(x_k))^{-1} \nabla f(x_k)$
- Quadratic convergence: $\exists R > 0, \exists M < 1/R$ such that

$$\|x_k - x_*\| \leq R \;\; \Rightarrow \;\; \|x_{k+1} - x_*\| \leq M \|x_k - x_*\|^2$$

- Line search to deal with the case $\|x_k - x_*\| > R$

$$x^+(\gamma) = x_k - \gamma (\nabla^2 f(x_k))^{-1} \nabla f(x_k)$$
$$x_{k+1} = x^+(\gamma_k)$$
$$f(x^+(\gamma_k)) \leq f(x_k) + \langle \nabla f(x_k), x^+(\gamma_k) - x_k \rangle + \frac{1}{2\gamma_k} \|x^+(\gamma_k) - x_k\|^2_{\nabla^2 f(x_k)}$$