First Order Descent Methods in Nonsmooth Convex Optimization

Juan PEYPOUQUET
Universidad Técnica Federico Santa María
Universidad de Chile

CIMPA School on Algorithmic Nonsmooth Optimization

Ibarra, September 19-21, 2017

The proximal-gradient method

Recall that the proximal-gradient method, defined by

$$
\begin{aligned}
x_{k+1} &= \mathrm{Prox}_{\lambda h}(\mathrm{Grad}_{\lambda g}(x_k)) \\
&= \operatorname{argmin}\left\{h(x) + \frac{1}{2\lambda}\left\|x - (x_k - \lambda \nabla g(x_k))\right\|^2\right\} \\
&= \operatorname{argmin}\left\{h(x) + g(x_k) + \langle \nabla g(x_k), x - x_k\rangle + \frac{1}{2\lambda}\|x - x_k\|^2\right\},
\end{aligned}
$$

converges weakly, and has a worst-case convergence rate of

$$(g+h)(x_k) - \min(g+h) \leq \frac{\mathrm{dist}(x_0, S)^2}{2\lambda k}.$$

Two improvements

- A slight variant produces a worst-case convergence rate of

$$(g+h)(x_k) - \min(g+h) \leq \frac{2\,\mathrm{dist}(x_0,S)^2}{\lambda(k+1)^2}.$$

- If the function $g+h$ satisfies a simple geometric condition, the algorithm converges strongly to some $x^* \in S$, and

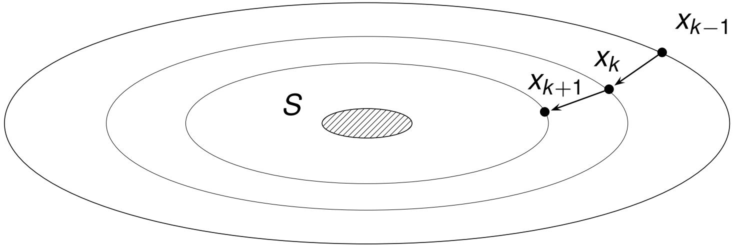
$$\|x_k - x^*\|^2 \leq \frac{4\,\mathrm{dist}(x_0,S)^2}{(1+2\lambda\eta)^k}$$

for some $\eta > 0$.

Accelerated Proximal-Gradient Method

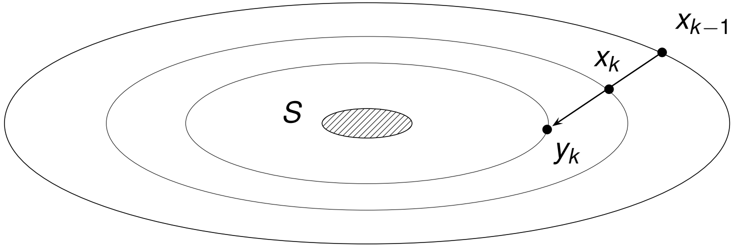
Acceleration

The idea is simple: Instead of doing this



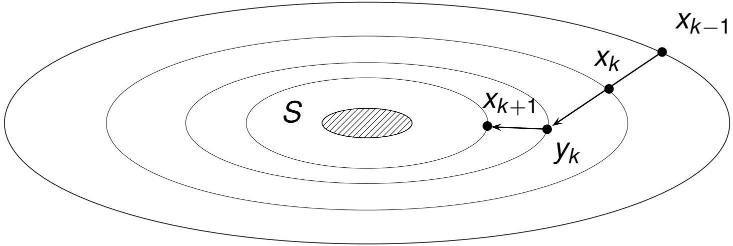
Acceleration

Let us try this:



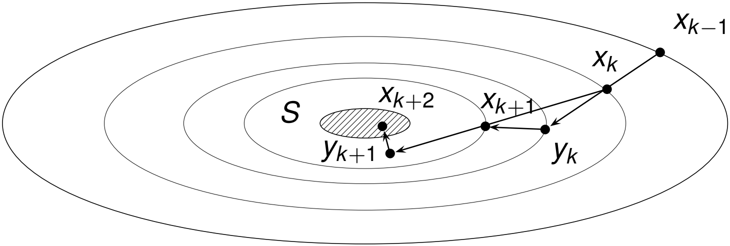
Acceleration

Let us try this:



Acceleration

Let us try this:



Acceleration

Accelerated proximal-gradient method. Start from $x_0 \in H$, pick $\alpha \geq 3$, and iterate:

$$\begin{cases} y_k = x_k + \frac{k}{k+\alpha}(x_k - x_{k-1}) \\ x_{k+1} = \text{Prox}_{\lambda h}(\text{Grad}_{\lambda g}(y_k)) \end{cases}$$

where ∇g is Lipschitz-continuous with constant L, and $\lambda L \leq 1$.

Theorem

$$f(x_k) - \min(f) \leq \frac{(\alpha-1)^2 dist(x_0, S)^2}{2\lambda(k+1)^2}, \quad k \geq 0.$$

Acceleration

Accelerated proximal-gradient method. Start from $x_0 \in H$, pick $\alpha \geq 3$, and iterate:

$$\begin{cases} y_k = x_k + \frac{k}{k+\alpha}(x_k - x_{k-1}) \\ x_{k+1} = \text{Prox}_{\lambda h}(\text{Grad}_{\lambda g}(y_k)) \end{cases}$$

where ∇g is Lipschitz-continuous with constant L, and $\lambda L \leq 1$.

Theorem

$$f(x_k) - \min(f) \leq \frac{(\alpha - 1)^2 dist(x_0, S)^2}{2\lambda(k+1)^2}, \quad k \geq 0.$$

Acceleration

The key is to use the fundamental property:

$$f(T_\lambda(x)) + \frac{\|y - T_\lambda(x)\|^2}{2\lambda} \leq f(y) + \frac{\|y - x\|^2}{2\lambda} \quad \forall x, y \in H,$$

with

$$\begin{aligned} x &= y_k &&= x_k + \tfrac{k}{k+\alpha}(x_k - x_{k-1}) \\ T_\lambda(x) &= T_\lambda(y_k) &&= x_{k+1} \\ && y &= \tfrac{\alpha-1}{k+\alpha}x^* + \tfrac{k+1}{k+\alpha}x_k \end{aligned}$$

Quadratic Growth and Exponential Convergence

Quadratic growth

Consider the minimization problem

$$\min\{f(x) : x \in H\},$$

where $f : H \to \mathbb{R} \cup \{+\infty\}$ is convex and lower-semicontinuous.

In many common applications, there exist $\eta > 0$ and $M \in \mathbb{R}$ such that

$$f(x) - \min(f) \geq \eta \, \mathrm{dist}(x, S)^2$$

for all $x \in [f \leq M]$.

Exponential convergence

Theorem

Let f satisfy the quadatic growth condition and let $\lambda > 0$. Suppose that (x_k) is a sequence in $[f \leq M]$ is such that, for each $u \in S$ and all k, we have

- $\|x_{k+1} - u\| \leq \|x_k - u\|$, *and*
- $2\lambda(f(x_{k+1}) - \min(f)) \leq \|x_k - u\|^2 - \|x_{k+1} - u\|^2$.

Then, x_k converges strongly to some $x^* \in S$, and

$$\|x_k - x^*\|^2 \leq \frac{4\, dist(x_0, S)^2}{(1 + 2\lambda\eta)^k}.$$

The proximal-gradient method

Proposition

Let $u \in S$ and let (x_k) be generated by the proximal-gradient method. Then, for all $k \geq 0$, we have

- $x_k \in [f \leq f(x_0)]$,
- $\|x_{k+1} - u\| \leq \|x_k - u\|$ *and*
- $2\lambda\big(f(x_{k+1}) - \min(f)\big) \leq \|x_k - u\|^2 - \|x_{k+1} - u\|^2$.

Corollary

If f has quadratic growth in $[f \leq f(x_0)]$, proximal-gradient sequences converge strongly and exponentially to a point in S.