Introduction to Optimization

Lecture 06: Gradient descent, continued.
Geometry and convergence rates. Computational exercise.

university of groningen

The gradient method

f is L-smooth and $x_{n+1} = x_n - \alpha \nabla f(x_n)$, with $0 < \alpha < \frac{2}{L}$

Proposition

1. *If* $\inf(f) > -\infty$, $\lim_{n\to\infty} \|\nabla f(x_n)\| = 0$, *and* $\min_{1\le i\le n} \{\|\nabla f(x_i)\|\} \le \frac{C}{\sqrt{n}}$.
2. *If* f *is convex,* $\lim_{n\to\infty} f(x_n) = \inf(f)$ *and every cluster point of* (x_n) *is a minimizer of* f.
3. *If, moreover,* f *has minimizers, then* x_n *converges to one of them,* $f(x_n) - \min(f) \le \dfrac{\mathrm{dist}(x_0, S)^2}{\alpha(2-\alpha L)n}$, *and* $\lim_{n\to\infty} n\left[f(x_n) - \min(f)\right] = 0$.

Strong convexity and the gradient method

The simplest example, one more time

For $f(x) = x^2$, we obtained $x_n = (1-2\alpha)^n x_0$, and so

$$f(x_n) - \min(f) = (1-2\alpha)^{2n}\big(f(x_0) - \min(f)\big).$$

Strong convexity and the gradient method

The simplest example, one more time

For $f(x) = x^2$, we obtained $x_n = (1-2\alpha)^n x_0$, and so

$$f(x_n) - \min(f) = (1-2\alpha)^{2n}\big(f(x_0) - \min(f)\big).$$

Part of an exercise from Lecture 04

If f is $\mu-$strongly convex, it has exactly on minimizer, and

$$f(y) \geq f(x) + \nabla f(x)\cdot(y-x) + \frac{\mu}{2}\|y-x\|^2 \qquad \text{for all } x,y\in\mathbb{R}^N.$$

Strong convexity and the gradient method

The simplest example, one more time

For $f(x) = x^2$, we obtained $x_n = (1-2\alpha)^n x_0$, and so

$$f(x_n) - \min(f) = (1-2\alpha)^{2n}\big(f(x_0) - \min(f)\big).$$

Part of an exercise from Lecture 04

If f is $\mu-$strongly convex, it has exactly on minimizer, and

$$f(y) \geq f(x) + \nabla f(x)\cdot(y-x) + \frac{\mu}{2}\|y-x\|^2 \qquad \text{for all } x, y \in \mathbb{R}^N.$$

An important consequence

If f is $\mu-$strongly convex, then $2\mu\big(f(x) - \min(f)\big) \leq \|\nabla f(x)\|^2$.

Linear convergence of the gradient method

f is L-smooth and $x_{n+1} = x_n - \alpha \nabla f(x_n)$, with $0 < \alpha < \frac{2}{L}$

Proposition

If f has minimizers and $2\mu\big(f(x) - \min(f)\big) \leq \|\nabla f(x)\|^2$ *for all* $x \in \mathbb{R}^N$*, then*

$$f(x_n) - \min(f) \leq \big[1 - \alpha\mu(2 - \alpha L)\big]^n \big(f(x_0) - \min(f)\big).$$

Linear convergence of the gradient method

f is L-smooth and $x_{n+1} = x_n - \alpha \nabla f(x_n)$, with $0 < \alpha < \frac{2}{L}$

Proposition

If f has minimizers and $2\mu\big(f(x) - \min(f)\big) \leq \|\nabla f(x)\|^2$ *for all* $x \in \mathbb{R}^N$, *then*

$$f(x_n) - \min(f) \leq \big[1 - \alpha\mu(2 - \alpha L)\big]^n \big(f(x_0) - \min(f)\big).$$

Questions

- How does this compare with the example?

Linear convergence of the gradient method

f is L-smooth and $x_{n+1} = x_n - \alpha \nabla f(x_n)$, with $0 < \alpha < \frac{2}{L}$

Proposition

If f has minimizers and $2\mu(f(x) - \min(f)) \leq \|\nabla f(x)\|^2$ *for all* $x \in \mathbb{R}^N$, *then*

$$f(x_n) - \min(f) \leq \left[1 - \alpha\mu(2 - \alpha L)\right]^n \big(f(x_0) - \min(f)\big).$$

Questions

- How does this compare with the example?
- What value of α gives the best rate?

Linear convergence of the gradient method

f is L-smooth and $x_{n+1} = x_n - \alpha \nabla f(x_n)$, with $0 < \alpha < \frac{2}{L}$

Proposition

If f has minimizers and $2\mu\big(f(x) - \min(f)\big) \leq \|\nabla f(x)\|^2$ *for all* $x \in \mathbb{R}^N$, *then*

$$f(x_n) - \min(f) \leq \big[1 - \alpha\mu(2 - \alpha L)\big]^n \big(f(x_0) - \min(f)\big).$$

Questions

- How does this compare with the example?
- What value of α gives the best rate?
- How does this relate to the homework assignment?

Uniform growth and Łojasiewicz inequalities

Proposition

If f is convex, has minimizers, and has quadratic growth:

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

for all $x \in \mathbb{R}^N$, then $c^2\big(f(x) - \min(f)\big) \leq \|\nabla f(x)\|^2$ for all $x \in \mathbb{R}^N$.

Uniform growth and Łojasiewicz inequalities

Proposition

If f is convex, has minimizers, and has quadratic growth:

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

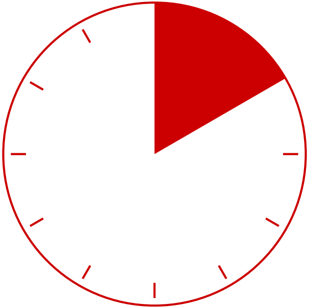
for all $x \in \mathbb{R}^N$, then $c^2\big(f(x) - \min(f)\big) \leq \|\nabla f(x)\|^2$ for all $x \in \mathbb{R}^N$.

Exercise

Suppose f has minimizers, and there is $p > 0$ such that, for all $x \in \mathbb{R}^N$, we have $c\,\mathrm{dist}(x,S)^p \leq f(x) - \min(f)$.

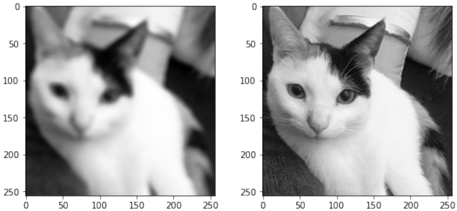
1. Which values of p are compatible with f being convex?
2. How fast does the gradient method converge in that case?

Break



Computational exercise

A simple example of image processing



Deblurring

In-painting

Warm up

We will solve a problem of the form

$$\min_{x\in\mathbb{R}^N}\left\{\frac{1}{2}\|Ax-b\|^2+\rho\|Lx\|_1\right\},$$

but we do not have the tools yet.

Warm up

We will solve a problem of the form

$$\min_{x\in\mathbb{R}^N} \left\{ \frac{1}{2}\|Ax-b\|^2 + \rho\|Lx\|_1 \right\},$$

but we do not have the tools yet.

Let us begin with something simpler: set $N, M \in \mathbb{N}$, $A \in \mathbb{R}^{M\times N}$, $b \in \mathbb{R}^M$, and consider the least squares minimization problem:

$$\min_{x\in\mathbb{R}^N} \left\{ \frac{1}{2}\|Ax-b\|^2 \right\}.$$

Exercise

1. Define a (pseudo)random matrix A of size $M \times N$.

Exercise

1. Define a (pseudo)random matrix A of size $M \times N$.
2. Write the optimality conditions and find a formula for a solution $\bar{x}$ (it depends on whether $b \in \operatorname{ran}(A)$ or not).

Exercise

1. Define a (pseudo)random matrix A of size $M \times N$.
2. Write the optimality conditions and find a formula for a solution $\bar{x}$ (it depends on whether $b \in \operatorname{ran}(A)$ or not).
3. Set $b = A\mathbf{1}$ and compute $\bar{x}$ with the formula.

Exercise

1. Define a (pseudo)random matrix A of size $M \times N$.
2. Write the optimality conditions and find a formula for a solution $\bar{x}$ (it depends on whether $b \in \operatorname{ran}(A)$ or not).
3. Set $b = A\mathbf{1}$ and compute $\bar{x}$ with the formula.
4. Now compute $\bar{x}$ by applying gradient descent with constant (in the appropriate range) and step sizes, as well as using exact minimization and backtracking.

Exercise

1. Define a (pseudo)random matrix A of size $M \times N$.
2. Write the optimality conditions and find a formula for a solution $\bar{x}$ (it depends on whether $b \in \operatorname{ran}(A)$ or not).
3. Set $b = A\mathbf{1}$ and compute $\bar{x}$ with the formula.
4. Now compute $\bar{x}$ by applying gradient descent with constant (in the appropriate range) and step sizes, as well as using exact minimization and backtracking.
5. Compare the execution times for different combinations of M and N.

- Constant: $\alpha_n \equiv \alpha \in (0, 2/L)$.
- Vanishing: $\alpha_n \to 0$, $\sum \alpha_n = \infty$.
- Exact minimization: $\min_{\alpha>0} f(x_n - \alpha \nabla f(x_n))$.
- Limited minimization: $\min_{\alpha\in(0,A]} f(x_n - \alpha \nabla f(x_n))$.
- Backtracking: Pick $\alpha_0 > 0$ and $\sigma, \beta \in (0,1)$, and set

$$m_n := \min \left\{ j \in \mathbb{N} \ : f\left(x_n - \alpha_0 \beta^j \nabla f(x_n)\right) \leq \alpha_0 \beta^j \sigma \|\nabla f(x_n)\|^2 \right\},$$

and then define $x_{n+1} = x_n - \alpha_0 \beta^{m_n} \nabla f(x_n)$.