Iterative Algorithms in Optimization, Variational Analysis and Fixed Point Theory

Unit 02: Smooth optimization and gradient descent.

university of groningen

Descent directions

Let $A \subset \mathbb{R}^N$ be nonempty and open, and let $f : A \to \mathbb{R}$. A vector $d \neq 0$ is a descent direction for f at x if there is $\Gamma > 0$ such that

$$f(x + \gamma d) < f(x) \qquad \text{for all } \gamma \in (0, \Gamma).$$

The numbers $\gamma \in (0, \Gamma)$ are descent step sizes.

Descent directions

Let $A \subset \mathbb{R}^N$ be nonempty and open, and let $f : A \to \mathbb{R}$. A vector $d \neq 0$ is a descent direction for f at x if there is $\Gamma > 0$ such that

$$f(x + \gamma d) < f(x) \qquad \text{for all } \gamma \in (0, \Gamma).$$

The numbers $\gamma \in (0, \Gamma)$ are descent step sizes.

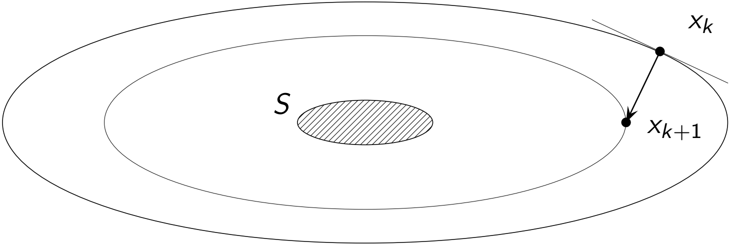
Remark

If f is differentiable at x and $\nabla f(x) \neq 0$, then $-\nabla f(x)$ is the steepest descent direction, and the set of all descent directions is the halfspace

$$\left\{ d \in \mathbb{R}^N : \nabla f(x) \cdot d < 0 \right\}.$$

Gradient descent

From $x_0 \in \mathbb{R}^N$, iterate $x_{k+1} = x_k - \gamma \nabla f(x_k)$



L-smoothness

A differentiable function $f : A \subset \mathbb{R}^N \to \mathbb{R}$ is L-smooth, with $L > 0$, if

$$\|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\|$$

for all $x, y \in A$.

L-smoothness

A differentiable function $f : A \subset \mathbb{R}^N \to \mathbb{R}$ is L-smooth, with $L > 0$, if

$$\|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\|$$

for all $x, y \in A$.

Proposition (Descent Lemma)

If f is L-smooth and A is convex, then

$$|f(y) - f(x) - \nabla f(x) \cdot (y - x)| \leq \frac{L}{2}\|x - y\|^2$$

for all $x, y \in A$.

Estimating descent step sizes

Proposition

Let $A \neq \emptyset$ be open and convex, and let $f : A \to \mathbb{R}$ be $L-$smooth. Then,

$$f\left(x - \gamma \nabla f(x)\right) \leq f(x) + \gamma \left(\frac{\gamma L}{2} - 1 \right) \|\nabla f(x)\|^2$$

for all sufficiently small $\gamma > 0$.[a]

Estimating descent step sizes

Proposition

Let $A \neq \emptyset$ be open and convex, and let $f : A \to \mathbb{R}$ be $L-$smooth. Then,

$$f\left(x - \gamma \nabla f(x)\right) \leq f(x) + \gamma \left(\frac{\gamma L}{2} - 1 \right) \|\nabla f(x)\|^2$$

for all sufficiently small $\gamma > 0$.[a] *In particular, if A is large enough, every $\gamma \in (0, \frac{2}{L})$ is a descent step size.*

---

[a] Sufficiently small to remain in A.

Estimating descent step sizes

Proposition

Let $A \neq \emptyset$ be open and convex, and let $f : A \to \mathbb{R}$ be $L-$smooth. Then,

$$f\left(x - \gamma \nabla f(x)\right) \leq f(x) + \gamma \left(\frac{\gamma L}{2} - 1 \right) \|\nabla f(x)\|^2$$

for all sufficiently small $\gamma > 0$.[a] In particular, if A is large enough, every $\gamma \in (0, \frac{2}{L})$ is a descent step size.

[a] Sufficiently small to remain in A.

There are reasons to take $\gamma = \frac{1}{L}$.

Convergence and complexity

Proposition

Let $f : \mathbb{R}^N \to \mathbb{R}$ be L-smooth and bounded from below. Iterate $x_{k+1} = x_k - \gamma_k \nabla f(x_k)$, where $\inf_{k\geq 0} \gamma_k(2 - \gamma_k L) > 0$.

Convergence and complexity

Proposition

Let $f : \mathbb{R}^N \to \mathbb{R}$ *be L-smooth and bounded from below. Iterate* $x_{k+1} = x_k - \gamma_k \nabla f(x_k)$, *where* $\inf_{k \geq 0} \gamma_k (2 - \gamma_k L) > 0$. *Then,*

1. $\exists \lim_{k \to \infty} f(x_k) \in \mathbb{R}$, *and* $\lim_{k \to \infty} \|\nabla f(x_k)\| = 0$.

Convergence and complexity

Proposition

Let $f : \mathbb{R}^N \to \mathbb{R}$ be L-smooth and bounded from below. Iterate $x_{k+1} = x_k - \gamma_k \nabla f(x_k)$, where $\inf_{k \geq 0} \gamma_k (2 - \gamma_k L) > 0$. Then,

1. $\exists \lim\limits_{k\to\infty} f(x_k) \in \mathbb{R}$, *and* $\lim\limits_{k\to\infty} \|\nabla f(x_k)\| = 0$.
2. *Cluster points are critical: if $x_{j_k} \to \hat{x}$, then $\nabla f(\hat{x}) = 0$.*

Convergence and complexity

Proposition

Let $f : \mathbb{R}^N \to \mathbb{R}$ be L-smooth and bounded from below. Iterate $x_{k+1} = x_k - \gamma_k \nabla f(x_k)$, *where* $\inf_{k \geq 0} \gamma_k (2 - \gamma_k L) > 0$. *Then,*

1. $\exists \lim_{k \to \infty} f(x_k) \in \mathbb{R}$, *and* $\lim_{k \to \infty} \|\nabla f(x_k)\| = 0$.
2. *Cluster points are critical: if* $x_{j_k} \to \hat{x}$, *then* $\nabla f(\hat{x}) = 0$.
3. *If f has no critical points, then* $\lim_{k \to \infty} \|x_k\| = +\infty$.

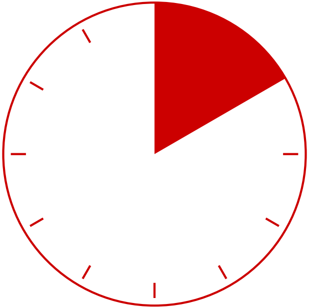
Convergence and complexity

Proposition

Let $f : \mathbb{R}^N \to \mathbb{R}$ *be L-smooth and bounded from below. Iterate* $x_{k+1} = x_k - \gamma_k \nabla f(x_k)$, *where* $\inf_{k \geq 0} \gamma_k (2 - \gamma_k L) > 0$. *Then,*

1. $\exists \lim\limits_{k \to \infty} f(x_k) \in \mathbb{R}$, *and* $\lim\limits_{k \to \infty} \|\nabla f(x_k)\| = 0$.
2. *Cluster points are critical: if* $x_{j_k} \to \hat{x}$, *then* $\nabla f(\hat{x}) = 0$.
3. *If* f *has no critical points, then* $\lim\limits_{k \to \infty} \|x_k\| = +\infty$.
4. *There is* $C > 0$ *such that* $\min \left\{ \|\nabla f(x_j)\| : 1 \leq j \leq k \right\} \leq \dfrac{C}{\sqrt{k}}$.

Break



Other step size selection criteria

Here, d_k is a descent direction for f at x_k

- Asymptotically vanishing step sizes: $\gamma_k \to 0$ and $\sum \gamma_k = +\infty$.

Other step size selection criteria

Here, d_k is a descent direction for f at x_k

- Asymptotically vanishing step sizes: $\gamma_k \to 0$ and $\sum \gamma_k = +\infty$.
- Exact or limited minimization:
  - $\gamma_k := \text{Argmin}\, f(x_k + \gamma d_k)$; with either $\gamma > 0$ or $\gamma \in (0, \gamma_+]$.

Other step size selection criteria

Here, d_k is a descent direction for f at x_k

- Asymptotically vanishing step sizes: $\gamma_k \to 0$ and $\sum \gamma_k = +\infty$.
- Exact or limited minimization:
  - $\gamma_k := \mathrm{Argmin}\, f(x_k + \gamma d_k)$; with either $\gamma > 0$ or $\gamma \in (0, \gamma_+]$.
- Line search:
  - Armijo: $f(x_k + \gamma_k d_k) \leq f(x_k) + \sigma \gamma_k \nabla f(x_k) \cdot d_k \quad (A)$.

Other step size selection criteria

Here, d_k is a descent direction for f at x_k

- Asymptotically vanishing step sizes: $\gamma_k \to 0$ and $\sum \gamma_k = +\infty$.
- Exact or limited minimization:
  - $\gamma_k := \text{Argmin}\, f(x_k + \gamma d_k)$; with either $\gamma > 0$ or $\gamma \in (0, \gamma_+]$.
- Line search:
  - Armijo: $f(x_k + \gamma_k d_k) \leq f(x_k) + \sigma \gamma_k \nabla f(x_k) \cdot d_k \quad (A)$.
  - Goldstein: (A), plus $f(x_k + \gamma_k d_k) \geq f(x_k) + (1 - \sigma)\gamma_k \nabla f(x_k) \cdot d_k$.

Other step size selection criteria

Here, d_k is a descent direction for f at x_k

- Asymptotically vanishing step sizes: $\gamma_k \to 0$ and $\sum \gamma_k = +\infty$.
- Exact or limited minimization:
  - $\gamma_k := \operatorname{Argmin} f(x_k + \gamma d_k)$; with either $\gamma > 0$ or $\gamma \in (0, \gamma_+]$.
- Line search:
  - Armijo: $f(x_k + \gamma_k d_k) \leq f(x_k) + \sigma \gamma_k \nabla f(x_k) \cdot d_k \quad (A)$.
  - Goldstein: (A), plus $f(x_k + \gamma_k d_k) \geq f(x_k) + (1 - \sigma)\gamma_k \nabla f(x_k) \cdot d_k$.
  - Wolfe: (A), plus $\nabla f(x_k + \gamma_k d_k) \cdot d_k \geq \tau \nabla f(x_k) \cdot d_k$.
  - Strong Wolfe: (A), plus $|\nabla f(x_k + \gamma_k d_k) \cdot d_k| \leq \tau |\nabla f(x_k) \cdot d_k|$.

Convex functions

A function $f : A \subset \mathbb{R}^N \to \mathbb{R}$ is convex if A is convex and

$$f\big(\lambda x + (1-\lambda)y\big) \leq \lambda f(x) + (1-\lambda)f(y) \quad \forall x, y \in A,\ \lambda \in (0,1).$$

Convex functions

A function $f : A \subset \mathbb{R}^N \to \mathbb{R}$ is convex if A is convex and

$$f\big(\lambda x + (1-\lambda)y\big) \leq \lambda f(x) + (1-\lambda) f(y) \quad \forall x, y \in A,\ \lambda \in (0,1).$$

Critical points of convex functions are global minimizers.

Convex functions

A function $f : A \subset \mathbb{R}^N \to \mathbb{R}$ is convex if A is convex and

$$f\big(\lambda x + (1-\lambda)y\big) \leq \lambda f(x) + (1-\lambda) f(y) \quad \forall x, y \in A,\ \ \lambda \in (0,1).$$

Critical points of convex functions are global minimizers.

If $f : A \to \mathbb{R}$ is convex and differentiable, then

$$f(y) \geq f(x) + \nabla f(x) \cdot (y - x) \qquad \forall x, y \in A.$$

Convex functions

A function $f : A \subset \mathbb{R}^N \to \mathbb{R}$ is convex if A is convex and

$$f\big(\lambda x + (1-\lambda)y\big) \leq \lambda f(x) + (1-\lambda)f(y) \quad \forall x, y \in A,\ \ \lambda \in (0,1).$$

Critical points of convex functions are global minimizers.

If $f : A \to \mathbb{R}$ is convex and differentiable, then

$$f(y) \geq f(x) + \nabla f(x) \cdot (y - x) \qquad \forall x, y \in A.$$

If $f : \mathbb{R}^N \to \mathbb{R}$ is convex and L-smooth, then

$$\frac{1}{L}\|\nabla f(y) - \nabla f(x)\|^2 \leq \big(\nabla f(y) - \nabla f(x)\big) \cdot (y - x) \qquad \forall x, y \in \mathbb{R}^N.$$

Convex functions and gradient descent

Theorem

Let $f : \mathbb{R}^N \to \mathbb{R}$ *be convex and L-smooth. Iterate* $x_{k+1} = x_k - \gamma \nabla f(x_k)$ *with* $0 < \gamma < \frac{2}{L}$.

Convex functions and gradient descent

Theorem

Let $f : \mathbb{R}^N \to \mathbb{R}$ be convex and L-smooth. Iterate $x_{k+1} = x_k - \gamma \nabla f(x_k)$ with $0 < \gamma < \frac{2}{L}$.

1. $\lim\limits_{k\to\infty} f(x_k) = \inf(f)$.

Convex functions and gradient descent

Theorem

Let $f : \mathbb{R}^N \to \mathbb{R}$ be convex and L-smooth. Iterate $x_{k+1} = x_k - \gamma \nabla f(x_k)$ *with* $0 < \gamma < \frac{2}{L}$.

1. $\lim_{k\to\infty} f(x_k) = \inf(f)$.
2. *If f has minimizers, then x_k converges to one of them, and*

$$f(x_k) - \min(f) \leq \frac{\mathrm{dist}(x_0, S)^2}{\gamma(2 - \gamma L)k}.$$

Moreover, $\lim_{k\to\infty} k\left[f(x_k) - \min(f)\right] = 0.$

Iterative Algorithms in Optimization, Variational Analysis and Fixed Point Theory

Unit 02: Smooth optimization and gradient descent.

university of groningen

Quadratic functions

In this course, we will study quadratic functions of the form

$$Q(x) = \frac{1}{2}\|Ax - b\|^2,$$

where $A \in \mathbb{R}^{M \times N}$ and $b \in \mathbb{R}^M$

Quadratic functions

In this course, we will study quadratic functions of the form

$$Q(x) = \frac{1}{2}\|Ax - b\|^2,$$

where $A \in \mathbb{R}^{M\times N}$ and $b \in \mathbb{R}^M$, which we can also write as

$$Q(x) = \frac{1}{2}x^T Px + c^T x + d,$$

where $P = A^T A \in \mathbb{R}^{N\times N}$ is symmetric and positive semidefinite, $c = -A^T b \in \mathbb{R}^N$ and $d = \frac{1}{2}\|b\|^2 \in \mathbb{R}$.

Quadratic functions

In this course, we will study quadratic functions of the form

$$Q(x) = \frac{1}{2}\|Ax - b\|^2,$$

where $A \in \mathbb{R}^{M \times N}$ and $b \in \mathbb{R}^M$, which we can also write as

$$Q(x) = \frac{1}{2}x^T Px + c^T x + d,$$

where $P = A^T A \in \mathbb{R}^{N \times N}$ is symmetric and positive semidefinite, $c = -A^T b \in \mathbb{R}^N$ and $d = \frac{1}{2}\|b\|^2 \in \mathbb{R}$.

Such functions are always convex and bounded from below.

A tiny bit of linear algebra and geometry

1. $\ker(P) = \ker(A)$.

A tiny bit of linear algebra and geometry

1. $\ker(P) = \ker(A)$.
2. $\operatorname{ran}(P) = \operatorname{ran}(A^T) = \ker(A)^\perp$ and $\mathbb{R}^N = \ker(P) \oplus \operatorname{ran}(P)$.

A tiny bit of linear algebra and geometry

1. $\ker(P) = \ker(A)$.
2. $\operatorname{ran}(P) = \operatorname{ran}(A^T) = \ker(A)^\perp$ and $\mathbb{R}^N = \ker(P) \oplus \operatorname{ran}(P)$.
3. P is diagonalizable and its eigenvalues are all (real and) nonnegative.

A tiny bit of linear algebra and geometry

1. $\ker(P) = \ker(A)$.
2. $\mathrm{ran}(P) = \mathrm{ran}(A^T) = \ker(A)^\perp$ and $\mathbb{R}^N = \ker(P) \oplus \mathrm{ran}(P)$.
3. P is diagonalizable and its eigenvalues are all (real and) nonnegative.
4. If μ is the smallest eigenvalue of P, then $\|Ax\|^2 = x^T P x \geq \mu \|x\|^2$ for all $x \in \mathbb{R}^N$. Moreover,

$$\ker(A) = \{0\} \Leftrightarrow \mu > 0 \Leftrightarrow P \text{ is invertible.}$$

A tiny bit of linear algebra and geometry

1. $\ker(P) = \ker(A)$.
2. $\mathrm{ran}(P) = \mathrm{ran}(A^T) = \ker(A)^\perp$ and $\mathbb{R}^N = \ker(P) \oplus \mathrm{ran}(P)$.
3. P is diagonalizable and its eigenvalues are all (real and) nonnegative.
4. If μ is the smallest eigenvalue of P, then $\|Ax\|^2 = x^T P x \geq \mu \|x\|^2$ for all $x \in \mathbb{R}^N$. Moreover,

$$\ker(A) = \{0\} \Leftrightarrow \mu > 0 \Leftrightarrow P \text{ is invertible.}$$

5. $\|Ax\|^2 = x^T P x \leq L\|x\|^2$, where L is the largest eigenvalue of P.

A tiny bit of linear algebra and geometry

1. $\ker(P) = \ker(A)$.
2. $\operatorname{ran}(P) = \operatorname{ran}(A^T) = \ker(A)^\perp$ and $\mathbb{R}^N = \ker(P) \oplus \operatorname{ran}(P)$.
3. P is diagonalizable and its eigenvalues are all (real and) nonnegative.
4. If μ is the smallest eigenvalue of P, then $\|Ax\|^2 = x^T P x \geq \mu \|x\|^2$ for all $x \in \mathbb{R}^N$. Moreover,

$$\ker(A) = \{0\} \Leftrightarrow \mu > 0 \Leftrightarrow P \text{ is invertible.}$$

5. $\|Ax\|^2 = x^T P x \leq L\|x\|^2$, where L is the largest eigenvalue of P.

If $\mu > 0$, the ratio $\kappa = \frac{L}{\mu}$ is the condition number of P.

Consequences for Q and for gradient descent

The gradient of Q is $\nabla Q(x) = A^T(Ax - b) = Px + c$.

Consequences for Q and for gradient descent

The gradient of Q is $\nabla Q(x) = A^T(Ax - b) = Px + c$.

If P is invertible, then $\min(Q) = 0$ and Q has exactly one minimizer

$$\hat{x} = -P^{-1}c = (A^T A)^{-1} A^T b = A^\dagger b.$$

The matrix $A^\dagger$ is called the Moore-Penrose pseudonverse of A.

Consequences for Q and for gradient descent

The gradient of Q is $\nabla Q(x) = A^T(Ax - b) = Px + c$.

If P is invertible, then $\min(Q) = 0$ and Q has exactly one minimizer

$$\hat{x} = -P^{-1}c = (A^TA)^{-1}A^Tb = A^\dagger b.$$

The matrix $A^\dagger$ is called the Moore-Penrose pseudonverse of A.

Proposition

Iterate $x_{k+1} = x_k - \gamma_k \nabla Q(x_k)$, *with* $\gamma_k \equiv \frac{2}{L+\mu}$. *For each* $k \geq 1$, *we have*

$$\|x_k - \hat{x}\| \leq \left(\frac{L-\mu}{L+\mu}\right)^k \|x_0 - \hat{x}\| = \left(\frac{\kappa - 1}{\kappa + 1}\right)^k \|x_0 - \hat{x}\|.$$

General case

If P is not invertible, then

$$\min(Q) = \frac{1}{2}\,\mathrm{dist}\left(b, \mathrm{ran}(A)\right)^2,$$

and $u \in S$ if, and only if, $Au = \mathrm{Proj}_{\mathrm{ran}(A)}\, b$.

General case

If P is not invertible, then

$$\min(Q) = \frac{1}{2} \operatorname{dist}\left(b, \operatorname{ran}(A)\right)^2,$$

and $u \in S$ if, and only if, $Au = \operatorname{Proj}_{\operatorname{ran}(A)} b$.

We denote by λ the smallest positive eigenvalue of $A^T A$.

General case

If P is not invertible, then

$$\min(Q) = \frac{1}{2}\operatorname{dist}\left(b, \operatorname{ran}(A)\right)^2,$$

and $u \in S$ if, and only if, $Au = \operatorname{Proj}_{\operatorname{ran}(A)} b$.

We denote by λ the smallest positive eigenvalue of $A^T A$.

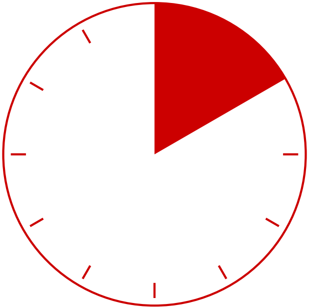
Proposition

Iterate $x_{k+1} = x_k - \gamma_k \nabla Q(x_k)$, *with* $\gamma_k \equiv \frac{2}{L+\lambda}$. *For each* $k \geq 1$, *we have*

$$\|x_k - \hat{u}\| \leq \left(\frac{L-\lambda}{L+\lambda}\right)^k \|x_0 - \hat{u}\|,$$

where $\hat{u} = \operatorname{Proj}_S x_0$.

Break



From quadratic to non quadratic functions I

The function $Q(x) = \frac{1}{2}\|Ax - b\|^2$ satisfies

$$Q\big(\theta x + (1-\theta)y\big) = \theta Q(x) + (1-\theta)Q(y) - \frac{\theta(1-\theta)}{2}\|A(x-y)\|^2,$$

for all $x, y \in \mathbb{R}^N$ and $\theta \in (0,1)$.

From quadratic to non quadratic functions I

The function $Q(x) = \frac{1}{2}\|Ax - b\|^2$ satisfies

$$Q\big(\theta x + (1-\theta)y\big) = \theta Q(x) + (1-\theta)Q(y) - \frac{\theta(1-\theta)}{2}\|A(x-y)\|^2,$$

for all $x, y \in \mathbb{R}^N$ and $\theta \in (0,1)$. If $\ker(A) = \{0\}$, then

$$Q\big(\theta x + (1-\theta)y\big) \leq \theta Q(x) + (1-\theta)Q(y) - \mu\frac{\theta(1-\theta)}{2}\|x-y\|^2.$$

A function $f : \mathbb{R}^N \to \mathbb{R}$ that satisfies such an inequality is called $\mu-$strongly convex. It must be continuous and must have exactly one minimizer.

Polyak-Łojasiewicz inequality

If f is $\mu-$ strongly convex, it satisfies the Polyak-Łojasiewicz inequality with constant μ:

$$2\mu\big(f(x) - \min(f)\big) \leq \|\nabla f(x)\|^2$$

for all $x \in \mathbb{R}^N$.

Polyak-Łojasiewicz inequality

If f is $\mu-$ strongly convex, it satisfies the Polyak-Łojasiewicz inequality with constant μ:

$$2\mu\big(f(x) - \min(f)\big) \leq \|\nabla f(x)\|^2$$

for all $x \in \mathbb{R}^N$.

A function satisfying the Polyak-Łojasiewicz inequality may have multiple minimizers.

Polyak-Łojasiewicz inequality

If f is $\mu-$ strongly convex, it satisfies the Polyak-Łojasiewicz inequality with constant μ:

$$2\mu\big(f(x) - \min(f)\big) \leq \|\nabla f(x)\|^2$$

for all $x \in \mathbb{R}^N$.

A function satisfying the Polyak-Łojasiewicz inequality may have multiple minimizers.

Example

Show that $Q(x) = \frac{1}{2}\|Ax - b\|^2$ satisfies a Polyak-Łojasiewicz inequality, even if $\ker(A)$ is nontrivial.

Polyak-Łojasiewicz inequality and gradient descent

Proposition

Let $f : \mathbb{R}^N \to \mathbb{R}$ be $L-$smooth and satisfy the Polyak-Łojasiewicz inequality with constant $\mu > 0$. Iterate $x_{k+1} = x_k - \frac{1}{L}\nabla f(x_k)$. Then,

$$f(x_k) - \min(f) \leq \left(1 - \frac{\mu}{L}\right)^k \left(f(x_0) - \min(f)\right),$$

for all $k \geq 1$.

Polyak-Łojasiewicz inequality and gradient descent

Proposition

Let $f : \mathbb{R}^N \to \mathbb{R}$ be $L-$smooth and satisfy the Polyak-Łojasiewicz inequality with constant $\mu > 0$. Iterate $x_{k+1} = x_k - \frac{1}{L}\nabla f(x_k)$. Then,

$$f(x_k) - \min(f) \leq \left(1 - \frac{\mu}{L}\right)^k \left(f(x_0) - \min(f)\right),$$

for all $k \geq 1$.

Remark

If we define the condition number of f as $\kappa = \frac{L}{\mu}$, then $1 - \frac{\mu}{L} = \frac{\kappa - 1}{\kappa}$. How does this compare with the quadratic case? How is the condition number related to the Hessian of f?

Iterative Algorithms in Optimization, Variational Analysis and Fixed Point Theory

Unit 02: Smooth optimization and gradient descent.

The conjugate gradient method

Consider, as before, the problem of minimizing

$$Q(x) = \frac{1}{2}\|Ax - b\|^2,$$

where $A \in \mathbb{R}^{M\times N}$ and $b \in \mathbb{R}^M$. Assume $\ker(A) = \{0\}$.

The conjugate gradient method

Consider, as before, the problem of minimizing

$$Q(x) = \frac{1}{2}\|Ax - b\|^2,$$

where $A \in \mathbb{R}^{M\times N}$ and $b \in \mathbb{R}^M$. Assume $\ker(A) = \{0\}$.

The conjugate gradient method iterates

$$x_{k+1} = x_k + \gamma_k d_k$$

for convenient choices of γ_k and d_k, in such a way that $\{d_0, \dots, d_{N-1}\}$ is a basis of $\mathbb{R}^N$ and x_k minimizes Q on the affine subspace

$$x_0 + \mathrm{span}\{d_0, \dots, d_{n-1}\}.$$

The (unique) solution is found in at most N steps.

The conjugate gradient method

Start with $x_0 \in \mathbb{R}^N$ and compute $d_0 = -\nabla Q(x_0) = -Px_0 - c$.

The conjugate gradient method

Start with $x_0 \in \mathbb{R}^N$ and compute $d_0 = -\nabla Q(x_0) = -Px_0 - c$.

At step k, we know x_k and d_k, and we compute $g_k = \nabla Q(x_k)$.

The conjugate gradient method

Start with $x_0 \in \mathbb{R}^N$ and compute $d_0 = -\nabla Q(x_0) = -Px_0 - c$.

At step k, we know x_k and d_k, and we compute $g_k = \nabla Q(x_k)$.

If $g_k = 0$, we stop because we have found the solution. Otherwise, we compute γ_k and then x_{k+1} by the exact minimization rule:

$$\gamma_k = \frac{\|g_k\|^2}{\|d_k\|_P^2} \qquad \text{and} \qquad x_{k+1} = x_k + \gamma_k d_k.$$

The conjugate gradient method

Start with $x_0 \in \mathbb{R}^N$ and compute $d_0 = -\nabla Q(x_0) = -Px_0 - c$.

At step k, we know x_k and d_k, and we compute $g_k = \nabla Q(x_k)$.

If $g_k = 0$, we stop because we have found the solution. Otherwise, we compute γ_k and then x_{k+1} by the exact minimization rule:

$$\gamma_k = \frac{\|g_k\|^2}{\|d_k\|_P^2} \qquad \text{and} \qquad x_{k+1} = x_k + \gamma_k d_k.$$

Finally, update

$$g_{k+1} = Px_{k+1} + c = g_k + P(x_{k+1} - x_k) = g_k + \gamma_k P d_k$$

The conjugate gradient method

Start with $x_0 \in \mathbb{R}^N$ and compute $d_0 = -\nabla Q(x_0) = -Px_0 - c$.

At step k, we know x_k and d_k, and we compute $g_k = \nabla Q(x_k)$.

If $g_k = 0$, we stop because we have found the solution. Otherwise, we compute γ_k and then x_{k+1} by the exact minimization rule:

$$\gamma_k = \frac{\|g_k\|^2}{\|d_k\|_P^2} \qquad \text{and} \qquad x_{k+1} = x_k + \gamma_k d_k.$$

Finally, update

$$g_{k+1} = Px_{k+1} + c = g_k + P(x_{k+1} - x_k) = g_k + \gamma_k P d_k,$$

and

$$d_{k+1} = -g_{k+1} + \beta_k d_k, \qquad \text{with} \qquad \beta_k = \frac{\|g_{k+1}\|^2}{\|g_k\|^2}.$$

Nonlinear extensions

Let $f : \mathbb{R}^N \to \mathbb{R}$ be differentiable, but not necessarily quadratic.

Nonlinear extensions

Let $f : \mathbb{R}^N \to \mathbb{R}$ be differentiable, but not necessarily quadratic.

Given $x_0 \in \mathbb{R}^N$, set $d_0 = -\nabla f(x_0)$.

Nonlinear extensions

Let $f : \mathbb{R}^N \to \mathbb{R}$ be differentiable, but not necessarily quadratic.

Given $x_0 \in \mathbb{R}^N$, set $d_0 = -\nabla f(x_0)$.

At step k, we know x_k and d_k, and we compute $g_k = \nabla Q(x_k)$.

Nonlinear extensions

Let $f : \mathbb{R}^N \to \mathbb{R}$ be differentiable, but not necessarily quadratic.

Given $x_0 \in \mathbb{R}^N$, set $d_0 = -\nabla f(x_0)$.

At step k, we know x_k and d_k, and we compute $g_k = \nabla Q(x_k)$.

If $g_k = 0$, we stop; otherwise, we compute (backtracking) γ_k satisfying

$$\begin{aligned} f(x_k + \gamma_k d_k) &\leq f(x_k) + \sigma \gamma_k \, g_k \cdot d_k \\ |\nabla f(x_k + \gamma_k d_k) \cdot d_k| &\leq \tau \, |g_k \cdot d_k|, \end{aligned}$$

with $0 < \sigma < \tau < 1$ (strong Wolfe conditions).

Nonlinear extensions, continued

Then, we update

$$\begin{aligned} x_{k+1} &= x_k + \gamma_k d_k \\ d_{k+1} &= -g_{k+1} + \beta_k d_k, \end{aligned}$$

where several choices for β_k are possible, such as:

- Fletcher-Reeves: $\frac{\|g_{k+1}\|^2}{\|g_k\|^2}$
- Polak-Ribière: $\frac{g_{k+1} \cdot (g_{k+1} - g_k)}{\|g_k\|^2}$
- Hestenes-Stiefel: $\frac{g_{k+1} \cdot (g_{k+1} - g_k)}{d_k \cdot (g_{k+1} - g_k)}$
- Dai-Yuan: $\frac{\|g_{k+1}\|^2}{d_k \cdot (g_{k+1} - g_k)}$