Introductory Course on Non-smooth Optimisation

Lecture 01

Gradient method

Outline

1. Unconstrained smooth optimisation
2. Descent methods
3. Gradient of convex functions
4. Gradient descent
5. Heavy-ball method
6. Nesterov's optimal schemes
7. Dynamical system

Outline

Convexity

Definition (Convex set)

A set $S \subset \mathbb{R}^n$ is convex if for any $\theta \in [0, 1]$ and two points $x, y \in S$,

$$\theta x + (1 - \theta)y \in S.$$

Definition (Convex function)

Function $F : \mathbb{R}^n \to \mathbb{R}$ is convex if $\mathrm{dom}(F)$ is convex and for all $x, y \in \mathrm{dom}(F)$ and $\theta \in [0, 1]$,

$$F(\theta x + (1 - \theta)y) \leq \theta F(x) + (1 - \theta)F(y).$$

Proper convex: $F(x) < +\infty$ at least for one x and $F(x) > -\infty$ for all x.

First-order condition: F is continuous differentiable

$$F(y) \geq F(x) + \langle \nabla F(x),\ y - x \rangle,\ \forall x, y \in \mathrm{dom}(F).$$

Second-order condition: if F is twice differentiable $\nabla^2 F(x) \succeq 0,\ \forall x \in \mathrm{dom}(F)$.

Unconstrained smooth optimisation

Problem

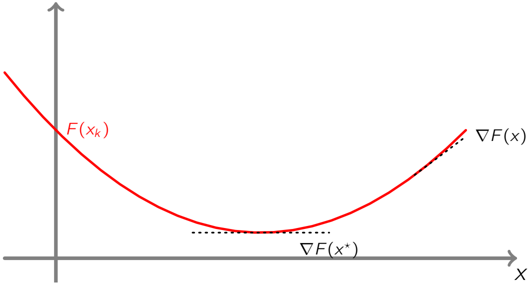
Unconstrained smooth optimisation

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

where $F : \mathbb{R}^n \to \mathbb{R}$ *is proper convex and smooth differentiable.*

Optimality condition: let $x^\star$ be an minimiser of $F(x)$, then

$$0 = \nabla F(x^\star).$$



Example (Quadratic programming)

General quadratic programming problem

$$\min_{x\in\mathbb{R}^n} \frac{1}{2}x^T Ax + b^T x + c,$$

where $A \in \mathbb{R}^{n\times n}$ is symmetric positive definite, $b \in \mathbb{R}^n$ and $c \in \mathbb{R}$.

Optimality condition:

$$0 = Ax^\star + b.$$

Special case: LSE

$$\|Ax - b\|^2 = x^T(A^T A)x - 2(A^T b)^T x + b^T b.$$

Optimality condition

$$A^T Ax^\star = A^T b.$$

Example: geometric programming

Example (Geometric programming)

General quadratic programming problem

$$\min_{x\in\mathbb{R}^n} \log\left(\textstyle\sum_{i=1}^m \exp(a_i^T x + b_i)\right).$$

Optimality condition:

$$0 = \frac{1}{\sum_{i=1}^m \exp(a_i^T x^\star + b_i)} \sum_{i=1}^m \exp(a_i^T x^\star + b_i) a_i.$$

Outline

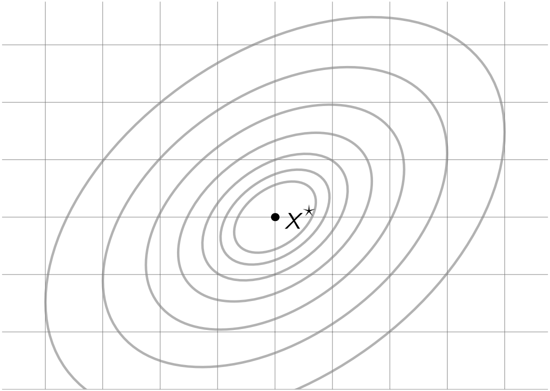
Problem

Problem (Unconstrained smooth optimisation)

Consider minising

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

where $F : \mathbb{R}^n \to \mathbb{R}$ is proper convex and smooth differentiable.



Problem

Problem (Unconstrained smooth optimisation)

Consider minising

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

where $F : \mathbb{R}^n \to \mathbb{R}$ *is proper convex and smooth differentiable.*

- The set of minimisers, *i.e.*

  $$\mathrm{Argmin}(F) = \{x \in \mathbb{R}^n : F(x) = \min_{x \in \mathbb{R}^n} F(x)\},$$

  is non-empty
- However, given $x^\star \in \mathrm{Argmin}(F)$, no closed form expression
- Iterative strategy to find one $x^\star \in \mathrm{Argmin}(F)$: start from x_0 and generate a train of sequsence $\{x_k\}_{k \in \mathbb{N}}$ such taht

  $$\lim_{k \to \infty} x_k = x^\star \in \mathrm{Argmin}(F)$$

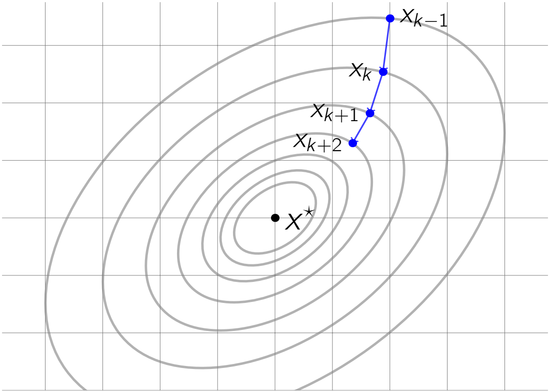
Problem

Problem (Unconstrained smooth optimisation)

Consider minising

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

where $F : \mathbb{R}^n \to \mathbb{R}$ *is proper convex and smooth differentiable.*



Descent methods

Iterative scheme

For each $k = 1, 2, ...,$ find $\gamma_k > 0$ and $d_k \in \mathbb{R}^n$ and then

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

where

- d_k is called search/descent direction
- γ_k is called step-size

Definition (Descent methods)

An algorithm is called descent method, if there holds

$$F(x_{k+1}) < F(x_k).$$

NB: if $x_k \in \text{Argmin}(F)$, then $x_{k+1} = x_k$...

Conditions

From convexity of F, we have

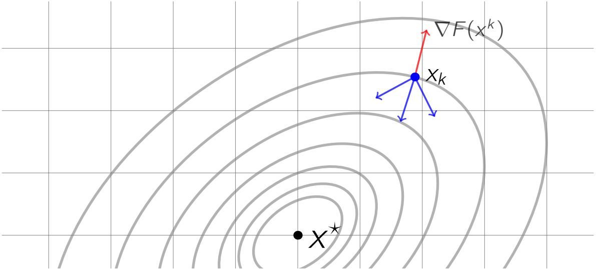
$$F(x_{k+1}) \geq F(x_k) + \langle \nabla F(x_k),\, x_{k+1} - x_k \rangle,$$

which gives

$$\langle \nabla F(x_k),\, x_{k+1} - x_k \rangle \geq 0 \implies F(x_{k+1}) \geq F(x_k).$$

Since $x_{k+1} - x_k = \gamma_k d_k$, the direction d_k should be such that

$$\langle \nabla F(x_k),\, d_k \rangle < 0.$$



General descent method

General descent method

initial: $x_0 \in \mathrm{dom}(F)$;

repeat:

1. Find a descent direction d_k
2. Choose a step-size γ_k: line search
3. Update $x_{k+1} = x_k + \gamma_k d_k$

until: stopping criterion is satisfied.

Stopping criterion: $\epsilon > 0$ is the tolerance,

- Function value: $F(x_{k+1}) - F(x_k) \leq \epsilon$ (can be time consuming)
- Sequence: $\|x_{k+1} - x_k\| \leq \epsilon$
- Optimality condition: $\|\nabla F(x_k)\| \leq \epsilon$

Exact line search

Exact line search

Choose γ_k such that $F(x)$ is minimised along the ray $x_k + \gamma d_k, \gamma > 0$:

$$\gamma_k = \mathrm{argmin}_{\gamma > 0} F(x_k + \gamma d_k).$$

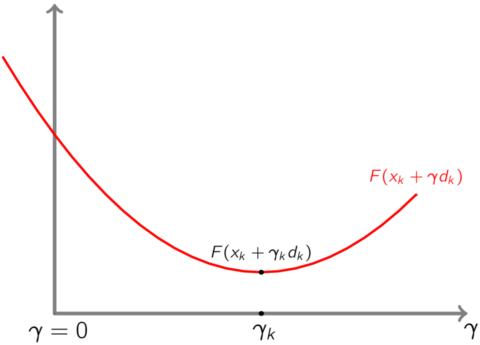
Exact line search

Exact line search

Choose γ_k such that $F(x)$ is minimised along the ray $x_k + \gamma d_k, \gamma > 0$:

$$\gamma_k = \text{argmin}_{\gamma > 0} F(x_k + \gamma d_k).$$

- Useful when the minimistion problem for γ_k is simple
- γ_k can be found analytically for special cases



Backtracking/inexact line search

Backtracking line search

Choose $\delta \in]0, 0.5[$ and $\beta \in]0, 1[$, let $\gamma = 1$

$$\text{while } F(x_k + \gamma d_k) > F(x_k) + \delta\gamma\langle \nabla F(x_k),\, d_k\rangle \,:\, \gamma = \beta\gamma.$$

Since d_k is a descent direction

$$\langle \nabla F(x_k),\, d_k\rangle < 0.$$

Stopping criterion for backtracking:

$$F(x_k + \gamma d_k) \le F(x_k) + \delta\gamma\langle \nabla F(x_k),\, d_k\rangle.$$

When γ is small enough

$$F(x_k + \gamma d_k) \approx F(x_k) + \gamma\langle \nabla F(x_k),\, d_k\rangle < F(x_k) + \delta\gamma\langle \nabla F(x_k),\, d_k\rangle,$$

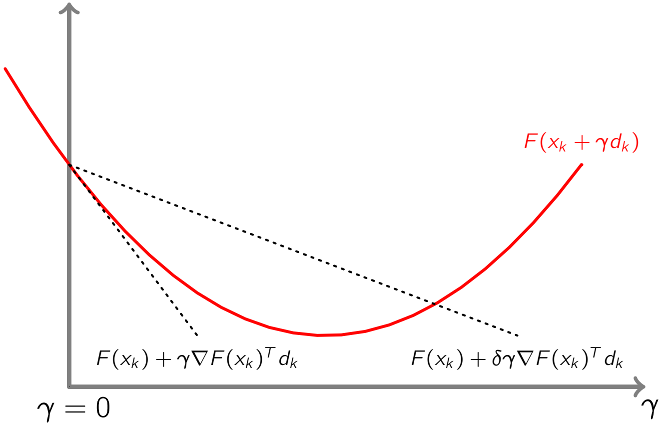
whcih means backtracking eventually will stop.

Backtracking/inexact line search

Backtracking line search

Choose $\delta \in]0, 0.5[$ and $\beta \in]0, 1[$, let $\gamma = 1$

$$\text{while } F(x_k + \gamma d_k) > F(x_k) + \delta\gamma\langle \nabla F(x_k), d_k\rangle : \gamma = \beta\gamma.$$



Outline

Monotonicity

Lemma (Monotonicity of gradient)

Let $F : \mathbb{R}^n \to \mathbb{R}$ be proper convex and smooth differentiable, then

$$\langle \nabla F(x) - \nabla F(y),\, x - y \rangle \geq 0,\ \forall x, y \in \mathrm{dom}(F).$$

C^1: proper convex and smooth differentiable functions on $\mathbb{R}^n$.

Monotonicity

Lemma (Monotonicity of gradient)

Let $F : \mathbb{R}^n \to \mathbb{R}$ be proper convex and smooth differentiable, then

$$\langle \nabla F(x) - \nabla F(y),\, x - y \rangle \geq 0,\ \forall x, y \in \mathrm{dom}(F).$$

C^1: proper convex and smooth differentiable functions on $\mathbb{R}^n$.

Proof Owing to convexity, given $x, y \in \mathrm{dom}(F)$, we have

$$F(y) \geq F(x) + \langle \nabla F(x),\, y - x \rangle$$

and

$$F(x) \geq F(y) + \langle \nabla F(y),\, x - y \rangle.$$

Summing them up yields

$$\langle \nabla F(x) - \nabla F(y),\, x - y \rangle \geq 0.$$

NB: Let $F \in C^1$, F is convex if and only if $\nabla F(x)$ is monotone.

Lipschitz continuous gradient

Definition (Lipschitz continuity)

The gradient of F is L-Lipschitz continuous if there exists $L > 0$ such that

$$\|\nabla F(x) - \nabla F(y)\| \leq L\|x - y\|, \ \forall x, y \in \mathrm{dom}(F).$$

C_L^1: proper convex functions with L-Lipschitz continuous gradient on $\mathbb{R}^n$.

Lipschitz continuous gradient

Definition (Lipschitz continuity)

The gradient of F is L-Lipschitz continuous if there exists $L > 0$ such that

$$\|\nabla F(x) - \nabla F(y)\| \leq L\|x - y\|, \ \forall x, y \in \mathrm{dom}(F).$$

C_L^1: proper convex functions with L-Lipschitz continuous gradient on $\mathbb{R}^n$.

If $F \in C_L^1$, then

$$H(x) \stackrel{\text{def}}{=} \frac{L}{2}\|x\|^2 - F(x)$$

is convex.

Hint: monotonicity of $\nabla H(x)$, *i.e.*

$$\begin{aligned}\langle \nabla H(x) - \nabla H(y),\, x - y\rangle &= L\|x - y\|^2 - \langle \nabla F(x) - \nabla F(y),\, x - y\rangle \\ &\geq L\|x - y\|^2 - L\|x - y\|^2 \\ &= 0.\end{aligned}$$

Descent lemma

Lemma (Descent lemma, quadratic upper bound)

Let $F \in C_L^1$, then there holds

$$F(y) \leq F(x) + \langle \nabla F(x),\, y - x \rangle + \frac{L}{2}\|y - x\|^2,\ \forall x, y \in \mathrm{dom}(F).$$

Descent lemma

Lemma (Descent lemma, quadratic upper bound)

Let $F \in C_L^1$, then there holds

$$F(y) \leq F(x) + \langle \nabla F(x), y - x \rangle + \frac{L}{2}\|y - x\|^2, \ \forall x, y \in \mathrm{dom}(F).$$

Proof. Define $H(t) = F(x + t(y - x))$, then

$$\begin{aligned}
F(y) - F(x) = H(1) - H(0) &= \int_0^1 \nabla H(t) \mathrm{d}t = \int_0^1 (y - x)^T \nabla F(x + t(y - x)) \mathrm{d}t \\
&\leq \int_0^1 (y - x)^T \nabla F(x) \mathrm{d}t + \int_0^1 \Big| (y - x)^T \big(\nabla F(x + t(y - x)) - \nabla F(x) \big) \Big| \mathrm{d}t \\
&\leq (y - x)^T \nabla F(x) + \int_0^1 \|y - x\| \|\nabla F(x + t(y - x)) - \nabla F(x)\| \mathrm{d}t \\
&\leq (y - x)^T \nabla F(x) + \|y - x\| \int_0^1 tL\|y - x\| \mathrm{d}t \\
&= (y - x)^T \nabla F(x) + \frac{L}{2}\|y - x\|^2.
\end{aligned}$$

NB: first-order condition of convexity for $H(x) \stackrel{\text{def}}{=} \frac{L}{2}\|x\|^2 - F(x)$.

Descent lemma: consequences

Corollary

Let $F \in C^1_L$ *and* $x^\star \in \mathrm{Argmin}(F)$*, then*

$$\frac{1}{2L}\|\nabla F(x)\|^2 \leq F(x) - F(x^\star) \leq \frac{L}{2}\|x - x^\star\|^2, \forall x \in \mathrm{dom}(F).$$

Descent lemma: consequences

Corollary

Let $F \in C_L^1$ and $x^\star \in \mathrm{Argmin}(F)$, then

$$\frac{1}{2L}\|\nabla F(x)\|^2 \leq F(x) - F(x^\star) \leq \frac{L}{2}\|x - x^\star\|^2, \forall x \in \mathrm{dom}(F).$$

Proof. Right-hand inequality: $\nabla F(x^\star) = 0$,

$$F(x) \leq F(x^\star) + \langle \nabla F(x^\star),\, x - x^\star \rangle + \frac{L}{2}\|x - x^\star\|^2,\ \forall x \in \mathrm{dom}(F).$$

Left-hand inequality:

$$\begin{aligned} F(x^\star) &\leq \min_{y \in \mathrm{dom}(F)} \left\{ F(x) + \langle \nabla F(x),\, y - x \rangle + \frac{L}{2}\|y - x\|^2 \right\} \\ &= F(x) - \frac{1}{2L}\|\nabla F(x)\|^2. \end{aligned}$$

The corresponding y is $y = x - \frac{1}{L}\nabla F(x)$.

Co-coercivity of gradient

Lemma (Co-coercivity)

Let $F \in C_L^1$, then

$$\langle x - y, \nabla F(x) - \nabla F(y) \rangle \geq \frac{1}{L} \|\nabla F(x) - \nabla F(y)\|^2.$$

- Co-coercivity implies Lipschitz continuity
- For $F \in C_L^1$, $H(x) \stackrel{\text{def}}{=} \frac{L}{2}\|x\|^2 - F(x)$

$$\begin{aligned} \text{Lipschitz continuity of } \nabla F &\implies \text{Convexity of } H(x) \\ &\implies \text{Co-coercivity of } \nabla F(x) \\ &\implies \text{Lipschitz continuity of } \nabla F \end{aligned}$$

Co-coercivity of gradient

Lemma (Co-coercivity)

Let $F \in C^1_L$, then

$$\langle x - y, \nabla F(x) - \nabla F(y) \rangle \geq \frac{1}{L} \|\nabla F(x) - \nabla F(y)\|^2.$$

Proof. Define $R(z) = F(z) - \langle \nabla F(x), z \rangle$, then $\nabla R(x) = 0$.

Recall the lemma

$$F \in C^1_L \text{ and } x^\star \in \mathrm{Argmin}(F)\text{: } \frac{1}{2L}\|\nabla F(x)\|^2 \leq F(x) - F(x^\star) \leq \frac{L}{2}\|x - x^\star\|^2.$$

Then we have

$$\begin{aligned} F(y) - F(x) - \langle \nabla F(x), y - x \rangle = R(y) - R(x) &\geq \frac{1}{2L}\|\nabla R(y)\|^2 \\ &= \frac{1}{2L}\|\nabla F(y) - \nabla F(x)\|^2. \end{aligned}$$

Similarly, define $S(z) = F(z) - \langle \nabla F(y), z \rangle$, then

$$F(x) - F(y) - \langle \nabla F(y), x - y \rangle = S(y) - S(x) \geq \frac{1}{2L}\|\nabla F(x) - \nabla F(y)\|^2.$$

Strongly convex function

Definition (Strongly convex function)

Function $F : \mathbb{R}^n \to \mathbb{R}$ is strongly convex if $\mathrm{dom}(F)$ is convex and for all $x, y \in \mathrm{dom}(F)$ and $\theta \in [0, 1]$, there exists $\alpha > 0$ such that

$$F(\theta x + (1-\theta)y) \leq \theta F(x) + (1-\theta)F(y) - \frac{\alpha}{2}\theta(1-\theta)\|x - y\|^2.$$

Strongly convex function

Definition (Strongly convex function)

Function $F : \mathbb{R}^n \to \mathbb{R}$ is strongly convex if $\mathrm{dom}(F)$ is convex and for all $x, y \in \mathrm{dom}(F)$ and $\theta \in [0, 1]$, there exists $\alpha > 0$ such that

$$F(\theta x + (1-\theta)y) \leq \theta F(x) + (1-\theta)F(y) - \frac{\alpha}{2}\theta(1-\theta)\|x-y\|^2.$$

F is strongly convex with parameter $\alpha > 0$ if

$$G(x) \stackrel{\text{def}}{=} F(x) - \frac{\alpha}{2}\|x\|^2$$

is convex.

Monotonicity:

$$\langle \nabla F(x) - \nabla F(y),\, x - y \rangle \geq \alpha \|x-y\|^2,\ \forall x, y \in \mathrm{dom}(F).$$

Second-order condition for strong convexity: if $F \in C^2$,

$$\nabla^2 F(x) \succeq \alpha \mathrm{Id},\ \forall x \in \mathrm{dom}(F).$$

Quadratic lower bound

Lemma (Quadratic lower bound)

Let $F \in C^1$ and strongly convex, then

$$F(y) \geq F(x) + \langle \nabla F(x),\, y - x \rangle + \frac{\alpha}{2}\|y - x\|^2,\ \forall x, y \in \mathrm{dom}(F).$$

Proof. first-order condition of convexity for $G(x) \stackrel{\text{def}}{=} F(x) - \frac{\alpha}{2}\|x\|^2$.

Quadratic lower bound

Lemma (Quadratic lower bound)

Let $F \in C^1$ and strongly convex, then

$$F(y) \geq F(x) + \langle \nabla F(x),\, y - x\rangle + \frac{\alpha}{2}\|y - x\|^2,\ \forall x, y \in \mathrm{dom}(F).$$

Proof. first-order condition of convexity for $G(x) \stackrel{\text{def}}{=} F(x) - \frac{\alpha}{2}\|x\|^2$.

Corollary

Let $F \in C^1$ be α-strongly convex and $x^\star \in \mathrm{Argmin}(F)$, then

$$\frac{\alpha}{2}\|x - x^\star\|^2 \leq F(x) - F(x^\star) \leq \frac{1}{2\alpha}\|\nabla F(x)\|^2,\forall x \in \mathrm{dom}(F).$$

Proof. Left-hand inequality: quadratic lower bound. Right-hand inequality:

$$\begin{aligned} F(x^\star) &\geq \min_{y \in \mathrm{dom}(F)} \left\{ F(x) + \langle \nabla F(x),\, y - x\rangle + \frac{\alpha}{2}\|y - x\|^2 \right\} \\ &= F(x) - \frac{1}{2\alpha}\|\nabla F(x)\|^2. \end{aligned}$$

Extension of co-coercivity

If $F \in C_L^1$ and α-strongly convex, then

$$G(x) \stackrel{\text{def}}{=} F(x) - \frac{\alpha}{2}\|x^2\|$$

is convex, and ∇G is $L - \alpha$-Lipschitz continuous.

The co-coercivity of ∇G yields

$$\langle \nabla F(x) - \nabla F(y),\, x - y \rangle \geq \frac{\alpha L}{\alpha + L}\|x - y\|^2 + \frac{1}{\alpha + L}\|\nabla F(x) - \nabla F(y)\|^2$$

for all $x, y \in \mathrm{dom}(F)$.

$S_{\alpha,L}^1$: functions in C_L^1 that are α-strongly convex.

Rate of convergence

Sequence x_k converges linearly to $x^\star$ if

$$\lim_{k \to +\infty} \frac{\|x_{k+1} - x^\star\|}{\|x_k - x^\star\|} = \rho$$

holds for $\rho \in]0, 1[$, and ρ is called the rate of convergence.

If x_k converges, let $\rho_k = \frac{\|x_{k+1} - x^\star\|}{\|x_k - x^\star\|}$,

- if $\lim_{k \to +\infty} \rho_k = 0$: super-linear convergence
- if $\lim_{k \to +\infty} \rho_k = 1$: sub-linear convergence

Rate of convergence

Sequence x_k converges linearly to $x^\star$ if

$$\lim_{k\to+\infty} \frac{\|x_{k+1} - x^\star\|}{\|x_k - x^\star\|} = \rho$$

holds for $\rho \in]0,1[$, and ρ is called the rate of convergence.

If x_k converges, let $\rho_k = \frac{\|x_{k+1}-x^\star\|}{\|x_k-x^\star\|}$,

- if $\lim_{k\to+\infty} \rho_k = 0$: super-linear convergence
- if $\lim_{k\to+\infty} \rho_k = 1$: sub-linear convergence

Superlinear convergence: $q > 1$

$$\lim_{k\to+\infty} \frac{\|x_{k+1} - x^\star\|}{\|x_k - x^\star\|^q} < \eta$$

for some $\eta \in]0,1[$.

- $q = 2$: quadratic convergence
- $q = 3$: cubic convergnce

Outline

Unconstrained smooth optimisation

Problem (Unconstrained smooth optimisation)

Consider minising

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

where $F : \mathbb{R}^n \to \mathbb{R}$ *is proper convex and smooth differentiable.*

Assumtions:

- $F \in C^1$ is convex
- $\nabla F(x)$ is L-Lipschitz continuous for some $L > 0$
- Set of minimisers is non-empty, *i.e.* $\mathrm{Argmin}(F) \neq \emptyset$

Gradient descent

Descent direction: let $d = -\nabla F(x)$, then

$$\langle \nabla F(x),\, d \rangle = -\|\nabla F(x)\|^2 \leq 0.$$

Gradient descent

initial: $x_0 \in \mathrm{dom}(F)$;

repeat:

1. Choose step-size $\gamma_k > 0$
2. Update $x_{k+1} = x_k - \gamma_k \nabla F(x_k)$

until: stopping criterion is satisfied.

Convergence analysis: constant step-size

- Owing to the quadratic upper bound

$$
\begin{aligned}
F(x_{k+1}) &\le F(x_k) + \langle \nabla F(x_k),\, x_{k+1} - x_k \rangle + \tfrac{L}{2}\|x_{k+1} - x_k\|^2 \\
&= F(x_k) - \gamma \|\nabla F(x_k)\|^2 + \tfrac{\gamma^2 L}{2}\|\nabla F(x_k)\|^2 \\
&= F(x_k) - \gamma(1 - \tfrac{\gamma L}{2})\|\nabla F(x_k)\|^2.
\end{aligned}
$$

  Hence

$$
F(x_k) - F(x_{k+1}) \ge \gamma(1 - \tfrac{\gamma L}{2})\|\nabla F(x_k)\|^2.
$$

Convergence analysis: constant step-size

- Owing to the quadratic upper bound

$$\begin{aligned} F(x_{k+1}) &\leq F(x_k) + \langle \nabla F(x_k),\, x_{k+1} - x_k \rangle + \tfrac{L}{2}\|x_{k+1} - x_k\|^2 \\ &= F(x_k) - \gamma\|\nabla F(x_k)\|^2 + \tfrac{\gamma^2 L}{2}\|\nabla F(x_k)\|^2 \\ &= F(x_k) - \gamma(1 - \tfrac{\gamma L}{2})\|\nabla F(x_k)\|^2. \end{aligned}$$

  Hence

$$F(x_k) - F(x_{k+1}) \geq \gamma(1 - \tfrac{\gamma L}{2})\|\nabla F(x_k)\|^2.$$

- Let $\gamma \in]0, 2/L[$,

$$\gamma(1 - \tfrac{\gamma L}{2}) \textstyle\sum_{i=0}^{k} \|\nabla F(x_i)\|^2 \leq F(x_0) - F(x_{k+1}) \leq F(x_0) - F(x^\star),$$

- $F(x^\star) > -\infty$, rhs is a positive constant
- for lhs, let $k \to +\infty$,

$$\lim_{k \to +\infty} \|\nabla F(x_k)\|^2 = 0.$$

NB: convexity is not required here.

Convergence analysis: constant step-size

- Let $\gamma \in]0, 1/L]$, then $\gamma(1 - \frac{\gamma L}{2}) \geq \frac{\gamma}{2}$, and

$$
\begin{aligned}
F(x_{k+1}) &\leq F(x_k) - \tfrac{\gamma}{2}\|\nabla F(x_k)\|^2 \\
\text{(cvx of } F \text{ at } x_k) &\leq F(x^\star) + \langle \nabla F(x_k),\, x_k - x^\star \rangle - \tfrac{\gamma}{2}\|\nabla F(x_k)\|^2 \\
&= F(x^\star) + \tfrac{1}{2\gamma}\big(\|x_k - x^\star\|^2 - \|x_k - x^\star - \gamma\nabla F(x_k)\|^2\big) \\
&= F(x^\star) + \tfrac{1}{2\gamma}\big(\|x_k - x^\star\|^2 - \|x_{k+1} - x^\star\|^2\big).
\end{aligned}
$$

Convergence analysis: constant step-size

- Let $\gamma \in]0, 1/L]$, then $\gamma(1 - \frac{\gamma L}{2}) \geq \frac{\gamma}{2}$, and

$$
\begin{aligned}
F(x_{k+1}) &\leq F(x_k) - \tfrac{\gamma}{2}\|\nabla F(x_k)\|^2 \\
\text{(cvx of } F \text{ at } x_k\text{)} &\leq F(x^\star) + \langle \nabla F(x_k),\, x_k - x^\star \rangle - \tfrac{\gamma}{2}\|\nabla F(x_k)\|^2 \\
&= F(x^\star) + \tfrac{1}{2\gamma}\big(\|x_k - x^\star\|^2 - \|x_k - x^\star - \gamma \nabla F(x_k)\|^2\big) \\
&= F(x^\star) + \tfrac{1}{2\gamma}\big(\|x_k - x^\star\|^2 - \|x_{k+1} - x^\star\|^2\big).
\end{aligned}
$$

- Summability of $F(x_k) - F(x^\star)$,

$$
\begin{aligned}
\textstyle\sum_{i=1}^{k}\big(F(x_k) - F(x^\star)\big) &\leq \tfrac{1}{2\gamma}\textstyle\sum_{i=1}^{k}\big(\|x_{i-1} - x^\star\|^2 - \|x_i - x^\star\|^2\big) \\
&= \tfrac{1}{2\gamma}\big(\|x_0 - x^\star\|^2 - \|x_{k+1} - x^\star\|^2\big) \\
&\leq \tfrac{1}{2\gamma}\|x_0 - x^\star\|^2.
\end{aligned}
$$

Convergence analysis: constant step-size

- Let $\gamma \in]0, 1/L]$, then $\gamma(1 - \frac{\gamma L}{2}) \geq \frac{\gamma}{2}$, and

$$\begin{aligned}
F(x_{k+1}) &\leq F(x_k) - \tfrac{\gamma}{2}\|\nabla F(x_k)\|^2 \\
\text{(cvx of } F \text{ at } x_k\text{)} &\leq F(x^\star) + \langle \nabla F(x_k),\, x_k - x^\star \rangle - \tfrac{\gamma}{2}\|\nabla F(x_k)\|^2 \\
&= F(x^\star) + \tfrac{1}{2\gamma}\big(\|x_k - x^\star\|^2 - \|x_k - x^\star - \gamma\nabla F(x_k)\|^2\big) \\
&= F(x^\star) + \tfrac{1}{2\gamma}\big(\|x_k - x^\star\|^2 - \|x_{k+1} - x^\star\|^2\big).
\end{aligned}$$

- Summability of $F(x_k) - F(x^\star)$,

$$\begin{aligned}
\textstyle\sum_{i=1}^{k}\big(F(x_k) - F(x^\star)\big) &\leq \tfrac{1}{2\gamma}\textstyle\sum_{i=1}^{k}\big(\|x_{i-1} - x^\star\|^2 - \|x_i - x^\star\|^2\big) \\
&= \tfrac{1}{2\gamma}\big(\|x_0 - x^\star\|^2 - \|x_{k+1} - x^\star\|^2\big) \\
&\leq \tfrac{1}{2\gamma}\|x_0 - x^\star\|^2.
\end{aligned}$$

- Since $F(x_k) - F(x^\star)$ is decreasing

$$F(x_k) - F(x^\star) \leq \frac{1}{k}\Big(\textstyle\sum_{i=1}^{k}\big(F(x_k) - F(x^\star)\big)\Big) \leq \frac{1}{2\gamma k}\|x_0 - x^\star\|^2.$$

Convergence analysis: strongly convex F

Besides the basic assumptions, let's further assume $F \in S^1_{\alpha,L}$.

Recall that, for all $x, y \in \mathrm{dom}(F)$

$$\langle \nabla F(x) - \nabla F(y),\, x - y \rangle \geq \frac{\alpha L}{\alpha + L} \|x - y\|^2 + \frac{1}{\alpha + L} \|\nabla F(x) - \nabla F(y)\|^2.$$

Convergence analysis: strongly convex F

Besides the basic assumptions, let's further assume $F \in S^1_{\alpha,L}$.

Recall that, for all $x, y \in \mathrm{dom}(F)$

$$\langle \nabla F(x) - \nabla F(y),\, x - y\rangle \geq \frac{\alpha L}{\alpha + L}\|x - y\|^2 + \frac{1}{\alpha + L}\|\nabla F(x) - \nabla F(y)\|^2.$$

Analysis for constant step-size: let $\gamma \in]0, 2/(\alpha + L)[$

$$\begin{aligned}
\|x_{k+1} - x^\star\|^2 &= \|x_k - \gamma\nabla F(x_k) - x^\star\|^2 \\
&= \|x_k - x^\star\|^2 - 2\gamma\langle \nabla F(x_k),\, x_k - x^\star\rangle + \gamma^2\|\nabla F(x_k)\|^2 \\
{\color{red}(\nabla F(x^\star) = 0)} &\leq \Big(1 - \frac{2\gamma\alpha L}{\alpha + L}\Big)\|x_k - x^\star\|^2 + \gamma\Big(\gamma - \frac{2}{\alpha + L}\Big)\|\nabla F(x_k)\|^2 \\
&\leq \Big(1 - \frac{2\gamma\alpha L}{\alpha + L}\Big)\|x_k - x^\star\|^2.
\end{aligned}$$

Convergence analysis: strongly convex F

Distance to minimiser: $\rho = 1 - \frac{2\gamma\alpha L}{\alpha+L}$

$$\|x_k - x^\star\|^2 \le \rho^k \|x_0 - x^\star\|^2.$$

- linear covnergence
- for $\gamma = \frac{2}{\alpha+L}$,

$$\rho = \Big(\frac{L-\alpha}{L+\alpha}\Big)^2.$$

Convergence analysis: strongly convex F

Distance to minimiser: $\rho = 1 - \frac{2\gamma\alpha L}{\alpha+L}$

$$\|x_k - x^\star\|^2 \leq \rho^k \|x_0 - x^\star\|^2.$$

- linear covnergence
- for $\gamma = \frac{2}{\alpha+L}$,

$$\rho = \Big(\frac{L-\alpha}{L+\alpha}\Big)^2.$$

Convergence rate of objective function value:

$$F(x_k) - F(x^\star) \leq \frac{L}{2}\|x_k - x^\star\|^2 \leq \frac{\rho^k L}{2}\|x_0 - x^\star\|^2.$$

Numer of iterations k needed for $F(x_k) - F(x^\star) \leq \epsilon$

- $F \in C_L^1$: $O(1/\epsilon)$
- $F \in S_{\alpha,L}^1$: $O(\log(1/\epsilon))$

Limits on convergence rate of gradient descent

First-order method: x_k is an element from the set

$$x_0 + \operatorname{span}\{\nabla F(x_0), ..., \nabla F(x_i), ..., \nabla F(x_{k-1})\}. \tag{4.1}$$

Problem class: C_L^1

Limits on convergence rate of gradient descent

First-order method: x_k is an element from the set

$$x_0 + \mathrm{span}\{\nabla F(x_0), ..., \nabla F(x_i), ..., \nabla F(x_{k-1})\}. \tag{4.1}$$

Problem class: C^1_L

Theorem (Nesterov)

For every integer $k \leq (n-1)/2$ and every x_0, there exist functions in the problem class such that for any first-order method satisfies (4.1),

$$F(x_k) - F(x^\star) \geq \frac{3}{32}\frac{L\|x_0 - x^\star\|^2}{(k+1)^2},$$

$$\|x_k - x^\star\|^2 \geq \frac{1}{8}\|x_0 - x^\star\|^2.$$

- Suggests $O(1/k)$ is not the optimal rate
- Accelerated gradient methods can achieve $O(1/k^2)$ rate

Outline

Observations

Gradient descent:

$$-\gamma \nabla F(x_k) = x_{k+1} - x_k.$$

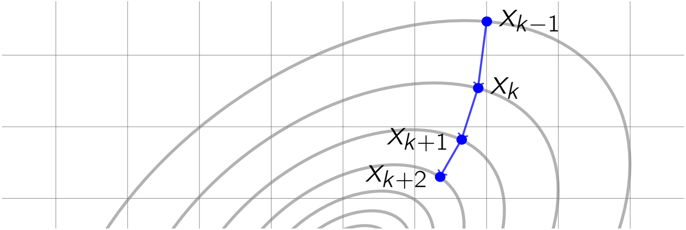
Consider the angle: $\theta_k \overset{\text{def}}{=} \mathrm{angle}(\nabla F(x_{k+1}), \nabla F(x_k))$,

$$\lim_{k \to +\infty} \theta_k = 0.$$

Prove this claim.

Let $a > 0$ be some constant,

$$-\nabla F(x_{k+1}) \approx a(x_{k+1} - x_k).$$

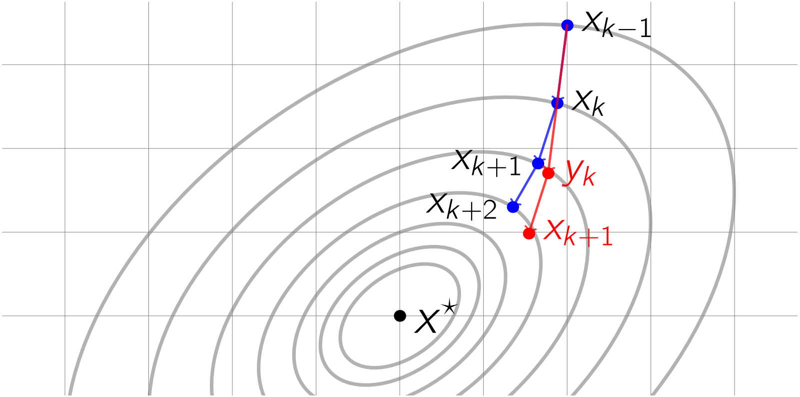


Heavy-ball method (Polyak)

Initial: $x_0 \in \mathrm{dom}(F)$ and $\gamma \in]0, 2/L[$;

$$y_k = x_k + a_k(x_k - x_{k-1}),\ a_k \in [0, 1],$$

$$x_{k+1} = y_k - \gamma \nabla F(x_k).$$



Heavy-ball method

Heavy-ball method (Polyak)

Initial: $x_0 \in \mathrm{dom}(F)$ and $\gamma \in]0, 2/L[$;

$$y_k = x_k + a_k(x_k - x_{k-1}),\ a_k \in [0, 1],$$
$$x_{k+1} = y_k - \gamma \nabla F(x_k).$$

- $x_k - x_{k-1}$ is called the inertial term or momentum term
- a_k is called the inertial parameter
- Convergence can be proved by studying the Lyapunov function

$$\mathcal{E}(x_k) \stackrel{\text{def}}{=} F(x_k) + \frac{a_k}{2\gamma}\|x_k - x_{k-1}\|^2.$$

- In general, no convergence rate for $F \in C_L^1$. Local rate for $F \in S_{\alpha,L}^2$

Convergence rate

Theorem

Let $x^\star$ be a (local) minimiser of F such that $\alpha\mathrm{Id} \preceq \nabla^2 F(x^\star) \preceq L\mathrm{Id}$ and choose a, γ with $a \in [0, 1[, \gamma \in]0, 2(1+a)/L[$. There exists $\underline{\rho} < 1$ such that if $\underline{\rho} < \rho < 1$ and if x_0, x_1 are close enougn to $x^\star$, one has

$$\|x_k - x^\star\| \leq C\rho^k.$$

Moreover, if

$$a = \left(\frac{\sqrt{L} - \sqrt{\alpha}}{\sqrt{L} + \sqrt{\alpha}}\right)^2, \ \gamma = \frac{4}{(\sqrt{L} + \sqrt{\alpha})^2} \ \text{ then } \ \underline{\rho} = \frac{\sqrt{L} - \sqrt{\alpha}}{\sqrt{L} + \sqrt{\alpha}}.$$

- Starting points need to close enough to $x^\star$
- Almost the optimal rate can be achieve by gradient method (or first-order method)
- Gradient descent

$$\underline{\rho} = \frac{L - \alpha}{L + \alpha}.$$

Convergence rate: proof

- Taylor expansion

$$x_{k+1} = x_k + a(x_k - x_{k-1}) - \gamma \nabla^2 F(x^\star)(x_k - x^\star) + o(\|x_k - x^\star\|).$$

- Let $z_k = (x_k - x^\star, x_{k-1} - x^\star)^T$ and $H = \nabla^2 F$, then

$$z_{k+1} = \underbrace{\begin{bmatrix} (1+a)\mathrm{Id} - aH & -a\mathrm{Id} \\ \mathrm{Id} & 0 \end{bmatrix}}_{M} z_k + o(\|z_k\|).$$

- Spectral radius $\rho(M)$, $\eta = 1 - \gamma\alpha$

$$0 = \rho^2 - (a + \eta)\rho + a\eta.$$

- $\rho(M)$ is a function of a and η (essentially γ).

Outline

Convergence rate of gradient descent

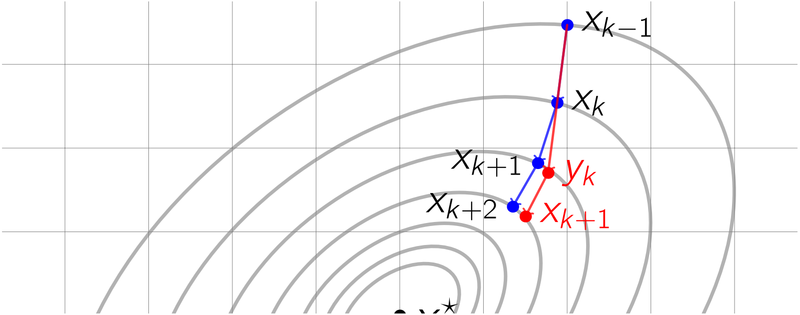
Gradient descent with constant step-size:

- $F \in C^1_L$

$$F(x_k) - F(x^\star) \le \frac{L\|x_0 - x^\star\|^2}{k+4}.$$

- $F \in S^1_{\alpha,L}$

$$F(x_k) - F(x^\star) \le \frac{L}{2}\left(\frac{L-\alpha}{L+\alpha}\right)^2 \|x_0 - x^\star\|^2.$$



Nesterov's optimal scheme

Optimal scheme with constant step-size

initial: Choose $x_0 \in \mathbb{R}^n, \phi_0 \in]0,1[$; Let $y_0 = x_0$ and $q = \alpha/L$.

repeat:

1. Compute $\phi_{k+1} \in]0,1[$ from equation
$$\phi_{k+1}^2 = (1-\phi_{k+1})\phi_k^2 + q\phi_{k+1}.$$
Let $a_k = \frac{\phi_k(1-\phi_k)}{\phi_k^2+\phi_{k+1}}$ and
$$y_k = x_k + a_k(x_k - x_{k-1}).$$
2. Update x_{k+1} by
$$x_{k+1} = y_k - \frac{1}{L}\nabla F(y_k).$$

until: stopping criterion is satisfied.

Convergence rate

Theorem (Convergence rate)

Let $\phi_0 \geq \sqrt{\alpha/L}$*, then*

$$F(x_k) - F(x^\star) \leq \min\left\{ \left(1 - \sqrt{\frac{\alpha}{L}}\right)^k, \frac{4L}{(2\sqrt{L} + k\sqrt{\nu})^2} \right\} \times \left(F(x_0) - F(x^\star) + \frac{\nu}{2}\|x_0 - x^\star\|^2\right),$$

where $\nu = \frac{\phi_0(\phi_0 L - \alpha)}{1 - \phi_0}$.

Convergence rate

Theorem (Convergence rate)

Let $\phi_0 \geq \sqrt{\alpha/L}$, then

$$F(x_k) - F(x^\star) \leq \min\left\{\left(1 - \sqrt{\frac{\alpha}{L}}\right)^k, \frac{4L}{(2\sqrt{L} + k\sqrt{\nu})^2}\right\} \times \left(F(x_0) - F(x^\star) + \frac{\nu}{2}\|x_0 - x^\star\|^2\right),$$

where $\nu = \frac{\phi_0(\phi_0 L - \alpha)}{1 - \phi_0}$.

Parameter choices:

- $F \in C_L^1$: $\phi_0 = 1$,
$$q = 0, \;\; \phi_k \approx \frac{2}{k+1} \to 0 \quad \text{and} \quad a_k \approx \frac{1 - \phi_k}{1 + \phi_k} \to 1.$$
- $F \in S_{\alpha,L}^1$: $\phi_0 = \sqrt{\alpha/L}$
$$q = \sqrt{\frac{\alpha}{L}}, \;\; \phi_k \equiv \sqrt{\frac{\alpha}{L}} \quad \text{and} \quad a_k \equiv \frac{\sqrt{L} - \sqrt{\alpha}}{\sqrt{L} + \sqrt{\alpha}}.$$

Outline

Dynamical system of gradient descent

From gradient descent

$$\frac{x_{k+1} - x_k}{\gamma} = -\nabla F(x_k).$$

Let γ be small enough

$$\dot{X}(t) + \nabla F(X(t)) = 0.$$

Dynamical system of gradient descent

From gradient descent

$$\frac{x_{k+1} - x_k}{\gamma} = -\nabla F(x_k).$$

Let γ be small enough

$$\dot{X}(t) + \nabla F(X(t)) = 0.$$

Discretisation

- Explicit Euler method

$$\dot{X}(t) = \frac{X(t+h) - X(t)}{h}.$$

- Implicit Euler method

$$\dot{X}(t) = \frac{X(t) - X(t-h)}{h}.$$

Dynamical system of inertial schemes

Given a 2nd order dynamical system

$$\ddot{X}(t) + \lambda(t)\dot{X}(t) + \nabla F(X(t)) = 0.$$

Dynamical system of inertial schemes

Given a 2nd order dynamical system

$$\ddot{X}(t) + \lambda(t)\dot{X}(t) + \nabla F(X(t)) = 0.$$

Discretisation:

- 2nd order term

$$\ddot{X}(t) = \frac{X(t+h) - 2X(t) + X(t-h)}{h^2}.$$

- Implicit Euler method

$$\dot{X}(t) = \frac{X(t) - X(t-h)}{h}.$$

Dynamical system of inertial schemes

Given a 2nd order dynamical system

$$\ddot{X}(t) + \lambda(t)\dot{X}(t) + \nabla F(X(t)) = 0.$$

Discretisation:

- 2nd order term

$$\ddot{X}(t) = \frac{X(t+h) - 2X(t) + X(t-h)}{h^2}.$$

- Implicit Euler method

$$\dot{X}(t) = \frac{X(t) - X(t-h)}{h}.$$

Combine together:

$$X(t+h) - X(t) - (1 - h\lambda(t))(X(t) - X(t-h)) + h^2\nabla F(X(t)) = 0.$$

Dynamical system of inertial schemes

Given a 2nd order dynamical system

$$\ddot{X}(t) + \lambda(t)\dot{X}(t) + \nabla F(X(t)) = 0.$$

Discretisation:

- 2nd order term
$$\ddot{X}(t) = \frac{X(t+h) - 2X(t) + X(t-h)}{h^2}.$$
- Implicit Euler method
$$\dot{X}(t) = \frac{X(t) - X(t-h)}{h}.$$

Combine together:

$$X(t+h) - X(t) - (1 - h\lambda(t))(X(t) - X(t-h)) + h^2 \nabla F(X(t)) = 0.$$

Choices:

- Heavy-ball: $h\lambda(t) \in]0, 1[$.
- Nesterov: $\lambda(t) = \frac{d}{t}$, $d > 3$.

Reference

- S. Boyd and L. Vandenberghe. Convex optimization. Cambridge university press, 2004.
- B. Polyak. Introduction to optimization. Optimization Software, 1987.
- Y. Nesterov. Introductory lectures on convex optimization: A basic course. Vol. 87. Springer Science & Business Media, 2013.