ROYAL INSTITUTE
OF TECHNOLOGY

EP3260: Machine Learning Over Networks

Lecture 2: Centralized Convex ML

(part 1)

Hossein S. Ghadikolaei, Hadi Ghauch, and Carlo Fischione

Division of Network and Systems Engineering
School of Electrical Engineering and Computer Science
KTH Royal Institute of Technology, Stockholm, Sweden

https://sites.google.com/view/mlons/home

January 2019

Learning outcomes

- Basic definitions of smooth and convex functions
- Important properties of smooth and convex functions
- Main (deterministic) iterative algorithms for convex problems
- Connections among them
- Pros and cons of them
- Convergence analysis

Outline

1. Student groups
2. Basic definitions and properties
3. Iterative solution approaches
4. Supplements

Outline

1. Student groups

2. Basic definitions and properties

3. Iterative solution approaches

4. Supplements

Student groups

Any question?

Outline

1. Student groups
2. Basic definitions and properties
3. Iterative solution approaches
4. Supplements

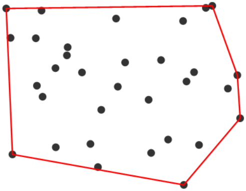
Basic definitions

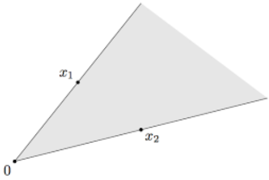
Convex combination of points $\mathcal{X} = \{\boldsymbol{x}_i\}_{i\in[n]}$ is $\sum_{i\in[n]} \theta_i \boldsymbol{x}_i$ where $\boldsymbol{\theta} = [\theta_1, \ldots, \theta_n]$ form a probability simplex; $\boldsymbol{\theta} \geq 0, \boldsymbol{\theta}^T \mathbf{1} = 1$

Conic combination of $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ is $\theta_1 \boldsymbol{x}_1 + \theta_2 \boldsymbol{x}_2$ where $\theta_i \geq 0$

Convex hull of $\mathcal{X}$: set of all convex combinations of points in $\mathcal{X}$

Convex cone of $\mathcal{X}$: set of all conic combinations of points in $\mathcal{X}$





Convexity: basic definitions

Convex set $\mathcal{X}$**:** $\forall \boldsymbol{x}_1, \boldsymbol{x}_2 \in \mathcal{X}$ and $\theta \in [0,1]$, $\theta \boldsymbol{x}_1 + (1-\theta)\boldsymbol{x}_2 \in \mathcal{X}$.

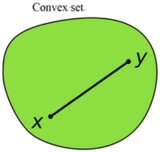
Euclidean/norm ball with radius r centered at $\boldsymbol{x}_c$:

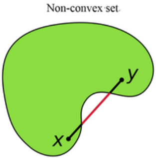
$$\{\boldsymbol{x} \mid \|\boldsymbol{x} - \boldsymbol{x}_c\|_2 \le r\} = \{\boldsymbol{x}_c + r\boldsymbol{u} \mid \|\boldsymbol{u}\|_2 \le 1\}$$

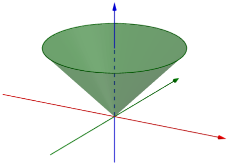
Norm cone: $\{(\boldsymbol{x}, r) \mid \|\boldsymbol{x}\| \le r\}$

Ellipsoid centered at $\boldsymbol{x}_c$: for symmetric positive definite $\boldsymbol{P}$ and square nonsingular $\boldsymbol{A}$:

$$\{(\boldsymbol{x} - \boldsymbol{x}_c)^T \boldsymbol{P}^{-1} (\boldsymbol{x} - \boldsymbol{x}_c) \le 1\} \equiv \{\boldsymbol{x}_c + \boldsymbol{A}\boldsymbol{u} \mid \|\boldsymbol{u}\|_2 \le 1\}$$







Convexity: basic definitions

Convex function $f : \mathcal{X} \to \mathbb{R}$**:** if $\mathcal{X}$ is convex and $\forall \boldsymbol{x}_1, \boldsymbol{x}_2 \in \mathcal{X}$ and $\theta \in [0, 1]$:

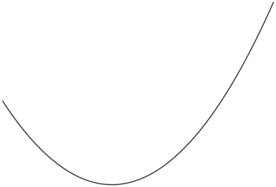
$$f\left(\theta \boldsymbol{x}_1 + (1-\theta)\boldsymbol{x}_2\right) \leq \theta f(\boldsymbol{x}_1) + (1-\theta) f(\boldsymbol{x}_2) \tag{1}$$

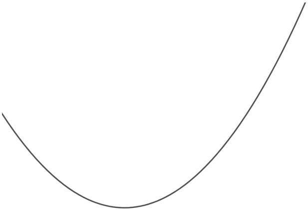
Assuming differentiability of f, (1) is equivalent to

$$f(\boldsymbol{x}_2) \geq f(\boldsymbol{x}_1) + \nabla f(\boldsymbol{x}_1)^T (\boldsymbol{x}_2 - \boldsymbol{x}_1) \tag{2}$$

Local information (gradient) determines a global lower bound

For twice differentiable f, (1) $\leftrightarrow \nabla^2 f(\boldsymbol{x}) \geq 0, \forall \boldsymbol{x} \in \mathcal{X}$: PSD Hessian, non-negative curvature everywhere.





Convexity: basic definitions

Convex function $f : \mathcal{X} \to \mathbb{R}$**:** if $\mathcal{X}$ is convex and $\forall \boldsymbol{x}_1, \boldsymbol{x}_2 \in \mathcal{X}$ and $\theta \in [0,1]$:

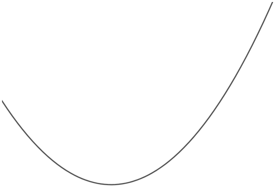
$$f\left(\theta \boldsymbol{x}_1 + (1-\theta)\boldsymbol{x}_2\right) \leq \theta f(\boldsymbol{x}_1) + (1-\theta) f(\boldsymbol{x}_2) \tag{1}$$

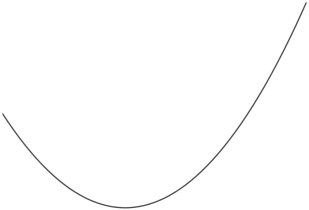
Assuming differentiability of f, (1) is equivalent to

$$f(\boldsymbol{x}_2) \geq f(\boldsymbol{x}_1) + \nabla f(\boldsymbol{x}_1)^T (\boldsymbol{x}_2 - \boldsymbol{x}_1) \tag{2}$$

Local information (gradient) determines a global lower bound

For twice differentiable f, (1) $\leftrightarrow \nabla^2 f(\boldsymbol{x}) \geq 0, \forall \boldsymbol{x} \in \mathcal{X}$: PSD Hessian, non-negative curvature everywhere.





Convexity: basic definitions

Convex function $f : \mathcal{X} \to \mathbb{R}$**:** if $\mathcal{X}$ is convex and $\forall \boldsymbol{x}_1, \boldsymbol{x}_2 \in \mathcal{X}$ and $\theta \in [0,1]$:

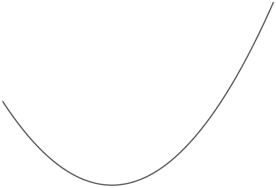
$$f\left(\theta \boldsymbol{x}_1 + (1-\theta)\boldsymbol{x}_2\right) \leq \theta f(\boldsymbol{x}_1) + (1-\theta) f(\boldsymbol{x}_2) \tag{1}$$

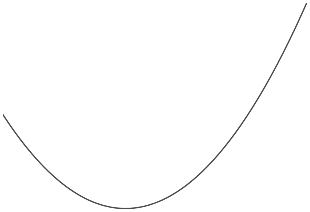
Assuming differentiability of f, (1) is equivalent to

$$f(\boldsymbol{x}_2) \geq f(\boldsymbol{x}_1) + \nabla f(\boldsymbol{x}_1)^T (\boldsymbol{x}_2 - \boldsymbol{x}_1) \tag{2}$$

Local information (gradient) determines a global lower bound

For twice differentiable f, (1) $\leftrightarrow \nabla^2 f(\boldsymbol{x}) \geq 0, \forall \boldsymbol{x} \in \mathcal{X}$: PSD Hessian, non-negative curvature everywhere.





Convexity: some examples

- $\|\boldsymbol{x}\|_p$ for any $p \geq 1$
- Quadratic function $f(\boldsymbol{x}) = \boldsymbol{x}^T \boldsymbol{A} \boldsymbol{x} + \boldsymbol{b}^T \boldsymbol{x} + \boldsymbol{c}$ for symmetric matrix $\boldsymbol{A}$

$$\nabla^2 f(\boldsymbol{x}) = 2\boldsymbol{A} \quad \text{convex iff} \quad \boldsymbol{A} \geq 0$$

- $f(\boldsymbol{x}) = \|\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b}\|_2^2$ is convex for any $\boldsymbol{A}$ (observe $\nabla^2 f(\boldsymbol{x}) = 2\boldsymbol{A}^T\boldsymbol{A}$)
- $\|\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b}\|_p$ for any $p \geq 1$
- $\max_{i \in [n]} \|\boldsymbol{A}_i^T \boldsymbol{x} - \boldsymbol{b}_i\|_p$ for any $p \geq 1$
- $\lambda_{\max} = \sup_{\|\boldsymbol{y}\|_2 = 1} \boldsymbol{y}^T \boldsymbol{X} \boldsymbol{y}$ for any symmetric $\boldsymbol{X}$
- $\mathrm{proj}(\boldsymbol{x}, \mathcal{C}) := \inf_{y \in \mathcal{C}} \|\boldsymbol{y} - \boldsymbol{x}\|$ for convex $\mathcal{C}$
- Check Boyd and Vandenberghe (2003) for more examples

  Nonnegative weighted sum, composition with affine function, pointwise maximum and supremum, composition, minimization, and perspective

Convexity: more definitions

- $f : \mathcal{X} \to \mathbb{R}$ is quasi-convex if $\mathcal{X}$ is convex and sub-level sets

$$\{\boldsymbol{x} \in \mathcal{X} \mid f(\boldsymbol{x}) \leq \alpha\}$$

  are convex for all α.

  Equivalently, if for all $\boldsymbol{x}_1, \boldsymbol{x}_2 \in \mathcal{X}$ and $\theta \in [0, 1]$,

$$f(\theta \boldsymbol{x}_1 + (1-\theta)\boldsymbol{x}_2) \leq \max\{f(\boldsymbol{x}_1), f(\boldsymbol{x}_2)\}$$

- A positive function is log-concave if $\forall \boldsymbol{x}_1, \boldsymbol{x}_2 \in \mathcal{X}$ and $\theta \in [0, 1]$

$$\log f(\theta \boldsymbol{x}_1 + (1-\theta)\boldsymbol{x}_2) \geq \theta \log f(\boldsymbol{x}_1) + (1-\theta) \log f(\boldsymbol{x}_2)$$

  Most of probability densities are log-concave

  For convex $\mathcal{S}$ and random variable $\boldsymbol{y}$ with log-concave PDF,
  $f(\boldsymbol{x}) = \Pr(\boldsymbol{x} + \boldsymbol{y} \in \mathcal{S})$ is log-concave and $\{\boldsymbol{x} \mid f(\boldsymbol{x}) \geq t\}$ is convex

- Majorization ($\boldsymbol{a} \succ \boldsymbol{b}$) and Schur convexity ($\boldsymbol{a} \succ \boldsymbol{b}$ implies $f(\boldsymbol{a}) \geq f(\boldsymbol{b})$)

Standard forms

$$\begin{aligned} \text{minimize} \quad & f_0(\boldsymbol{x}) \\ \text{s.t.} \quad & f_i(\boldsymbol{x}) \leq 0, i \in [m] \\ & h_i(\boldsymbol{x}) = 0, i \in [p] \end{aligned} \tag{3}$$

Linear program: affine objective over a (open/closed) polyhedron

Quadratic program: convex quadratic objective over a polyhedron

Quadratically constrained quadratic program

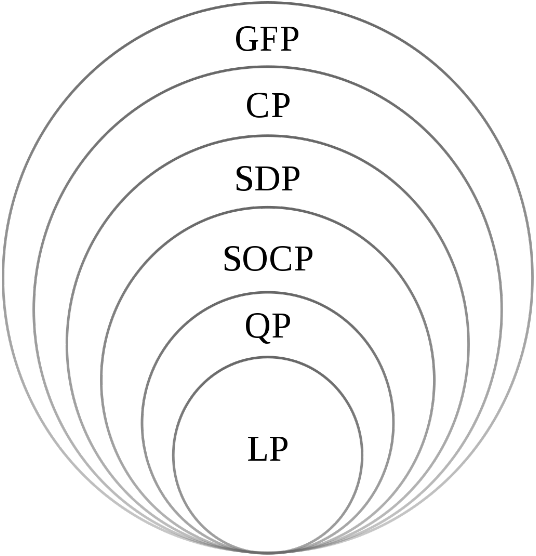
Second-order cone program:

$$\begin{aligned} \text{minimize} \quad & \boldsymbol{c}^T \boldsymbol{x} \\ \text{s.t.} \quad & \|\boldsymbol{A}_i \boldsymbol{x} + \boldsymbol{b}_i\|_2 \leq \boldsymbol{d}_i^T \boldsymbol{x} + \boldsymbol{e}_i, i \in [m], \quad \boldsymbol{F}\boldsymbol{x} = \boldsymbol{g} \end{aligned}$$

Semidefinite program for symmetric $\boldsymbol{A}$ and $\boldsymbol{B}$:

$$\begin{aligned} \text{minimize} \quad & \boldsymbol{c}^T \boldsymbol{x} \\ \text{s.t} \quad & \boldsymbol{A} + \sum_{i \in [n]} \boldsymbol{x}_i \boldsymbol{B}_i \leq 0, \quad \boldsymbol{D}\boldsymbol{x} = \boldsymbol{e} \end{aligned}$$

Standard forms



Duality

Consider (3) with objective f_0, inequality and equality constraints f_i and h_i, and optimal solution $(x^\star, f^\star = f_0(\boldsymbol{x}^\star))$

Lagrange dual function: $g : \mathbb{R}^m \times \mathbb{R}^p \to \mathbb{R}$

$$g(\lambda, \nu) = \inf_{x \in \mathcal{X}} \; L(x, \lambda, \nu) := f_0(\boldsymbol{x}) + \sum_{i \in [m]} \lambda_i f_i(\boldsymbol{x}) + \sum_{i \in [p]} \nu_i h_i(\boldsymbol{x})$$

if $\lambda \geq 0$, then $g(\lambda, \nu) \leq f^\star$.

Proof: given non-negative λ_i, $h_i(\boldsymbol{x}) = 0$, and $\lambda_i f_i(\boldsymbol{x}) \leq 0$ for any feasible point $\boldsymbol{x}$. Therefore,

$$f^\star = f_0(\boldsymbol{x}^\star) \geq L(\boldsymbol{x}^\star, \lambda, \nu) \geq \inf_{x \in \mathcal{X}} L(\boldsymbol{x}, \lambda, \nu) = g(\lambda, \nu)$$

Lagrange dual problem: with solution $d^\star (\leq f^\star, =$ with strong duality)

$$\begin{aligned} \text{maximize} \quad & g(\lambda, \nu) \\ \text{s.t.} \quad & \lambda \geq 0 \end{aligned}$$

Duality

Primal: minimize $\boldsymbol{c}^T\boldsymbol{x}$

s.t. $\boldsymbol{x} \geq 0$

$\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}$

Dual: maximize $-b^T\nu$

s.t. $\boldsymbol{A}^T\nu + \boldsymbol{c} = 0$

In a network with one unit of communication per constraint, dual is more communication-efficient for tall $\boldsymbol{A}$

- Check Boyd and Vandenberghe (2003) for more details

  Weak and strong duality

  Constraint qualifications

  Slater's constraint qualification

  Complementary slackness

  Karush-Kuhn-Tucker (KKT) conditions

Strong convexity

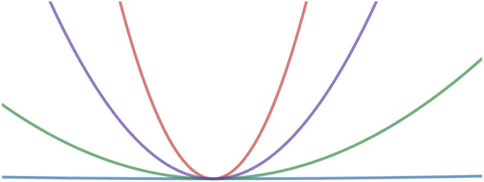
Differentiable function f is μ-strongly convex iff $\forall \boldsymbol{x}_1, \boldsymbol{x}_2 \in \mathcal{X}, \mu > 0$

$$f(\boldsymbol{x}_2) \geq f(\boldsymbol{x}_1) + \nabla f(\boldsymbol{x}_1)^T (\boldsymbol{x}_2 - \boldsymbol{x}_1) + \frac{\mu}{2} \|\boldsymbol{x}_2 - \boldsymbol{x}_1\|_2^2$$

Gradient can be replaced by sub-gradient for non-smooth functions

Main intuition: linear lower bound with convexity, quadratic lower bound with strong convexity

Global definitions not local ($\forall \boldsymbol{x} \in \mathcal{X}$)



Strong convexity

$$f(\boldsymbol{x}_2) \geq f(\boldsymbol{x}_1) + \nabla f(\boldsymbol{x}_1)^T(\boldsymbol{x}_2 - \boldsymbol{x}_1) + \frac{\mu}{2}\|\boldsymbol{x}_2 - \boldsymbol{x}_1\|_2^2 \tag{4}$$

- (4) is equivalent to a minimum positive curvature $\nabla^2 f(\boldsymbol{x}) \geq \mu \boldsymbol{I}_d, \forall \boldsymbol{x} \in \mathcal{X}$
- (4) is equivalent to $\left(\nabla f(\boldsymbol{x}_2) - \nabla f(\boldsymbol{x}_1)\right)^T (\boldsymbol{x}_2 - \boldsymbol{x}_1) \geq \mu\|\boldsymbol{x}_2 - \boldsymbol{x}_1\|_2^2$
- (4) implies
  (a) $f(\boldsymbol{x}) - f^\star \leq \dfrac{1}{2\mu}\|\nabla f(\boldsymbol{x})\|_2^2, \forall \boldsymbol{x}$
  (b) $\|\boldsymbol{x}_2 - \boldsymbol{x}_1\|_2 \leq \dfrac{1}{\mu}\|\nabla f(\boldsymbol{x}_2) - \nabla f(\boldsymbol{x}_1)\|_2, \forall \boldsymbol{x}_1, \boldsymbol{x}_2$
  (c) $\left(\nabla f(\boldsymbol{x}_2) - \nabla f(\boldsymbol{x}_1)\right)^T (\boldsymbol{x}_2 - \boldsymbol{x}_1) \leq \dfrac{1}{\mu}\|\nabla f(\boldsymbol{x}_2) - \nabla f(\boldsymbol{x}_1)\|_2^2, \forall \boldsymbol{x}_1, \boldsymbol{x}_2$
  (d) $f(\boldsymbol{x}) + r(\boldsymbol{x})$ is strongly convex for any convex f and strongly convex r

HW1(a): prove all the statements of this slide

Smoothness

A function $f : \mathbb{R}^d \to \mathbb{R}$ is L-smooth iff it is differentiable and its gradient is L-Lipschitz-continuous (usually w.r.t. norm-2):

$$\forall \boldsymbol{x}_1, \boldsymbol{x}_2 \in \mathbb{R}^d, \|\nabla f(\boldsymbol{x}_2) - \nabla f(\boldsymbol{x}_1)\|_2 \leq L\|\boldsymbol{x}_2 - \boldsymbol{x}_1\|_2 \tag{5}$$

Recall strong convexity result: $\|\nabla f(\boldsymbol{x}_2) - \nabla f(\boldsymbol{x}_1)\|_2 \geq \mu\|\boldsymbol{x}_2 - \boldsymbol{x}_1\|_2$

For twice differentiable f, (5) $\leftrightarrow \nabla^2 f(\boldsymbol{x}) \leq L\boldsymbol{I}_d$

Smoothness: $f(\boldsymbol{x}_2) - f(\boldsymbol{x}_1)$ can be over-estimated by a quadratic function

- (5) implies for all $\boldsymbol{x}_1, \boldsymbol{x}_2$ (**HW1(b):** prove them. Assume convexity if needed)

  (a) $f(\boldsymbol{x}_2) \leq f(\boldsymbol{x}_1) + \nabla f(\boldsymbol{x}_1)^T(\boldsymbol{x}_2 - \boldsymbol{x}_1) + \dfrac{L}{2}\|\boldsymbol{x}_2 - \boldsymbol{x}_1\|_2^2$

  (b) $f(\boldsymbol{x}_2) \geq f(\boldsymbol{x}_1) + \nabla f(\boldsymbol{x}_1)^T(\boldsymbol{x}_2 - \boldsymbol{x}_1) + \dfrac{1}{2L}\|\nabla f(\boldsymbol{x}_2) - \nabla f(\boldsymbol{x}_1)\|_2^2$

  (c) $\left(\nabla f(\boldsymbol{x}_2) - \nabla f(\boldsymbol{x}_1)\right)^T(\boldsymbol{x}_2 - \boldsymbol{x}_1) \geq \dfrac{1}{L}\|\nabla f(\boldsymbol{x}_2) - \nabla f(\boldsymbol{x}_1)\|_2^2$

Outline

1. Student groups

2. Basic definitions and properties

3. Iterative solution approaches

4. Supplements

Gradient descent

- Problem: $\underset{\boldsymbol{x}}{\text{minimize}}\ f(\boldsymbol{x})$ for some differentiable $f : \mathbb{R}^d \to \mathbb{R} \cup \{\pm\infty\}$

Gradient descent (GD), also called batch GD, full GD, ...

$$\boldsymbol{x}_{k+1} = \boldsymbol{x}_k - \alpha_k \nabla f(\boldsymbol{x}_k) \tag{6}$$

for some sequence of non-negative step sizes $(\alpha_k)_{k\in\mathbb{N}}$.

Theorem 1: Convergence of GD with constant step size

Convex and L-smooth f with $\alpha \le 1/L$ satisfies $f(\boldsymbol{x}_k) - f^\star \le \dfrac{\|\boldsymbol{x}_0 - \boldsymbol{x}^\star\|_2^2}{2k\alpha}$.

μ**-strongly convex and L-smooth** f with $\alpha \le 2/(\mu + L)$ satisfies $f(\boldsymbol{x}_k) - f^\star \le e^{-ck} L \|\boldsymbol{x}_0 - \boldsymbol{x}^\star\|_2^2/2$ for $c = 2\alpha\mu L/(\mu + L)$. With $\alpha = 2/(\mu + L)$, we have $\|\boldsymbol{x}_k - \boldsymbol{x}^\star\|_2^2 \le \left(1 - \frac{2}{1+L/\mu}\right)^{2k} \|\boldsymbol{x}_0 - \boldsymbol{x}^\star\|_2^2$.

Smooth convex: $\mathcal{O}(1/\epsilon)$ iterations for ϵ-optimality

Smooth strongly-convex: $\mathcal{O}(\log(1/\epsilon))$ iterations for ϵ-optimality

Proofs

Foods for thought

1. Define the conjugate function as

$$f^*(\boldsymbol{y}) = \sup_{y \in \mathbb{R}^d} \boldsymbol{y}^T\boldsymbol{x} - f(\boldsymbol{x}).$$

   Observe that f^* is convex even when f is not (why?). When f is μ-strongly convex and L-smooth, f^* is $\frac{1}{L}$-strongly convex and $\frac{1}{\mu}$-smooth (why?).

2. Define projection operator for convex set $\mathcal{X}$ as

$$\text{proj}(\boldsymbol{x}, \mathcal{X}) := \text{argmin}_{y \in \mathcal{X}} \|\boldsymbol{y} - \boldsymbol{x}\|.$$

   Observe $\|\text{proj}(\boldsymbol{y}, \mathcal{X}) - \boldsymbol{x}\|^2 \le \|\boldsymbol{y} - \boldsymbol{x}\|^2$ for any $\boldsymbol{x} \in \mathcal{X}$ and any $\boldsymbol{y}$. Modify (6) to solve $\min_{\boldsymbol{x} \in \mathcal{X}} f(\boldsymbol{x})$. What is the convergence of this "projected GD" algorithm?

3. Define the set of subgradients of $f : \mathcal{X} \to \mathbb{R}$ at $\boldsymbol{x} \in \mathcal{X}$ as

$$\partial f(\boldsymbol{x}) := \{\boldsymbol{s} \mid f(\boldsymbol{x}) - f(\boldsymbol{y}) \le \boldsymbol{s}^T(\boldsymbol{x} - \boldsymbol{y}) \ \forall \boldsymbol{y} \in \mathcal{X}\}.$$

   Modify (6) to solve $\min_{\boldsymbol{x} \in \mathcal{X}} f(\boldsymbol{x})$ for non-smooth (non-differentiable) functions that are Lipschitz ($|f(\boldsymbol{x}) - f(\boldsymbol{y})| \le \beta\|\boldsymbol{x} - \boldsymbol{y}\|_2$). What is the convergence of this "projected sub-GD" for Lipschitz functions? What is the optimal step size?

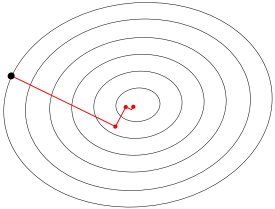
4. Check `https://ee227c.github.io/code/lecture4.html`

GD for smooth and strongly convex functions

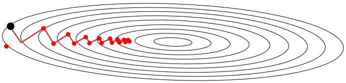
Linear convergence rate $1 - \dfrac{2}{1 + L/\mu}$

GD may need many iterations to converge

Preconditioning to change the space geometry, to make sub-levels similar in all coordinates



small L/μ



large L/μ

Descent methods

Descent methods: $$\boldsymbol{x}_{k+1} = \boldsymbol{x}_k + \alpha_k \boldsymbol{d}(\boldsymbol{x}_k) \;\text{ s.t. } f(\boldsymbol{x}_{k+1}) < f(\boldsymbol{x}_k) \tag{7}$$

for some sequence of non-negative step sizes $(\alpha_k)_{k\in\mathbb{N}}$ and decent direction $\boldsymbol{d}$

$f(\boldsymbol{x}_{k+1}) < f(\boldsymbol{x}_k)$ implies $-\nabla f(\boldsymbol{x}_k)^T \boldsymbol{d}(\boldsymbol{x}_k) > 0$, same half-space as the negative gradient

Steepest descent: $\boldsymbol{d}(\boldsymbol{x}) = \operatorname{argmin}\{\nabla f(\boldsymbol{x})^T \nu \mid \|\nu\| = 1\}$ in some norm $\|\cdot\|$

* Note that we need to unnormalize the descent direction using the dual-norm

Maximizes the first order prediction of decrease (for small ν):
$f(\boldsymbol{x}+\nu) - f(\boldsymbol{x}) \approx \nabla f(\boldsymbol{x})^T \nu$

Define $\|\boldsymbol{x}\|_{\boldsymbol{P}} = (\boldsymbol{x}^T \boldsymbol{P} \boldsymbol{x})^{1/2}$ for positive definite $\boldsymbol{P}$ (this is called Mahalanobis distance): $\boldsymbol{d}(\boldsymbol{x}) = -\boldsymbol{P}^{-1} \nabla f(\boldsymbol{x})$

Reduces to GD on Euclidian norm $(\boldsymbol{P} = \boldsymbol{I}, \boldsymbol{d}(\boldsymbol{x}) = -\nabla f(\boldsymbol{x}_k)^T)$

Good norm: should be consistent with the geometry of sublevel sets

Same theoretical convergence as of GD, much better in practice

Newton methods

Around optimal point:

$$f(\boldsymbol{x}) \approx f(\boldsymbol{x}^\star) + \cancelto{0}{\nabla f(\boldsymbol{x}^\star)^T(\boldsymbol{x} - \boldsymbol{x}^\star)} + \frac{1}{2}(\boldsymbol{x} - \boldsymbol{x}^\star)^T \nabla^2 f(\boldsymbol{x}^\star)(\boldsymbol{x} - \boldsymbol{x}^\star)$$

Sublevel sets are like ellipsoids (determined by Hessian) near the minimum

Recall $\|\boldsymbol{x}\|_{\boldsymbol{P}} = (\boldsymbol{x}^T \boldsymbol{P} \boldsymbol{x})^{1/2}$ and its descent direction $\boldsymbol{d} = \boldsymbol{P}^{-1} \nabla f(\boldsymbol{x})$

How about steepest descent on the norm induced by Hessian $\nabla^2 f(\boldsymbol{x}^\star)$?

Oops! we do not know $\boldsymbol{x}^\star$

Newton method:

$$\boldsymbol{x}_{k+1} = \boldsymbol{x}_k - \alpha_k \nabla^2 f(\boldsymbol{x}_k)^{-1} \nabla f(\boldsymbol{x}_k). \tag{8}$$

Newton methods

$$\boldsymbol{x}_{k+1} = \boldsymbol{x}_k - \alpha_k \nabla^2 f(\boldsymbol{x}_k)^{-1} \nabla f(\boldsymbol{x}_k)$$

Theorem 2: Quadratic convergence of Newton's method

Assume f is a twice continuously differentiable and set $\alpha_k = 1$. If $\|\boldsymbol{x}_k - \boldsymbol{x}^\star\|$ is small enough, there exist a positive constant c such that $\|\boldsymbol{x}_{k+1} - \boldsymbol{x}^\star\| \leq c(\|\boldsymbol{x}_k - \boldsymbol{x}^\star\|_2)^2$.

constant $+ \mathcal{O}(\log\log(1/\epsilon))$ iterations for ϵ-optimality

Expensive iterations due to $\nabla^2 f(\boldsymbol{x}_k)$

- **Affine invariance:** apply coordinate change for non-singular matrix $\boldsymbol{A}$. Newton's method have same iterations for $\min_{\boldsymbol{x}} f(\boldsymbol{x})$ and $\min_{\boldsymbol{y}} f(\boldsymbol{A}\boldsymbol{y})$ (namely $\boldsymbol{x}_k = \boldsymbol{A}\boldsymbol{y}_k$), whereas GD has $\nabla f(\boldsymbol{x}) = \boldsymbol{A}^T \nabla(\boldsymbol{A}\boldsymbol{y})$.

  If we change coordinate/metric, GD iterates change

  Finding a good coordinate for GD is usually very hard in high-dimension!

Proximal methods

Smoothness implies $f(\boldsymbol{x}_{k+1}) \leq f(\boldsymbol{x}_k) + \nabla f(\boldsymbol{x}_k)(\boldsymbol{x} - \boldsymbol{x}_k) + \frac{L}{2}\|\boldsymbol{x} - \boldsymbol{x}_k\|_2^2, \forall \boldsymbol{x}$

GD iterations to minimize differentiable f is like *successive quadratic upper-bound minimization*:

$$\boldsymbol{x}_{k+1} = \underset{\boldsymbol{x} \in \mathbb{R}^d}{\operatorname{argmin}}\ f(\boldsymbol{x}_k) + \nabla f(\boldsymbol{x}_k)^T(\boldsymbol{x} - \boldsymbol{x}_k) + \frac{L}{2}\|\boldsymbol{x} - \boldsymbol{x}_k\|_2^2$$

New objective: minimize $f(\boldsymbol{x}) = g(\boldsymbol{x}) + h(\boldsymbol{x})$ for convex differentiable g and convex (possibly) non-differentiable h

Define **proximal mapping** as

$$\operatorname{prox}_{\alpha h}(\boldsymbol{x}) = \operatorname{argmin}_{\boldsymbol{u}}\ h(\boldsymbol{u}) + \frac{1}{2\alpha}\|\boldsymbol{u} - \boldsymbol{x}\|_2^2$$

Proximal method:

$$\boldsymbol{x}_k = \operatorname{prox}_{\alpha_k h}\left(\boldsymbol{x}_{k-1} - \alpha_k \nabla g(\boldsymbol{x}_k)\right)$$

Proximal methods

Smoothness implies $f(\boldsymbol{x}_{k+1}) \leq f(\boldsymbol{x}_k) + \nabla f(\boldsymbol{x}_k)(\boldsymbol{x} - \boldsymbol{x}_k) + \frac{L}{2}\|\boldsymbol{x} - \boldsymbol{x}_k\|_2^2, \forall \boldsymbol{x}$

GD iterations to minimize differentiable f is like *successive quadratic upper-bound minimization*:

$$\boldsymbol{x}_{k+1} = \operatorname*{argmin}_{\boldsymbol{x} \in \mathbb{R}^d} \; f(\boldsymbol{x}_k) + \nabla f(\boldsymbol{x}_k)^T(\boldsymbol{x} - \boldsymbol{x}_k) + \frac{L}{2}\|\boldsymbol{x} - \boldsymbol{x}_k\|_2^2$$

New objective: minimize $f(\boldsymbol{x}) = g(\boldsymbol{x}) + h(\boldsymbol{x})$ for convex differentiable g and convex (possibly) non-differentiable h

Define **proximal mapping** as

$$\operatorname{prox}_{\alpha h}(\boldsymbol{x}) = \operatorname{argmin}_{\boldsymbol{u}} \; h(\boldsymbol{u}) + \frac{1}{2\alpha}\|\boldsymbol{u} - \boldsymbol{x}\|_2^2$$

Proximal method:

$$\boldsymbol{x}_k = \operatorname{prox}_{\alpha_k h}\left(\boldsymbol{x}_{k-1} - \alpha_k \nabla g(\boldsymbol{x}_k)\right)$$

Proximal methods

- Observe from the definition of the proximal method

$$
\begin{aligned}
\boldsymbol{x}_k &= \operatorname{argmin}_{\boldsymbol{x}} \ h(\boldsymbol{x}) + \frac{1}{2\alpha_k}\|\boldsymbol{x} - \boldsymbol{x}_{k-1} + \alpha_k \nabla g(\boldsymbol{x}_k)\|_2^2 \\
&= \operatorname{argmin}_{\boldsymbol{x}} \ h(\boldsymbol{x}) + g(\boldsymbol{x}_k) + \nabla g(\boldsymbol{x}_k)^T(\boldsymbol{x} - \boldsymbol{x}_k) + \frac{1}{2\alpha_k}\|\boldsymbol{x} - \boldsymbol{x}_k\|_2^2
\end{aligned}
$$

GD $\leftrightarrow$ proximal method with $h(\boldsymbol{x}) = 0$ and $\alpha = 1/L$

Projected GD $\leftrightarrow$ proximal method with $h(\boldsymbol{x}) = \begin{cases} 0, & \text{if } \ \boldsymbol{x} \in \mathcal{X} \\ \infty, & \text{otherwise.} \end{cases}$

Soft thresholding for ℓ_1 regularization $\leftrightarrow$ proximal method with $h(\boldsymbol{x}) = \lambda\|\boldsymbol{x}\|_1$.

$$
[\operatorname{prox}_h(\boldsymbol{x})]_i = \begin{cases} x_i - \lambda, & \text{if } \ x_i \geq \lambda \\ 0, & \text{if } \ -\lambda \leq x_i \leq \lambda \\ x_i + \lambda, & \text{if } x_i \leq -\lambda. \end{cases}
$$

Resource allocation

HW1(c): Consider

$$\begin{aligned} \text{minimize} \quad & \frac{1}{N} \sum\nolimits_{i \in [N]} f_i(x_i) \\ \text{s.t.} \quad & \boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}. \end{aligned}$$

for $\boldsymbol{A} \in \mathbb{R}^{p \times N}$ and $\boldsymbol{x} = [x_1, \ldots, x_N]^T$.

(a) Assume strong-convexity and smoothness on f. How would you solve this problem when $N = 1000$?

(b) What if $N = 10^9$?

(c) Can we use Newton's method for $N = 10^9$? Try efficient method for computing $\nabla^2 f(\boldsymbol{x}_k)$ for $p = 1$ and $b = 1$ (probability simplex constraint). Extend it to $1 \leq p \ll N$.

(d) Now, add twice differentiable $r(\boldsymbol{x})$ to the objective and solve (a)-(c).

Some references

- S. Bubeck, “Convex optimization: Algorithms and complexity,” Foundations and Trends in Machine Learning, 2015.

- Y. Nesterov, “Introductory lectures on convex optimization: A basic course, ” Springer Science & Business Media, 2004.

- F. Bach, “Large-scale machine learning and convex optimization,” Machine Learning Summer School, 2018.

- L. Bottou, F. Curtis, and J. Norcedal, “Optimization methods for large-scale machine learning,” SIAM Rev., 2018.

- N. Parikh and S. Boyd, “Proximal algorithms”, gradient methods,” Foundations and Trends in Optimization, 2013.

- A. Beck and M. Teboulle, “A fast iterative shrinkage-thresholding algorithm for linear inverse problems,” SIAM Journal on Imaging Sciences, 2009.

Outline

Proof sketch for convex and L-smooth f

By convexity of f, $f(\boldsymbol{x}_i) \leq f(\boldsymbol{x}^\star) + \langle \nabla f(\boldsymbol{x}_i), \boldsymbol{x}_i - \boldsymbol{x}^\star \rangle$ $\qquad (*)$

Use smoothness property $f(\boldsymbol{x}_{i+1}) \leq f(\boldsymbol{x}_i) + \nabla f(\boldsymbol{x}_i)^T(\boldsymbol{x}_{i+1} - \boldsymbol{x}_i) + \dfrac{L}{2}\|\boldsymbol{x}_{i+1} - \boldsymbol{x}_i\|_2^2$ and $\alpha \leq 1/L$ to conclude $f(\boldsymbol{x}_{i+1}) \leq f(\boldsymbol{x}_i) - \frac{\alpha}{2}\|\nabla f(\boldsymbol{x}_i)\|_2^2$ $\qquad (**)$

Substitute $(*)$ into $(**)$ and note that from (6): $\nabla f(\boldsymbol{x}_i) = \dfrac{\boldsymbol{x}_i - \boldsymbol{x}_{i+1}}{\alpha}$

Observe $f(\boldsymbol{x}_{i+1}) \leq f^\star + \dfrac{1}{2\alpha}\left(\|\boldsymbol{x}_i - x^\star\|_2^2 - \|\boldsymbol{x}_{i+1} - x^\star\|_2^2\right)$

Summing over iterations: $\dfrac{1}{k}\sum_{i\in[k]}\left(f(\boldsymbol{x}_i) \leq f^\star\right) \leq \dfrac{1}{2\alpha k}\|\boldsymbol{x}_0 - x^\star\|_2^2$

Conclude from the non-increasing property of GD iterates:

$$f(\boldsymbol{x}_k) - f^\star \leq \frac{1}{k}\sum_{i\in[k]} f(\boldsymbol{x}_i) - f^\star \leq \frac{1}{2\alpha k}\|\boldsymbol{x}_0 - x^\star\|_2^2.$$

Return

Proof sketch for strongly-convex and L-smooth f

From smoothness and vanishing gradient of the optimal point, conclude
$f(\boldsymbol{x}_i) - f(\boldsymbol{x}^\star) \le \frac{L}{2}\|\boldsymbol{x}_k - \boldsymbol{x}^\star\|_2^2 \qquad (*)$

Use the coercivity of the gradient (**HW1(d):** prove it)

$$(\nabla f(\boldsymbol{x}) - \nabla f(\boldsymbol{y}))^T(\boldsymbol{x} - \boldsymbol{y}) \ge \frac{\mu L}{\mu + L}\|\boldsymbol{x} - \boldsymbol{y}\|_2^2 + \frac{1}{\mu + L}\|\nabla f(\boldsymbol{x}) - \nabla f(\boldsymbol{y})\|_2^2$$

Use $\alpha < 2/(L+\mu)$ to obtain $\boldsymbol{x}_i - \boldsymbol{x}^\star\|^2 \le \left(1 - 2t\alpha\frac{\mu L}{\mu + L}\right)\|\boldsymbol{x}_i - \boldsymbol{x}^\star\|_2^2$

Iterate over i and use $(*)$ to obtain

$$f(\boldsymbol{x}_k) - f(\boldsymbol{x}^\star) \le \frac{L}{2}\prod_{i\in[k]}\left(1 - 2t\alpha\frac{\mu L}{\mu + L}\right)\|\boldsymbol{x}_0 - \boldsymbol{x}^\star\|_2^2$$

Use $e^{-x} \ge 1 - x$ to conclude $f(\boldsymbol{x}_k) - f(\boldsymbol{x}^\star) \le \dfrac{L}{2}e^{-ck}\|\boldsymbol{x}_0 - \boldsymbol{x}^\star\|_2^2$ for $c = \frac{2\alpha\mu L}{\mu+L}$

Return