Mathematics of Data: From Theory to Computation

Prof. Volkan Cevher
volkan.cevher@epfl.ch

Lecture 3: Unconstrained, smooth minimization I

Laboratory for Information and Inference Systems (LIONS)
École Polytechnique Fédérale de Lausanne (EPFL)

EE-556 (Fall 2019)

License Information for Mathematics of Data Slides

Outline

- This lecture
  1. Unconstrained convex & non-convex optimization: the basics
  2. Gradient descent methods
- Next lecture
  1. Gradient and accelerated gradient descent methods

Recommended reading

- Chapters 2, 3, 5, 6 in Nocedal, Jorge, and Wright, Stephen J., *Numerical Optimization*, Springer, 2006.
- Chapter 9 in Boyd, Stephen, and Vandenberghe, Lieven, *Convex optimization*, Cambridge university press, 2009.
- Chapter 1 in Bertsekas, Dimitris, *Nonlinear Programming*, Athena Scientific, 1999.
- Chapters 1, 2 and 4 in Nesterov, Yurii, *Introductory Lectures on Convex Optimization: A Basic Course*, Vol. 87, Springer, 2004.

Motivation

Motivation

This lecture covers the basics of numerical methods for *unconstrained* and *smooth* minimization for **convex** and **non-convex** problem settings in a unified manner.

Smooth unconstrained convex minimization

Problem (Mathematical formulation)

The unconstrained convex minimization problem is defined as:

$$f^\star := \min_{\mathbf{x}\in\mathbb{R}^p} f(\mathbf{x})$$

- f *is a proper, closed and smooth convex function,* $-\infty < f^\star < +\infty$.
- *The solution set* $\mathcal{S}^\star := \{\mathbf{x}^\star \in \mathrm{dom}\,(f) : f(\mathbf{x}^\star) = f^\star\}$ *is nonempty.*

Example: Maximum likelihood estimation and M-estimators

Problem

*Let $\mathbf{x}^\natural \in \mathbb{R}^p$ be unknown and $b_1, ..., b_n$ be i.i.d. samples of a random variable B with p.d.f. $p_{\mathbf{x}^\natural}(b) \in \mathcal{P} := \{p_{\mathbf{x}}(b) : \mathbf{x} \in \mathbb{R}^p\}$. **Goal:** Estimate $\mathbf{x}^\natural$ from $b_1, \ldots, b_n$.*

Optimization formulation (ML estimator)

$$\hat{\mathbf{x}}_{\mathsf{ML}} := \arg\min_{\mathbf{x}\in\mathbb{R}^p} \left\{ -\frac{1}{n}\sum_{i=1}^{n} \ln\left[p_{\mathbf{x}}(b_i)\right] \right\} = \arg\min_{\mathbf{x}\in\mathbb{R}^p} f(\mathbf{x})$$

Theorem (Performance of the ML estimator [2, 5])

The random variable $\hat{\mathbf{x}}_{ML}$ satisfies

$$\lim_{n\to\infty} \sqrt{n}\,\mathbf{J}^{-1/2}\left(\hat{\mathbf{x}}_{ML} - \mathbf{x}^\natural\right) \overset{d}{=} Z \sim \mathcal{N}(\mathbf{0}, \mathbf{I}),$$

where

$$\mathbf{J} := -\mathbb{E}\left[\nabla^2_{\mathbf{x}} \ln\left[p_{\mathbf{x}}(B)\right]\right]\Big|_{\mathbf{x}=\mathbf{x}^\natural}.$$

is the Fisher information matrix associated with one sample. Roughly speaking,

$$\left\|\sqrt{n}\,\mathbf{J}^{-1/2}\left(\hat{\mathbf{x}}_{ML} - \mathbf{x}^\natural\right)\right\|_2^2 \sim \mathrm{Tr}\left(\mathbf{I}\right) = p \quad \Rightarrow \quad \boxed{\left\|\hat{\mathbf{x}}_{ML} - \mathbf{x}^\natural\right\|_2^2 = \mathcal{O}(p/n)}.$$

Example: Maximum likelihood estimation and M-estimators

Problem

Let $\mathbf{x}^\natural \in \mathbb{R}^p$ *be unknown and* $b_1, ..., b_n$ *be i.i.d. samples of a random variable* B *with p.d.f.* $p_{\mathbf{x}^\natural}(b) \in \mathcal{P} := \{p_{\mathbf{x}}(b) : \mathbf{x} \in \mathbb{R}^p\}$. ***Goal:*** *Estimate* $\mathbf{x}^\natural$ *from* $b_1, \ldots, b_n$.

Optimization formulation (ML estimator)

$$\hat{\mathbf{x}}_{\text{ML}} := \arg\min_{\mathbf{x} \in \mathbb{R}^p} \left\{ -\frac{1}{n} \sum_{i=1}^{n} \ln \left[p_{\mathbf{x}}(b_i) \right] \right\} = \arg\min_{\mathbf{x} \in \mathbb{R}^p} f(\mathbf{x})$$

Optimization formulation (M-estimator)

In general, we can replace the negative log-likelihoods by any appropriate, convex g_i's

$$\min_{x \in \mathcal{X}} \underbrace{\frac{1}{n} \sum_{i=1}^{n} g_i(b_i; \mathbf{x})}_{f(\mathbf{x})}.$$

Approximate vs. exact optimality

Is it possible to solve a convex optimization problem?

"In general, optimization problems are **unsolvable***"* - Y. Nesterov [3]

- Even when a closed-form solution exists, numerical accuracy may still be an issue.
- We must be content with **approximately** optimal solutions.

Definition

We say that $\mathbf{x}^\star_\epsilon$ is ϵ-optimal in **objective value** if

$$f(\mathbf{x}^\star_\epsilon) - f^\star \leq \epsilon\ .$$

Definition

We say that $\mathbf{x}^\star_\epsilon$ is ϵ-optimal in **sequence** if, for some norm $\|\cdot\|$,

$$\|\mathbf{x}^\star_\epsilon - \mathbf{x}^\star\| \leq \epsilon\ ,$$

- The **latter** approximation guarantee is considered **stronger**.

A gradient method

Lemma (First-order necessary optimality condition)

Let $\mathbf{x}^\star$ be a global minimum of a differentiable convex function f. Then, it holds that

$$\nabla f(\mathbf{x}^\star) = \mathbf{0}.$$

Fixed-point characterization

Multiply by -1 and add $\mathbf{x}^\star$ to both sides to obtain a fixed point condition,

$$\mathbf{x}^\star = \mathbf{x}^\star - \alpha \nabla f(\mathbf{x}^\star) \qquad \text{for all } 0 \neq \alpha \in \mathbb{R}$$

Gradient method

Choose a starting point $\mathbf{x}^0$ and iterate

$$\mathbf{x}^{k+1} = \mathbf{x}^k - \alpha_k \nabla f(\mathbf{x}^k)$$

where α_k is a step-size to be chosen so that $\mathbf{x}^k$ converges to $\mathbf{x}^\star$.

When does the gradient method converge?

Lemma

Let f be a closed, convex function. Assume that

1. *There exists* $\mathbf{x}^\star \in \mathrm{dom}(f)$ *such that* $\nabla f(\mathbf{x}^\star) = 0$.
2. *The mapping* $\psi(\mathbf{x}) = \mathbf{x} - \alpha \nabla f(\mathbf{x})$ *is contractive for some* α*: i.e., there exists* $\gamma \in [0, 1)$ *such that*

$$\|\psi(\mathbf{x}) - \psi(\mathbf{z})\| \leq \gamma \|\mathbf{x} - \mathbf{z}\| \quad \text{for all } \mathbf{x}, \mathbf{z} \in \mathrm{dom}(f)$$

Then, for any starting point $\mathbf{x}^0 \in \mathrm{dom}(f)$*, the gradient method converges to* $\mathbf{x}^\star$.

When does the gradient method converge?

Lemma

Let f be a closed, convex function. Assume that

1. *There exists $\mathbf{x}^\star \in \mathrm{dom}(f)$ such that $\nabla f(\mathbf{x}^\star) = 0$.*
2. *The mapping $\psi(\mathbf{x}) = \mathbf{x} - \alpha\nabla f(\mathbf{x})$ is contractive for some α: i.e., there exists $\gamma \in [0, 1)$ such that*

$$\|\psi(\mathbf{x}) - \psi(\mathbf{z})\| \leq \gamma\|\mathbf{x} - \mathbf{z}\| \quad \textit{for all } \mathbf{x}, \mathbf{z} \in \mathrm{dom}(f)$$

Then, for any starting point $\mathbf{x}^0 \in \mathrm{dom}(f)$, the gradient method converges to $\mathbf{x}^\star$.

Proof.

If we start the gradient method at $\mathbf{x}^0 \in \mathrm{dom}(f)$, then we have

$$\begin{aligned}
\|\mathbf{x}^{k+1} - \mathbf{x}^\star\| &= \|\mathbf{x}^k - \alpha\nabla f(\mathbf{x}^k) - \mathbf{x}^\star\| \\
&= \|\psi(\mathbf{x}^k) - \psi(\mathbf{x}^\star)\| && \text{(fixed-point characterization)} \\
&\leq \gamma\|\mathbf{x}^k - \mathbf{x}^\star\| && \text{(contractiveness)} \\
&\leq \gamma^{k+1}\|\mathbf{x}^0 - \mathbf{x}^\star\| .
\end{aligned}$$

We then have that the sequence $\{\mathbf{x}^k\}$ converges globally to $\mathbf{x}^\star$ at a **linear** rate.

□

Short (but important) detour: convergence rates

Definition (Convergence of a sequence)

The sequence $\mathbf{u}^1, \mathbf{u}^2, ..., \mathbf{u}^k, ...$ converges to $\mathbf{u}^\star$ (denoted $\lim_{k\to\infty} \mathbf{u}^k = \mathbf{u}^\star$), if

$$\forall\ \varepsilon > 0, \exists\ K \in \mathbb{N} : k \geq K \Rightarrow \|\mathbf{u}^k - \mathbf{u}^\star\| \leq \varepsilon$$

Convergence rates: the *"speed"* at which a sequence converges

- **sublinear:** if there exists $c > 0$ such that

$$\|\mathbf{u}^k - \mathbf{u}^\star\| = O(k^{-c})$$

- **linear:** if there exists $\alpha \in (0, 1)$ such that

$$\|\mathbf{u}^k - \mathbf{u}^\star\| = O(\alpha^k)$$

- **Q-linear:** if there exists a constant $r \in (0, 1)$ such that

$$\lim_{k\to\infty} \frac{\|\mathbf{u}^{k+1} - \mathbf{u}^\star\|}{\|\mathbf{u}^k - \mathbf{u}^\star\|} = r$$

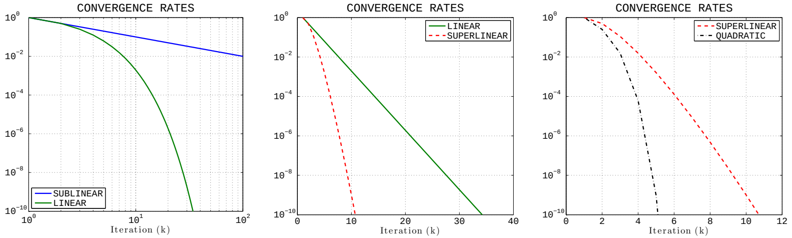
- **superlinear:** If $r = 0$, we say that the sequence converges *superlinearly*.
- **quadratic:** if there exists a constant $\mu > 0$ such that

$$\lim_{k\to\infty} \frac{\|\mathbf{u}^{k+1} - \mathbf{u}^\star\|}{\|\mathbf{u}^k - \mathbf{u}^\star\|^2} = \mu$$

Example: Convergence rates

Examples of sequences that all converge to $u^\star = 0$:

- Sublinear: $u^k = 1/k$
- Linear: $u^k = 0.5^k$
- Superlinear: $u^k = k^{-k}$
- Quadratic: $u^k = 0.5^{2^k}$



Remark

For **unconstrained** convex minimization as in (1), we always have $f(\mathbf{x}^k) - f^\star \geq 0$. Hence, we do not need to use the absolute value when we show convergence results based on the objective value, such as $f(\mathbf{x}^k) - f^\star \leq O(1/k^2)$, which is sublinear.

Contractive maps and convexity

Proposition (Contractivity implies convexity with structure)

Let $f \in \mathcal{C}^2$ *and define* $\psi(\mathbf{x}) = \mathbf{x} - \alpha \nabla f(\mathbf{x})$, *with* $\alpha > 0$.
If $\psi(\mathbf{x})$ *is contractive, with a constant contraction factor* $\gamma < 1$, *then* $f \in \mathcal{F}_{L,\mu}^{2,1}$.

Proof.

Consider $\mathbf{y} = \mathbf{x} + t\Delta\mathbf{x}$. By the contractivity assumption it must hold that

$$\|\psi(\mathbf{x} + t\Delta\mathbf{x}) - \psi(\mathbf{x})\| \leq t\gamma\|\Delta\mathbf{x}\| \quad \forall t \,.$$

We also have that

$$\begin{aligned}
\lim_{t\to 0} \frac{1}{t}\|\psi(\mathbf{x} + t\Delta\mathbf{x}) - \psi(\mathbf{x})\| &= \lim_{t\to 0} \left\|\Delta\mathbf{x} - \frac{\alpha}{t}\left(\nabla f(\mathbf{x} + t\Delta\mathbf{x}) - \nabla f(\mathbf{x})\right)\right\| \\
&= \left\|\left(\mathbf{I} - \alpha\nabla^2 f(\mathbf{x})\right)\Delta\mathbf{x}\right\| \\
&\leq \gamma\|\Delta\mathbf{x}\| \qquad \text{(by assumption)}
\end{aligned}$$

The inequality implies (derivation on the board) that

$$\mathbf{0} \prec \frac{1-\gamma}{\alpha}\mathbf{I} \preceq \nabla^2 f(\mathbf{x}) \preceq \frac{1+\gamma}{\alpha}\mathbf{I},$$

which can be reinterpreted as $f \in \mathcal{F}_{L,\mu}^{2,1}$ with $L = \frac{1+\gamma}{\alpha}$ and $\mu = \frac{1-\gamma}{\alpha}$ (next!). □

Gradient descent methods

Definition

Gradient descent (GD) Starting from $\mathbf{x}^0 \in \mathrm{dom}(f)$, update $\{\mathbf{x}^k\}_{k \geq 0}$ as

$$\mathbf{x}^{k+1} = \mathbf{x}^k - \alpha_k \nabla f(\mathbf{x}^k) = \mathbf{x}^k + \alpha_k \mathbf{p}^k.$$

Notice that $\mathbf{p}^k := -\nabla f(\mathbf{x}^k)$ is the steepest descent (anti-gradient) search direction.

Key question: how to choose α_k to have descent/contraction?

Gradient descent methods

Definition

Gradient descent (GD) Starting from $\mathbf{x}^0 \in \mathrm{dom}(f)$, update $\{\mathbf{x}^k\}_{k\geq 0}$ as

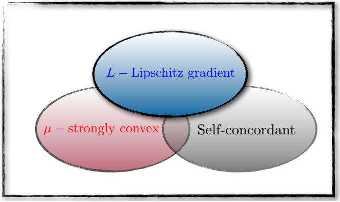
$$\mathbf{x}^{k+1} = \mathbf{x}^k - \alpha_k \nabla f(\mathbf{x}^k) = \mathbf{x}^k + \alpha_k \mathbf{p}^k.$$

Notice that $\mathbf{p}^k := -\nabla f(\mathbf{x}^k)$ is the steepest descent (anti-gradient) search direction.

Key question: how to choose α_k to have descent/contraction?

We need structure!

We use $\mathcal{F}$ to denote the class of smooth convex functions.
(The domain of each function will be apparent from the context.)



Next few slides: structural assumptions

L-Lipschitz gradient class of functions

Definition (L-Lipschitz gradient convex functions)

Let $f : \mathcal{Q} \to \mathbb{R}$ be differentiable and convex, i.e., $f \in \mathcal{F}^1(\mathcal{Q})$. Then, f has a Lipschitz gradient if there exists $L > 0$ (the Lipschitz constant) s.t.

$$\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\|_2 \leq L\|\mathbf{x} - \mathbf{y}\|_2, \quad \forall \mathbf{x}, \mathbf{y} \in \mathcal{Q}.$$

Proposition (L-Lipschitz gradient convex functions)

$f \in \mathcal{F}^1(\mathcal{Q})$ has L-Lipschitz gradient if and only if the following function is convex:

$$h(\mathbf{x}) = \frac{L}{2}\|\mathbf{x}\|_2^2 - f(\mathbf{x}) \quad \forall \mathbf{x} \in \mathcal{Q}.$$

Definition (Class of 2-nd order Lipschitz functions)

The class of twice continuously differentiable functions f on $\mathcal{Q}$ with Lipschitz continuous Hessian is denoted as $\mathcal{F}_L^{2,2}(\mathcal{Q})$ (with $2 \to 2$ denoting the spectral norm)

$$\|\nabla^2 f(\mathbf{x}) - \nabla^2 f(\mathbf{y})\|_{2\to 2} \leq L\|\mathbf{x} - \mathbf{y}\|_2, \quad \forall \mathbf{x}, \mathbf{y} \in Q,$$

- $\mathcal{F}_L^{l,m}$: functions that are l-times differentiable with m-th order Lipschitz property.

Example: Logistic regression

Problem (Logistic regression)

Given a sample vector $\mathbf{a}_i \in \mathbb{R}^p$ *and a binary class label* $b_i \in \{-1, +1\}$ $(i = 1, \ldots, n)$*, we define the conditional probability of* b_i *given* $\mathbf{a}_i$ *as:*

$$\mathbb{P}(b_i | \mathbf{a}_i, \mathbf{x}^\natural, \mu) \propto 1/(1 + e^{-b_i(\langle \mathbf{x}^\natural, \mathbf{a}_i \rangle + \mu)}),$$

where $\mathbf{x}^\natural \in \mathbb{R}^p$ *is some true weight vector,* $\mu \in \mathbb{R}$ *is called the intercept. How to estimate* $\mathbf{x}^\natural$ *given the sample vectors, the binary labels, and* μ*?*

Optimization formulation

$$\min_{\mathbf{x} \in \mathbb{R}^p} \underbrace{\frac{1}{n} \sum_{i=1}^{n} \log(1 + \exp(-b_i(\mathbf{a}_i^T \mathbf{x} + \mu)))}_{f(\mathbf{x})}$$

Structural properties

Let $\mathbf{A} = [\mathbf{a}_1, \ldots, \mathbf{a}_n]^T$ (design matrix), then $f \in \mathcal{F}_L^{2,1}$, with $L = \frac{1}{4}\|\mathbf{A}^T\mathbf{A}\|$

μ-strongly convex functions

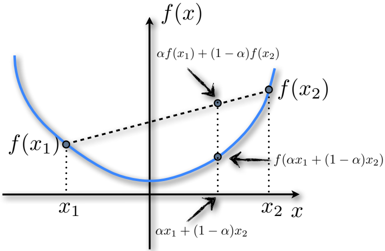
Definition

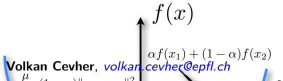
A function $f : \mathcal{Q} \to \mathbb{R} \cup \{+\infty\}, \mathcal{Q} \subseteq \mathbb{R}^p$ is called μ-*strongly* convex on its domain if and only if for any $\mathbf{x},\ \mathbf{y} \in \mathcal{Q}$ and $\alpha \in [0, 1]$ we have:

$$f(\alpha\mathbf{x} + (1-\alpha)\mathbf{y}) \leq \alpha f(\mathbf{x}) + (1-\alpha) f(\mathbf{y}) - \frac{\mu}{2}\alpha(1-\alpha)\|\mathbf{x} - \mathbf{y}\|_2^2.$$

The constant μ is called the convexity parameter of function f.

- The class of k-differentiable μ-strongly functions is denoted as $\mathcal{F}_\mu^k(\mathcal{Q})$.
- Strong convexity $\Rightarrow$ strict convexity, BUT strict convexity $\not\Rightarrow$ strong convexity





Alternative: μ-strongly convex functions

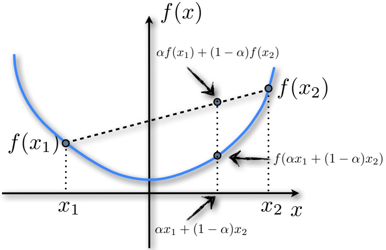
Definition

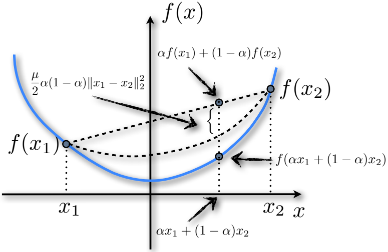
A convex function $f : \mathcal{Q} \to \mathbb{R}$ is said to be μ-strongly convex if

$$h(\mathbf{x}) = f(\mathbf{x}) - \frac{\mu}{2}\|\mathbf{x}\|_2^2$$

is convex, where μ is called the strong convexity parameter.

- The class of k-differentiable μ-strongly functions is denoted as $\mathcal{F}_\mu^k(\mathcal{Q})$.
- Non-smooth functions can be μ-strongly convex: e.g., $f(\mathbf{x}) = \|\mathbf{x}\|_1 + \frac{\mu}{2}\|\mathbf{x}\|_2^2$.





Example: Least-squares estimation

Problem

Let $\mathbf{x}^\natural \in \mathbb{R}^p$ *and* $\mathbf{A} \in \mathbb{R}^{n\times p}$ *(full column rank). Goal: estimate* $\mathbf{x}^\natural$*, given* $\mathbf{A}$ *and*

$$\mathbf{b} = \mathbf{A}\mathbf{x}^\natural + \mathbf{w},$$

where $\mathbf{w}$ *denotes unknown noise.*

Optimization formulation (Least-squares estimator)

$$\min_{\mathbf{x}\in\mathbb{R}^p} \underbrace{\frac{1}{2}\|\mathbf{b} - \mathbf{A}\mathbf{x}\|_2^2}_{f(\mathbf{x})} .$$

Structural properties

- $\nabla f(\mathbf{x}) = \mathbf{A}^T(\mathbf{A}\mathbf{x} - \mathbf{b})$, and $\nabla^2 f(\mathbf{x}) = \mathbf{A}^T\mathbf{A}$.
- $\lambda_p\mathbf{I} \preceq \nabla^2 f(\mathbf{x}) \preceq \lambda_1\mathbf{I}$, where $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p$ are the eigenvalues of $\mathbf{A}^T\mathbf{A}$.
- It follows that $L = \lambda_1$ and $\mu = \lambda_p$. If $\lambda_p > 0$, then $f \in \mathcal{F}^{2,1}_{L,\mu}$, otherwise $f \in \mathcal{F}^{2,1}_{L}$.
- Since $\mathrm{rank}(\mathbf{A}^T\mathbf{A}) \leq \min\{n, p\}$, if $n < p$, then $\lambda_p = 0$.

⋆Self-concordant functions

Definition (Self-concordant functions in 1-dimension)

A convex function $\varphi : \mathbb{R} \to \mathbb{R}$ is self-concordant if

$$|\varphi'''(t)| \leq 2\varphi''(t)^{3/2}, \quad \forall t \in \mathbb{R}.$$

*Self-concordant functions

Definition (Self-concordant functions in 1-dimension)

A convex function $\varphi : \mathbb{R} \to \mathbb{R}$ is self-concordant if

$$|\varphi'''(t)| \leq 2\varphi''(t)^{3/2}, \quad \forall t \in \mathbb{R}.$$

Affine Invariance of self-concordant functions

Let $\tilde{\varphi}(t) = \varphi(\alpha t + \beta)$ where $\alpha \neq 0$. Then, $\tilde{\varphi}$ is self-concordant iff φ is.

*Self-concordant functions

Definition (Self-concordant functions in 1-dimension)

A convex function $\varphi : \mathbb{R} \to \mathbb{R}$ is self-concordant if

$$|\varphi'''(t)| \leq 2\varphi''(t)^{3/2}, \quad \forall t \in \mathbb{R}.$$

Affine Invariance of self-concordant functions

Let $\tilde{\varphi}(t) = \varphi(\alpha t + \beta)$ where $\alpha \neq 0$. Then, $\tilde{\varphi}$ is self-concordant iff φ is.

Important remarks of self-concordance

1. Generalize to higher dimension: A convex function $f : \mathbb{R}^n \to \mathbb{R}$ is said to be (standard) self-concordant if $|\varphi'''(t)| \leq 2\varphi''(t)^{3/2}$, where $\varphi(t) := f(\mathbf{x} + t\mathbf{v})$ for all $t \in \mathbb{R}$, $\mathbf{x} \in \operatorname{dom} f$ and $\mathbf{v} \in \mathbb{R}^n$ such that $\mathbf{x} + t\mathbf{v} \in \operatorname{dom} f$.
2. Affine invariance still holds in high dimension.
3. Self-concordant functions are efficiently minimized by the Newton method and its variants (see Lecture 6).

Back to gradient descent methods

Gradient descent (GD) algorithm

Starting from $\mathbf{x}^0 \in \mathrm{dom}(f)$, produce the sequence $\mathbf{x}^1, ..., \mathbf{x}^k, ...$ according to

$$\mathbf{x}^{k+1} = \mathbf{x}^k - \alpha_k \nabla f(\mathbf{x}^k) = \mathbf{x}^k + \alpha_k \mathbf{p}^k.$$

Notice that $\mathbf{p}^k := -\nabla f(\mathbf{x}^k)$ is the steepest descent (anti-gradient) direction.
Key question: how do we choose α_k to have descent/contraction?

Back to gradient descent methods

Gradient descent (GD) algorithm

Starting from $\mathbf{x}^0 \in \mathrm{dom}(f)$, produce the sequence $\mathbf{x}^1, ..., \mathbf{x}^k, ...$ according to

$$\mathbf{x}^{k+1} = \mathbf{x}^k - \alpha_k \nabla f(\mathbf{x}^k) = \mathbf{x}^k + \alpha_k \mathbf{p}^k.$$

Notice that $\mathbf{p}^k := -\nabla f(\mathbf{x}^k)$ is the steepest descent (anti-gradient) direction.
Key question: how do we choose α_k to have descent/contraction?

Step-size selection

Case 1: If $f \in \mathcal{F}_L^{1,1}(\mathbb{R}^p)$, then:

- We can choose $0 < \alpha_k < \frac{2}{L}$. The optimal choice is $\alpha_k := \frac{1}{L}$.
- α_k can be determined by a line-search procedure:
  1. **Exact line search**: $\alpha_k := \arg\min_{\alpha > 0} f(\mathbf{x}^k - \alpha \nabla f(\mathbf{x}^k))$.
  2. **Back-tracking line search** with Armijo-Goldstein's condition:

$$f(\mathbf{x}^k - \alpha \nabla f(\mathbf{x}^k)) \leq f(\mathbf{x}^k) - c\alpha \|\nabla f(\mathbf{x}^k)\|^2, \quad c \in (0, 1/2].$$

Case 2: If $f \in \mathcal{F}_{L,\mu}^{1,1}(\mathbb{R}^p)$, then:

- We can choose $0 < \alpha_k \leq \frac{2}{L+\mu}$. The optimal choice is $\alpha_k := \frac{2}{L+\mu}$.

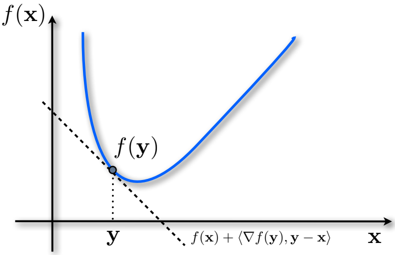
Case 3: If $f \in \mathcal{F}_2(\mathcal{Q})$, then, a bit more complicated (more later).

Towards a geometric interpretation I

Recall:

- Let $f \in \mathcal{F}_L^2(\mathbb{R}^p)$ with gradient $\nabla f(\mathbf{x})$ and Hessian $\nabla^2 f(\mathbf{x})$.
- First-order Taylor approximation of f at $\mathbf{y}$:

$$f(\mathbf{x}) \geq f(\mathbf{y}) + \langle \nabla f(\mathbf{y}), \mathbf{x} - \mathbf{y} \rangle$$



- Convex functions: 1^{st}**-order Taylor approximation is a global lower surrogate.**

Towards a geometric interpretation II

Lemma

Let $f \in \mathcal{F}_L^{1,1}(\mathcal{Q})$. *Then, we have:*

$$f(\mathbf{y}) - f(\mathbf{x}) - \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle \leq \frac{L}{2} \|\mathbf{y} - \mathbf{x}\|_2^2, \quad \forall \mathbf{x}, \mathbf{y} \in \mathcal{Q}.$$

Proof.

By the Taylor's theorem:

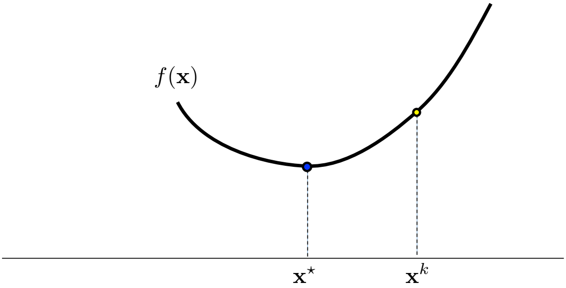
$$f(\mathbf{y}) = f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle + \int_0^1 \langle \nabla f(\mathbf{x} + \tau(\mathbf{y} - \mathbf{x})) - \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle d\tau.$$

Therefore,

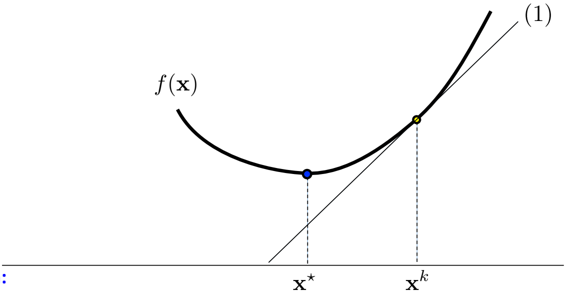
$$\begin{aligned} f(\mathbf{y}) - f(\mathbf{x}) - \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle &\leq \int_0^1 \|\nabla f(\mathbf{x} + \tau(\mathbf{y} - \mathbf{x})) - \nabla f(\mathbf{x})\|^* \cdot \|\mathbf{y} - \mathbf{x}\| d\tau \\ &\leq L \|\mathbf{y} - \mathbf{x}\|_2^2 \int_0^1 \tau d\tau = \frac{L}{2} \|\mathbf{y} - \mathbf{x}\|_2^2 \end{aligned}$$

□

Gradient descent methods: geometrical intuition



Gradient descent methods: geometrical intuition



Structure in optimization:

$(1) \qquad f(\mathbf{x}) \geq f(\mathbf{x}^k) + \langle \nabla f(\mathbf{x}^k), \mathbf{x} - \mathbf{x}^k \rangle$

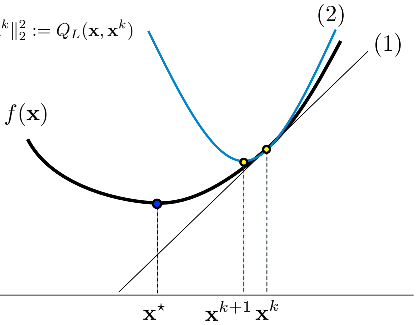
Gradient descent methods: geometrical intuition

Majorize:

$$f(\mathbf{x}) \leq f(\mathbf{x}^k) + \langle \nabla f(\mathbf{x}^k), \mathbf{x} - \mathbf{x}^k \rangle + \frac{L}{2}\|\mathbf{x} - \mathbf{x}^k\|_2^2 := Q_L(\mathbf{x}, \mathbf{x}^k)$$

Minimize:

$$\begin{aligned}\mathbf{x}^{k+1} &= \arg\min_{\mathbf{x}} Q_L(\mathbf{x}, \mathbf{x}^k) \\ &= \arg\min_{\mathbf{x}} \left\| \mathbf{x} - \left(\mathbf{x}^k - \frac{1}{L}\nabla f(\mathbf{x}^k) \right) \right\|^2 \\ &= \mathbf{x}^k - \frac{1}{L}\nabla f(\mathbf{x}^k)\end{aligned}$$



Structure in optimization:

(1) $\quad f(\mathbf{x}) \geq f(\mathbf{x}^k) + \langle \nabla f(\mathbf{x}^k), \mathbf{x} - \mathbf{x}^k \rangle$

(2) $\quad f(\mathbf{x}) \leq f(\mathbf{x}^k) + \langle \nabla f(\mathbf{x}^k), \mathbf{x} - \mathbf{x}^k \rangle + \frac{L}{2}\|\mathbf{x} - \mathbf{x}^k\|_2^2$

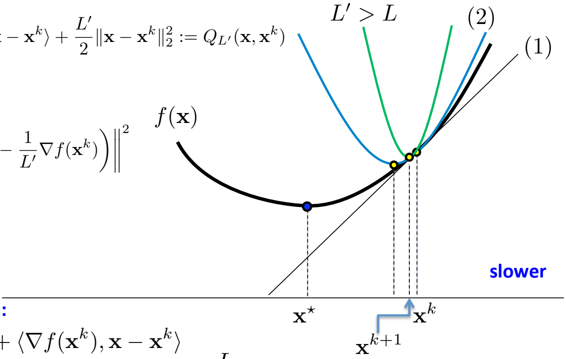
Gradient descent methods: geometrical intuition

Majorize:

$$f(\mathbf{x}) \leq f(\mathbf{x}^k) + \langle \nabla f(\mathbf{x}^k), \mathbf{x} - \mathbf{x}^k \rangle + \frac{L'}{2}\|\mathbf{x} - \mathbf{x}^k\|_2^2 := Q_{L'}(\mathbf{x}, \mathbf{x}^k)$$

Minimize:

$$\begin{aligned}\mathbf{x}^{k+1} &= \arg\min_{\mathbf{x}} Q_{L'}(\mathbf{x}, \mathbf{x}^k) \\ &= \arg\min_{\mathbf{x}} \left\| \mathbf{x} - \left(\mathbf{x}^k - \frac{1}{L'}\nabla f(\mathbf{x}^k) \right) \right\|^2 \\ &= \mathbf{x}^k - \frac{1}{L'}\nabla f(\mathbf{x}^k)\end{aligned}$$



Structure in optimization:

(1) $\quad f(\mathbf{x}) \geq f(\mathbf{x}^k) + \langle \nabla f(\mathbf{x}^k), \mathbf{x} - \mathbf{x}^k \rangle$

(2) $\quad f(\mathbf{x}) \leq f(\mathbf{x}^k) + \langle \nabla f(\mathbf{x}^k), \mathbf{x} - \mathbf{x}^k \rangle + \frac{L}{2}\|\mathbf{x} - \mathbf{x}^k\|_2^2$

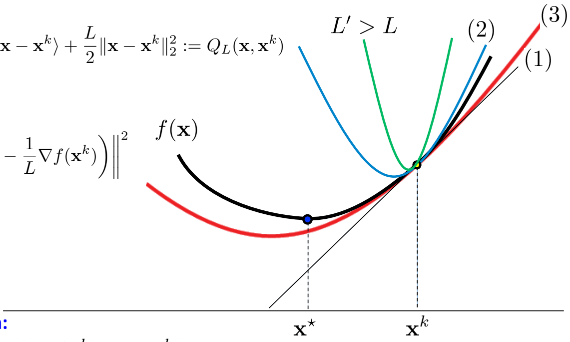
Gradient descent methods: geometrical intuition

Majorize:

$$f(\mathbf{x}) \le f(\mathbf{x}^k) + \langle \nabla f(\mathbf{x}^k), \mathbf{x} - \mathbf{x}^k \rangle + \frac{L}{2}\|\mathbf{x} - \mathbf{x}^k\|_2^2 := Q_L(\mathbf{x}, \mathbf{x}^k)$$

Minimize:

$$\begin{aligned}\mathbf{x}^{k+1} &= \arg\min_{\mathbf{x}} Q_L(\mathbf{x}, \mathbf{x}^k) \\ &= \arg\min_{\mathbf{x}} \left\| \mathbf{x} - \left(\mathbf{x}^k - \frac{1}{L}\nabla f(\mathbf{x}^k) \right) \right\|^2 \\ &= \mathbf{x}^k - \frac{1}{L}\nabla f(\mathbf{x}^k)\end{aligned}$$



Structure in optimization:

(1) $f(\mathbf{x}) \ge f(\mathbf{x}^k) + \langle \nabla f(\mathbf{x}^k), \mathbf{x} - \mathbf{x}^k \rangle$

(2) $f(\mathbf{x}) \le f(\mathbf{x}^k) + \langle \nabla f(\mathbf{x}^k), \mathbf{x} - \mathbf{x}^k \rangle + \frac{L}{2}\|\mathbf{x} - \mathbf{x}^k\|_2^2$

(3) $f(\mathbf{x}) \ge f(\mathbf{x}^k) + \langle \nabla f(\mathbf{x}^k), \mathbf{x} - \mathbf{x}^k \rangle + \frac{\mu}{2}\|\mathbf{x} - \mathbf{x}^k\|_2^2$

Convergence rate of gradient descent

Theorem

$$f \in \mathcal{F}_L^{2,1}, \quad \alpha = \frac{1}{L}: \qquad f(\mathbf{x}^k) - f(\mathbf{x}^\star) \leq \frac{2L}{k+4}\|\mathbf{x}^0 - \mathbf{x}^\star\|_2^2$$

$$f \in \mathcal{F}_{L,\mu}^{2,1}, \quad \alpha = \frac{2}{L+\mu}: \qquad \|\mathbf{x}^k - \mathbf{x}^\star\|_2 \leq \left(\frac{L-\mu}{L+\mu}\right)^k \|\mathbf{x}^0 - \mathbf{x}^\star\|_2$$

$$f \in \mathcal{F}_{L,\mu}^{2,1}, \quad \alpha = \frac{1}{L}: \qquad \|\mathbf{x}^k - \mathbf{x}^\star\|_2 \leq \left(\frac{L-\mu}{L+\mu}\right)^{\frac{k}{2}} \|\mathbf{x}^0 - \mathbf{x}^\star\|_2$$

Note that $\frac{L-\mu}{L+\mu} = \frac{\kappa-1}{\kappa+1}$, where $\kappa := \frac{L}{\mu}$ is the condition number of $\nabla^2 f$.

Convergence rate of gradient descent

Theorem

$$f \in \mathcal{F}_L^{2,1}, \quad \alpha = \frac{1}{L}: \qquad f(\mathbf{x}^k) - f(\mathbf{x}^\star) \leq \frac{2L}{k+4}\|\mathbf{x}^0 - \mathbf{x}^\star\|_2^2$$

$$f \in \mathcal{F}_{L,\mu}^{2,1}, \quad \alpha = \frac{2}{L+\mu}: \qquad \|\mathbf{x}^k - \mathbf{x}^\star\|_2 \leq \left(\frac{L-\mu}{L+\mu}\right)^k \|\mathbf{x}^0 - \mathbf{x}^\star\|_2$$

$$f \in \mathcal{F}_{L,\mu}^{2,1}, \quad \alpha = \frac{1}{L}: \qquad \|\mathbf{x}^k - \mathbf{x}^\star\|_2 \leq \left(\frac{L-\mu}{L+\mu}\right)^{\frac{k}{2}} \|\mathbf{x}^0 - \mathbf{x}^\star\|_2$$

Note that $\frac{L-\mu}{L+\mu} = \frac{\kappa-1}{\kappa+1}$, where $\kappa := \frac{L}{\mu}$ is the condition number of $\nabla^2 f$.

Remarks

- **Assumption:** Lipschitz gradient. **Result:** convergence rate in **objective values**.
- **Assumption:** Strong convexity. **Result:** convergence rate in **sequence** of the iterates and in **objective values**.
- Note that the suboptimal step-size choice $\alpha = \frac{1}{L}$ adapts to the strongly convex case (i.e., it features a linear rate vs. the standard sublinear rate).

Example: Ridge regression

Optimization formulation

- Let $\mathbf{A} \in \mathbb{R}^{n \times p}$ and $\mathbf{b} \in \mathbb{R}^n$ given by $\mathbf{b} = \mathbf{A}\mathbf{x}^\natural + \mathbf{w}$, where $\mathbf{w} \in \mathbb{R}^n$ is some noise.
- A classical estimator of $\mathbf{x}^\natural$, known as ridge regression, is

$$\min_{\mathbf{x} \in \mathbb{R}^p} f(\mathbf{x}) := \frac{1}{2}\|\mathbf{b} - \mathbf{A}\mathbf{x}\|_2^2 + \frac{\rho}{2}\|\mathbf{x}\|_2^2.$$

where $\rho \geq 0$ is a regularization parameter

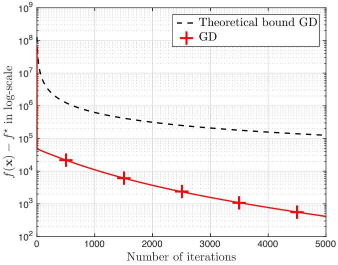
Remarks

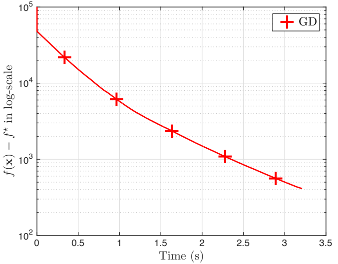
- $f \in \mathcal{F}_{L,\mu}^{2,1}$ with:
  - $L = \lambda_1(\mathbf{A}^T\mathbf{A}) + \rho$;
  - $\mu = \lambda_p(\mathbf{A}^T\mathbf{A}) + \rho$;
  - where $\lambda_1 \geq \ldots \geq \lambda_p$ are the eigenvalues of $\mathbf{A}^T\mathbf{A}$.
- The ratio $\kappa = \frac{L}{\mu}$ decreases as ρ increases, leading to faster linear convergence.
- Note that if $n < p$ and $\rho = 0$, we have $\mu = 0$, hence $f \in \mathcal{F}_L^{2,1}$ and we can expect only $\mathcal{O}(1/k)$ convergence from the gradient descent method.

Example: Ridge regression

Case 1:

$n = 500, p = 2000, \rho = 0$





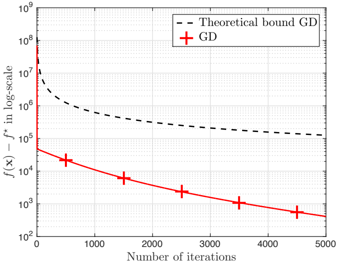
Example: Ridge regression

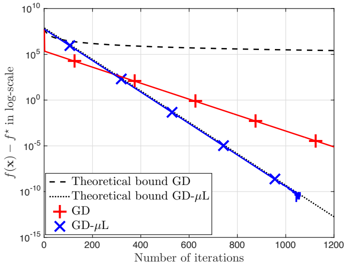
Case 1:

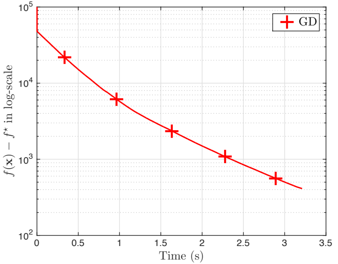
$n = 500, p = 2000, \rho = 0$

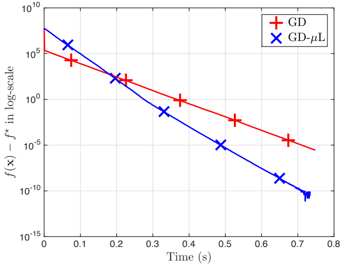
Case 2:

$n = 500, p = 2000, \rho = 0.01\lambda_p(\mathbf{A}^T\mathbf{A})$









*From gradient descent to mirror descent

Gradient descent as a majorization-minimization scheme

- **Majorize** f at $\mathbf{x}^k$ by using L-Lipschitz gradient continuity

$$f(\mathbf{x}) \leq f(\mathbf{x}^k) + \langle \nabla f(\mathbf{x}^k), \mathbf{x} - \mathbf{x}^k \rangle + \frac{L}{2}\|\mathbf{x} - \mathbf{x}^k\|_2^2 := Q(\mathbf{x}, \mathbf{x}^k)$$

- **Minimize** $Q(\mathbf{x}, \mathbf{x}^k)$ to obtain the next iterate $\mathbf{x}^{k+1}$

$$\mathbf{x}^{k+1} = \arg\min_{\mathbf{x}} Q(\mathbf{x}, \mathbf{x}^k) \Rightarrow \nabla f(\mathbf{x}^k) + L(\mathbf{x}^{k+1} - \mathbf{x}^k) = 0$$
$$\mathbf{x}^{k+1} = \mathbf{x}^k - \frac{1}{L}\nabla f(\mathbf{x}^k)$$

Other majorizers

We can re-write the majorization step as

$$f(\mathbf{x}) \leq f(\mathbf{x}^k) + \langle \nabla f(\mathbf{x}^k), \mathbf{x} - \mathbf{x}^k \rangle + \alpha d(\mathbf{x}, \mathbf{x}^k)$$

where $d(\mathbf{x}, \mathbf{x}^k) = \frac{1}{2}\|\mathbf{x} - \mathbf{x}^k\|_2^2$ is the Euclidean distance and $\alpha = L$.

- Can we use a different function $d(\mathbf{x}, \mathbf{x}^k)$ that is better suited to minimizing f?

*Bregman divergences

Definition (Bregman divergence)

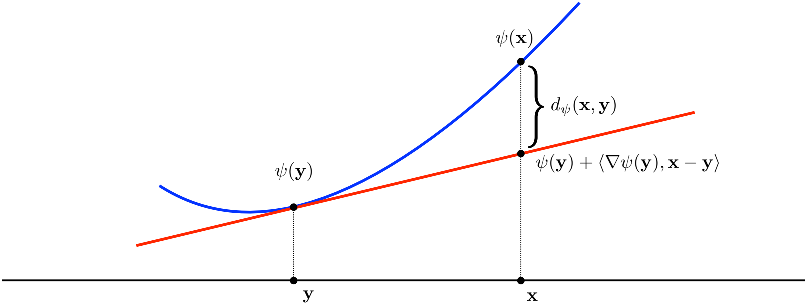
Let $\psi : \mathcal{S} \to \mathbb{R}$ be a continuously-differentiable and strictly convex function defined on a closed convex set $\mathcal{S}$. The **Bregman divergence** (d_ψ) associated with ψ for points $\mathbf{x}$ and $\mathbf{y}$ is:

$$d_\psi(\mathbf{x}, \mathbf{y}) = \psi(\mathbf{x}) - \psi(\mathbf{y}) - \langle \nabla \psi(\mathbf{y}), \mathbf{x} - \mathbf{y} \rangle$$

- $\psi(\cdot)$ is referred to as the Bregman or proximity function.
- The Bregman divergence satisfies the following properties:
  (a) $d_\psi(\mathbf{x}, \mathbf{y}) \geq 0$ for all $\mathbf{x}$ and $\mathbf{y}$ with equality if and only if $\mathbf{x} = \mathbf{y}$
  (b) Define $q(\mathbf{x}) := d_\psi(\mathbf{x}, \mathbf{y})$ for a fixed $\mathbf{y}$, then $\nabla q(\mathbf{x}) = \nabla \psi(\mathbf{x}) - \nabla \psi(\mathbf{y})$
  (c) For all $\mathbf{x}, \mathbf{y}, \mathbf{z} \in \mathcal{S}$, $d_\psi(\mathbf{x}, \mathbf{y}) = d_\psi(\mathbf{x}, \mathbf{z}) + d_\psi(\mathbf{z}, \mathbf{y}) + \langle (\mathbf{x} - \mathbf{z}), \nabla \psi(\mathbf{y}) - \nabla \psi(\mathbf{z}) \rangle$
  (d) For all $\mathbf{x}, \mathbf{y} \in \mathcal{S}$, $d_\psi(\mathbf{x}, \mathbf{y}) + d_\psi(\mathbf{y}, \mathbf{x}) = \langle (\mathbf{x} - \mathbf{y}), \nabla \psi(\mathbf{x}) - \nabla \psi(\mathbf{y}) \rangle$
- The Bregman divergence becomes a Bregman distance when it is *symmetric* (i.e. $d_\psi(\mathbf{x}, \mathbf{y}) = d_\psi(\mathbf{y}, \mathbf{x})$) and satisfies the *triangle inequality*.
- "*All Bregman distances are Bregman divergences but the reverse is not true!*"

*Bregman divergences

- The Bregman divergence is the vertical distance at $\mathbf{x}$ between ψ and the tangent of ψ at $\mathbf{y}$, see figure below



- The Bregman divergence measures the strictness of convexity of $\psi(\cdot)$.

*Bregman divergences

Table: **Bregman functions** $\psi(\mathbf{x})$ & corresponding Bregman divergences/distances $d_\psi(\mathbf{x},\mathbf{y})^a$.

Name (or Loss)	Domain^b	$\psi(\mathbf{x})$	$d_\psi(\mathbf{x},\mathbf{y})$
Squared loss	$\mathbb{R}$	x^2	$(x-y)^2$
Itakura-Saito divergence	$\mathbb{R}_{++}$	$-\log x$	$\frac{x}{y} - \log\left(\frac{x}{y}\right) - 1$
Squared Euclidean distance	$\mathbb{R}^p$	$\|\mathbf{x}\|_2^2$	$\|\mathbf{x}-\mathbf{y}\|_2^2$
Squared Mahalanobis distance	$\mathbb{R}^p$	$\langle \mathbf{x}, \mathbf{A}\mathbf{x}\rangle$	$\langle (\mathbf{x}-\mathbf{y}), \mathbf{A}(\mathbf{x}-\mathbf{y})\rangle^c$
Entropy distance	p-simplex^d	$\sum_i x_i \log x_i$	$\sum_i x_i \log\left(\frac{x_i}{y_i}\right)$
Generalized I-divergence	$\mathbb{R}^p_+$	$\sum_i x_i \log x_i$	$\sum_i \left(\log\left(\frac{x_i}{y_i}\right) - \left(x_i - y_i\right)\right)$
von Neumann divergence	$\mathbb{S}^{p\times p}_+$	$\mathbf{X}\log\mathbf{X} - \mathbf{X}$	$\mathrm{tr}\left(\mathbf{X}(\log\mathbf{X} - \log\mathbf{Y}) - \mathbf{X} + \mathbf{Y}\right)^e$
logdet divergence	$\mathbb{S}^{p\times p}_+$	$-\log\det\mathbf{X}$	$\mathrm{tr}\left(\mathbf{X}\mathbf{Y}^{-1}\right) - \log\det\left(\mathbf{X}\mathbf{Y}^{-1}\right) - p$

^{a} $x, y \in \mathbb{R}$, $\mathbf{x},\mathbf{y} \in \mathbb{R}^p$ and $\mathbf{X},\mathbf{Y} \in \mathbb{R}^{p\times p}$.

^{b} $\mathbb{R}_+$ and $\mathbb{R}_{++}$ denote non-negative and positive real numbers respectively.

^{c} $\mathbf{A} \in \mathbb{S}^{p\times p}_+$, the set of symmetric positive semidefinite matrix.

^{d} p-simplex:= $\{\mathbf{x} \in \mathbb{R}^p : \sum_{i=1}^p x_i = 1, x_i \geq 0, i = 1, \ldots, p\}$

^{e} $\mathrm{tr}(\mathbf{A})$ is the trace of $\mathbf{A}$.

*Mirror descent [1]

What happens if we use a Bregman distance d_ψ in gradient descent?

Let $\psi : \mathbb{R}^p \to \mathbb{R}$ be a μ-strongly convex and continuously differentiable function and let the associated Bregman distance be $d_\psi(\mathbf{x}, \mathbf{y}) = \psi(\mathbf{x}) - \psi(\mathbf{y}) - \langle \mathbf{x} - \mathbf{y}, \nabla\psi(\mathbf{y})\rangle$. Assume that the inverse mapping $\psi^\star$ of ψ is easily computable (i.e., its convex conjugate).

- **Majorize**: Find α_k such that

$$f(\mathbf{x}) \leq f(\mathbf{x}^k) + \langle \nabla f(\mathbf{x}^k), \mathbf{x} - \mathbf{x}^k\rangle + \frac{1}{\alpha_k} d_\psi(\mathbf{x}, \mathbf{x}^k) := Q_\psi^k(\mathbf{x}, \mathbf{x}^k)$$

- **Minimize**

$$\mathbf{x}^{k+1} = \arg\min_{\mathbf{x}} Q_\psi^k(\mathbf{x}, \mathbf{x}^k) \Rightarrow \nabla f(\mathbf{x}^k) + \frac{1}{\alpha_k}\Big(\nabla\psi(\mathbf{x}^{k+1}) - \nabla\psi(\mathbf{x}^k)\Big) = 0$$

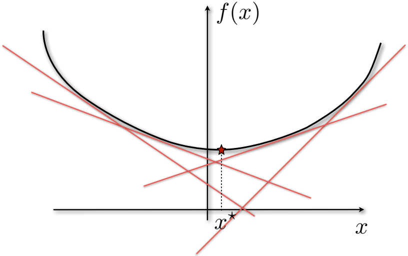
$$\nabla\psi(\mathbf{x}^{k+1}) = \nabla\psi(\mathbf{x}^k) - \alpha_k \nabla f(\mathbf{x}^k)$$

$$\mathbf{x}^{k+1} = \nabla\psi^*(\nabla\psi(\mathbf{x}^k) - \alpha_k\nabla f(\mathbf{x}^k)) \qquad (\nabla\psi(\cdot))^{-1} = \nabla\psi^*(\cdot)[4].$$

- Mirror descent is a **generalization** of gradient descent for functions that are Lipschitz-gradient in norms other than the Euclidean.
- MD allows to deal with some **constraints** via a proper choice of ψ.

⋆What to keep in mind about mirror descent?

- Approximates the optimum by lower bounding the function via hyperplanes at $\mathbf{x}_t$



- The smaller the gradients, the better the approximation!

*Mirror descent example

How can we minimize a convex function over the unit simplex?

$$\min_{\mathbf{x}\in\Delta} f(\mathbf{x}),$$

where

- $\Delta := \{\mathbf{x} \in \mathbb{R}^p \ : \ \sum_{j=1}^p x_j = 1, \mathbf{x} \geq 0\}$ is the **unit simplex**;
- f is convex L_f-Lipschitz continuous with respect to some norm $\|\cdot\|$.

Entropy function

- Define the entropy function

$$\psi_e(\mathbf{x}) = \sum_{j=1}^{p} x_j \ln x_j \quad \text{if } \mathbf{x} \in \Delta, \quad +\infty \text{ otherwise.}$$

- ψ_e is 1-strongly convex over $\mathrm{int}\Delta$ with respect to $\|\cdot\|_1$.
- $\psi_e^\star(\mathbf{z}) = \ln \sum_{j=1}^p e^{z_j}$ and $\|\nabla \psi_e(\mathbf{x})\| \to \infty$ as $\mathbf{x} \to \tilde{\mathbf{x}} \in \Delta$.
- Let $\mathbf{x}^0 = p^{-1}\mathbf{1}$, then $d_\psi(\mathbf{x}, \mathbf{x}^0) \leq \ln p$ for all $\mathbf{x} \in \Delta$.

*Entropic descent algorithm [1]

Entropic descent algorithm (EDA)

Let $\mathbf{x}^0 = p^{-1}\mathbf{1}$ and generate the following sequence

$$x_j^{k+1} = \frac{x_j^k e^{-t_k f_j'(\mathbf{x}^k)}}{\sum_{j=1}^p x_j^k e^{-t_k f_j'(\mathbf{x}^k)}}, \quad t_k = \frac{\sqrt{2\ln p}}{L_f}\frac{1}{\sqrt{k}},$$

where $f'(\mathbf{x}) = (f_1(\mathbf{x})', \ldots, f_p(\mathbf{x})')^T \in \partial f(\mathbf{x})$, which is the **subdifferential** of f at $\mathbf{x}$.

- This is an example of **non-smooth** and **constrained** optimization;
- The updates are multiplicative.

⋆Convergence of mirror descent

Problem

$$\min_{\mathbf{x}\in\mathcal{X}} f(\mathbf{x}) \tag{1}$$

where

- $\mathcal{X}$ *is a closed convex subset of* $\mathbb{R}^p$;
- f *is convex* L_f*-Lipschitz continuous with respect to some norm* $\|\cdot\|$.

Theorem ([1])

Let $\{\mathbf{x}^k\}$ *be the sequence generated by mirror descent with* $\mathbf{x}^0 \in \mathrm{int}\mathcal{X}$.
If the step-sizes are chosen as

$$\alpha_k = \frac{\sqrt{2\mu d_\psi(\mathbf{x}^\star, \mathbf{x}^0)}}{L_f}\frac{1}{\sqrt{k}}$$

the following convergence rate holds

$$\min_{0\leq s\leq k} f(\mathbf{x}^k) - f^\star \leq L_f\sqrt{\frac{2d_\psi(\mathbf{x}^\star, \mathbf{x}^0)}{\mu}}\frac{1}{\sqrt{k}}$$

- *This convergence rate is* **optimal** *for solving* (1) *with a first-order method.*

Smooth unconstrained non-convex minimization

Problem (Mathematical formulation)

Let us consider the following problem formulation:

$$\min_{\mathbf{x}\in\mathbb{R}^p} f(\mathbf{x})$$

- f *is a smooth and possibly non-convex function.*
- *Recall that finding the global minimizer, i.e.,* $f^\star := \min_{\mathbf{x}\in\mathbb{R}^p} f(\mathbf{x})$, *is NP-hard*

Example: Recall M-estimation

Problem: Binary classification

Let $(a_1, b_1), ..., (a_n, b_n)$ be i.i.d. samples where a_i represent a feature vector and b_i its corresponding label (either 0 or 1). **Goal:** Learn a model that gives probability of assigning a particular label, given the feature vector.

Formulation (*M-estimator*)

Find the minimizing parameter for some "loss function", averaged over multiple samples.

$$\min_{x \in \mathcal{X}} \frac{1}{n} \sum_{i=1}^{n} l_i(a_i, b_i; \mathbf{x})$$

- What if we want to select a non-convex loss function?
- Why would we prefer a non-convex problem when convexity has nice properties?

Example: Recall M-estimation

Why non-convex?

- Inherent properties of optimization problem, e.g., phase retrieval
- Robustness or better estimation, e.g., binary classification with non-convex losses

Optimization Formulation: Phase Retrieval

$$\min_x \| |Ax|^2 - b \|_2^2$$

Optimization Formulation: Binary Classification

$$\min_x \left\{ \frac{1}{n} \sum_{i=1}^{n} (b_i - g(a_i, x))^2 \right\}$$

where $g(\cdot, \cdot)$ is non-linear, and hence, the loss function is non-convex.

Notion of convergence: Stationarity

- Let $f(\mathbf{x}) : \mathbb{R}^d \to \mathbb{R}$ be a twice-differentiable and $\mathbf{x}^\star = \min_{x \in \mathbb{R}^d} f(\mathbf{x})$

Definition (Recall - First order stationary point)

A point $\bar{\mathbf{x}}$ is a first order stationary point of a twice differentiable function $f(\mathbf{x})$ if

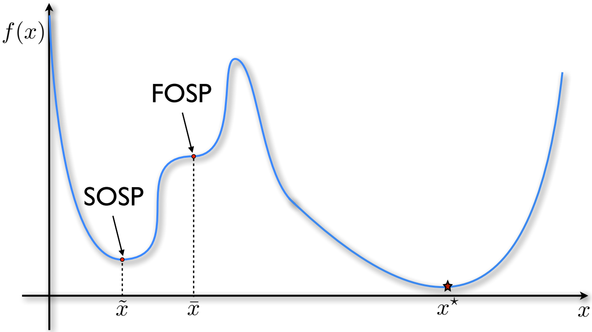
$$\nabla f(\bar{\mathbf{x}}) = \mathbf{0}.$$

Definition (Recall - Second order stationary point)

A point $\tilde{\mathbf{x}}$ is a second order stationary point of a twice differentiable function $f(\mathbf{x})$ if

$$\nabla f(\tilde{\mathbf{x}}) = \mathbf{0} \quad \text{and} \quad \nabla^2 f(\tilde{\mathbf{x}}) \succeq \mathbf{0}.$$

Geometric interpretation of stationarity



- Note that neither $\bar{\mathbf{x}}$, nor $\tilde{\mathbf{x}}$ is **not necessarily** equal to $\mathbf{x}^\star$!!

Assumptions and the gradient method

Assumption: Smoothness

Let $f \in \mathcal{F}_L^{2,1}$ and has L-Lipschitz gradient with respect to ℓ_2 norm, such that,

$$||\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})||_2 \leq L||\mathbf{x} - \mathbf{y}||$$

Gradient descent

Let $\alpha \leq \frac{1}{L}$ be the constant step size and $\mathbf{x}^0 \in \mathrm{dom}(f)$ be the initial point. Then, gradient method produces iterates using the following iterative update,

$$\mathbf{x}^{t+1} = \mathbf{x}^t - \alpha \nabla f(\mathbf{x}^t)$$

Convergence rate and iteration complexity

Theorem

Let $f \in \mathcal{F}_L^{2,1}$, and $\alpha \leq \frac{1}{L}$. Then, gradient method converges to the FOSP with the following properties:

Convergence rate to an ϵ-FOSP:

$$\|\nabla f(\mathbf{x}^t)\| = O\left(\frac{1}{\sqrt{t}}\right)$$

Iteration complexity to reach an ϵ-FOSP:

$$O\left(\frac{1}{\epsilon^2}\right)$$

Example: Phase retrieval for fourier pytchography

Definition (Phase retrieval)

Given a set of measurements of the amplitude of a signal, phase retrieval is the task of finding the phase for the original signal that satisfies certain constraints/properties.

Definition (Fourier ptychography)

Fourier ptychography is the task of reconstructing high-resolution images from low resolution samples, based on optical microscopy. It is a special case of phase retrieval problem.

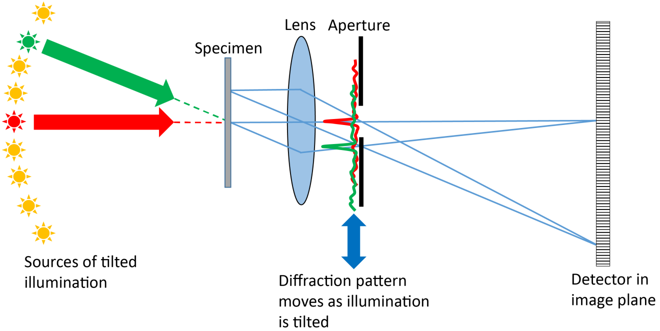
Example: Phase retrieval for fourier pytchography

Definition (Phase retrieval)

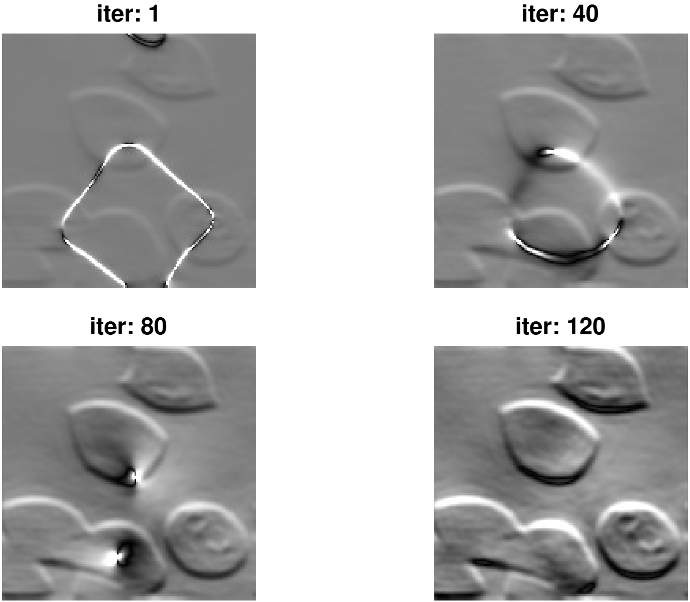
Given a set of measurements of the amplitude of a signal, phase retrieval is the task of finding the phase for the original signal that satisfies certain constraints/properties.

Definition (Fourier ptychography)

Fourier ptychography is the task of reconstructing high-resolution images from low resolution samples, based on optical microscopy. It is a special case of phase retrieval problem.



Example: Malaria infection detection



Example: Image classification using neural networks

Recall: Neural network formulation

- $(\mathbf{a}_i, b_i)$: sample points, $\sigma(\cdot)$: non-linear activation function
- the function class $\mathcal{F}$ is given by $\mathcal{F} := \left\{ g(\cdot, \cdot), \mathbf{x} \in \mathbb{R}^d \right\}$, where

$$\mathbf{x} = (W_1, \boldsymbol{c}_1, W_2, \boldsymbol{c}_2, \ldots, W_k, \boldsymbol{c}_k), \quad W_i \in \mathbb{R}^{d_i \times d_{i-1}}, \quad \boldsymbol{c}_i \in \mathbb{R}^{d_i},$$
$$g(\mathbf{a}, \mathbf{x}) = \sigma\left(W_k \sigma\left(\cdots \sigma\left(W_2 \sigma\left(W_1 \mathbf{a} + \boldsymbol{c}_1\right) + \boldsymbol{c}_2\right) \cdots\right) + \boldsymbol{c}_k\right)$$

- the loss function is given by $L(g(\mathbf{a}, \mathbf{x}), b) := (b - g(\mathbf{a}, \mathbf{x}))^2$.

Example: Image classification

Imagenet: 1000 object classes.
1.2M/100K train/test images
Below human level error rates!

References I

[1] Amir Beck and Marc Teboulle.
Mirror descent and nonlinear projected subgradient methods for convex optimization.
Operations Research Letters, 31(3):167–175, 2003.

[2] Lucien Le Cam.
Asymptotic methods in Statistical Decision Theory.
Springer-Verl., New York, NY, 1986.

[3] Yu. Nesterov.
Introductory Lectures on Convex Optimization: A Basic Course.
Kluwer, Boston, MA, 2004.

[4] R.T. Rockafellar.
Convex analysis.
Princeton University Press (Princeton, NJ), 1970.

[5] A. W. van der Vaart.
Asymptotic Statistics.
Cambridge Univ. Press, Cambridge, UK, 1998.