

Mathematics of Data: From Theory to Computation

Prof. Volkan Cevher
volkan.cevher@epfl.ch

Lecture 4: Unconstrained, smooth minimization I

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

EE-556 (Fall 2017)

lions@epfl



License Information for Mathematics of Data Slides

- ▶ This work is released under a [Creative Commons License](#) with the following terms:
- ▶ **Attribution**
 - ▶ The licensor permits others to copy, distribute, display, and perform the work. In return, licensees must give the original authors credit.
- ▶ **Non-Commercial**
 - ▶ The licensor permits others to copy, distribute, display, and perform the work. In return, licensees may not use the work for commercial purposes – unless they get the licensor's permission.
- ▶ **Share Alike**
 - ▶ The licensor permits others to distribute derivative works only under a license identical to the one that governs the licensor's work.
- ▶ [Full Text of the License](#)

Outline

- ▶ This lecture
 1. Unconstrained 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* convex minimization.

Smooth unconstrained convex minimization

Problem (Mathematical formulation)

The unconstrained convex minimization problem is defined as:

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

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

Example: Maximum likelihood estimation and M-estimators

Problem

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

Goal: estimate \mathbf{x}^\dagger from b_1, \dots, b_n .

Optimization formulation (ML estimator)

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

Theorem (Performance of the ML estimator [4, 8])

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

$$\lim_{n \rightarrow \infty} \sqrt{n} \mathbf{J}^{-1/2} (\hat{\mathbf{x}}_{ML} - \mathbf{x}^\dagger) \stackrel{d}{=} Z \sim \mathcal{N}(\mathbf{0}, \mathbf{I}),$$

where

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

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

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

Example: Maximum likelihood estimation and M-estimators

Problem

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

Goal: estimate \mathbf{x}^\dagger from b_1, \dots, b_n .


Optimization formulation (ML estimator)

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

Optimization formulation (M-estimator)

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

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


 $f(\mathbf{x})$

Approximate vs. exact optimality

Is it possible to solve a convex optimization problem?

*"In general, optimization problems are **unsolvable**" - Y. Nesterov [5]*

- ▶ 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}_ϵ^* is ϵ -optimal in **objective value** if

$$f(\mathbf{x}_\epsilon^*) - f^* \leq \epsilon .$$

Definition

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

$$\|\mathbf{x}_\epsilon^* - \mathbf{x}^*\| \leq \epsilon ,$$

- ▶ The latter approximation guarantee is considered stronger.

A gradient method

Lemma (First-order necessary optimality condition)

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

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

Fixed-point characterization

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

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

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}^* .

When does the gradient method converge?

Lemma

Assume that

1. There exists $\mathbf{x}^* \in \text{dom}(f)$ such that $\nabla f(\mathbf{x}^*) = 0$.
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

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

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

When does the gradient method converge?

Lemma

Assume that

1. There exists $\mathbf{x}^* \in \text{dom}(f)$ such that $\nabla f(\mathbf{x}^*) = 0$.
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

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

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

Proof.

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

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

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

Short (but important) detour: convergence rates

Definition (Convergence of a sequence)

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

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

Convergence rates: the “speed” at which a sequence converges

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

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

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

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

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

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

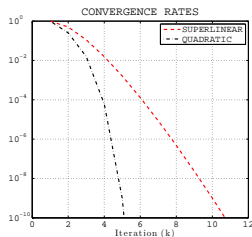
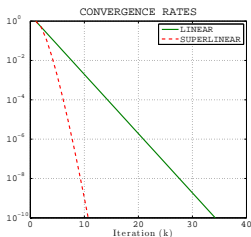
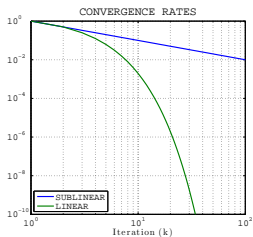
- ▶ **superlinear**: If $r = 0$, we say that the sequence converges *superlinearly*.
- ▶ **quadratic**: if there exists a constant $\mu > 0$ such that

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

Example: Convergence rates

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

- ▶ 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 $f(\mathbf{x}^k) - f^* \geq 0$. Hence, we do not need to use the absolute value when we show convergence results based on the objective value, such as $f(\mathbf{x}^k) - f^* \leq O(1/k^2)$, 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

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

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 $\mathbf{x}^0 \in \text{dom}(f)$, 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 $\mathbf{x}^0 \in \text{dom}(f)$, 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.

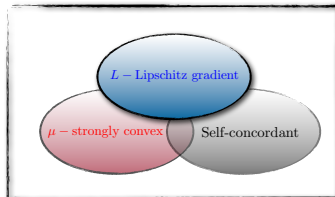
Key question: how to choose α_k to have descent/contraction?

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 $f : \mathcal{Q} \rightarrow \mathbb{R}$ 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 $2 \rightarrow 2$ denoting the spectral norm)

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

- ▶ $\mathcal{F}_L^{l,m}$: 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 $b_i \in \{-1, +1\}$ ($i = 1, \dots, n$), we define the conditional probability of b_i given \mathbf{a}_i as:

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

where $\mathbf{x}^h \in \mathbb{R}^p$ is some true weight vector, $\mu \in \mathbb{R}$ is called the intercept. How to estimate \mathbf{x}^h 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 $\mathbf{A} = [\mathbf{a}_1, \dots, \mathbf{a}_n]^T$ (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 $f : \mathcal{Q} \rightarrow \mathbb{R} \cup \{+\infty\}$, $\mathcal{Q} \subseteq \mathbb{R}^p$ 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

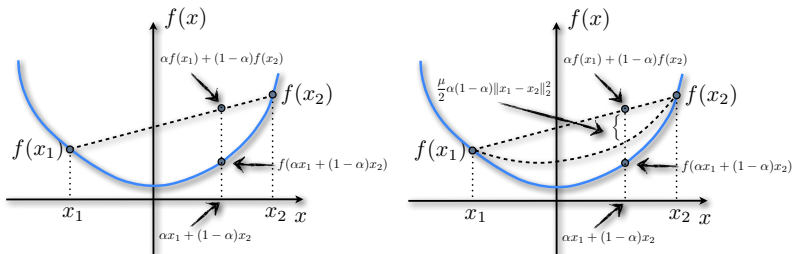


Figure: (Left) Convex (Right) Strongly convex

μ -strongly convex functions (Alternative)

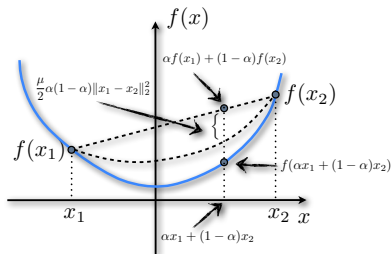
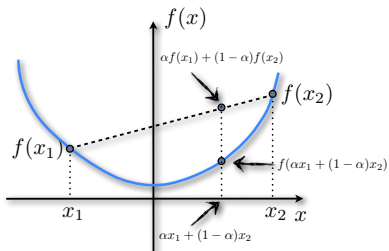
Definition

A convex function $f : \mathcal{Q} \rightarrow \mathbb{R}$ 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 $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_p$ are the eigenvalues of $\mathbf{A}^T \mathbf{A}$.
- ▶ It follows that $L = \lambda_1$ and $\mu = \lambda_p$. If $\lambda_p > 0$, then $f \in \mathcal{F}_{L,\mu}^{2,1}$, otherwise $f \in \mathcal{F}_L^{2,1}$.
- ▶ Since $\text{rank}(\mathbf{A}^T \mathbf{A}) \leq \min\{n, p\}$, if $n < p$, then $\lambda_p = 0$.

Self-concordant functions

Definition (Self-concordant functions in 1-dimension)

A convex function $\varphi : \mathbb{R} \rightarrow \mathbb{R}$ 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 $\varphi : \mathbb{R} \rightarrow \mathbb{R}$ 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 $\varphi : \mathbb{R} \rightarrow \mathbb{R}$ 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 $f : \mathbb{R}^n \rightarrow \mathbb{R}$ 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 $t \in \mathbb{R}$, $\mathbf{x} \in \text{dom}f$ and $\mathbf{v} \in \mathbb{R}^n$ such that $\mathbf{x} + t\mathbf{v} \in \text{dom}f$.
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 $\mathbf{x}^0 \in \text{dom}(f)$, produce the sequence $\mathbf{x}^1, \dots, \mathbf{x}^k, \dots$ 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 $\mathbf{x}^0 \in \text{dom}(f)$, produce the sequence $\mathbf{x}^1, \dots, \mathbf{x}^k, \dots$ 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 $\alpha_k := \frac{2}{L+\mu}$.

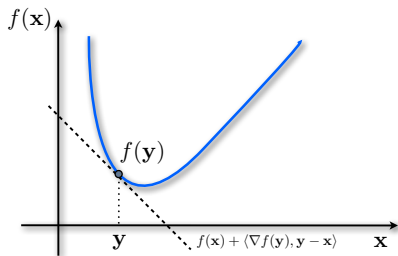
Case 3: If $f \in \mathcal{F}_2(\mathcal{Q})$, then, a bit more complicated (more later).

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: **1st-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:

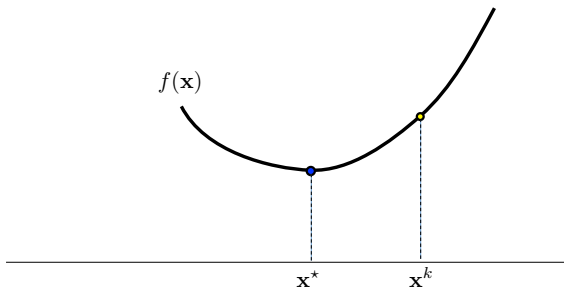
$$f(\mathbf{y}) = f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle + \int_0^1 \langle \nabla f(\mathbf{x} + \tau(\mathbf{y} - \mathbf{x})) - \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle d\tau.$$

Therefore,

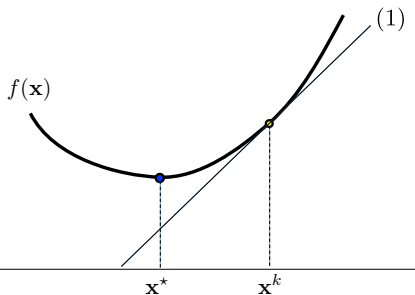
$$\begin{aligned} f(\mathbf{y}) - f(\mathbf{x}) - \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle &\leq \int_0^1 \|\nabla f(\mathbf{x} + \tau(\mathbf{y} - \mathbf{x})) - \nabla f(\mathbf{x})\|^* \cdot \|\mathbf{y} - \mathbf{x}\| d\tau \\ &\leq L \|\mathbf{y} - \mathbf{x}\|_2^2 \int_0^1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

□

Gradient descent methods: geometrical intuition



Gradient descent methods: geometrical intuition



Structure in optimization:

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

Gradient descent methods: geometrical intuition

Majorize:

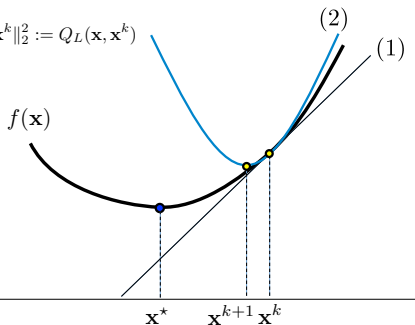
$$f(\mathbf{x}) \leq f(\mathbf{x}^k) + \langle \nabla f(\mathbf{x}^k), \mathbf{x} - \mathbf{x}^k \rangle + \frac{L}{2} \|\mathbf{x} - \mathbf{x}^k\|_2^2 := Q_L(\mathbf{x}, \mathbf{x}^k)$$

Minimize:

$$\mathbf{x}^{k+1} = \arg \min_{\mathbf{x}} Q_L(\mathbf{x}, \mathbf{x}^k)$$

$$= \arg \min_{\mathbf{x}} \left\| \mathbf{x} - \left(\mathbf{x}^k - \frac{1}{L} \nabla f(\mathbf{x}^k) \right) \right\|^2$$

$$= \mathbf{x}^k - \frac{1}{L} \nabla f(\mathbf{x}^k)$$



Structure in optimization:

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

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

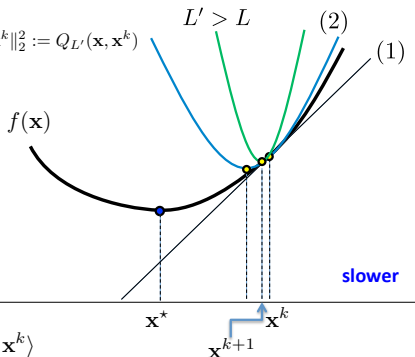
Gradient descent methods: geometrical intuition

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

Gradient descent methods: geometrical intuition

Majorize:

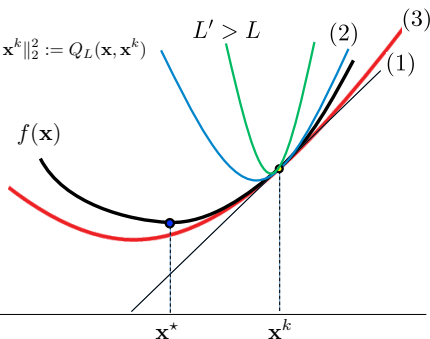
$$f(\mathbf{x}) \leq f(\mathbf{x}^k) + \langle \nabla f(\mathbf{x}^k), \mathbf{x} - \mathbf{x}^k \rangle + \frac{L}{2} \|\mathbf{x} - \mathbf{x}^k\|_2^2 := Q_L(\mathbf{x}, \mathbf{x}^k)$$

Minimize:

$$\mathbf{x}^{k+1} = \arg \min_{\mathbf{x}} Q_L(\mathbf{x}, \mathbf{x}^k)$$

$$= \arg \min_{\mathbf{x}} \left\| \mathbf{x} - \left(\mathbf{x}^k - \frac{1}{L} \nabla f(\mathbf{x}^k) \right) \right\|^2$$

$$= \mathbf{x}^k - \frac{1}{L} \nabla f(\mathbf{x}^k)$$



Structure in optimization:

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

Convergence rate of gradient descent

Theorem

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

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

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

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

$$\begin{aligned} f \in \mathcal{F}_{L,\mu}^{2,1}, \quad \alpha = \frac{1}{L} : & \quad f(\mathbf{x}^k) - f(\mathbf{x}^*) \leq \frac{2L}{k+4} \|\mathbf{x}^0 - \mathbf{x}^*\|_2^2 \\ f \in \mathcal{F}_{L,\mu}^{2,1}, \quad \alpha = \frac{2}{L+\mu} : & \quad \|\mathbf{x}^k - \mathbf{x}^*\|_2 \leq \left(\frac{L-\mu}{L+\mu}\right)^k \|\mathbf{x}^0 - \mathbf{x}^*\|_2 \\ f \in \mathcal{F}_{L,\mu}^{2,1}, \quad \alpha = \frac{1}{L} : & \quad \|\mathbf{x}^k - \mathbf{x}^*\|_2 \leq \left(\frac{L-\mu}{L+\mu}\right)^{\frac{k}{2}} \|\mathbf{x}^0 - \mathbf{x}^*\|_2 \end{aligned}$$

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 $\mathbf{b} = \mathbf{A}\mathbf{x}^\dagger + \mathbf{w}$, where $\mathbf{w} \in \mathbb{R}^n$ is some noise.
- ▶ A classical estimator of \mathbf{x}^\dagger , 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

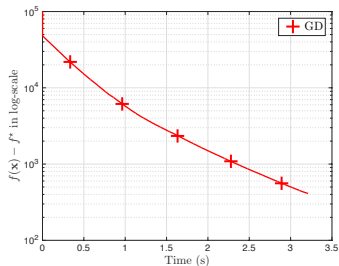
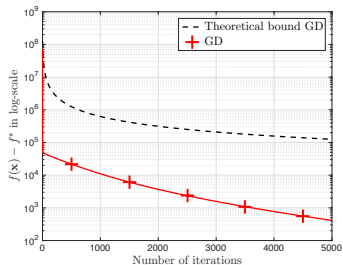
Remarks

- ▶ $f \in \mathcal{F}_{L,\mu}^{2,1}$ with:
 - ▶ $L = \lambda_1(\mathbf{A}^T \mathbf{A}) + \rho$;
 - ▶ $\mu = \lambda_p(\mathbf{A}^T \mathbf{A}) + \rho$;
 - ▶ where $\lambda_1 \geq \dots \geq \lambda_p$ 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:

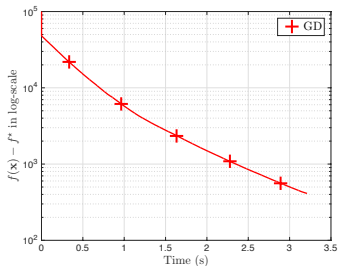
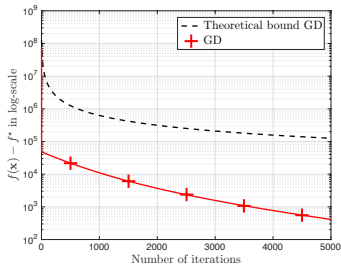
$$n = 500, p = 2000, \rho = 0$$



Example: Ridge regression

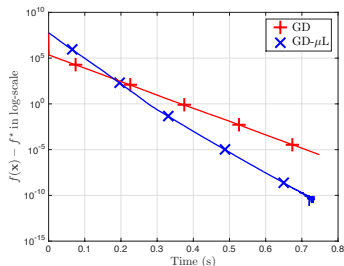
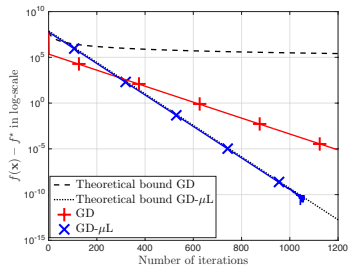
Case 1:

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



Case 2:

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



* Adagrad: An adaptive step-size gradient method

Recall the gradient descent:

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

where $\eta > 0$ is the step-size.

Two potential improvements

1. Instead of fixing an η for all k , we may consider η_k .
2. Instead of applying η to all coordinates of $\nabla f(\mathbf{x}^k)$, we may consider $[\eta_i \nabla f(\mathbf{x}^k)_i]_i$ (coordinate-wise step-size).

* **Adagrad: An adaptive step-size gradient method**

Recall the gradient descent:

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

where $\eta > 0$ is the step-size.

Two potential improvements

1. Instead of fixing an η for all k , we may consider η_k .
2. Instead of applying η to all coordinates of $\nabla f(\mathbf{x}^k)$, we may consider $[\eta_i \nabla f(\mathbf{x}^k)_i]_i$ (coordinate-wise step-size).

Example (Adaptive gradient methods)

Many algorithms build upon this idea, for instance

1. Adagrad [2].
2. Adam [3]
3. RMSprop [7].
4. Adadelta [9].

We present the simplest version of **Adagrad** below.

* Adagrad: An adaptive step-size gradient method

Definition (Adagrad)

Define

$$G_i^k = \sum_{t=1}^k [\nabla f(\mathbf{x}^t)]_i^2.$$

The Adagrad iterate is defined by, for each coordinate i ,

$$\mathbf{x}_i^{k+1} = \mathbf{x}_i^k - \frac{\eta}{\sqrt{G_i^k}} [\nabla f(\mathbf{x}^t)]_i.$$

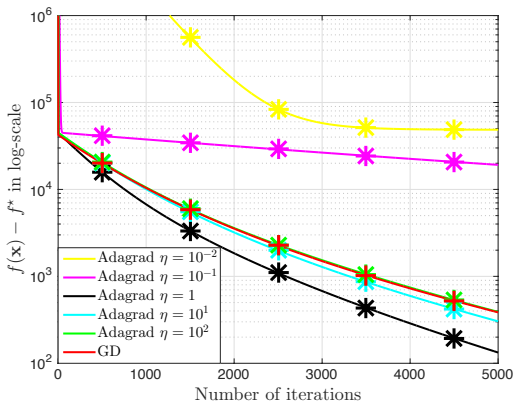
Intuition:

1. G_i^k is increasing in k for all i , and hence the step-sizes for all coordinates are decreasing in k .
2. The step-size for each coordinate is different. Smaller *accumulated* gradient (G_i^k) indicates the requirement for a larger step-size for more progress.
3. Slower convergence rate ($O\left(\frac{1}{\sqrt{k}}\right)$ [2]), but very effective in practice.

Example: Effect of η in Adagrad

Ridge regression ($n = 500, p = 2000, \rho = 0$)

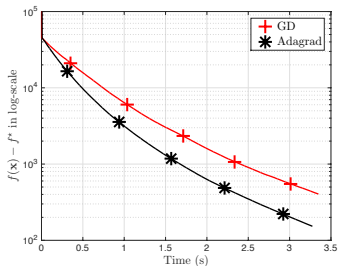
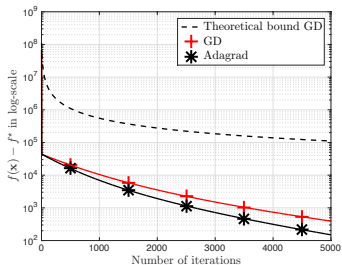
$$\min_{\mathbf{x} \in \mathbb{R}^p} f(\mathbf{x}) := \frac{1}{2} \|\mathbf{b} - \mathbf{A}\mathbf{x}\|_2^2 + \frac{\rho}{2} \|\mathbf{x}\|_2^2.$$



Example: Ridge regression

Case 1:

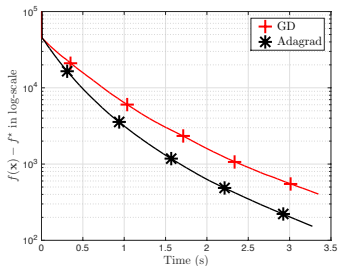
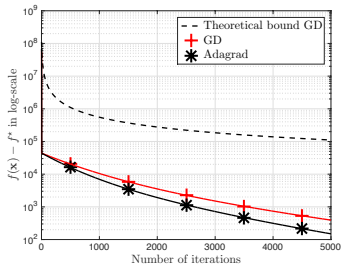
$$n = 500, p = 2000, \rho = 0$$



Example: Ridge regression

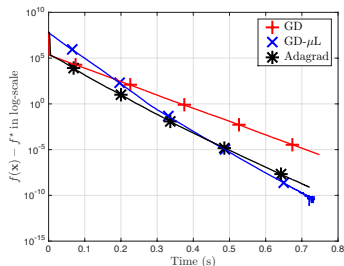
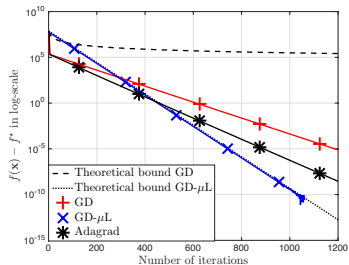
Case 1:

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



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)

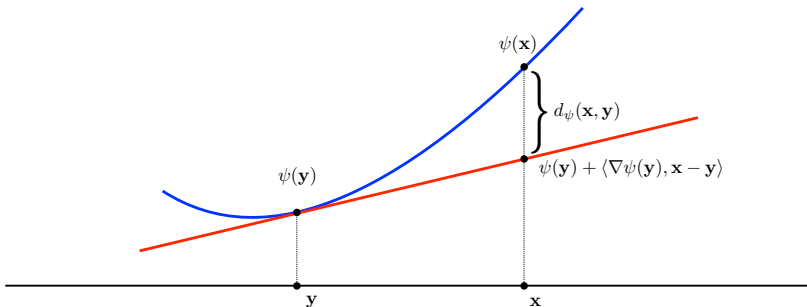
Let $\psi : \mathcal{S} \rightarrow \mathbb{R}$ be a continuously-differentiable and strictly convex function defined on a closed convex set \mathcal{S} . The **Bregman divergence** (d_ψ) associated with ψ for points \mathbf{x} and \mathbf{y} is:

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

- ▶ $\psi(\cdot)$ 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	$\sum_i x_i \log x_i$	$\sum_i \left(\log\left(\frac{x_i}{y_i}\right) - (x_i - y_i) \right)$
von Neumann divergence	$\mathbb{S}_+^{p \times p}$	$\mathbf{X} \log \mathbf{X} - \mathbf{X}$	$\text{tr}(\mathbf{X}(\log \mathbf{X} - \log \mathbf{Y}) - \mathbf{X} + \mathbf{Y})^e$
logdet divergence	$\mathbb{S}_+^{p \times p}$	$-\log \det \mathbf{X}$	$\text{tr}(\mathbf{X}\mathbf{Y}^{-1}) - \log \det(\mathbf{X}\mathbf{Y}^{-1}) - p$

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

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

*Mirror descent [1]

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

Let $\psi : \mathbb{R}^p \rightarrow \mathbb{R}$ 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 ψ^* 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**

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

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

$$\mathbf{x}^{k+1} = \nabla\psi^*(\nabla\psi(\mathbf{x}^k) - \alpha_k \nabla f(\mathbf{x}^k)) \quad (\nabla\psi(\cdot))^{-1} = \nabla\psi^*(\cdot) [6].$$

- ▶ 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 .

*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 $\text{int}\Delta$ with respect to $\|\cdot\|_1$.
- ▶ $\psi_e^*(\mathbf{z}) = \ln \sum_{j=1}^P e^{z_j}$ and $\|\nabla \psi_e(\mathbf{x})\| \rightarrow \infty$ as $\mathbf{x} \rightarrow \tilde{\mathbf{x}} \in \Delta$.
- ▶ 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 $f'(\mathbf{x}) = (f_1(\mathbf{x})', \dots, f_p(\mathbf{x})')^T \in \partial f(\mathbf{x})$, which is the **subdifferential** of f at \mathbf{x} .

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

*Convergence analysis of mirror descent

Problem

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

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 $\mathbf{x}^0 \in \text{int}\mathcal{X}$.
If the step-sizes are chosen as

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

the following convergence rate holds

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

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

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] John Duchi, Elad Hazan, and Yoram Singer.
Adaptive subgradient methods for online learning and stochastic optimization.
Journal of Machine Learning Research, 12(Jul):2121–2159, 2011.
- [3] Diederik Kingma and Jimmy Ba.
Adam: A method for stochastic optimization.
arXiv preprint arXiv:1412.6980, 2014.
- [4] Lucien Le Cam.
Asymptotic methods in Statistical Decision Theory.
Springer-Verl., New York, NY, 1986.
- [5] Yu. Nesterov.
Introductory Lectures on Convex Optimization: A Basic Course.
Kluwer, Boston, MA, 2004.
- [6] R.T. Rockafellar.
Convex analysis.
Princeton University Press (Princeton, NJ), 1970.

References II

- [7] Tijmen Tieleman and Geoffrey Hinton.
Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude.
COURSERA: Neural networks for machine learning, 4(2):26–31, 2012.
- [8] A. W. van der Vaart.
Asymptotic Statistics.
Cambridge Univ. Press, Cambridge, UK, 1998.
- [9] Matthew D Zeiler.
Adadelta: an adaptive learning rate method.
arXiv preprint arXiv:1212.5701, 2012.