Notice

- Homework assignments Week 12:
  - 1 multiple choice question (2%) and 1 programming work (6%).
  - Due date: 10 a.m. on December 15, 2021.

Outline

- **Lecture 4: Introduction to convex analysis**
  - Optimization problems in statistics and machine learning
  - Properties and examples of convex functions
- **Lecture 5: Introduction to convex optimization and the gradient descent algorithm**
  - Convex optimization
  - The gradient descent algorithm
- **Solutions to quizzes and homework assignments in Week 11**

Computation in Data Science: Week 12 Lecture 4

Tso-Jung Yen

Institute of Statistical Science
Academia Sinica

tjyen@stat.sinica.edu.tw

Data Science Degree Program

National Taiwan University

December 8, 2021

Optimization Problems in ST and ML

- **Famous optimization problems in statistics and machine learning:**
  - Most parameter estimation problems in statistics and machine learning can be formulated as optimization problems.
  - Some examples include:
    - **Regularized estimation:** Observations $\{y_i, \mathbf{x}_i\}_{i=1}^n$, the response y_i is a scalar-valued quantity and the covariates $\mathbf{x}_i$ is a p-dimensional vector.
    - **Support vector machines:** Observations $\{y_i, \mathbf{x}_i\}_{i=1}^n$, the label $y_i \in \{-1, +1\}$ and the covariates $\mathbf{x}_i$ is a p-dimensional vector.
    - **Non-negative matrix factorization:** Observations $\mathbf{X}$, a $p \times n$ matrix containing n documents described by p terms.

Optimization Problems in ST and ML

- **Regularized parameter estimation:**
  - Linear regression model: $\mathbf{y} = \mathbf{X}\boldsymbol{\theta} + \boldsymbol{\epsilon}$.
  - **Motivation:** Control behavior of estimated $\boldsymbol{\theta}$, e.g.
    I. Keeping the estimate as simple as possible (**sparsity**).
    II. Preventing the estimate from singularity (**smoothness**).
    III. Reducing difference between the estimate and a reference point (**similarity**).
  - **Optimization problem:**
    - Estimate parameters $\boldsymbol{\theta}$ by solving

$$\min_{\boldsymbol{\theta}} \qquad \frac{1}{2n}||\mathbf{y} - \mathbf{X}\boldsymbol{\theta}||_2^2 + \text{pen}(\boldsymbol{\theta}),$$

      where $\text{pen}(\boldsymbol{\theta})$ is a function that regularizes behavior of the estimated $\boldsymbol{\theta}$.

Optimization Problems in ST and ML

- **Regularized parameter estimation (contd):**
  - Commonly-seen examples in ST and ML include:
    - **Lasso-type estimation (sparsity):** $\text{pen}(\boldsymbol{\theta}) = (\lambda_1/2)||\boldsymbol{\theta}||_1 + (\lambda_2/2)||\boldsymbol{\theta}||_2$, where $\lambda_1, \lambda_2 \geq 0$.
    - **Ridge regression (smoothness):** $\text{pen}(\boldsymbol{\theta}) = (\lambda/2)||\boldsymbol{\theta}||_2^2$, where $\lambda \geq 0$.
    - **Distance-based regularization (similarity):** With $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \cdots)$, $\text{pen}(\boldsymbol{\theta}) = \lambda \sum_{(i,j)} ||\boldsymbol{\theta}_i - \boldsymbol{\theta}_j||_p$, where $\lambda \geq 0$ and $1 \leq p \leq \infty$.

Optimization Problems in ST and ML

- **Support vector machines (SVM):**
  - **Motivation:** Binary classification for an object according to its observation $\mathbf{x}$.
  - **How?** Define a classifier $f(\mathbf{x}) = \boldsymbol{\theta}^T \mathbf{x} + \alpha$ such that for some $d \geq 0$, if $f(\mathbf{x}) \geq d \Rightarrow \mathbf{x}$ belongs to one class; If $f(\mathbf{x}) < -d \Rightarrow \mathbf{x}$ belongs to the other class.
  - For observations $\mathbf{x}'$ and $\mathbf{x}''$, if they belong to *different* classes, then they should satisfy

$$\begin{aligned} \boldsymbol{\theta}^T \mathbf{x}' + \alpha &\geq d \\ \boldsymbol{\theta}^T \mathbf{x}'' + \alpha &< -d. \end{aligned} \tag{1}$$

    for some $d \geq 0$
  - **Why?** $f(\mathbf{x})$ *separates* $\mathbf{x}'$ and $\mathbf{x}''$ since

$$\begin{aligned} 2d &\leq |\boldsymbol{\theta}^T(\mathbf{x}' - \mathbf{x}'')| \\ &\leq ||\mathbf{x}' - \mathbf{x}''||_2 ||\boldsymbol{\theta}||_2 \\ \Rightarrow &\quad ||\mathbf{x}' - \mathbf{x}''||_2 \geq \frac{2d}{||\boldsymbol{\theta}||_2}. \end{aligned}$$

Optimization Problems in ST and ML

- **Support vector machines (contd):**
  - Our aim is to find a function $f(\mathbf{x}) = \boldsymbol{\theta}^T \mathbf{x} + \alpha$ such that
    I. For different observations $\mathbf{x}'$ and $\mathbf{x}''$ belong to different classes, $f(\mathbf{x})$ can separate them like (1).
    II. The margin $d/||\boldsymbol{\theta}||_2$ is at its maximum, i.e. $||\boldsymbol{\theta}||_2$ should be as small as possible.
  - **Optimization problem:**
    - **Supervised learning:** Observations $\{\mathbf{x}_i, y_i\}_{i=1}^n$, where $y_i \in \{-1, +1\}$ is the class label of i.
    - Both $\boldsymbol{\theta}$ and α can be estimated by solving

$$\begin{aligned} \min_{\boldsymbol{\theta}, \alpha} \quad & \frac{1}{2}||\boldsymbol{\theta}||_2^2 \\ \text{subject to} \quad & -y_i(\mathbf{x}_i^T \boldsymbol{\theta} + \alpha) \leq -d \text{ for } i = 1, 2, \cdots, n. \end{aligned}$$

    - *Remark:* The problem is called the **hyperplane separation** problem.
  - **Inference:** $f(\mathbf{x}^{\text{new}}) \geq d \Rightarrow y^{\text{new}} = +1$; $f(\mathbf{x}^{\text{new}}) < -d \Rightarrow y^{\text{new}} = -1$.

Optimization Problems in ST and ML

- **Non-negative matrix factorization:**
  - A $p \times n$ matrix $\mathbf{X}$ in which n documents that are described by p terms.
  - **Non-negativeness:** We assume $(\mathbf{X})_{ij} \geq 0$ for all i and j.

    Note that the n documents can be seen as a **linear combination of** p **dummy vectors** of the terms.
  - **Motivation:** Find a simpler representation for $\mathbf{X}$ such that

    I. Each document can be expressed in terms of r **topics**:

    $$\mathbf{x}_{[j]} \approx \sum_{i=1}^{r} \theta_{ij} \boldsymbol{\gamma}_{[i]} + \mathbf{e}_{[j]},$$

    where θ_{ij} is a weight of the **topic vector** $\boldsymbol{\gamma}_{[i]}$ and $\mathbf{e}_{[j]}$ is an error vector.

    II. The **topic vector** $\boldsymbol{\gamma}_{[i]}$ is non-negative valued.

    III. $\theta_{ij} \geq 0$ and $\sum_{i=1}^{r} \theta_{ij} = 1$.

    IV. The number of topics $r \leq$ the number of terms p (optional).

Optimization Problems in ST and ML

- **Non-negative matrix factorization (contd):**
  - We can express $\mathbf{X}$ as

$$\mathbf{X} = \mathbf{\Gamma\Theta} + \mathbf{E},$$

    where $\mathbf{\Gamma}$ is a $p \times r$ matrix, $\mathbf{\Theta}$ is an $r \times n$ matrix, and $\mathbf{E}$ is a $p \times n$ error matrix.

  - **Optimization problem:**
    - We estimate $\mathbf{\Gamma}$ and $\mathbf{\Theta}$ by solving

$$\begin{aligned} \min_{\mathbf{\Gamma},\mathbf{\Theta}} \quad & \frac{1}{2n}||\mathbf{X} - \mathbf{\Gamma\Theta}||_F^2 \\ \text{subject to} \quad & \gamma_{ij} \geq 0 \text{ for } i = 1, 2, \cdots, p \text{ and } j = 1, 2 \cdots, r, \\ & \theta_{ij} \geq 0 \text{ and } \sum_{i=1}^{r} \theta_{ij} = 1 \text{ for } i = 1, 2, \cdots, r \text{ and } j = 1, 2 \cdots, \end{aligned}$$

    - This problem is called the **non-negative matrix factorization problem**.

Optimization Problems in ST and ML

- We focus on the optimization problem in which the objective function is a **convex function** and the solution set is a **convex set**.
- **Convex optimization problems:**
  - Ridge regression, lasso-type estimation and distance-based regularization.
  - The hyperplane separate problem.
- **Non-convex optimization problems:**
  - The non-negative matrix factorization problem.

Optimization Problems in ST and ML

- **Notation use:**
  - Assume $f(\mathbf{x})$ is differentiable, and the gradient of $f(\mathbf{x})$ is denoted by $\nabla f(\mathbf{x})$.
  - Assume $f(\mathbf{x})$ is twice differentiable, and the Hessian of $f(\mathbf{x})$ is denoted by $\nabla^2 f(\mathbf{x})$.
  - We use the following notation

$$\mathbf{A} \succeq \mathbf{0}$$

    to denote the situation in which the square matrix **A** is **positive semidefinite**. When $\succeq$ becomes $\succ$, the matrix **A** is **positive definite**.

Convex Functions

- **Definition of the convex set:**
  - A set $\mathcal{C}$ is **convex** if for any $\mathbf{x}, \mathbf{y} \in \mathcal{C}$ and any $\alpha \in [0, 1]$ the following relation holds:

$$\alpha\mathbf{x} + (1 - \alpha)\mathbf{y} \in \mathcal{C}$$

- In the following discussion we assume $\mathcal{C}$ is a convex set.
- **Definition of the convex function:**
  - A function $f : \mathcal{C} \mapsto \mathbb{R}$ is **convex** if for any $\mathbf{x}, \mathbf{y} \in \mathcal{C}$ and any $\alpha \in [0, 1]$ the following inequality holds:

$$f(\alpha\mathbf{x} + (1 - \alpha)\mathbf{y}) \leq \alpha f(\mathbf{x}) + (1 - \alpha) f(\mathbf{y}).$$

Convex Functions

- **Properties of convex functions:**

  i). **The first-order characterization:** Assume $f : \mathcal{C} \mapsto \mathbb{R}$ is differentiable. f is **convex** if and only if

$$f(\mathbf{x}) \geq f(\mathbf{y}) + \nabla f(\mathbf{y})^T (\mathbf{x} - \mathbf{y}) \tag{2}$$

  for any $\mathbf{x}, \mathbf{y} \in \mathcal{C}$.

  - **Proof:**

    - **The only if part:** Assume f is convex. For $\alpha \in [0, 1]$ we have

$$\begin{aligned} & f(\mathbf{y} + \alpha(\mathbf{x} - \mathbf{y})) \leq (1 - \alpha) f(\mathbf{y}) + \alpha f(\mathbf{x}) \\ \Rightarrow \quad & \frac{f(\mathbf{y} + \alpha(\mathbf{x} - \mathbf{y})) - f(\mathbf{y})}{\alpha} \leq f(\mathbf{x}) - f(\mathbf{y}). \end{aligned}$$

      By letting $\alpha \to 0$ we obtain

$$\nabla f(\mathbf{y})^T (\mathbf{x} - \mathbf{y}) \leq f(\mathbf{x}) - f(\mathbf{y}) \quad \Rightarrow \quad f(\mathbf{x}) \geq f(\mathbf{y}) + \nabla f(\mathbf{y})^T (\mathbf{x} - \mathbf{y}).$$

    - **The if part:** Assume f satisfies (2). Define $\mathbf{u} = \alpha\mathbf{x} + (1 - \alpha)\mathbf{y}$ with $\alpha \in [0, 1]$. We have $f(\mathbf{x}) \geq f(\mathbf{u}) + \nabla f(\mathbf{u})^T (\mathbf{x} - \mathbf{u})$ and $f(\mathbf{y}) \geq f(\mathbf{u}) + \nabla f(\mathbf{u})^T (\mathbf{y} - \mathbf{u})$. Then

$$\begin{aligned} \alpha f(\mathbf{x}) + (1 - \alpha) f(\mathbf{y}) \quad & \geq \quad f(\mathbf{u}) + \nabla f(\mathbf{u})^T [\alpha\mathbf{x} + (1 - \alpha)\mathbf{y} - \mathbf{u}] \\ & = \quad f(\mathbf{u}). \end{aligned}$$

Convex Functions

- **Properties of convex functions (contd):**

ii). **The second-order characterization:** Assume $f : \mathcal{C} \mapsto \mathbb{R}$ is twice differentiable. f is **convex** if and only if

$$\nabla^2 f(\mathbf{x}) \succeq \mathbf{0} \tag{3}$$

for any $\mathbf{x}, \mathbf{y} \in \mathcal{C}$, where $\nabla^2 f(\mathbf{x})$ is the Hessian of $f(\mathbf{x})$ and $\succeq 0$ means $\nabla^2 f(\mathbf{x})$ is positive semidefinite.

- **Proof:**

  - **The only if part:** Assume f is convex. For $\mathbf{y} = \mathbf{x} + \alpha\mathbf{u}$ with $\alpha \in [0, 1]$, we have

$$\begin{aligned} f(\mathbf{y}) &= f(\mathbf{x}) + \nabla f(\mathbf{x})^T(\mathbf{y} - \mathbf{x}) + \frac{1}{2}(\mathbf{y} - \mathbf{x})^T \nabla^2 f(\mathbf{x})(\mathbf{y} - \mathbf{x}) \\ &\quad + O(||\mathbf{y} - \mathbf{x}||_2^3) \\ &= f(\mathbf{x}) + \alpha \nabla f(\mathbf{x})^T \mathbf{u} + \frac{1}{2}\alpha^2 \mathbf{u}^T \nabla^2 f(\mathbf{x})\mathbf{u} + O(\alpha^3 ||\mathbf{u}||_2^3), \end{aligned}$$

    which implies

$$\frac{1}{2}\alpha^2 \mathbf{u}^T \nabla^2 f(\mathbf{x})\mathbf{u} + O(\alpha^3 ||\mathbf{u}||_2^3) = f(\mathbf{y}) - f(\mathbf{x}) - \alpha \nabla f(\mathbf{x})^T \mathbf{u}.$$

Convex Functions

- **Properties of convex functions (contd):**
  - **Proof (contd):**
    - Further since f is convex we have

$$\begin{aligned} f(\mathbf{y}) \geq f(\mathbf{x}) + \nabla f(\mathbf{x})^T(\mathbf{y}-\mathbf{x}) \quad &\Rightarrow \quad f(\mathbf{y}) - f(\mathbf{x}) - \nabla f(\mathbf{x})^T(\mathbf{y}-\mathbf{x}) \geq 0 \\ &\Rightarrow \quad f(\mathbf{y}) - f(\mathbf{x}) - \alpha \nabla f(\mathbf{x})^T \mathbf{u} \geq 0, \end{aligned}$$

      which implies

$$\frac{1}{2}\alpha^2 \mathbf{u}^T \nabla^2 f(\mathbf{x})\mathbf{u} + O(\alpha^3 ||\mathbf{u}||_2^3) \geq 0 \quad \Rightarrow \quad \mathbf{u}^T \nabla^2 f(\mathbf{x})\mathbf{u} \geq 0$$

      as $\alpha \to 0$.

    - **The if part:** If $\nabla^2 f(\mathbf{x}) \succeq \mathbf{0}$, then there exists a point $\mathbf{z}$ between $\mathbf{x}$ and $\mathbf{y}$ such that

$$\begin{aligned} f(\mathbf{y}) &= f(\mathbf{x}) + \nabla f(\mathbf{x})^T(\mathbf{y}-\mathbf{x}) + \frac{1}{2}(\mathbf{y}-\mathbf{x})^T \nabla^2 f(\mathbf{z})(\mathbf{y}-\mathbf{x}) \\ &\geq f(\mathbf{x}) + \nabla f(\mathbf{x})^T(\mathbf{y}-\mathbf{x}), \end{aligned}$$

      which implies f is convex.

Convex Functions

- **Properties of convex functions (contd):**
  - **Definition of the epigraph of a function:** For f, the corresponding epigraph is defined as

$$\text{epi}(f) = \{(\mathbf{x}, t) : \mathbf{x} \in \mathcal{C}, t \in \mathbb{R}, f(\mathbf{x}) \leq t\},$$

    where $\mathcal{C}$ is a convex set. The epigraph epi(f) is a subset of $\mathbb{R}^{p+1}$ given that $\mathcal{C} \in \mathbb{R}^p$.

iii). **The epigraph characterization:** f is **convex** if and only if epi(f) is a convex set.

- **Proof:**
  - **The only if part:** Assume f is convex. For $(\mathbf{x}, t_1) \in \text{epi}(f)$ and $(\mathbf{y}, t_2) \in \text{epi}(f)$, we have

$$\alpha(\mathbf{x}, t_1) + (1-\alpha)(\mathbf{y}, t_2) = (\alpha\mathbf{x} + (1-\alpha)\mathbf{y}, \alpha t_1 + (1-\alpha)t_2) = (\mathbf{u}, t).$$

    We need to show $(\mathbf{u}, t) \in \text{epi}(f)$. Obviously $\mathbf{u} \in \mathcal{C}$ and $t \in \mathbb{R}$. In addition, since f is convex we have

$$\begin{aligned} f(\mathbf{u}) &= f(\alpha\mathbf{x} + (1-\alpha)\mathbf{y}) \\ &\leq \alpha f(\mathbf{x}) + (1-\alpha) f(\mathbf{y}) \\ &\leq \alpha t_1 + (1-\alpha) t_2 \\ &= t, \end{aligned}$$

    which implies $(\mathbf{u}, t) \in \text{epi}(f)$.

Convex Functions

- **Properties of convex functions (contd):**
  - **Proof (contd):**
    - **The if part:** Assume epi(f) is a convex set. For $(\mathbf{x}, f(\mathbf{x})) \in \text{epi}(f)$, $(\mathbf{y}, f(\mathbf{y})) \in \text{epi}(f)$ and $\alpha \in [0, 1]$, we have

$$\begin{aligned} \alpha(\mathbf{x}, f(\mathbf{x})) + (1-\alpha)(\mathbf{y}, f(\mathbf{y})) &= (\alpha\mathbf{x} + (1-\alpha)\mathbf{y}, \alpha f(\mathbf{x}) + (1-\alpha) f(\mathbf{y})) \\ &\in \text{epi}(f). \end{aligned}$$

      As a result of that

$$f(\alpha\mathbf{x} + (1-\alpha)\mathbf{y}) \leq \alpha f(\mathbf{x}) + (1-\alpha) f(\mathbf{y}).$$

Convex Functions

- **Summary of convex functions:**
  - **Definition:** $f : \mathcal{C} \mapsto \mathbb{R}$ is **convex** if and only if

$$f(\alpha \mathbf{x} + (1-\alpha)\mathbf{y}) \leq \alpha f(\mathbf{x}) + (1-\alpha) f(\mathbf{y})$$

    for any $\mathbf{x}, \mathbf{y} \in \mathcal{C}$ and any $\alpha \in [0, 1]$.

  - **The first-order characterization:** Assume f is differentiable. f is **convex** if and only if

$$f(\mathbf{x}) \geq f(\mathbf{y}) + \nabla f(\mathbf{y})^T (\mathbf{x} - \mathbf{y}).$$

  - **The second-order characterization:** Assume f is twice differentiable. f is **convex** if and only if

$$\nabla^2 f(\mathbf{x}) \succeq \mathbf{0}.$$

  - **The epigraph characterization:** f is **convex** if and only if the eipgraph of f

$$\text{epi}(f) = \{(\mathbf{x}, t) : \mathbf{x} \in \mathcal{C}, t \in \mathbb{R}, f(\mathbf{x}) \leq t\}$$

    is a convex set.

Convex Functions

- **Quiz:**

  1. Let $\mathcal{C}$ be a convex set. Assume $\mathbf{x}, \mathbf{y} \in \mathcal{C}$ and $f : \mathcal{C} \mapsto \mathbb{R}$ is convex and is differentiable. Which of the following statements are *true*?

  a. We have

  $$(\nabla f(\mathbf{x}) - \nabla f(\mathbf{y}))^T (\mathbf{y} - \mathbf{x}) \geq 0.$$

  b. We have

  $$(\nabla f(\mathbf{y}) - \nabla f(\mathbf{x}))^T (\mathbf{y} - \mathbf{x}) \geq 0.$$

  c. We have

  $$(f(\mathbf{y}) - f(\mathbf{x}))^T (\mathbf{y} - \mathbf{x}) \geq 0,$$

Convex Functions

- **Examples of convex functions:**
  - **Example 1:** Assume $x \in \mathbb{R}$. Define $f(x) = x^2$. $f(x)$ is convex since $f''(x) \geq 0$ for all $x \in \mathbb{R}$.

    Moreover, we have

    $$[\mathbb{E}(X)]^2 \leq \mathbb{E}[X^2].$$

  - **Example 2 (Jensen's inequality):** Assume that the probability $\mathbb{P}(X = x_j) = w_j \geq 0$ and $\sum_{j=1}^{k} w_j = 1$. For a scalar-valued convex function $f(x)$, we have

    $$f[\mathbb{E}(X)] = f\left(\sum_{j=1}^{k} w_j x_j\right) \leq \sum_{j=1}^{k} w_j f(x_j) = \mathbb{E}[f(X)].$$

Convex Functions

- **Examples of convex functions (contd):**
  - **Example 3 (Norms):** Since both vector norms and matrix norms satisfy the triangle inequality and homogeneity, they are convex functions.

    For example, with $\alpha \in [0, 1]$, one has

$$
\begin{aligned}
||\alpha \mathbf{A}_1 + (1-\alpha)\mathbf{A}_2||_2 &= \max_{||\mathbf{x}||_2=1} ||[\alpha \mathbf{A}_1 + (1-\alpha)\mathbf{A}_2]\mathbf{x}||_2 \\
&\leq \max_{||\mathbf{x}||_2=1} \{\alpha ||\mathbf{A}_1\mathbf{x}||_2 + (1-\alpha)||\mathbf{A}_2\mathbf{x}||_2\} \\
&\leq \alpha \max_{||\mathbf{x}||_2=1} ||\mathbf{A}_1\mathbf{x}||_2 + (1-\alpha) \max_{||\mathbf{x}'||_2=1} ||\mathbf{A}_2\mathbf{x}'||_2 \\
&= \alpha ||\mathbf{A}_1||_2 + (1-\alpha)||\mathbf{A}_2||_2.
\end{aligned}
$$

Convex Functions

- **Examples of convex functions (contd):**
  - **Example 4 (Composition with an affine mapping):** Let $l : \mathbb{R}^n \mapsto \mathbb{R}$ and $\mathbf{f} \in \mathbb{R}^n$. Assume $l(\mathbf{f})$ is convex.

    Now let $\mathbf{f} = \mathbf{A}\mathbf{x}$, where $\mathbf{A}$ is an $n \times p$ matrix and $\mathbf{x} \in \mathbb{R}^p$ is a p-dimensional vector. Then the function $h(\mathbf{x}) = l(\mathbf{A}\mathbf{x})$ is a convex function of $\mathbf{x}$.

    Verification: Note that

    $$\frac{\partial^2 h(\mathbf{x})}{\partial x_j \partial x_k} = \mathbf{e}_j^T \mathbf{A}^T \nabla^2 l(\mathbf{f}) \mathbf{A} \mathbf{e}_k \Rightarrow \nabla^2 h(\mathbf{x}) = \mathbf{A}^T \nabla^2 l(\mathbf{f}) \mathbf{A}.$$

    For an arbitrary vector $\mathbf{u} \in \mathbb{R}^p$, define $\mathbf{w} = \mathbf{A}\mathbf{u}$. Then we have

    $$\mathbf{u}^T \nabla^2 h(\mathbf{x}) \mathbf{u} = \mathbf{u}^T \mathbf{A}^T \nabla^2 l(\mathbf{f}) \mathbf{A}\mathbf{u} = \mathbf{w}^T \nabla^2 l(\mathbf{f}) \mathbf{w}.$$

    Since $l(\mathbf{f})$ is a convex function of $\mathbf{f}$, we have $\nabla^2 l(\mathbf{f}) \succeq 0$, which implies

    $$\mathbf{w}^T \nabla^2 l(\mathbf{f}) \mathbf{w} \geq 0 \Rightarrow \mathbf{u}^T \nabla^2 h(\mathbf{x}) \mathbf{u} \geq 0,$$

    which further implies $\nabla^2 h(\mathbf{x}) \succeq 0$, and therefore $h(\mathbf{x})$ is a convex function of $\mathbf{x}$.

Convex Functions

- **Examples of convex functions (contd):**
  - **Example 5 (log-sum-exp function):** Now we will show the following function is convex:

$$g(\mathbf{x}) = \log\left[\sum_{l=1}^{m} \exp\left(\sum_{i=1}^{p} a_{li}x_i\right)\right].$$

First note that the (i,j)th entry of the Hessian of $g(\mathbf{x})$ is

$$\begin{aligned}[\nabla^2 g(\mathbf{x})]_{ij} \quad = \quad & \frac{[\sum_{l=1}^{m} a_{li}a_{lj}\exp(\sum_{i=1}^{p} a_{li}x_i)][\sum_{l=1}^{m}\exp(\sum_{i=1}^{p} a_{li}x_i)]}{[\sum_{l=1}^{m}\exp(\sum_{i=1}^{p} a_{li}x_i)]^2} \\ & - \frac{[\sum_{l=1}^{m} a_{li}\exp(\sum_{i=1}^{p} a_{li}x_i)][\sum_{l=1}^{m} a_{lj}\exp(\sum_{i=1}^{p} a_{li}x_i)]}{[\sum_{l=1}^{m}\exp(\sum_{i=1}^{p} a_{li}x_i)]^2}. \end{aligned} \tag{4}$$

- Now define $u_l = \exp(\sum_{i=1}^{p} a_{li}x_i)/[\sum_{l=1}^{m}\exp(\sum_{i=1}^{p} a_{li}x_i)]$. Then we can express (4) as

$$[\nabla^2 g(\mathbf{x})]_{ij} = \sum_{l=1}^{m} a_{li}a_{lj}u_l - \left(\sum_{l=1}^{m} a_{li}u_l\right)\left(\sum_{l=1}^{m} a_{lj}u_l\right).$$

Convex Functions

- **Positive semidefiniteness of the Hessian:**

  Example 5 (contd): Then for $\mathbf{w} \in \mathbb{R}^p$, we have

$$\begin{aligned}
\mathbf{w}^T \nabla^2 g(\mathbf{x}) \mathbf{w} &= \sum_{i,j} w_i w_j [\nabla^2 g(\mathbf{x})]_{ij} \\
&= \sum_{i,j} \Big\{ \sum_{l=1}^m w_i w_j a_{li} a_{lj} u_l - \Big(\sum_{l=1}^m w_i a_{li} u_l \Big) \Big(\sum_{l=1}^m w_j a_{lj} u_l \Big) \Big\} \\
&= \Big[\sum_{i,j} \sum_{l=1}^m w_i w_j a_{li} a_{lj} u_l \Big] \\
&\quad - \Big[\sum_i \Big(\sum_{l=1}^m w_i a_{li} u_l \Big) \Big] \Big[\sum_j \Big(\sum_{l=1}^m w_j a_{lj} u_l \Big) \Big] \\
&= \Big[\sum_{l=1}^m \Big(\sum_i w_i a_{li} \Big)^2 u_l \Big] - \Big[\sum_{l=1}^m \Big(\sum_i w_i a_{li} \Big) u_l \Big]^2 .
\end{aligned}$$

Convex Functions

- **Positive semidefiniteness of the Hessian:**

  Example 5 (contd): Here we have

$$\begin{aligned}
\sum_{l=1}^{m}\Big(\sum_{i} w_i a_{li}\Big)u_l &= \sum_{l=1}^{m}\Big[\Big(\sum_{i} w_i a_{li}\Big)\sqrt{u_l}\Big]\sqrt{u_l} \\
&\leq \Big[\sum_{l=1}^{m}\Big(\sum_{i} w_i a_{li}\Big)^2 u_l\Big]^{1/2}\Big(\sum_{l=1}^{m} u_l\Big)^{1/2} \\
&= \Big[\sum_{l=1}^{m}\Big(\sum_{i} w_i a_{li}\Big)^2 u_l\Big]^{1/2},
\end{aligned}$$

  which implies

$$\mathbf{w}^T \nabla^2 g(\mathbf{x})\mathbf{w} \geq 0,$$

  i.e. $\nabla^2 g(\mathbf{x})$ is positive semidefinite and therefore $g(\mathbf{x})$ is a convex function of $\mathbf{x}$.

Convex Functions

- **Examples of convex functions (contd):**
  - **A line criterion for convexity:** Now assume $f(\mathbf{x})$ is twice differentiable. Define $g(t) = f(\mathbf{x} + t\mathbf{v})$. Now if $g(t)$ is a convex function of $t \in [0, \infty)$, i.e. $g''(t) \geq 0$ for all $t \in [0, \infty)$, then one can show that $f(\mathbf{x})$ is a convex function of $\mathbf{x}$.

    To see why it is, first note that

$$\begin{aligned} g'(t) &= \mathbf{v}^T \nabla f(\mathbf{x} + t\mathbf{v}) \\ g''(t) &= \mathbf{v}^T \nabla^2 f(\mathbf{x} + t\mathbf{v}) \mathbf{v}. \end{aligned}$$

    Since $g(t)$ is a convex function of t, we have

$$0 \leq g''(0) = \mathbf{v}^T \nabla^2 f(\mathbf{x}) \mathbf{v},$$

    for an arbitrary $\mathbf{v} \in \mathbb{R}^p$.

    This implies $\nabla^2 f(\mathbf{x})$ is positive semidefinite, and therefore $f(\mathbf{x})$ is a convex function of $\mathbf{x}$.

    The above result allows us to check whether $f(\mathbf{x})$ is a convex function of $\mathbf{x}$ by checking whether $g(t) = f(\mathbf{x} + t\mathbf{v})$ is a convex function of t.

Convex Functions

- **Examples of convex functions (contd):**
  - **Example 6 (The log-barrier function):** For $\mathbf{X} \in \mathcal{S}^p_{++}$, i.e. $\mathbf{X}$ is *positive definite*, we claim the function

$$f(\mathbf{X}) = -\log\det(\mathbf{X}).$$

is a convex function of $\mathbf{X}$.

Proof: Let $\mathbf{V} \in \mathcal{S}^p_{++}$. Note that for $\mathbf{X} \in \mathcal{S}^p_{++}$, we have $\mathbf{X} + t\mathbf{V} \in \mathcal{S}^p_{++}$ for $t \in [0, \infty)$. In addition,

$$\begin{aligned} f(\mathbf{X} + t\mathbf{V}) &= -\log\det(\mathbf{X} + t\mathbf{V}) \\ &= -\log\det(\mathbf{X}^{1/2}(\mathbf{I} + t\mathbf{X}^{-1/2}\mathbf{V}\mathbf{X}^{-1/2})\mathbf{X}^{1/2}) \\ &= -2\log\det(\mathbf{X}^{1/2}) - \log\det(\mathbf{I} + t\mathbf{X}^{-1/2}\mathbf{V}\mathbf{X}^{-1/2}). \end{aligned}$$

Convex Functions

- **Examples of convex functions (contd):**

  Example 6 (contd): Now by using the eigenvalue decomposition, we can express $\mathbf{X}^{-1/2}\mathbf{V}\mathbf{X}^{-1/2} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^T$, where $\mathbf{U}$ is an orthogonal matrix and $\mathbf{\Lambda} \succ 0$ is a diagonal matrix. The second term can be expressed as

$$\begin{aligned} \log\det(\mathbf{I} + t\mathbf{X}^{-1/2}\mathbf{V}\mathbf{X}^{-1/2}) &= \log\det(\mathbf{U}(\mathbf{I} + t\mathbf{\Lambda})\mathbf{U}^T) \\ &= 2\log\det(\mathbf{U}) + \log\det(\mathbf{I} + t\mathbf{\Lambda}) \\ &= \sum_{j=1}^{p} \log(1 + t\lambda_j). \end{aligned}$$

  Define

$$g(t) = f(\mathbf{X} + t\mathbf{V}) = -2\log\det(\mathbf{X}^{1/2}) - \sum_{j=1}^{p} \log(1 + t\lambda_j).$$

  The function $g(t)$ is a convex function of t since $g''(t) \geq 0$ for all $t \geq 0$. Therefore $f(\mathbf{X})$ is a convex function of $\mathbf{X}$.

Convex Functions

- **Quiz:**

  1. Assume both $g : \mathbb{R} \mapsto \mathbb{R}$ and $h : \mathbb{R} \mapsto \mathbb{R}$ are convex and twice differentiable. Further assume g is nondecreasing on $\mathbb{R}$. Which of the following statements are *true*?

  a. The function composition $h(g(x))$ is a convex function of x.

  b. The function composition $-g(h(x))$ is a convex function of x.

  c. The function composition $g(h(x))$ is a convex function of x.

Computation in Data Science: Week 12 Lecture 5

Tso-Jung Yen

Institute of Statistical Science
Academia Sinica

tjyen@stat.sinica.edu.tw

Data Science Degree Program

National Taiwan University

December 8, 2021

Convex Optimization

- **Constrained optimization problems:**
  - We can write a constrained optimization problem as:

$$\begin{aligned} &\text{minimize}_{\mathbf{x}} && l(\mathbf{x}) \\ &\text{subject to} && f_i(\mathbf{x}) \leq 0 \text{ for } i = 1, 2, \cdots, m \\ & && h_j(\mathbf{x}) = 0 \text{ for } j = 1, 2, \cdots, n, \end{aligned} \tag{5}$$

    where $l : \mathbb{R}^p \mapsto \mathbb{R}$ is the objective function corresponding to the constrained optimization problem (5).
  - The constrained optimization problem contains m inequality constraints and n equality constraints.

Convex Optimization

- **Constrained optimization problems (contd):**
  - Now define

$$\mathcal{C} = \Big\{ \mathbf{x} : \mathbf{x} \in \mathbb{R}^p, f_i(\mathbf{x}) \leq 0 \text{ for } i = 1, 2, \cdots, m, \text{ and } h_j(\mathbf{x}) = 0 \text{ for } j = 1, 2, \cdots, n \Big\}. \quad (6)$$

  - We further define the indicator function ι as

$$\iota\{\mathcal{A}\} = \begin{cases} 0 \text{ if } \mathcal{A} \text{ is true} \\ \infty \text{ otherwise} \end{cases} . \quad (7)$$

  - With (6) and (7) we may write (5) as

$$\text{minimize} \qquad l(\mathbf{x}) + \iota\{\mathbf{x} \in \mathcal{C}\}, \quad (8)$$

    where $\mathcal{C}$ is defined in (6).

Convex Optimization

- **Convex optimization problems:**
  - **Definition:** When $\mathcal{C}$ is a **convex set** and $l : \mathcal{C} \mapsto \mathbb{R}$ is a **convex function**, the constrained optimization problem (5) is called the **convex optimization problem**.
  - Note that if $\mathcal{C}$ is a convex set, then for $\mathbf{x}, \mathbf{y} \in \mathcal{C}$ and $\alpha \in [0, 1]$, we have

$$\iota\{\alpha\mathbf{x} + (1-\alpha)\mathbf{y} \in \mathcal{C}\} \leq \alpha\iota\{\mathbf{x} \in \mathcal{C}\} + (1-\alpha)\iota\{\mathbf{y} \in \mathcal{C}\},$$

    which implies $\iota\{\mathbf{x} \in \mathcal{C}\}$ is a convex function.
  - Therefore when l is convex, (8) is a convex optimization problem.
  - Note that if $\mathbf{x}^{**}$ minimizes $l(\mathbf{x})$ but $\mathbf{x}^{**} \notin \mathcal{C}$, then we have $l(\mathbf{x}^{**}) + \iota\{\mathbf{x}^{**} \in \mathcal{C}\} = \infty$. Obviously $\mathbf{x}^{**}$ is not a solution to (5) and (8).
  - If x^* is a solution to (5), then $\mathbf{x}^* \in \mathcal{C}$ and $\iota\{\mathbf{x}^* \in \mathcal{C}\} = 0$, and

$$l(\mathbf{x}^*) + \iota\{\mathbf{x}^* \in \mathcal{C}\} = l(\mathbf{x}^*),$$

    which implies $\mathbf{x}^*$ is also a solution to (8).

Convex Optimization

- **Convex optimization problems (contd):**
  - **Sufficient conditions for a solution:** Assume $\mathcal{C}$ is a convex set and l is convex and differentiable. If $\mathbf{x}^*$ satisfies the following conditions

$$\begin{aligned} \nabla l(\mathbf{x}^*) &= 0, \\ \iota\{\mathbf{x}^* \in \mathcal{C}\} &= 0, \end{aligned}$$

then $\mathbf{x}^*$ is a solution to the constrained optimization problem (5). To verify the claim, note that

$$\iota\{\mathbf{x}^* \in \mathcal{C}\} = 0 \Leftrightarrow \mathbf{x}^* \in \mathcal{C}.$$

Since l is convex we have

$$l(\mathbf{y}) \geq l(\mathbf{x}^*) + \nabla l(\mathbf{x}^*)^T(\mathbf{y} - \mathbf{x}) = l(\mathbf{x}^*)$$

for any $\mathbf{y} \in \mathcal{C}$.

Gradient Descent Algorithms

- **Basic idea:**
  - When $\mathcal{C} = \mathbb{R}^p$, the optimization problem (8) becomes:

$$\min_{\boldsymbol{\theta} \in \mathbb{R}^p} l(\boldsymbol{\theta}). \tag{9}$$

    This is an unconstrained optimization problem.

  - **Differentiability assumption:** Further if l is differentiable and has the gradient vector evaluated at $\boldsymbol{\theta}'$, we may write $l(\boldsymbol{\theta})$ via Taylor's expansion around $\boldsymbol{\theta}'$ as

$$\begin{aligned} l(\boldsymbol{\theta}) &= l(\boldsymbol{\theta}') + \nabla l(\boldsymbol{\theta}')^T(\boldsymbol{\theta} - \boldsymbol{\theta}') + O(||\boldsymbol{\theta} - \boldsymbol{\theta}'||_2^2) \\ &\approx l(\boldsymbol{\theta}') + \nabla l(\boldsymbol{\theta}')^T(\boldsymbol{\theta} - \boldsymbol{\theta}') + \frac{1}{2c}||\boldsymbol{\theta} - \boldsymbol{\theta}'||_2^2. \end{aligned} \tag{10}$$

    where $c > 0$ is a constant.

  - With (10), we may "approximate" (9) as

$$\min_{\boldsymbol{\theta} \in \mathbb{R}^p} \left\{ l(\boldsymbol{\theta}') + \nabla l(\boldsymbol{\theta}')^T(\boldsymbol{\theta} - \boldsymbol{\theta}') + \frac{1}{2c}||\boldsymbol{\theta} - \boldsymbol{\theta}'||_2^2 \right\} \tag{11}$$

    Differentiating (11) with respect to $\boldsymbol{\theta}$ and equaling the derivative to zero yields

$$\nabla l(\boldsymbol{\theta}') + \frac{\boldsymbol{\theta} - \boldsymbol{\theta}'}{c} = 0. \tag{12}$$

    Solving (12) yields $\boldsymbol{\theta} = \boldsymbol{\theta}' - c\nabla l(\boldsymbol{\theta}')$.

Gradient Descent Algorithms

- **Basic idea (contd):**
  - **Iterative scheme:** This inspires us to build an iterative scheme to solve (9) via

$$\begin{aligned} \boldsymbol{\theta}^{r+1} &= \arg\min_{\boldsymbol{\theta}\in\mathbb{R}^p} \left\{ l(\boldsymbol{\theta}^r) + \nabla l(\boldsymbol{\theta}^r)^T(\boldsymbol{\theta}-\boldsymbol{\theta}^r) + \frac{1}{2c_r}||\boldsymbol{\theta}-\boldsymbol{\theta}^r||_2^2 \right\} \\ &= \boldsymbol{\theta}^r - c_r \nabla l(\boldsymbol{\theta}^r). \end{aligned} \tag{13}$$

  - The iterative scheme (13) is an example of the **gradient descent algorithm**.
  - Here c_r is called the **stepsize** at iteration r.
  - **Stopping criterion:** From (13) we know that

$$||\boldsymbol{\theta}^{r+1} - \boldsymbol{\theta}^r||_2 = c_r ||\nabla l(\boldsymbol{\theta}^r)||_2,$$

    which suggests that we may stop the iterative scheme when $||\nabla l(\boldsymbol{\theta}^r)||_2$ is small.
  - In practice to stop the algorithm, we use the following **stopping criterion**:

$$||\nabla l(\boldsymbol{\theta}^r)||_2 \leq \epsilon. \tag{14}$$

    where ϵ is the tolerance for the error specified by researchers. At the rth step, if (14) is satisfied, we stop the iterative scheme (13).

Gradient Descent Algorithms

- **The descent property:**
  - By letting $\boldsymbol{\theta} = \boldsymbol{\theta}^{r+1}$, $\boldsymbol{\theta}' = \boldsymbol{\theta}^{r}$ in (10) we have

$$\begin{aligned} l(\boldsymbol{\theta}^{r+1}) &\approx l(\boldsymbol{\theta}^{r}) - c_r ||\nabla l(\boldsymbol{\theta}^{r})||_2^2 + \frac{c_r^2}{2c} ||\nabla l(\boldsymbol{\theta}^{r})||_2^2 \\ &= l(\boldsymbol{\theta}^{r}) - \left(c_r - \frac{c_r^2}{2c} \right) ||\nabla l(\boldsymbol{\theta}^{r})||_2^2. \end{aligned}$$

  - From the above result we can see if

$$c_r - \frac{c_r^2}{2c} \geq 0, \tag{15}$$

    then we will have

$$l(\boldsymbol{\theta}^{r+1}) \leq l(\boldsymbol{\theta}^{r}). \tag{16}$$

Gradient Descent Algorithms

- **The descent property (contd):**
  - That means, the sequence $\{\boldsymbol{\theta}^r\}_r$ generated by the iterative scheme (13) leads to a decrease in the loss function l.
  - The inequality (16) is called the **descent property** associated with the sequence $\{\boldsymbol{\theta}^r\}_r$.
  - Therefore to make a sequence generated by (13) to satisfy the descent property, we need to ensure that (16) holds. Now maximizing (15) (with respect to c_r) yields

$$c_r = c,$$

    which is the optimal choice for the stepsize c_r. This ensures that

$$c_r - \frac{c_r^2}{2c} = \frac{c}{2} \geq 0.$$

Gradient Descent Algorithms

- **Choices of the stepsize:**
  - To ensure the descent property (16) holds, we need to choose a proper c_r.
  - In the above case we may let $c_r = c$. However in most situations c is not available.
  - **Line search:** At each r, we choose a candidate value $\tilde{c}$ for the stepsize c_r and check the following condition:

$$l(\boldsymbol{\theta}^r - \tilde{c}\nabla l(\boldsymbol{\theta}^r)) \leq l(\boldsymbol{\theta}^r) - \frac{\tilde{c}}{2}||\nabla l(\boldsymbol{\theta}^r)||_2^2. \tag{17}$$

    If (17) is satisfied, we let $c_r = \tilde{c}$, otherwise we replace $\tilde{c}$ with $\eta\tilde{c}$, where $\eta \in (0, 1)$ (e.g. $\eta = 0.95$ or 0.9), and evaluate (17) again.

    Carry out the above procedure until a proper c_r is found.

Gradient Descent Algorithms

- **Choices of the stepsize (contd):**
  - **The Lipschitz constant for the stepsize:** If the gradient vector of $l(\boldsymbol{\theta})$ satisfies the following inequality:

$$||\nabla l(\boldsymbol{\theta}) - \nabla l(\boldsymbol{\theta}')||_2 \leq M||\boldsymbol{\theta} - \boldsymbol{\theta}'||_2, \tag{18}$$

    then the following inequality holds:

$$l(\boldsymbol{\theta}) \leq l(\boldsymbol{\theta}') + \nabla l(\boldsymbol{\theta}')^T(\boldsymbol{\theta} - \boldsymbol{\theta}') + \frac{M}{2}||\boldsymbol{\theta} - \boldsymbol{\theta}'||_2^2$$

    for any $\boldsymbol{\theta}, \boldsymbol{\theta}' \in \mathbb{R}^p$.
  - The condition (18) is called the **Lipschitz continuous gradient condition**, and M is called the **Lipschitz (gradient) constant**, which plays a crucial role in determining the stepsize c_r.

Gradient Descent Algorithms

- **Choices of the stepsize (contd):**
  - From our previous result we have

$$\begin{aligned} l(\boldsymbol{\theta}^{r+1}) &\leq l(\boldsymbol{\theta}^r) + \nabla l(\boldsymbol{\theta}^r)^T(\boldsymbol{\theta}^{r+1} - \boldsymbol{\theta}^r) + \frac{M}{2}||\boldsymbol{\theta}^{r+1} - \boldsymbol{\theta}^r||_2^2 \\ &= l(\boldsymbol{\theta}^r) - c_r||\nabla l(\boldsymbol{\theta}^r)||_2^2 + \frac{c_r^2 M}{2}||\nabla l(\boldsymbol{\theta}^r)||_2^2 \\ &= l(\boldsymbol{\theta}^r) - \left(c_r - \frac{c_r^2 M}{2}\right)||\nabla l(\boldsymbol{\theta}^r)||_2^2. \end{aligned}$$

  - To make the descent property hold, the optimal choice of c_r is

$$c_r = \frac{1}{M}.$$

Gradient Descent Algorithms

- **Choices of the stepsize (contd):**
  - **Example 1:** In least squares estimation, where $l(\boldsymbol{\beta}) = (2n)^{-1}||\mathbf{y} - \mathbf{X}\boldsymbol{\beta}||_2^2$, we have

    $$||\nabla l(\boldsymbol{\beta}_1) - \nabla l(\boldsymbol{\beta}_2)||_2 = n^{-1}||\mathbf{X}^T\mathbf{X}(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2)||_2 \leq n^{-1}||\mathbf{X}^T\mathbf{X}||_2||\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2||_2,$$

    and therefore we may let $M = n^{-1}||\mathbf{X}^T\mathbf{X}||_2 = \lambda_1(\mathbf{X}^T\mathbf{X}/n)$, the largest eigenvalue of $\mathbf{X}^T\mathbf{X}/n$.
  - **Example 2:** Now if $l(\boldsymbol{\beta})$ is twice differentiable on $\mathcal{C}$, then one can show that

    $$||\nabla l(\boldsymbol{\beta}_1) - \nabla l(\boldsymbol{\beta}_2)||_2 \leq ||\nabla^2 l(\tilde{\boldsymbol{\beta}})||_2||\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2||_2$$

    for some $\tilde{\boldsymbol{\beta}} \in \mathcal{C}$. We may obtain M by letting

    $$M = \max_{\boldsymbol{\beta}} ||\nabla^2 l(\boldsymbol{\beta})||_2.$$

    For example, for the squared l_2 loss in the least squares estimation, we have

    $$||\nabla^2 l(\boldsymbol{\beta})||_2 = n^{-1}||\mathbf{X}^T\mathbf{X}||_2 = \lambda_1(\mathbf{X}^T\mathbf{X}/n).$$

Gradient Descent Algorithms

- **Convergence analysis:**
  - Let $\boldsymbol{\theta}^*$ be a solution to (9).
  - **Convex cases:** If l is convex, in general, we have

$$l(\boldsymbol{\theta}^r) - l(\boldsymbol{\theta}^*) = O(r^{-1}),$$
$$\min_{r=0,1,2,\cdots} ||\nabla l(\boldsymbol{\theta}^r)||_2 = O(r^{-1}).$$

  - If l satisfies the Lipschitz continuous gradient condition (18) and has the Lipschitz constant M, then we have

$$l(\boldsymbol{\theta}^r) - l(\boldsymbol{\theta}^*) \leq \frac{M||\boldsymbol{\theta}^0 - \boldsymbol{\theta}^*||_2^2}{2r}.$$

  - **Strong convex cases:** If l is γ-strongly convex, i.e. $\nabla^2 l(\boldsymbol{\theta}) \succeq \gamma \mathbf{I}_{p\times p}$, then

$$l(\boldsymbol{\theta}^r) - l(\boldsymbol{\theta}^*) \leq \frac{M}{2}\left(1 - \frac{\gamma}{M}\right)^r ||\boldsymbol{\theta}^0 - \boldsymbol{\theta}^*||_2^2 \tag{19}$$

    In addition, $\boldsymbol{\theta}^*$ is unique and

$$||\boldsymbol{\theta}^r - \boldsymbol{\theta}^*||_2^2 \leq \left(1 - \frac{\gamma}{M}\right)^r ||\boldsymbol{\theta}^0 - \boldsymbol{\theta}^*||_2^2.$$

Gradient Descent Algorithms

- **Quiz:**

1. To stop the gradient algorithm, we need to specify some criterion. Which of the following statements are *true*?

a. We use

$$\frac{||\boldsymbol{\theta}^{r+1} - \boldsymbol{\theta}^{r}||_2}{||\nabla l(\boldsymbol{\theta}^{r})||_2} \leq \epsilon$$

as the stoppig criterion.

b. We use

$$\frac{||\boldsymbol{\theta}^{r+1} - \boldsymbol{\theta}^{r}||_2}{|l(\boldsymbol{\theta}^{r+1}) - l(\boldsymbol{\theta}^{r})|} \leq \epsilon$$

as the stoppig criterion.

c. We use

$$||\nabla l(\boldsymbol{\theta}^{r})||_2 \leq \epsilon$$

as the stoppig criterion.

Week 12

- **References:**
  - S. Boyd and L. Vandenberghe (2004). *Convex Optimization*. Cambridge University Press. This classic book provides a lot of examples of convex sets and convex functions. Most examples of convex sets and functions in the slides are from this book. This book can be downloaded from the Internet.
  - G. C. Calafiore and L. E. Ghaoui (2014). *Optimization Models*. Cambridge University Press. The materials on the gradient algorithm in the slides are from this book.
  - B. Schlkopf and A. J. Smola (2001). *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. The MIT Press. The materials on support vector machines in the slides are from Chapter 1 of the book.