Lecture 1

Mathematically Preliminaries

Optimization in Data Science

$$\textbf{Learning} = \textbf{Representation} + \textbf{Evaluation} + \textbf{Optimization}^{1}$$

Linear Regression Given data points $\mathcal{D} = \{\boldsymbol{x}_i, y_i\}_{i=1}^n$ drawn from an unknown distribution $\mathcal{P}(\mathcal{X}, \mathcal{Y})$ over the support $\mathcal{X} \times \mathcal{Y} \subset \mathbb{R}^d \times \mathbb{R}$. Linear regression aims to find a linear function $f : \mathcal{X} \to \mathcal{Y}$ such that for any $(\boldsymbol{x}, y) \sim \mathcal{P}(\mathcal{X}, \mathcal{Y})$ we have $f(\boldsymbol{x}) = y$. Suppose the linear function has the form $f(\boldsymbol{x}) = \boldsymbol{w}^T\boldsymbol{x} + b$, the parameters $\boldsymbol{w} \in \mathbb{R}^d$ and $b \in \mathbb{R}$ can be found by solving the following optimization problem:

$$\min_{\boldsymbol{w},b} \frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{w}^T\boldsymbol{x}_i + b - y_i)^2. \tag{1.1}$$

We denote the objective function as $\mathcal{L}(\boldsymbol{w}, b; \mathcal{D})$ known as the mean-squared error (MSE) loss in machine learning.

There are multiple common forms for linear regression:

$$\min_{\boldsymbol{w},b} \mathcal{L}(\boldsymbol{w}, b; \mathcal{D}) \qquad \text{(Least Squares Regression)}$$

$$\min_{\boldsymbol{w},b} \mathcal{L}(\boldsymbol{w}, b; \mathcal{D}) + \underbrace{\lambda\sum_{k=1}^{d} w_k^2}_{\text{Smooth penalty}} \qquad \text{(Ridge Regression)}$$

$$\min_{\boldsymbol{w},b} \mathcal{L}(\boldsymbol{w}, b; \mathcal{D}) + \underbrace{\lambda\sum_{k=1}^{d} |w_k|}_{\text{Non-smooth penalty}} \qquad \text{(Lasso Regression)}$$

$$\min_{\boldsymbol{w},b} \mathcal{L}(\boldsymbol{w}, b; \mathcal{D}) + \lambda_1\sum_{k=1}^{d} |w_k| + \lambda_2\sum_{k=1}^{d} w_k^2 \qquad \text{(Elastic-Net)}$$

[1] Pedro Domingos. 2012. A few useful things to know about machine learning. Commun. ACM 55, 10 (October 2012), 78–87. https://doi.org/10.1145/2347736.2347755

Question 1.0.1

- Why the intercept b is not included in the penalty?
- How does the linear parameter $\boldsymbol{w}$ behave differently in different forms of regression? Alternatively, why should we add such penalties in regression?

One-Time Graph Cut Given a graph $\mathcal{G}(V, E)$ and its adjacent matrix $\boldsymbol{W} = (w_{ij}) \in \{0, 1\}^{n\times n}$, the one-time graph cut partitions vertices of $\mathcal{G}$ into to disjoint subsets A and $\bar{A}$ such that they have minimum number of edges in between.

Definition 1.0.2: Cut

Given two disjoint subsets of vertices A and B, the cut between A and B is defined as

$$\text{cut}(A, B) = \sum_{u\in A, v\in B} w_{uv}. \qquad \text{(Graph Cut)}$$

By definition of graph cut, the one-time graph cut problem becomes

$$\text{Find a subset } A \subset V \text{ sucht that } \text{cut}(A, \bar{A}) \text{ attains its minimum.}$$

We define the a vector $\boldsymbol{f} = (f_i)_{i=1}^n$ as

$$f_i = \begin{cases} 1 & \text{, if } i \in A \\ -1 & \text{, if } i \in \bar{A} \end{cases} \tag{1.2}$$

Define a diagonal matrix $\boldsymbol{D} \in \mathbb{R}^{n\times n}$ with its element as $d_{ii} = \sum_{j=1}^{n} w_{ij}$. Hence

$$\#\text{ Edges within } V = \sum_{i,j=1}^{n} w_{ij} = \sum_{i=1}^{n} d_{ii} = \sum_{i=1}^{n} f_i^2 d_{ii} = \boldsymbol{f}^T \boldsymbol{D}\boldsymbol{f}.$$

For any vertex i, the number of edges between vertex i and its counterpart subset $\bar{i}$(i.e. $i \notin \bar{i}$) can be expressed

$$\frac{1}{2}(d_{ii} - \sum_{j=1}^{n} f_i f_j w_{ij}).$$

So we can write the cut between A and $\bar{A}$ as

$$\begin{aligned}\text{cut}(A,\bar{A}) &= \sum_{u\in A, v\in\bar{A}} w_{uv} = \frac{1}{2}\left(\sum_{u\in A}\sum_{v\in\bar{A}} w_{uv} + \sum_{u\in\bar{A}}\sum_{v\in A} w_{uv}\right)\\ &= \frac{1}{2}\sum_{u\in V}\sum_{v\in\bar{u}} w_{uv} = \frac{1}{4}\sum_{i=1}^{n}\left(d_{ii} - \sum_{j=1}^{n} f_i f_j w_{ij}\right)\\ &= \frac{1}{4}\left(\boldsymbol{f}^T\boldsymbol{D}\boldsymbol{f} - \sum_{i=1}^{n}\sum_{j=1}^{n} f_i f_j w_{ij}\right) = \frac{1}{4}\left(\boldsymbol{f}^T\boldsymbol{D}\boldsymbol{f} - \boldsymbol{f}^T\boldsymbol{W}\boldsymbol{f}\right)\\ &= \frac{1}{4}\boldsymbol{f}^T(\boldsymbol{D}-\boldsymbol{W})\boldsymbol{f}\end{aligned}$$

Let $\boldsymbol{L} := \boldsymbol{D} - \boldsymbol{W}$, the one-time graph cut problem can be formulated as

$$\begin{aligned}\min_{\boldsymbol{f}}\ & \boldsymbol{f}^T\boldsymbol{L}\boldsymbol{f}\\ \text{s.t.}\ & \boldsymbol{f}\in\{-1,1\}^n\end{aligned} \qquad \text{(One-Time Graph Cut)}$$

The matrix $\boldsymbol{L}$ is known as the Laplacian matrix of the graph $\mathcal{G}$.

Question 1.0.3

- How can we generalize it to multiple cuts? Check out ratio cuts and normalized cuts.
- How can we connect the graph cuts with spectral clustering?

Optimal Transport Consider a market with m sellers and n buyers. Let the supply vector of sellers be denoted by $\boldsymbol{a}\in\mathbb{R}^m$, where each element represents the supply of a particular seller, and the demand vector of buyers be denoted by $\boldsymbol{b}\in\mathbb{R}^n$, where each element corresponds to the demand of a particular buyer. The transportation cost between seller i and buyer j is given by c_{ij}. Suppose the quantity transported from seller i to buyer j is given by p_{ij}, optimal transport seeks to determine a transport plan $\boldsymbol{P} = (p_{ij}) \in \mathbb{R}_+^{m\times n}$ that minimizes the total transportation cost while satisfying the supply-demand constraints.

The total cost is $\sum_{i=1}^{m}\sum_{j=1}^{n} c_{ij}p_{ij}$. The optimal transport problem can be expressed as in matrix form

$$\begin{aligned}\min_{\boldsymbol{P}\in\mathbb{R}_+^{m\times n}}\ & \langle\boldsymbol{P},\boldsymbol{C}\rangle\\ \text{s.t.}\ & \boldsymbol{P}\mathbf{1}_n = \boldsymbol{a},\\ & \boldsymbol{P}^T\mathbf{1}_m = \boldsymbol{b}.\end{aligned} \qquad \text{(Discrete Optimal Transport)}$$

where $\mathbf{1}_n$ and $\mathbf{1}_m$ denote all-one vectors of length n and m, respectively. The inner product between matrices is defined as $\langle\boldsymbol{P},\boldsymbol{C}\rangle = \text{tr}(\boldsymbol{P}^T\boldsymbol{C}) = \sum_{i=1}^{m}\sum_{j=1}^{n} c_{ij}p_{ij}$.

Question 1.0.4

- Find you find some applications of optimal transport in economics?
- Optimal transport is a valuable tool in non i.i.d machine learning applications, such as transfer learning and multi-modal learning, for data alignment. Consider reading some of the papers that are most relevant to your interests.

Inner Products and Norms

Definition 1.0.5: Inner Product

An inner product on $\mathbb{R}^d$ is map $\langle \cdot, \cdot \rangle : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ with the following properties:

1. (**symmetry**) $\langle \boldsymbol{x}, \boldsymbol{y} \rangle = \langle \boldsymbol{y}, \boldsymbol{x} \rangle$ for any $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^d$.
2. (**additivity**) $\langle \boldsymbol{x}, \boldsymbol{y} + \boldsymbol{z} \rangle = \langle \boldsymbol{x}, \boldsymbol{y} \rangle + \langle \boldsymbol{x}, \boldsymbol{z} \rangle$ for any $\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z} \in \mathbb{R}^d$.
3. (**homogeneity**) $\langle \lambda \boldsymbol{x}, \boldsymbol{y} \rangle = \lambda \langle \boldsymbol{x}, \boldsymbol{y} \rangle$ for any $\lambda \in \mathbb{R}$ and $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^d$.
4. (**positive definitiness**) $\langle \boldsymbol{x}, \boldsymbol{x} \rangle \geq 0$ for any $\boldsymbol{x} \in \mathbb{R}^d$ and $\langle \boldsymbol{x}, \boldsymbol{x} \rangle = 0$ iff. $\boldsymbol{x} = \boldsymbol{0}$.

The dot-product is defined as $\langle \boldsymbol{x}, \boldsymbol{y} \rangle = \sum_{i=1}^{d} x_i y_i$ for any $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^d$.

The definition of inner product only provide properties that it should be satisfied, but it does not involve how we calculate an inner product. Hence, the dot product is not the only possible product.

Question 1.0.6

Given a matrix $\boldsymbol{A} \in \mathbb{R}^{d \times d}$, we define a map $\langle \cdot, \cdot \rangle_{\boldsymbol{A}} : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ as

$$\langle \boldsymbol{x}, \boldsymbol{y} \rangle_{\boldsymbol{A}} = \boldsymbol{x}^T \boldsymbol{A} \boldsymbol{y}.$$

Is the map an inner product? Can you define an inner product in a similar form?

Definition 1.0.7: Norm

A norm on $\mathbb{R}^d$ is a function $\| \cdot \| : \mathbb{R}^d \to \mathbb{R}$ satisfying the following

1. (**nonnegativity**) $\|\boldsymbol{x}\| \geq 0$ for any $\boldsymbol{x} \in \mathbb{R}^d$.
2. (**positive homogeneity**) $\|\lambda \boldsymbol{x}\| = |\lambda| \boldsymbol{x}$ for any $\boldsymbol{x} \in \mathbb{R}^d$ and $\lambda \in \mathbb{R}$.
3. (**triangle inequality**) $\|\boldsymbol{x} + \boldsymbol{y}\| \leq \|\boldsymbol{x}\| + \|\boldsymbol{y}\|$ for any $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^d$.

Example.

Euclidean norm: $\|\boldsymbol{x}\|_2 = \sqrt{\sum_{i=1}^{d} x_i^2} = \sqrt{\langle \boldsymbol{x}, \boldsymbol{x} \rangle}$.

ℓ_p-norms: $\|\boldsymbol{x}\|_p = \sqrt[p]{\sum_{i=1}^{d} |x_i|^p} (p \geq 1)$

ℓ_∞-norm: $\|\boldsymbol{x}\|_\infty = \max_{i=1,\ldots,d} |x_i|$.

Question 1.0.8

Show that $\|\boldsymbol{x}\|_\infty = \lim_{p \to \infty} \|\boldsymbol{x}\|_p$ (Hint: Reduce from right)

The Cauchy-Schwartz inequality shows the relationship between a inner product and its induced norm.

Lemma 1.0.9: Cauchy-Schwartz inequality

For any $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^d$,

$$|\langle \boldsymbol{x}, \boldsymbol{y} \rangle| \leq \|\boldsymbol{x}\| \|\boldsymbol{y}\|.$$

Equality holds iff. $\boldsymbol{x}$ and $\boldsymbol{y}$ are linear dependent.

Similar to vector norms, we define that matrix norms as

Definition 1.0.10:

A norm on $\mathbb{R}^{m \times n}$ is a function $\|\cdot\| : \mathbb{R}^{m \times n} \to \mathbb{R}$ satisfying the following

1. (**nonnegativity**) $\|\boldsymbol{A}\| \geq 0$ for any $\boldsymbol{A} \in \mathbb{R}^{m \times n}$.
2. (**positive homogeneity**) $\|\lambda \boldsymbol{A}\| = |\lambda| \boldsymbol{A}$ for any $\boldsymbol{A} \in \mathbb{R}^{m \times n}$ and $\lambda \in \mathbb{R}$.
3. (**triangle inequality**) $\|\boldsymbol{A} + \boldsymbol{B}\| \leq \|\boldsymbol{A}\| + \|\boldsymbol{B}\|$ for any $\boldsymbol{A}, \boldsymbol{B} \in \mathbb{R}^{m \times n}$.

Matrix norms can be generated from vector norms. Specifically, given two norms $\|\cdot\|_a$ and $\|\cdot\|_b$ on $\mathbb{R}^n$ and $\mathbb{R}^m$, respectively, the induced matrix norm on $\mathbb{R}^{m \times n}$ is defined as

$$\|\boldsymbol{A}\|_{a,b} = \max_{\boldsymbol{x} \in \mathbb{R}^n} \{\|\boldsymbol{A}\boldsymbol{x}\|_b : \|\boldsymbol{x}\|_a \leq 1\}. \qquad \text{(Induced Matrix Norm)}$$

If $a = b$, we simplify the notation $\|\cdot\|_{a,a}$ as $\|\cdot\|_a$.

From (Induced Matrix Norm), we can show that

$$\|\boldsymbol{A}\boldsymbol{x}\|_b \leq \|\boldsymbol{A}\|_{a,b} \|\boldsymbol{x}\|_a.$$

Example.

1. (**spectral norm**)
$$\|\boldsymbol{A}\|_2 = \max_{\boldsymbol{x}\in\mathbb{R}^n}\{\|\boldsymbol{A}\boldsymbol{x}\|_2 : \|\boldsymbol{x}\|_2 \leq 1\} = \sqrt{\lambda_{max}(\boldsymbol{A}^T\boldsymbol{A})}$$
2. (**1-norm: Maximum absolute column sum norm**)
$$\|\boldsymbol{A}\|_1 = \max_{\boldsymbol{x}\in\mathbb{R}^n}\{\|\boldsymbol{A}\boldsymbol{x}\|_1 : \|\boldsymbol{x}\|_1 \leq 1\} = \max_{j=1,2,\ldots,n}\sum_{i=1}|A_{ij}|$$
3. (∞**-norm: Maximum absolute row sum norm**)
$$\|\boldsymbol{A}\|_\infty = \max_{\boldsymbol{x}\in\mathbb{R}^n}\{\|\boldsymbol{A}\boldsymbol{x}\|_\infty : \|\boldsymbol{x}\|_\infty \leq 1\} = \max_{i=1,2,\ldots,n}\sum_{j=1}|A_{ij}|$$

There are several norms that are not induced matrix norms in the literature.

Example.

1. (**Frobenius norm**)
$$\|\boldsymbol{A}\|_F = \sqrt{\sum_{i=1}^m\sum_{j=1}^n A_{ij}^2} = \sqrt{\mathrm{tr}(\boldsymbol{A}^T\boldsymbol{A})}$$
2. (L_1 **norm**)
$$\|\boldsymbol{A}\|_1 = \sum_{i=1}^m\sum_{j=1}^n |A_{ij}|$$
3. ($L_{2,1}$ **norm**)
$$\|\boldsymbol{A}\|_{2,1} = \sum_{j=1}^n\sqrt{\sum_{i=1}^m A_{ij}^2}$$
4. (**Nuclear norm**)
$$\|\boldsymbol{A}\|_* = \mathrm{tr}(\boldsymbol{A}^T\boldsymbol{A}) = \sum_{i=1}^{\min\{m,n\}}\sigma_i(\boldsymbol{A}),$$
where $\sigma_i(\boldsymbol{A})$ is the singular value of $\boldsymbol{A}$.

$L_{2,1}$-norm and nuclear norm control the structure of a matrix, hence, they are favorable in high-dimensional analysis and machine learning for structure learning , e.g. Robust PCA, matrix completion.

Question 1.0.11

For a matrix $\boldsymbol{A}$ of rank at most r, show that
$$\|\boldsymbol{A}\|_2 \leq \|\boldsymbol{A}\|_F \leq \|\boldsymbol{A}\|_* \leq \sqrt{r}\|\boldsymbol{A}\|_F \leq r\|\boldsymbol{A}\|_2.$$

Differentiability

Definition 1.0.12: Directional derivative

Let f be a real-valued function defined on a set $\mathcal{S} \subseteq \mathbb{R}^d$. Let $\boldsymbol{x} \in \text{int}(\mathcal{S})$ and let $\boldsymbol{d}$ be any non-zero vector. If the limit

$$\lim_{t \to 0^+} \frac{f(\boldsymbol{x} + t\boldsymbol{d}) - f(\boldsymbol{x})}{t}$$

exists, then it is called the directional derivation of f at $\boldsymbol{x}$ along the direction $\boldsymbol{d}$ and is denoted by $f'(\boldsymbol{x}; \boldsymbol{d})$.

If the limit

$$\lim_{t \to 0} \frac{f(\boldsymbol{x} + t\boldsymbol{e}_i) - f(\boldsymbol{x})}{t}$$

exists, we can it the i-th partial derivative of f at $\boldsymbol{x}$ and write it as $\frac{\partial f}{\partial x_i}(\boldsymbol{x})$.

If all the partial derivatives exist of f at $\boldsymbol{x}$, we call the their column vector gradient and denote as $\nabla f(\boldsymbol{x})$, i.e.,

$$\nabla f(\boldsymbol{x}) = \left(\frac{\partial f}{\partial x_1}(\boldsymbol{x}), \frac{\partial f}{\partial x_2}(\boldsymbol{x}), \cdots, \frac{\partial f}{\partial x_d}(\boldsymbol{x}) \right)^T$$

A function f defined on an open set $\mathcal{U} \subseteq \mathbb{R}^d$ is called continuously differentiable over $\mathcal{U}$ is all its partial derivatives exist and are continuous on $\mathcal{U}$ for any $\boldsymbol{x} \in \mathbb{R}^d$. In such case, we have

$$f'(\boldsymbol{x}; \boldsymbol{d}) = \nabla f(\boldsymbol{x})^T \boldsymbol{d},$$

from that fact that

$$f'(\boldsymbol{x}; \boldsymbol{d}) = \frac{\partial f(x + \tau \boldsymbol{d})}{\partial \tau}\Big|_{\tau=0} = \sum_{i=1}^{d} \frac{\partial f(\boldsymbol{x} + \tau \boldsymbol{d})}{\partial x_i} d_i \Big|_{\tau=0} = \langle \nabla f(\boldsymbol{x}), \boldsymbol{d} \rangle.$$

Theorem 1.0.13: Taylor's Theorem

Suppose that $f : \mathcal{U} \to \mathbb{R}$ is continuously differential over an open set $\mathcal{U} \subseteq \mathbb{R}^d$ and for and $\boldsymbol{d} \in \mathbb{R}^d$ we have

$$f(\boldsymbol{x} + \boldsymbol{d}) = f(\boldsymbol{x}) + \nabla f(\boldsymbol{x} + t\boldsymbol{d})^T \boldsymbol{d},$$

for some $t \in (0, 1)$ and any $\boldsymbol{x} \in \mathcal{U}$

Proof. Let $g(\tau) = f(\boldsymbol{x} + \tau \boldsymbol{d})$. By the mean-value theorem we have

$$g(1) - g(0) = g'(t),$$

for some $t \in (0, 1)$.

Since

$$g'(t) = f'(\boldsymbol{x} + \tau \boldsymbol{d})|_{\tau=t} = \nabla f(\boldsymbol{x} + t\boldsymbol{d})^T \boldsymbol{d}$$

by definition of gradient we complete the proof. □

Proposition 1.0.14

Suppose that $f : \mathcal{U} \to \mathbb{R}$ is continuously differential over an open set $\mathcal{U} \subseteq \mathbb{R}^d$. Then

$$\lim_{\boldsymbol{d}\to\mathbf{0}} \frac{f(\boldsymbol{x}+\boldsymbol{d}) - f(\boldsymbol{x}) - \nabla f(\boldsymbol{x})^T \boldsymbol{d}}{\|\boldsymbol{d}\|} = 0,$$

for all $\boldsymbol{x} \in \mathcal{U}$.

Proof. By assumption, for any unit vector $\boldsymbol{v} \in \mathbb{R}^n$ we have

$$f'(\boldsymbol{x}; \boldsymbol{v}) = \nabla f(\boldsymbol{x})^T \boldsymbol{v},$$

Hence

$$\lim_{t\to 0^+} \frac{f(\boldsymbol{x}+t\boldsymbol{v}) - f(\boldsymbol{x}) - t\nabla f(\boldsymbol{x})^T \boldsymbol{v}}{t} = 0,$$

Let $\boldsymbol{d} = t\boldsymbol{v}$. The above equality is equivalent to

$$\lim_{\boldsymbol{d}\to\mathbf{0}} \frac{f(\boldsymbol{x}+\boldsymbol{d}) - f(\boldsymbol{x}) - \nabla f(\boldsymbol{x})^T \boldsymbol{d}}{\|\boldsymbol{d}\|} = 0,$$

□

The proposition implies that

$$f(\boldsymbol{y}) = f(\boldsymbol{x}) + \nabla f(\boldsymbol{x})^T (\boldsymbol{y} - \boldsymbol{x}) + o(\|\boldsymbol{x} - \boldsymbol{y}\|)$$

for any $\boldsymbol{x}, \boldsymbol{y} \in \mathcal{U}$, where $o(\|\boldsymbol{x} - \boldsymbol{y}\|)$ is a higher order infinitesimal of $\|\boldsymbol{x} - \boldsymbol{y}\|$.

The (i, j)-th partial derivatives of f at $\boldsymbol{x} \in \mathbb{R}^d$ is defined by

$$\frac{\partial^2 f}{\partial x_i \partial x_j}(\boldsymbol{x}) = \frac{\partial \left(\frac{\partial f}{\partial x_j}\right)}{\partial x_i}(\boldsymbol{x}).$$

We call a real-valued function twice continuously differentiable if all its partial derivatives exist and are continuous. In such case, we have

$$\frac{\partial^2 f}{\partial x_i \partial x_j}(\boldsymbol{x}) = \frac{\partial^2 f}{\partial x_j \partial x_i}(\boldsymbol{x}).$$

The Hessian matrix of f at $\boldsymbol{x}$ is denoted by

$$\nabla^2 f(\boldsymbol{x}) = \left(\frac{\partial^2 f}{\partial x_i \partial x_j}(\boldsymbol{x})\right)_{i,j}.$$

Theorem 1.0.15: Taylor's Theorem (Continued)

If f is twice continuously differentiable, we have

$$\nabla f(\boldsymbol{x}+\boldsymbol{d}) = \nabla f(\boldsymbol{x}) + \int_0^1 \nabla^2 f(\boldsymbol{x}+\tau\boldsymbol{d})\boldsymbol{d}d\tau$$

and

$$f(\boldsymbol{x}+\boldsymbol{d}) = f(\boldsymbol{x}) + \nabla f(\boldsymbol{x})^T\boldsymbol{d} + \frac{1}{2}\boldsymbol{d}^T\nabla^2 f(\boldsymbol{x}+t\boldsymbol{d})\boldsymbol{d}$$

for some $t \in (0,1)$.

The gradient can also be defined from Fréchet differentiability.

Definition 1.0.16: Fréchet differentiability

Let $f : \mathcal{S} \to \mathbb{R}$ be defined over $\mathcal{S} \subseteq \mathbb{R}$, and $\boldsymbol{x} \in \text{int}(\mathcal{S})$. Then function is said to be differentiable at $\boldsymbol{x}$ if there exists a vector $\boldsymbol{g}$ such that

$$\lim_{\boldsymbol{d}\to\boldsymbol{0}} \frac{f(\boldsymbol{x}+\boldsymbol{d}) - f(\boldsymbol{x}) - \boldsymbol{g}^T\boldsymbol{d}}{\|\boldsymbol{d}\|} = 0.$$

The unique vector $\boldsymbol{g}$ satisfying the equality is call gradient f at $\boldsymbol{x}$ and it is denoted by $\nabla f(\boldsymbol{x})$.

The definition of the gradient is consistent with the one we defined earlier.

Optimality Conditions for Unconstrained Optimization

Definition 1.0.17: Global and local minimum

Let $f : \mathcal{S} \to \mathbb{R}$ be defined over a set $\mathcal{S} \subseteq \mathbb{R}^d$. Then

1. We call $\boldsymbol{x}^*$ a (strict) global minimum point of f over $\mathcal{S}$ if $f(\boldsymbol{x}^*) \leq (<) f(\boldsymbol{x})$ for any $\boldsymbol{x} \in \mathcal{S}$.
2. We call $\boldsymbol{x}^*$ a (strict) local minimum point of f over $\mathcal{S}$ if there exists $r > 0$ such that $f(\boldsymbol{x}^*) \leq (<) f(\boldsymbol{x})$ for any $\boldsymbol{x} \in \mathcal{B}(\boldsymbol{x}^*, r)$.

The set of global minimum point of f over $\mathcal{S}$ is represented by

$$\arg\min_{\boldsymbol{x}\in\mathcal{S}} f(\boldsymbol{x}).$$

The open ball $\mathcal{B}(\boldsymbol{x}^*, r)$ denotes the set $\{\boldsymbol{x} \in \mathbb{R}^d \big| \|\boldsymbol{x}^* - \boldsymbol{x}\| \leq r\}$.

Theorem 1.0.18: First order optimality condition for local minimum points

Let $f : \mathcal{S} \to \mathbb{R}$ be defined over $\mathcal{S} \subseteq \mathbb{R}^d$. Suppose that $\boldsymbol{x}^* \in \text{int}(\mathcal{S})$ is a local minimum point and that $\nabla f(\boldsymbol{x}^*)$ exists. Then $\nabla f(\boldsymbol{x}^*) = \mathbf{0}$.

Proof. By assumption, $\frac{\partial f(\boldsymbol{x}^*)}{\partial x_i}$ exists for all $i = 1, \ldots, d$. Define $g(t) = f(\boldsymbol{x}^* + t\boldsymbol{e}_i)$ and since $\boldsymbol{x}^* \in \text{int}(\mathcal{S})$, there exists $r > 0$ such that $t \in (-r, r)$. We have $t = 0$ is a local minimum point of $g(t)$ by local optima of $\boldsymbol{x}^*$. As $g'(0)$ exits, hence $g'(0) = \frac{\partial f(\boldsymbol{x}^*)}{\partial x_i}$. Such claim holds for all i, we have $\nabla f(\boldsymbol{x}^*) = \mathbf{0}$. □

Definition 1.0.19: Stationary points

Let $f : \mathcal{S} \to \mathbb{R}$ be defined over $\mathcal{S} \subseteq \mathbb{R}^d$. Suppose that $\boldsymbol{x}^* \in \text{int}(\mathcal{S})$ and that f is differentiable over some neighborhood of $\boldsymbol{x}^*$. Then $\boldsymbol{x}^*$ is called a stationary point of f if $\nabla f(\boldsymbol{x}^*) = \mathbf{0}$.

Theorem 1.0.20: Necessary second order optimality conditions

Let $f : \mathcal{S} \to \mathbb{R}$ be defined over $\mathcal{S} \subseteq \mathbb{R}^d$. Suppose that f is twice continuously differentiable over $\mathcal{S}$ and $\boldsymbol{x}^*$ is a stationary point. Then the following hold

(a) If $\boldsymbol{x}^*$ is a local minimum point of f over $\mathcal{S}$, then $\nabla^2 f(\boldsymbol{x}^*) \succcurlyeq \mathbf{0}$;

(b) If $\boldsymbol{x}^*$ is a local maximum point of f over $\mathcal{S}$, then $\nabla^2 f(\boldsymbol{x}^*) \preccurlyeq \mathbf{0}$;

Theorem 1.0.21: Sufficient second order optimality condition

Let $f : \mathcal{S} \to \mathbb{R}$ be defined over $\mathcal{S} \subseteq \mathbb{R}^d$. Suppose that f is twice continuously differentiable over $\mathcal{S}$ and $\boldsymbol{x}^*$ is a stationary point. Then the following hold:

(a) If $\nabla^2 f(\boldsymbol{x}^*) \succ \mathbf{0}$, then $\boldsymbol{x}^*$ is a strict local minimum point of f over $\mathcal{S}$.

(b) If $\nabla^2 f(\boldsymbol{x}^*) \prec \mathbf{0}$, then $\boldsymbol{x}^*$ is a strict local maximum point of f over $\mathcal{S}$.

Definition 1.0.22: Saddle point

Let $f : \mathcal{S} \to \mathbb{R}$ be defined over $\mathcal{S} \subseteq \mathbb{R}^d$. Suppose that f is continuously differentiable over $\mathcal{S}$. A stationary point $\boldsymbol{x}^*$ is called a saddle point of f over $\mathcal{S}$ if its not a local optimal point of f over $\mathcal{S}$

Theorem 1.0.23: Sufficient condition for a saddle point

Let $f : \mathcal{S} \to \mathbb{R}$ be defined over $\mathcal{S} \subseteq \mathbb{R}^d$. Suppose that f is twice continuously differentiable over $\mathcal{S}$ and $\boldsymbol{x}^*$ is a stationary point. If $\nabla^2 f(\boldsymbol{x}^*)$ is an indefinite matrix, then $\boldsymbol{x}^*$ is a saddle point f over $\mathcal{S}$.

Attainable of Optimal

When we try find a global minimum/maximum point of a function, we need to be sure of its existence. Weierstrass theorem provides such guarantee for a continuous function over a compact set.

Theorem 1.0.24: Weierstrass theorem

Let f be a continuous function defined over a nonempty and compact set $\mathcal{C} \subseteq \mathbb{R}^d$. Then global optimal points over $\mathcal{C}$ are attainable.

The compact property is not necessary if the function f satisfies coerciveness.

Definition 1.0.25: Coerciveness

Let $f : \mathbb{R}^d \to \mathbb{R}$ be a continuous function defined over $\mathbb{R}^d$. Then the function f is called coercive if

$$\lim_{\|\boldsymbol{x}\| \to \infty} f(\boldsymbol{x}) = \infty.$$

A coercive function always achieves its global minimum point over any closed set.

Theorem 1.0.26: Attainment under coerciveness

Let $f : \mathbb{R}^d \to \mathbb{R}$ be a continuous and coercive function and let $\mathcal{S} \subseteq \mathbb{R}^d$ be a nonempty closed set. Then f has a global minimum point over $\mathcal{S}$.

Proof. Given $\boldsymbol{x}_0 \in \mathcal{S}$, there exists $M > 0$ such that

$$f(\boldsymbol{x}) > f(\boldsymbol{x}_0) \text{ for any } \boldsymbol{x} \text{ satisfying } \|\boldsymbol{x}\| > M,$$

by the coerciveness of f. Hence the level set $\mathcal{L}_{\boldsymbol{x}_0} = \{\boldsymbol{x} \in \mathbb{R}^d | f(\boldsymbol{x}) \leq f(\boldsymbol{x}_0)\}$ is bounded and also is closed by the continuity of f, i.e., $\mathcal{L}_{\boldsymbol{x}_0}$ is a non-empty and compact set. And so is $\mathcal{S} \bigcap \mathcal{L}_{\boldsymbol{x}_0} = \{\boldsymbol{x} \in \mathcal{S} | f(\boldsymbol{x}) \leq f(\boldsymbol{x}_0)\}$. From Weierstrass theorem we know that there exists a global minimum point of f over $\mathcal{S} \bigcap \mathcal{L}_{\boldsymbol{x}_0}$. This point is also a global minimum point of f over $\mathcal{S}$.

□