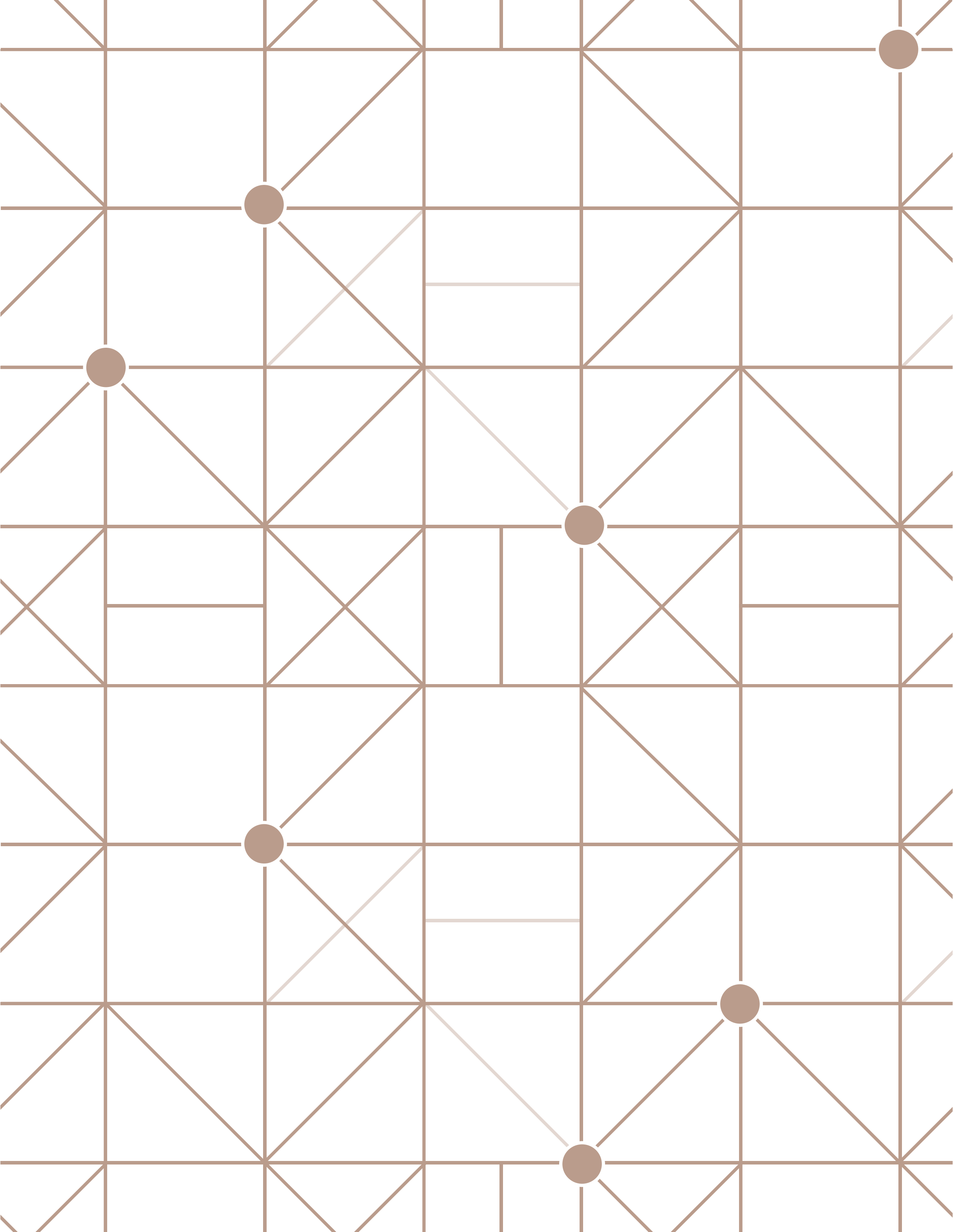

Real Vector Spaces

- nonempty set $(+, \cdot)$

closure property

$x+y \in \mathbb{R}^n$; $\alpha X \in R^n$

- eg. $C^k(\mathbb{R})$ - k-differentiable functions

$L^2(\mathbb{R})$ - square-integrable functions

8 properties that all real vector spaces satisfy

EN.530.641: Statistical Learning for Engineers
Mathematics For Statistical Learning

Jin Seob Kim, Ph.D.
Senior Lecturer, ME dept., LCSR, JHU

1 Real Vector Spaces And Matrices

This section provides a brief review of real vector spaces and matrices.

1.1 Real vector spaces

Formally, a *real vector* is defined as an element of a real vector space. Here, we consider a real vector of dimension n, which is denoted by

$$\mathbf{x} = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix}$$

where the real number $x_i \in \mathbb{R}$ is called the i-th component (or coordinate) of the vector $\mathbf{x}$. We will denote a vector as a bold-faced letter $\mathbf{x}$. Note that a vector is expressed as a column vector. The corresponding vector space for n-tuple real vectors is denoted as $\mathbb{R}^n$, an n-dimensional real vector space (or Euclidean space), i.e., $\mathbf{x} \in \mathbb{R}^n$.

The *transpose* of a column vector $\mathbf{x}$, denoted as $\mathbf{x}^T$, becomes a row vector which is written as

$$\mathbf{x}^T = [x_1, x_2, \cdots, x_n].$$

Using this notation, we can write the vector $\mathbf{x}$ as

$$\mathbf{x} = [x_1, x_2, \cdots, x_n]^T.$$

We will also use the notation with parentheses

$$\mathbf{x} = (x_1, x_2, \cdots, x_n)^T.$$

In general, a vector space is defined with a nonempty set on which addition "+" of elements and scalar multiplication with elements are defined. Here "defined" means that the operations of element addition and scalar multiplications have closure property. Note that vector spaces are a general mathematical space that even includes a space that consists of functions (e.g., $C^k(\mathbb{R})$: the set of k-differentiable functions, or $L^2(\mathbb{R})$: the set of square-integrable functions, $f : \mathbb{R} \to \mathbb{R}$). We will consider this function space later. In this note, let us consider geometric vectors.

Let $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$ (i.e., they are n-dimensional real vectors). Also Let $\alpha \in \mathbb{R}$. Then closure means that

$$\mathbf{x} + \mathbf{y} \in \mathbb{R}^n\, ;\ \alpha\, \mathbf{x} \in \mathbb{R}^n$$

Here given $\mathbf{x} = [x_1, x_2, \cdots, x_n]^T$ and $\mathbf{y} = [y_1, y_2, \cdots, y_n]^T$, the addition of the two vectors (or the sum of $\mathbf{x}$ and $\mathbf{y}$) is expressed as

$$\mathbf{x} + \mathbf{y} = [x_1 + y_1, x_2 + y_2, \cdots, x_n + y_n]^T.$$

Also the two vectors $\mathbf{x}$ and $\mathbf{y}$ are equal if $x_i = y_i$ $(i = 1, 2, \cdots, n)$.

There are important properties that all real vector spaces satisfy: $\forall \mathbf{x}, \mathbf{y}, \mathbf{z} \in \mathbb{R}^n$ and $\alpha, \beta \in \mathbb{R}$:

1. $\mathbf{x} + \mathbf{y} = \mathbf{y} + \mathbf{x}$, commutativity of vector addition
2. $(\mathbf{x} + \mathbf{y}) + \mathbf{z} = \mathbf{x} + (\mathbf{y} + \mathbf{z})$, associativity of vector addition
3. There is a null vector in $\mathbb{R}^n$, $\mathbf{0}$, such that $\mathbf{0} + \mathbf{x} = \mathbf{x}$
4. There exists $-\mathbf{x} \in \mathbb{R}^n$ for each $x \in \mathbb{R}^n$ such that $\mathbf{x} + (-\mathbf{x}) = \mathbf{0}$.
5. $\alpha\,(\mathbf{x} + \mathbf{y}) = \alpha\,\mathbf{x} + \alpha\mathbf{y}$, distributivity over vector addition
6. $(\alpha + \beta)\,\mathbf{x} = \alpha\mathbf{x} + \beta\mathbf{x}$, distributivity over scalar addition
7. $(\alpha\beta)\,\mathbf{x} = \alpha\,(\beta\,\mathbf{x})$, associativity of scalar-vector multiplication
8. $1\,\mathbf{x} = \mathbf{x}$

Let $\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_k$ be arbitrary vectors in a vector space. The set of all linear combinations is called the *span* of $\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_k$:

$$\text{span}\,\{\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_k\} \doteq \left\{ \sum_{i=1}^{k} \alpha_i \mathbf{x}_i \mid \alpha_1, \cdots, \alpha_k \in \mathbb{R} \right\}.$$

Note that if $\mathbf{y}$ can be expressed as a linear combination of $\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_k$, then

$$\text{span}\,\{\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_k, \mathbf{y}\} = \text{span}\,\{\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_k\}\,.$$

A set of vectors $\{\mathbf{a}_1, \cdots \mathbf{a}_k\}$ is linearly independent if the equality containing a linear combination form $\sum_{i=1}^{k} \alpha_i \mathbf{a}_i = \mathbf{0}$ implies $\alpha_i = 0$ $(\forall i = 1, \cdots, k)$. With this set, the *span* of these vectors is defined as

$$\text{span}\,\{\mathbf{a}_1, \mathbf{a}_2, \cdots, \mathbf{a}_k\} = \left\{ \sum_{i=1}^{k} \alpha_i \mathbf{a}_i \,|\, \alpha_1, \cdots, \alpha_k \in \mathbb{R} \right\}$$

and these vectors become the *basis vectors* of the spanned space. For $\mathbb{R}^n$, the *standard basis vectors* (or natural basis) are:

$$\mathbf{e}_1 = \begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix}, \mathbf{e}_2 = \begin{pmatrix} 0 \\ 1 \\ \vdots \\ 0 \end{pmatrix}, \cdots, \mathbf{e}_n = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 1 \end{pmatrix}.$$

Note that there are another possible choices of basis vectors for $\mathbb{R}^n$, but the number of the basis vectors is always the same. In fact, this number of the basis vectors defines the *dimension* of the vector space. For example, $\dim(\mathbb{R}^n) = n$.

2D and 3D Cartesian space are respectively denoted as $\mathbb{R}^2$ and $\mathbb{R}^3$. For example, let $\mathbf{p} \in \mathbb{R}^2$. Then when we write it as

$$\mathbf{p} = \begin{pmatrix} p_1 \\ p_2 \end{pmatrix} = [p_1 \ p_2]^T ,$$

it means that above component form is obtained by using the standard basis vectors $\mathbf{e}_1 = [1 \ 0]^T$ and $\mathbf{e}_2 = [0 \ 1]^T$, i.e., $\mathbf{p} = p_1\mathbf{e}_1 + p_2\mathbf{e}_2$. 3D space $\mathbb{R}^3$ can be expressed similarly. In general, for $\mathbf{p} \in \mathbb{R}^n$, it can be expressed as $\mathbf{p} = \sum_{i=1}^n p_i \mathbf{e}_i$ or $\mathbf{p} = [p_1 \ p_2 \ \cdots \ p_n]^T$.

The dot product of $\mathbf{x} = [x_1 \ x_2 \ \cdots \ x_n]^T \in \mathbb{R}^n$ and $\mathbf{y} = [y_1 \ y_2 \ \cdots \ y_n]^T \in \mathbb{R}^n$ is defined as

$$\mathbf{x} \cdot \mathbf{y} = \sum_{i=1}^n x_i y_i = \mathbf{x}^T \mathbf{y}.$$

Sometimes we will use the notation

$$\langle \mathbf{x}, \mathbf{y} \rangle = \mathbf{x} \cdot \mathbf{y}$$

to denote the dot product (or inner product).

The Euclidean norm of a vector $\mathbf{x} = [x_1 \ x_2 \ \cdots \ x_n]^T \in \mathbb{R}^n$ is defined as

$$\|\mathbf{x}\| = \sqrt{\mathbf{x} \cdot \mathbf{x}} = \sqrt{\sum_{i=1}^n x_i^2}.$$

By using the Euclidean norm, the distance between two point vectors $\mathbf{x}$ and $\mathbf{y}$ is computed as

$$\|\mathbf{x} - \mathbf{y}\| = \sqrt{\sum_{i=1}^n (x_i - y_i)^2}.$$

Generally, the dot product above is an example of the *inner product*. Let $V = (\mathcal{X}, +, \cdot)$ be a vector space over $\mathbb{R}$. An *inner product*, $\langle x, y \rangle$ $(x, y \in V)$, is defined as a function $\langle \cdot, \cdot \rangle : V \times V \to \mathbb{R}$ that satisfies, for every $x, y, z \in V$:

- $\langle x, x \rangle \geq 0$ and $\langle x, x \rangle = 0$ iff $x = 0$
- $\langle x, y + z \rangle = \langle x, y \rangle + \langle x, z \rangle$
- $\langle x, \alpha y \rangle = \alpha \langle x, y \rangle$, $\forall \alpha \in \mathbb{R}$
- $\langle x, y \rangle = \langle y, x \rangle$

Then the norm can be defined as

$$\|x\| \doteq \sqrt{\langle x, x \rangle}.$$

1.2 Matrices

Let $A \in \mathbb{R}^{m\times n}$, i.e., the set of $m \times n$ real matrices. We can write

$$A = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{pmatrix} = (a_{ij}) \;\; (i = 1, \cdots, m, \;\; j = 1, \cdots, n).$$

In our class, $\mathbb{I}_n$ denotes the $n \times n$ identity matrix.

I_n

The transpose of $A = (a_{ij})$ is defined as: $A^T = (a_{ji})$.

The column rank of $A \in \mathbb{R}^{m\times n}$ is the dimension of the column space. Let $A = [\mathbf{a}_1, \cdots, \mathbf{a}_n]$ where $\mathbf{a}_i \in \mathbb{R}^m$ $(i = 1, 2, \cdots, n)$. Then the column space of A is

$$\mathcal{R}(A) = \text{span}\{\mathbf{a}_1, \cdots ., \mathbf{a}_n\}.$$

It is also called as the *range* of A. Likewise row rank is the dimension of the row space of A, or equivalently, the column space of A^T. Fundamental fact is that row rank and column rank of A are the same, which uniquely defines the *rank* of A. For $A \in \mathbb{R}^{m\times n}$, it follows that

$$\text{rank}(A) \doteq \dim\left(\text{span}\{\mathbf{a}_1, \cdots, \mathbf{a}_n\}\right) \leq \min(m, n).$$

If $\text{rank}(A) = \min(m, n)$ then A is called full-rank, and if $\text{rank}(A) < \min(m, n)$ then A is called rank-deficient.

For square matrices, that $A \in \mathbb{R}^{n\times n}$ is full-rank is equivalent to saying that A is non-singular (i.e., A is invertible, or $\det(A) \neq 0$).

The null space of a matrix $A \in \mathbb{R}^{m\times n}$ is defined as

$$\mathcal{N}(A) \doteq \{\mathbf{x} \in \mathbb{R}^n | A\mathbf{x} = \mathbf{0}\}.$$

A useful example of a matrix norm, given $A \in \mathbb{R}^{m\times n}$, is called the *Frobenius norm*, which is defined as

$$\|A\| = \left(\sum_{i=1}^{m}\sum_{j=1}^{n} a_{ij}^2\right)^{1/2} = \sqrt{\text{tr}(AA^T)}$$

where $\text{tr}(A)$ denotes the trace of a matrix A which is defined as the sum of all diagonal elements of A.

Given $A \in \mathbb{R}^{n\times n}$, the eigenvalue decomposition gives

$$A = U\Lambda U^{-1}$$

where $U = [\mathbf{u}_1, \cdots, \mathbf{u}_n]$ is a matrix of eigenvectors $\mathbf{u}_i \in \mathbb{R}^n$. Λ is a diagonal matrix of which the diagonal elements are the eigenvalues of A. Note that if $A^T = A$, then $U^{-1} = U^T$, i.e., U is an $n \times n$ orthogonal matrix $(U \in O(n))$.

Given $A \in \mathbb{R}^{m\times n}$, the singular value decomposition gives a unique form as

$$A = U\Sigma V^T$$

where $U \in O(m)$, $V \in O(n)$, and $\Sigma \in \mathbb{R}^{m \times n}$ with $\Sigma_{ii} = \sigma_i \geq 0$ (singular values) and $\Sigma_{ij} = 0$ if $i \neq j$.

Given $A \in \mathbb{R}^{n \times n}$, a quadratic form is defined as $\mathbf{x}^T A \mathbf{x} \in \mathbb{R}$ where $\mathbf{x} \in \mathbb{R}^n$. If the quadratic form is always greater than zero $\forall \mathbf{x} \in \mathbb{R}^n$, and equal to zero iff $\mathbf{x} = \mathbf{0}$, then A is called as positive definite. If the quadratic form is always greater than or equal to zero $\forall \mathbf{x} \in \mathbb{R}^n$, then A is positive semi-definite.

1.3 Gradient of a function

Let $f : \mathbb{R}^n \to \mathbb{R}$ be a continuously differentiable function. Note that, in our class, a "nice" function means continuously differentiable and integrable. Then we can write it as $y = f(\mathbf{x})$ where $y \in \mathbb{R}$ and $\mathbf{x} = [x_1, x_2, \cdots, x_n]^T \in \mathbb{R}^n$ with the standard basis elements $\mathbf{e}_i$ of $\mathbb{R}^n$. Then the gradient of f at $\mathbf{x}$ is defined as

$$\nabla f \doteq \sum_{i=1}^{n} \frac{\partial f}{\partial x_i} \mathbf{e}_i = \left[\frac{\partial f}{\partial x_1}, \frac{\partial f}{\partial x_2}, \cdots, \frac{\partial f}{\partial x_n} \right]^T$$

↳The collection of all its partial derivatives into a vector.

where $\frac{\partial f}{\partial x_i}$ denotes the partial derivative of f along the $\mathbf{e}_i$ direction. Note that in some literature, the following notation is also used

$$Df = (\nabla f)^T = \left[\frac{\partial f}{\partial x_1}, \frac{\partial f}{\partial x_2}, \cdots, \frac{\partial f}{\partial x_n} \right].$$

In addition, the Hessian matrix of f at $\mathbf{x}$ is defined as

$$D^2 f = \begin{pmatrix} \frac{\partial^2 f}{\partial x_1^2} & \frac{\partial^2 f}{\partial x_2 \partial x_1} & \cdots & \frac{\partial^2 f}{\partial x_n \partial x_1} \\ \frac{\partial^2 f}{\partial x_1 \partial x_2} & \frac{\partial^2 f}{\partial x_2^2} & \cdots & \frac{\partial^2 f}{\partial x_n \partial x_2} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 f}{\partial x_1 \partial x_n} & \frac{\partial^2 f}{\partial x_2 \partial x_n} & \cdots & \frac{\partial^2 f}{\partial x_n^2} \end{pmatrix}.$$

In short, we can write it as

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

$(i, j = 1, \cdots, n)$.

Using the above notations, the Taylor series expansion of $f : \mathbb{R}^n \to \mathbb{R}$ about $\mathbf{x}$ is expressed as

$$f(\mathbf{x} + \mathbf{v}) \approx f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{v} \rangle + \frac{1}{2} \langle \mathbf{v}, D^2 f(\mathbf{x}) \mathbf{v} \rangle$$

up to the 2nd order.

Note that in the textbook ESL [1, Ch.4], the gradient vector and the Hessian matrix are expressed as

$$\nabla f = \frac{\partial f}{\partial \mathbf{x}}; \; D^2 f = \frac{\partial^2 f}{\partial \mathbf{x} \partial \mathbf{x}^T}.$$

2 Optimization

This section provides a brief review on gradient-based optimization methods. Let $\mathbf{f} : \mathbb{R}^n \to \mathbb{R}$ be a continuously differentiable function. Basically we want to solve the optimization problem as

$$\min_{\mathbf{x}\in S} f(\mathbf{x})$$

which minimizes $f(\mathbf{x})$. Note that this minimization problem equivalent to

$$\max_{\mathbf{x}\in S} -f(\mathbf{x}).$$

Here if $S = \mathbb{R}^n$, then we call it as an unconstrained optimization problem, which is the main topic in this section. Constrained optimization will be discussed later in the class.

Let us consider a minimization problem. We have to find the points in $S = \mathbb{R}^n$ that correspond to the minima. There are two kinds of minima.

Definition 2.1 (Local minimizer) *Given a function $f : S \subset \mathbb{R}^n \to \mathbb{R}$, a point $\mathbf{x}^* \in S$ is called a local minimizer of f over S if $\exists \epsilon > 0$ such that $f(\mathbf{x}) \geq f(\mathbf{x}^*)$, $\forall \mathbf{x} \in S \setminus \{\mathbf{x}^*\}$ and $\|\mathbf{x} - \mathbf{x}^*\| < \epsilon$.*

Definition 2.2 (Global minimizer) *Given a function $f : S \subset \mathbb{R}^n \to \mathbb{R}$, a point $\mathbf{x}^* \in S$ is called a global minimizer of f over S if $f(\mathbf{x}) \geq f(\mathbf{x}^*)$, $\forall \mathbf{x} \in S \setminus \{\mathbf{x}^*\}$.*

Here $S \setminus \{\mathbf{x}^*\}$ denotes the set of all elements in S that are not in the set $\{\mathbf{x}^*\}$.

In practice, what we can find is local minimizers in most of optimization problems. Finding (local) minimizers is the same as solving minimization problems. There are many different methods to solve the optimization problems (see [2] for reference). Here we discuss one type of method which is widely and commonly used in statistical learning.

2.1 Gradient-based optimization

First, let us talk about conditions for a point $\mathbf{x}^*$ to be a local minimizer. Again let us consider a nice function $f : S \subset \mathbb{R}^n \to \mathbb{R}$.

Definition 2.3 (Feasible direction) *A vector $\mathbf{v} \in \mathbb{R}^n$ $(\mathbf{v} \neq \mathbf{0})$ is called a feasible direction at $\mathbf{x} \in S$ if $\exists \alpha_0 > 0$ such that $\mathbf{x} + \alpha\mathbf{v} \in S$, $\forall \alpha \in [0, \alpha_0]$.*

Now let us consider the Taylor series expansion of f around $\mathbf{x}^*$ in any feasible direction $\mathbf{v}$. Then we have

$$f(\mathbf{x}^* + \alpha\mathbf{v}) = f(\mathbf{x}^*) + \alpha\langle \nabla f(\mathbf{x}^*), \mathbf{v}\rangle + O(\alpha^2).$$

Based on this, the first-order necessary condition is as follows:

Theorem 1 (First-order necessary condition) *If $\mathbf{x}^*$ is a local minimizer of f, then for any feasible direction $\mathbf{v}$ at $\mathbf{x}^*$, the following is satisfied*

$$\langle \mathbf{v}, \nabla f(\mathbf{x}^*)\rangle = \mathbf{v}^T \nabla f(\mathbf{x}^*) \geq 0.$$

Corollary 2 (First-order necessary condition) *Furthermore, if the local minimizer* $\mathbf{x}^*$ *is inside* S*, then it follows that*

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

Since in unconstrained optimization problems, all local minimizers are located inside the domain, hence the corollary is the most important necessary condition. Then we can see one possible method to solve minimization problems: *find a point* $\mathbf{x}^*$ *that satisfies* $\nabla f(\mathbf{x}^*) = \mathbf{0}$. If this can be solved analytically, then we have a set of good candidates relatively easily. In most cases, however, we have to resort to numerical approaches.

Now let us discuss (numerical) gradient-based methods. Again we consider a nice function $f : \mathbb{R}^n \to \mathbb{R}$. Note that the direction of maximum rate increase of f at a point $\mathbf{x}$ is $\nabla f(\mathbf{x})$. The idea of gradient-based methods is that we move along the negative direction of the gradient to reach a local minimizer. This is called a *gradient descent algorithm* (or simply gradient algorithm). Numerically we need an iterative approach.The procedure is as follows.

Let $\mathbf{x}^{(0)}$ be a initial point. Now suppose we are at $\mathbf{x}^{(k)}$ (k denotes the step index). Then the update of the point is done by

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

where α_k denotes the step size (or learning rate) at the k-th step. The gradient becomes zero as we move closer to a local minimizer. In practice, we have to impose some numerical conditions to check the updated point is considered a local minimizer.

An important question is how to determine α_k. We will talk about three approaches. Note that there are many different algorithms in this category.

2.1.1 Gradient descent algorithm

The first is to choose a constant α_k (usually called as a *gradient descent algorithm*). In this case, α_k should not be too small (leading to too much computation time) nor too large (leading to zig-zag paths near the minimizer; not reaching the minimizer). Note that this algorithm (with constant learning rate) has been widely used in machine learning fields. Common examples of learning rate can be chosen as a small number such as 0.001 to relatively large one such as 0.3 [3].

2.1.2 Steepest descent algorithm

The second is called the *steepest gradient algorithm*. Mathematically, we solve

$$\alpha_k = \arg\min_{\alpha \geq 0} f(\mathbf{x}^{(k)} - \alpha \nabla f(\mathbf{x}^{(k)}))$$

to determine α_k for the k-th step. In order to solve this, usually any line search algorithm can be applied (see [2] for details).

seeks the minimum of a defined nonlinear function by selecting a reasonable

2.1.3 Newton's method

Another example of gradient-based optimization algorithm is *Newton's method.* It involves the second order approximation (the Hessian matrix) for the update. Specifically at the k-th step, we

direction vector that, when computed with a reasonable stepsize, will provide a function value closer to the absolute minimum.

compute the gradient vector $\mathbf{g}_k = \nabla f(\mathbf{x}^{(k)})$ and the Hessian matrix $H_k = \nabla^2 f(\mathbf{x}^{(k)}) = D^2 f(\mathbf{x}^{(k)})$. Then the update is done by

$$\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} - H_k^{-1}\mathbf{g}_k.$$

Basically, the method uses a function approximation up to the second order.

3 Review of Probability Theory

First of all, what is probability? Let us consider the following statement: "the probability of a coin landing heads when a coin is tossed is 0.5". What does this mean? There are two possible interpretations. The one is that when you flip a coin many times (i.e., sufficiently large number of trials: see [4] for detailed and interesting explanation), then we expect that the coin lands heads about half the time. This is called the frequentist interpretation. The other interpretation is that the probability is to quantify the uncertainty. This is called the Bayesian interpretation (or the classical definition of the probability). Regardless of this, in this section, we provide a brief review on the probability theory. For more detailed review, refer to [5, 6].

3.1 Definition of probability

Let us discuss the classical definition of probability. Let S be the set that contains all the events (or outcomes) possible. And let A be the set that contains the events of interests. For now, we consider discrete cases. Then number of elements are denoted as $N_S = |S|$ and $N_A = |A|$, respectively. Then the probability of A occurring, denoted as $P(A)$ (or $\Pr(A)$ as in our main text [1]), is computed as

$$P(A) = \frac{N_A}{N_S}$$

assuming that all events are equally likely to occur. For example, let us consider a dice. Let $S = \{1, 2, 3, 4, 5, 6\}$ and $A = \{1, 2\}$. The probability of having 1 or 2 when casting a dice once is $P(A) = \frac{2}{6} = \frac{1}{3}$. Also we can see that $0 \leq P(A) \leq 1$.

Let A and B be subsets of S. The union of two events A and B is denoted as $A \cup B$. Also the intersection of the events A and B is denoted as $A \cap B$. The probability that event A or event B or both occur is denoted as $P(A \cup B)$ and is computed via set theory as

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

where $P(A \cap B)$ denotes the probability that event A and B occur at the same time. It is also called the *joint probability* (see the later sections). It is often dented as $P(A, B)$. If the two events are *mutually exclusive*, then $P(A \cap B) = 0$ ($\because A \cap B = \emptyset$). Hence

$$P(A \cup B) = P(A) + P(B).$$

The events A and B are *independent* if and only if

$$P(A \cap B) = P(A)P(B).$$

Since $P(A \cap B) \neq 0$, independent events are not mutually exclusive.

The *conditional probability* $P(B \mid A)$ is the probability that event B occurs given that event A has occurred. It is defined as

$$P(B \mid A) = \frac{P(A \cap B)}{P(A)}$$

as long as $P(A) \neq 0$. Since $P(A \cap B) = P(A)P(B \mid A) = P(B \cap A) = P(B)P(A \mid B)$, We can also find that

$$P(B \mid A) = \frac{P(A \mid B)P(B)}{P(A)}$$

which is essentially the *Bayes rule* (see the later sections).

3.2 Random variables and stochastic processes

Simply speaking, a *random variable* can be thought of as an uncertain number (mostly in $\mathbb{R}$). Consider jet engine noise as an example. Suppose you are measuring the intensity of the noise at the same position. Every time you record the intensity as a function of time, the record will be different from one another, which means it cannot be measured/predicted deterministically. In other words, it changes randomly. Now imagine you collect its all time histories of the record. This ensemble of all possible time histories becomes a *random process* or *stochastic process*.

Every random process has its own probability distribution. For example, let us consider binary random variable A. The probability that event A is true is denoted as $P(A)$. On the other hand, $P(\bar{A})$ denotes the probability that event A is not true (or false), and it follows that $P(\bar{A}) = 1 - P(A)$. This is an example of discrete random process. In general, for a discrete random process $\{x\}$, there is a probability distribution $\{P(x)\}$ such that $\sum_x P(x) = 1$. Formally, with representing X as discrete random variable, we write $P(X = x)$ (or $\Pr(X = x)$) as the probability that X is equal to x.

For continuous random processes, the corresponding random variable X is continuous. In this case, the probability distribution is expressed with a probability density function, which we denote here as $f(X)$, with $\int_x f(x)dx = 1$. Here as an example, let us consider $X \in \mathbb{R}$. Since X is continuous, the probability is defined with respect to the interval as

$$P(a < X \leq b) = \int_a^b f(x)\, dx.$$

The probability satisfies the aforementioned properties exactly. To list a few important properties:

- Joint probability, which is the probability of the joint event A and B (A and B are also random variables), denoted as $P(A, B) = P(A \mid B)P(B)$ (called the *product rule*). We define the *marginal probability* as

  $$P(A) = \sum_b P(A \mid B = b)P(B = b)$$

  i.e., summing over all possible states of B. It is also called the *sum rule.*

- Conditional probability, the conditional probability of event A, given that event B is true, is defined as

  $$P(A \mid B) = \frac{P(A, B)}{P(B)}$$

if $P(B) > 0$.

3.2.1 Independence and conditional independence

- Random variables X and Y are *unconditionally independent*, denoted as $X \perp Y$, if $P(X, Y) = P(X)P(Y)$.
- Random variables X and Y are *conditionally independent* given another random variable Z, denoted as $X \perp Y \mid Z$, iff $P(X, Y \mid Z) = P(X \mid Z)P(Y \mid Z)$.

3.2.2 Mean and variance

For discrete random variable X with the probability distribution $\{P(x)\}$, the mean is defined as

$$E[X] \doteq \sum_x xP(x).$$

If X is a continuous random variable, then the mean is defined as

$$E[X] \doteq \int_x x\, f(x)\, dx$$

where $f(X)$ denotes the probability density function.

The variance of discrete random variable X, with $\mu = E[X]$, is defined as

$$\mathrm{Var}(X) \doteq E\big[(X-\mu)^2\big] = \sum_x (x-\mu)^2 P(x),$$

while for continuous random variable X, it is defined as

$$\mathrm{Var}[X] \doteq E\big[(X-\mu)^2\big] = \int_x (x-\mu)^2\, f(x)\, dx.$$

Either case, it follows that

$$\mathrm{Var}[X] = E[X^2] - \mu^2.$$

The standard deviation is defined as

$$\mathrm{std}[X] \doteq \sqrt{\mathrm{Var}[X]}.$$

3.2.3 Examples of probability distributions

- **The binomial distribution**
  Let the discrete random variable X follow the binomial distribution with parameters $n \in \mathbb{N}$ and $p \in [0, 1]$ (i.e., probability that X is true or probability of success in one trial). Then the probability of having k successes in n trials is

$$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$$

where the binomial coefficient is defined as

$$\binom{n}{k} = \frac{n!}{(n-k)!k!}.$$

It follows that $E[X] = np$ and $\mathrm{Var}[X] = np(1-p)$.

- **Bernoulli distribution**
  Let $X = \{0, 1\}$ be a binary random variable. Also as before, let p be the probability of success $(X = 1)$. We say that X has a Bernoulli distribution, $X \sim \mathrm{Ber}(p)$ where the probability is

  $$P(X = x; p) = p^{\mathbb{1}(x=1)}(1-p)^{\mathbb{1}(x=0)}$$

  where $\mathbb{1}(x = 1)$ denotes the indicator function (i.e., it is 1 when $x = 1$ and 0 otherwise). In other words,

  $$P(X = x; p) = \begin{cases} p & \text{if } x = 1 \\ 1-p & \text{if } x = 0 \end{cases}$$

- **Gaussian distribution**
  Let X be continuous random variable. Gaussian distribution is such that the PDF is Gaussian as

  $$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}}\, e^{-\frac{1}{2\sigma^2}(x-\mu)^2}$$

  where $\mu = E[X]$ and $\sigma^2 = \mathrm{Var}[X]$.

3.2.4 Multivariate distributions

So far, we have talked about univariate distributions (i.e., single random variable). As an example, when we consider two random variables X and Y, we need the joint probability distribution $P(X, Y)$. when X is a multivariate random variable, we can represent it with a vector $\mathbf{x}$. Let $\mathbf{x} \in \mathbb{R}^d$ where d denotes the dimension of X. Then the probability distribution is $P(\mathbf{x})$ in discrete form, or $f_X(\mathbf{x}) \geq 0$ as a PDF (continuous). The mean is defined in the same form as

$$E[\mathbf{x}] = \sum_x \mathbf{x} P(\mathbf{x})$$

in discrete form, and

$$E[\mathbf{x}] = \int_{\mathbb{R}^d} \mathbf{x}\, f_X(\mathbf{x})\, d\mathbf{x}$$

in continuous form.

Instead of variance, we define the *covariance* as

$$\begin{aligned}\mathrm{Cov}[\mathbf{x}] &\doteq E\left[(\mathbf{x} - E[\mathbf{x}])(\mathbf{x} - E[\mathbf{x}])^T\right] \\ &= \begin{pmatrix} \mathrm{Var}[X_1] & \mathrm{Cov}[X_1, X_2] & \cdots & \mathrm{Cov}[X_1, X_d] \\ \mathrm{Cov}[X_2, X_1] & \mathrm{Var}[X_2] & \cdots & \mathrm{Cov}[X_2, X_d] \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{Cov}[X_d, X_1] & \mathrm{Cov}[X_d, X_2] & \cdots & \mathrm{Var}[X_d] \end{pmatrix}\end{aligned}$$

where $\mathrm{Cov}[X_i, X_j] \doteq E\left[(X_i - E[X_i])(X_j - E[X_j])\right] = E[X_i X_j] - E[X_i]E[X_j]$. Note that if random variables X and Y are independent, $P(X,Y) = P(X)P(Y)$ and $\mathrm{Cov}[X,Y] = 0$. Note that in the textbook, Var is used to mean Cov.

Let $\boldsymbol{\mu} = E[\mathbf{x}] \in \mathbb{R}^d$ and $\Sigma = \mathrm{cov}[\mathbf{x}] \in \mathbb{R}^{d\times d}$. The multivariate Gaussian is defined as

$$f(\mathbf{x}; \boldsymbol{\mu}, \Sigma) \doteq \frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}} \exp\left[-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu})\right].$$

3.3 Transformations of random variables

Let $\mathbf{x}$ be an n-dimension random variable with its PDF as $f(\mathbf{x})$. Let $\boldsymbol{\mu} = E[\mathbf{x}]$ and $\Sigma = \mathrm{Cov}[\mathbf{x}]$. We want to consider another m-dimensional random variable $\mathbf{y}$ as

$$\mathbf{y} = A\mathbf{x} + \mathbf{b}$$

where $A \in \mathbb{R}^{m\times n}$ and $\mathbf{b} \in \mathbb{R}^m$. Then it follows that

$$E[\mathbf{y}] = A\,\boldsymbol{\mu} + \mathbf{b}$$

and

$$\mathrm{Cov}[\mathbf{y}] = A\Sigma A^T.$$

3.4 Bayes rule

Bayes rule (or Bayes theorem) plays a very important role in statistics and statistical learning. It combines the conditional probability with the product and sum rule, expressed as

$$P(X=x \mid Y=y) = \frac{P(X=x, Y=y)}{P(Y=y)} = \frac{P(X=x)P(Y=y \mid X=x)}{\sum_{x'} P(X=x')P(Y=y \mid X=x')}$$

for discrete random variables. For continuous random variables, we need to consider PDF as

$$f_{X|Y}(x \mid y) = \frac{f_{XY}(x,y)}{f_Y(y)} = \frac{f_X(x) f_{Y|X}(y \mid x)}{\int_{x'} f_X(x') f_{Y|X}(y \mid x')\, dx'}.$$

With another random variable included, Bayes rule can be written as

$$P(X=x \mid Y=y, Z=z) = \frac{P(Y=y \mid X=x, Z=z)P(X=x \mid Z=z)}{P(Y=y \mid Z=z)}$$

or simply

$$P(x \mid y, z) = \frac{P(y \mid x, z)P(x \mid z)}{P(y \mid z)}.$$

This Bayes rule can be found in many places. As an example, we want to compute the possibility of each case occurrence c in a set $\mathcal{C}$, given a training data $\mathcal{T}$. This is an example of classification. The *posterior probability*, which is important in the classification, is expressed as

$$P(c \mid \mathcal{T}) = \frac{P(\mathcal{T} \mid c)\, P(c)}{\sum_{c' \in \mathcal{C}} P(\mathcal{T}, c')}$$

where $P(c)$ is called a prior probability, and $P(\mathcal{T} \mid c)$ is called a likelihood. In other words. posterior $\sim$ likelihood $\times$ prior.

Another ample examples can be found in Bayesian and Gaussian filters (e.g., Kalman filters). See [7] for more details.

References

[1] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning.* Springer, 2009.

[2] Edwin K.P. Chong and Stanislaw H. Żak. *An Introduction to Optimization.* John Wiley & Sons, Inc., 1996.

[3] https://towardsdatascience.com/gradient-descent-algorithm-and-its-variants-10f652806a3.

[4] Athanasios Papoulis. *Probability, Random Variables, and Stochastic Processes.* McGraw-Hill, Inc., 3rd edition, 1991.

[5] Kevin P. Murphy. *Machine Learning: A Probabilistic Perspective.* The MIT Press, 2012.

[6] Christopher M. Bishop. *Pattern Recognition and Machine Learning.* Springer, 2006.

[7] Sebastian Thrun, Wolfram Burgard, and Dieter Fox. *Probabilistic Robotics.* The MIT Press, 2006.