Some Mathematical Basics

Kaikai Zhao

First draft: January 1, 2022 Last update: January 27, 2023

Contents

1 To-do list

1.1 Some results about inequalities

1. Make notes of the result about absolute-value inequalities. See p. 50 in Exploring Inequalities.
2. Make notes of (weighted) Power average inequalities. See pp. 106-108 in Exploring Inequalities.

2 Some equalities

1. $(a+b+c)^2 = a^2+b^2+c^2+2(ab+bc+cd)$.
2. $(a+b+c)^3 = a^3+b^3+c^3+3(a^2b+ab^2+b^2c+bc^2+a^2c+ac^2)+6abc$.
3. $(a+b+c)^3+3abc = a^3+b^3+c^3+3(a+b+c)(ab+bc+cd)$.

   Here's an example: given $a+b+c=1, a^2+b^2+c^2=2, a^3+b^3+c^3=3$, find abc.

Solution: Obviously, abc can be obtained through the foregoing third equality. Before that, we need to calculate $ab+bc+cd$. By the first equality, $ab+bc+cd=\frac{1}{2}(1^2-2)=-\frac{1}{2}$. Thus, $abc=\frac{1}{3}\left(3+3\times 1\times(-\frac{1}{2})-1^3\right)=\frac{1}{6}$. □

1. $\max\{f,g\}=\frac{1}{2}[f+g+|f-g|]$.
2. $\min\{f,g\}=\frac{1}{2}[f+g-|f-g|]$.
3. $x^3+1=(x+1)(x^2-x+1)$; $x^3-1=(x-1)(x^2+x+1)$.
4. $(x+1)^3=x^3+3x^2+3x+1$; $(x-1)^3=x^3-3x^2+3x-1$.
5. $x^4+x^2+1=(x^2+1+x)(x^2+1-x)$.
6. If $a\le x\le b$, then $x=a\sin^2 t+b\cos^2 t$ as follows.

$$\begin{aligned} & x=a\sin^2 t+b\cos^2 t=a(1-\cos^2 t)+b\cos^2 t=a+(b-a)\cos^2 t \\ & 0\le\cos^2 t\le 1 \\ \Leftrightarrow & 0\le(b-a)\cos^2 t\le b-a \\ \Leftrightarrow & a\le a+(b-a)\cos^2 t\le b \\ \Leftrightarrow & a\le x\le b \end{aligned}$$

   where $x=a$ when $t=\frac{k\pi}{2}$ and $x=b$ when $t=k\pi$ with $k\in\mathbf{Z}$.

1. $2^n\cdot n!=(2\cdot 2\cdot 2\cdots 2)\cdot(1\cdot 2\cdot 3\cdots n)=2\times 4\times 6\times\cdots\times 2n=(2n)!!$
2. $\frac{(2n)!!}{(2n)!}=\frac{2\times 4\times 6\times\cdots\times 2n}{1\times 2\times 3\times 4\times\cdots\times(2n-1)\times 2n}=\frac{1}{(2n-1)!!}$
3. $3^{\ln n}=(e^{\ln 3})^{\ln n}=(e^{\ln n})^{\ln 3}=n^{\ln 3}\Longleftrightarrow 3^{\ln n}=n^{\ln 3}$
4. $(\ln n)^{\ln n}=(e^{\ln\ln n})^{\ln n}=(e^{\ln n})^{\ln\ln n}=n^{\ln\ln n}\Longleftrightarrow(\ln n)^{\ln n}=n^{\ln\ln n}>n^2$ $(n\ge\mathrm{ceil}(e^{e^2}))$
5. $(\ln n)^{\ln\ln n}=(e^{(\ln\ln n)})^{\ln\ln n}=e^{(\ln\ln n)^2}\Longleftrightarrow(\ln n)^{\ln\ln n}=e^{(\ln\ln n)^2}<e^{\ln n}=n$ (n is big enough)

1. $xy=\frac{1}{4}[(x+y)^2-(x-y)^2]$, where x,y can be vectors. Hence, if $x+y$ is fixed, the smaller $|x-y|$ is, the greater xy will be. To put this rule in math,

   Proposition 1. *If $a+d=b+c$ and $|a-d|>|b-c|$, then $ad<bc$.*

2. $|x+y|=\sqrt{|x-y|^2+4xy}$, $|x-y|=\sqrt{|x+y|^2-4xy}$, where x,y can be vectors. If xy is fixed, then the lesser $|x-y|$ is, the smaller $|x+y|$ will be, namely,

   Proposition 2. *If $ad=bc$ and $|a-d|<|b-c|$, then $|a+d|<|b+c|$.*

3. Let $x+\frac{1}{x}=t$, then

$$\begin{aligned} x^2+\frac{1}{x^2}&=(x+\frac{1}{x})^2-2=t^2-2 \\ x^3+\frac{1}{x^3}&=(x+\frac{1}{x})^3-3(x+\frac{1}{x})=t^3-3t \\ x^4+\frac{1}{x^4}&=(x^2+\frac{1}{x^2})^2-2=(t^2-2)^2-2 \end{aligned}$$

3 Some inequalities

3.1 The integral average inequalities of a function

Definition 1 (Integral average of a function)**.** $N_p(f)$ denotes the average of f w.r.t the set E and exponent p,

$$N_p(f) = \left(\frac{1}{\mu(E)}\int_E |f|^p \mathrm{d}\,\mu\right)^{\frac{1}{p}}$$

where $0 < \mu(E) < +\infty, 0 \le p < +\infty$.

$N_p(f)$ satisfies the following properties:

1. Hölder inequality: $N_1(fg) \le N_p(f)N_q(g)$ where $1/p + 1/q = 1, 1 < p < +\infty$.
2. Minkowski inequality: $N_p(f+g) \le N_p(f) + N_p(g)$ where $1 \le p < +\infty$.
3. $N_p(f)$ is increasing w.r.t p: $1 \le p_1 \le p_2 \Rightarrow N_{p_1}(f) \le N_{p_2}(f)$.

The definition and properties are taken from the Section 1.2.4.7 of Jichang Kuang's Applied Inequalites, 5^{th} edition in Chinese, 2021.

We'll present the specific forms of the first two properities later. To prove the integral form of Hölder inequality, we first need to introduce a lemma.

Lemma 1. *Given $a, b \ge 0$ and p, q satisfying $1/p + 1/q = 1, 1 \le p < +\infty$, then $ab \le \frac{1}{p}a^p + \frac{1}{q}b^q$ holds.*

Proof. Either a or b is 0, the claim is obvious. When $a, b > 0$, $f(x) = \ln x$ is strictly concave on $(0, +\infty)$ due to the fact that $f'(x) = -\frac{1}{x^2} < 0$. By Jensen's inequality, we have

$$\frac{1}{p}f(a^p) + \frac{1}{q}f(b^q) \le f(\frac{1}{p}a^p + \frac{1}{q}b^q)$$

where the equality holds when $a = b$ and $p = q = 2$. Thus,

$$\ln(ab) = \frac{1}{p}\ln(a^p) + \frac{1}{q}\ln(b^q) \le \ln(\frac{1}{p}a^p + \frac{1}{q}b^q)$$

Since $\ln x$ is monotonically increasing in x, we obtain the claim

$$ab \le \frac{1}{p}a^p + \frac{1}{q}b^q \quad (a, b > 0)$$

□

Theorem 1 (**Hölder inequality in integral form**)**.** *Provided $f(x), g(x)$ are continuous functions on $[a, b]$,*

$$\int_a^b |f(x)g(x)|\mathrm{d}\,x \le \left(\int_a^b |f(x)|^p \mathrm{d}\,x\right)^{\frac{1}{p}} \left(\int_a^b |g(x)|^q \mathrm{d}\,x\right)^{\frac{1}{q}}$$

holds, where $1/p + 1/q = 1, p, q \le 1$.

Proof. If $f(x) \equiv 0$ or $g(x) \equiv 0$, the statement is obvious. Otherwise, let

$$\varphi(x) = \frac{|f(x)|}{\left(\int_a^b |f(x)|^p \mathrm{d}\,x\right)^{\frac{1}{p}}},\ \psi(x) = \frac{|g(x)|}{\left(\int_a^b |g(x)|^q \mathrm{d}\,x\right)^{\frac{1}{q}}},\ x \in [a, b]$$

By Lemma 1, we get

$$\varphi(x)\psi(x) = \frac{1}{p}\varphi(x)^p + \frac{1}{q}\psi(x)^q,$$

i.e.,

$$\frac{|f(x)g(x)|}{\left(\int_a^b |f(x)|^p \mathrm{d}\,x\right)^{\frac{1}{p}} \left(\int_a^b |g(x)|^q \mathrm{d}\,x\right)^{\frac{1}{q}}} \leq \frac{|f(x)|^p}{p\int_a^b |f(x)|^p \mathrm{d}\,x} + \frac{|g(x)|^q}{q\int_a^b |g(x)|^q \mathrm{d}\,x},\ x \in [a,b].$$

Integrating both sides over x, we have

$$\frac{\int_a^b |f(x)g(x)| \mathrm{d}\,x}{\left(\int_a^b |f(x)|^p \mathrm{d}\,x\right)^{\frac{1}{p}} \left(\int_a^b |g(x)|^q \mathrm{d}\,x\right)^{\frac{1}{q}}} \leq \frac{\int_a^b |f(x)|^p \mathrm{d}\,x}{p\int_a^b |f(x)|^p \mathrm{d}\,x} + \frac{\int_a^b |g(x)|^q \mathrm{d}\,x}{q\int_a^b |g(x)|^q \mathrm{d}\,x} = \frac{1}{p} + \frac{1}{q} = 1.$$

Multiplying both sides by $\left(\int_a^b |f(x)|^p \mathrm{d}\,x\right)^{\frac{1}{p}} \left(\int_a^b |g(x)|^q \mathrm{d}\,x\right)^{\frac{1}{q}}$, we have

$$\int_a^b |f(x)g(x)| \mathrm{d}\,x \leq \left(\int_a^b |f(x)|^p \mathrm{d}\,x\right)^{\frac{1}{p}} \left(\int_a^b |g(x)|^q \mathrm{d}\,x\right)^{\frac{1}{q}}$$

This completes the proof. □

When $p = q = 2$, an important special case is obtained as the following corollary says.

Corollary 1 (**Schwarz inequality in integral form**). *If both $f(x)$ and $f(x)$ are integrable on $x \in [a,b]$, then*

$$\left(\int_a^b |f(x)g(x)| \mathrm{d}\,x\right)^2 \leq \int_a^b |f(x)|^2 \mathrm{d}\,x \int_a^b |g(x)|^2 \mathrm{d}\,x.$$

Since it is a special case of Theorem 1, we do not have to prove it any more. However, a concise proof is available in Exercise 6.2.12(1) the solution manual of Jixiu Chen's Mathematical Analysis, 3^{rd} edition in Chinese, 2019.5, and we present it here.

Proof. For any real t, we always have $(tf(x) + g(x))^2 \geq 0$. Thus,

$$t^2 f(x)^2 + 2tf(x)g(x) + g(x)^2 \geq 0$$

Taking integrals on both sides over x,

$$t^2 \int_a^b f(x)^2 \,\mathrm{d}\,x + 2t \int_a^b f(x)g(x) \,\mathrm{d}\,x + \int_a^b g(x)^2 \,\mathrm{d}\,x \geq 0$$

Since the above holds for any real t, we get

$$4t^2 \left(\int_a^b f(x)g(x) \,\mathrm{d}\,x\right)^2 - 4t^2 \int_a^b f(x)^2 \,\mathrm{d}\,x \int_a^b g(x)^2 \,\mathrm{d}\,x \leq 0 \Rightarrow \left(\int_a^b f(x)g(x) \,\mathrm{d}\,x\right)^2 \leq \int_a^b f(x)^2 \,\mathrm{d}\,x \int_a^b g(x)^2 \,\mathrm{d}\,x$$

This completes the proof. □

An immediate result from this is Minkowski inequality in integral form.

Corollary 2 (**Minkowski inequality in integral form(when $p = 2$)**). *If both $f(x)$ and $f(x)$ are integrable on $x \in [a,b]$, then*

$$\left(\int_a^b [f(x) + g(x)]^2 \,\mathrm{d}\,x\right)^{\frac{1}{2}} \leq \left(\int_a^b f(x)^2 \,\mathrm{d}\,x\right)^{\frac{1}{2}} + \left(\int_a^b g(x)^2 \,\mathrm{d}\,x\right)^{\frac{1}{2}}.$$

Proof. By the Schwarz inequality in integral form, we get

$$2\int_a^b |f(x)g(x)|\mathrm{d}\,x \le 2\left(\int_a^b |f(x)|^2\mathrm{d}\,x\right)^{\frac{1}{2}}\left(\int_a^b |g(x)|^2\mathrm{d}\,x\right)^{\frac{1}{2}}$$

$$\int_a^b \left(f(x)^2\,\mathrm{d}\,x + g(x)^2\,\mathrm{d}\,x + 2|f(x)g(x)|\right)\mathrm{d}\,x \le \int_a^b f(x)^2\,\mathrm{d}\,x + \int_a^b g(x)^2\,\mathrm{d}\,x + 2\left(\int_a^b |f(x)|^2\mathrm{d}\,x\right)^{\frac{1}{2}}\left(\int_a^b |g(x)|^2\mathrm{d}\,x\right)^{\frac{1}{2}}$$

$$\int_a^b [f(x)+g(x)]^2\,\mathrm{d}\,x \le \left(\left(\int_a^b |f(x)|^2\mathrm{d}\,x\right)^{\frac{1}{2}} + \left(\int_a^b |g(x)|^2\mathrm{d}\,x\right)^{\frac{1}{2}}\right)^2$$

$$\left(\int_a^b [f(x)+g(x)]^2\,\mathrm{d}\,x\right)^{\frac{1}{2}} \le \left(\int_a^b f(x)^2\,\mathrm{d}\,x\right)^{\frac{1}{2}} + \left(\int_a^b g(x)^2\,\mathrm{d}\,x\right)^{\frac{1}{2}}.$$

This completes our proof. □

The following part concerning vector norm properties is taken from Matrix Computations, 4th edition, page 69.

3.2 Hölder inequality

A classic result concerning p-norms is the **Hölder inequality in inner-product form**:

$$|x^{\mathrm{T}}y| \le \|x\|_p\|y\|_q, \quad \frac{1}{p}+\frac{1}{q}=1.$$

Proof. Recall the definition of dual norm,

$$\|x\|_p = \max_{\|z\|_q\le 1} x^T z.$$

Thus, $x^T z \le \|x\|_p$ holds for any z satisfying $\|z\|_q \le 1$, including $\|z\|_q = 1$. When $\|z\|_q = 1$, we have

$$x^T z \le \|x\|_p\|z\|_q$$

Now let $z = ty$ with $t > 0$. Thus, we have

$$x^T(ty) \le \|x\|_p\|ty\|_q \Longleftrightarrow x^T y \le \|x\|_p\|y\|_q$$

where $\|y\|_q = \|z/t\|_q = 1/t\|z\|_q = 1/t > 0$. This completes the proof. □

The key part of the above proof is using the dual representation of the ℓ_p norm, namely, $\|x\|_p = \max_{\|z\|_q\le 1} x^T z$.

A very important special case of Hölder inequality is **Cauchy-Schwarz inequality**:

$$|x^{\mathrm{T}}y| \le \|x\|_2\|y\|_2.$$

which can also be expressed as $(x^T y)^2 \le \|x\|_2^2\|y\|_2^2$. Here, $x = (x_1, x_2, \ldots, x_n)$ and $y = (y_1, y_2, \ldots, y_n)$. Thus,

$$(x_1y_1 + x_2y_2 + \ldots + x_ny_n)^2 \le (x_1^2 + x_2^2 + \ldots + x_n^2)(y_1^2 + y_2^2 + \ldots + y_n^2)$$

$$(\sum_{i=1}^n x_iy_i)^2 = \sum_{i=1}^n x_i^2 \sum_{i=1}^n y_i^2$$

Let $y_i = \sqrt{b_i}$ and $x_i = \frac{a_i}{\sqrt{b_i}}$. We have an important **variant** of Cauchy-Schwarz inequality.

$$(\sum_{i=1}^n \frac{a_i}{\sqrt{b_i}}\sqrt{b_i})^2 = (\sum_{i=1}^n a_i)^2 \le (\sum_{i=1}^n \frac{a_i^2}{b_i})\sum_{i=1}^n b_i \Longleftrightarrow \frac{(\sum_{i=1}^n a_i)^2}{\sum_{i=1}^n b_i} \le \sum_{i=1}^n \frac{a_i^2}{b_i}$$

3.3 Norm equivalence

All norms on $\mathbb{R}^n$ are equivalent, i.e., if $\|\cdot\|_\alpha$ and $\|\cdot\|_\beta$ are norms on $\mathbb{R}^n$, then there exist positive constants c_1 and c_2 such that

$$c_1\|x\|_\alpha \leq \|x\|_\beta \leq c_2\|x\|_\alpha$$

for all $x \in \mathbb{R}^n$. For example, if $x \in \mathbb{R}^n$, then

$$\|x\|_2 \leq \|x\|_1 \leq \sqrt{n}\|x\|_2$$
$$\|x\|_\infty \leq \|x\|_2 \leq \sqrt{n}\|x\|_\infty$$
$$\|x\|_\infty \leq \|x\|_1 \leq n\|x\|_\infty$$

Finally, we mention that the **ℓ_2-norm is preserved under orthogonal transformation**. Indeed, if $Q \in \mathbb{R}^{n\times n}$ is orthogonal and $x \in \mathbb{R}^n$, then

$$\|Qx\|_2^2 = (Qx)^{\mathrm{T}}Qx = x^{\mathrm{T}}Q^{\mathrm{T}}Qx = x^{\mathrm{T}}x = \|x\|_2^2$$

3.4 Inequalities

1. If $a > b > 0, m > 0$, then
$$\frac{b}{a} < \frac{b+m}{a+m}$$
If $a > b > m > 0$, then
$$\frac{b-m}{a-m} < \frac{b}{a} < \frac{b+m}{a+m}$$
If $k > 0$,
$$\frac{bk-m}{ak-m} < \frac{b}{a} < \frac{bk+m}{ak+m}$$
If $a_1 > b_1 > 0, a_2 > b_2 > 0$ and $\frac{b_1}{a_1} < \frac{b_2}{a_2}$, then
$$\frac{b_1}{a_1} < \frac{b_1+b_2}{a_1+a_2} < \frac{b_2}{a_2}$$
If $k_1, k_2 > 0$,
$$\frac{b_1}{a_1} < \frac{b_1k_1+b_2k_2}{a_1k_1+a_2k_2} < \frac{b_2}{a_2}$$
2. If $b > a > 0, m > 0$, then
$$\frac{b}{a} > \frac{b+m}{a+m}$$
3. For any positive a, b, if $x > 0$,
$$ax + \frac{b}{x} \geq 2\sqrt{ab}$$
where the equality holds if and only if $x = \sqrt{\frac{b}{a}}$. If $x < 0$,
$$ax + \frac{b}{x} \leq -2\sqrt{ab}$$
where the equality holds if and only if $x = -\sqrt{\frac{b}{a}}$. To sum up, the following always hold
$$a|x| + \frac{b}{|x|} \geq 2\sqrt{ab}, \quad |ax + \frac{b}{x}| \geq 2\sqrt{ab}$$
where the equality holds if and only if $|x| = \sqrt{\frac{b}{a}}$.
4. if $ab > 0$, then $\frac{a}{b} + \frac{b}{a} \geq 2$; if $ab < 0$, then $\frac{a}{b} + \frac{b}{a} \leq -2$. In a nutshell, if $ab \neq 0$, then $|\frac{a}{b} + \frac{b}{a}| \geq 2$

3.5 Triangle inequality

1. For any reals a, b, the following holds,

$$||a|-|b||\leq |a \pm b|\leq |a|+|b|$$

2. For any vectors $\mathbf{v}, \mathbf{w}$,

$$|\|\mathbf{v}\|_2-\|\mathbf{w}\|_2|\leq \|\mathbf{v} \pm \mathbf{w}\|_2\leq \|\mathbf{v}\|_2+\|\mathbf{w}\|_2$$

3.

$$| \underbrace{\sqrt{x_1}}_{\text{hypotenuse}} - \underbrace{\sqrt{x_2}}_{\text{one leg}} |\leq \underbrace{\sqrt{|x_1-x_2|}}_{\text{the other leg}} \leq \underbrace{\sqrt{|x_1-x_2|}+\sqrt{x_2}}_{\text{one leg + another leg}} \leq \underbrace{\sqrt{x_1}+\sqrt{x_2}}_{\text{hypotenuse + one leg}}$$

where the last inequality only when $\sqrt{x_1}$ is the hypotenuse.

Proof. In terms of geometry, we can think of $\sqrt{x_1}, \sqrt{x_2}$ as the hypotenuse and one leg of a right triangle, respectively. Then the difference between two sides of a triangle is always less than the length of the third side.

In terms of algebra,

$$|\sqrt{x_1}-\sqrt{x_2}|\leq \frac{|x_1-x_2|}{\sqrt{x_1}+\sqrt{x_2}} \leq \frac{\sqrt{|x_1-x_2|}\sqrt{|x_1-x_2|}}{\sqrt{x_1}+\sqrt{x_2}} \leq \sqrt{|x_1-x_2|}$$

□

A more general formulation is

$$\underbrace{|\sqrt{x_1^2+x_2^2}-\sqrt{y_1^2+y_2^2}|}_{\text{difference between two sides}} \leq \underbrace{\sqrt{(x_1-y_1)^2+(x_2-y_2)^2}}_{\text{the third side}} \leq \underbrace{|x_1-y_1|+|x_2-y_2|}_{\text{sum of two legs}}$$

4 Trigonometric inequalities

1.

$$|\sin x|\leq |x|, \quad x \in \mathbb{R}$$

2.

$$|\sin x-\sin y|= |2\cos(\frac{x+y}{2})\sin(\frac{x-y}{2})|\leq |x-y|$$

3.

$$|\cos x-\cos y|= |-2\sin(\frac{x+y}{2})\sin(\frac{x-y}{2})|\leq |x-y|$$

4. Since $f(x)=\frac{\sin x}{x}$ is decreasing on $(0, \frac{\pi}{2}]$, then $f(x)\geq f(\frac{\pi}{2})=\frac{2}{\pi}$, i.e., $\frac{\sin x}{x}\geq \frac{2}{\pi}$.

$$\sin x \geq \frac{2}{\pi}x, \quad x \in [0, \frac{\pi}{2}].$$

5 Trigonometric identities

1. $\sin(x \pm y) = \sin x \cos y \pm \cos x \sin y$
2. $\cos(x \pm y) = \cos x \cos y \mp \sin x \sin y$
3. $\tan(x \pm y) = \frac{\tan x \pm \tan y}{1 \mp \tan x \cdot \tan y}, \quad \arctan x - \arctan y = \arctan \frac{x-y}{1+xy}$
4. $\sin x - \sin y = \sin(\frac{x+y}{2} + \frac{x-y}{2}) - \sin(\frac{x+y}{2} - \frac{x-y}{2}) = 2\cos(\frac{x+y}{2})\sin(\frac{x-y}{2})$
5. $\cos x - \cos y = \cos(\frac{x+y}{2} + \frac{x-y}{2}) - \cos(\frac{x+y}{2} - \frac{x-y}{2}) = -2\sin(\frac{x+y}{2})\sin(\frac{x-y}{2})$
6. $\cos 2x = \cos^2 x - \sin^2 x = 1 - 2\sin^2 x$
7. $\sec x = \frac{1}{\cos x},\ \csc x = \frac{1}{\sin x},\ \sec^2 x = 1 + \tan^2 x,\ \csc^2 x = 1 + \cot^2 x$
8. $\sinh x = \frac{e^x - e^{-x}}{2},\ \cosh x = \frac{e^x + e^{-x}}{2},\ \tanh x = \frac{\sinh x}{\cosh x} = \frac{e^x - e^{-x}}{e^x + e^{-x}},\ \coth x = \frac{1}{\tanh} = \frac{e^x + e^{-x}}{e^x - e^{-x}}$
9. $\cosh^2 x - \sinh^2 x = 1,\ \mathrm{sech} x = \frac{1}{\sinh},\ \mathrm{csch} x = \frac{1}{\cosh},\ \mathrm{sech}^2 x = 1 - \tanh^2 x,\ \mathrm{csch}^2 x = \coth^2 - 1$
10. $\tan x - \cot x = \frac{\sin x}{\cos x} + \frac{\cos x}{\sin x} = \frac{1}{\sin x \cos x} = \frac{2}{\sin 2x} = 2\csc 2x$
11. $\tan x - \cot x = \frac{\sin x}{\cos x} - \frac{\cos x}{\sin x} = \frac{\sin^2 x - \cos^2 x}{\sin x \cos x} = \frac{-\cos 2x}{\frac{1}{2}\sin 2x} = -2\cot 2x$
12. $$\arctan x + \operatorname{arccot} x = \begin{cases} \frac{\pi}{2}, & x \geq 0, \\ -\frac{\pi}{2}, & x < 0, \end{cases}$$
    where $\lim_{x \to 0^+}\{\arctan x + \operatorname{arccot} x\} = \frac{\pi}{2}$ and $\lim_{x \to 0^-}\{\arctan x + \operatorname{arccot} x\} = -\frac{\pi}{2}$.
13. Using Euler formula yields
$$\cos\theta = \frac{e^{i\theta} + e^{-i\theta}}{2}, \quad \sin\theta = \frac{e^{i\theta} - e^{-i\theta}}{2i}$$
$$\cos iy = \frac{e^{i\cdot iy} + e^{-i\cdot iy}}{2} = \frac{e^{-y} + e^{y}}{2} = \cosh y$$
$$\sin iy = \frac{e^{i\cdot iy} - e^{-i\cdot iy}}{2i} = \frac{e^{-y} - e^{y}}{2i} = i\sinh y$$
$$\begin{aligned} \cos(x + iy) &= \cos x \cos iy - \sin x \sin iy \\ &= \cos x \cosh y - i \sin x \sinh y \end{aligned}$$
$$\begin{aligned} \sin(x + iy) &= \sin x \cos iy + \cos x \sin iy \\ &= \sin x \cosh y + i \cos x \sinh y \end{aligned}$$

6 Infinitesimal of same order

1. $\ln(1+x) \sim x \quad (x \to 0)$.
2. $e^x - 1 \sim x \quad (x \to 0)$.
3. $(1+x)^a - 1 \sim ax \quad (x \to 0)$.
4. $1 - \cos x \sim \frac{x^2}{2} \quad (x \to 0)$
5. Sterling formula: $n! \sim \sqrt{2\pi} n^{n+\frac{1}{2}} e^{-n} \quad (n \to \infty)$

$$n^n \gg n! \gg a^n \ (a > 1) \gg n^\alpha \ (\alpha > 0) \gg \log^\beta n \ (\beta > 0)$$

7 Combinatorics

1. $\mathrm{C}_m^k + \mathrm{C}_m^{k-1} = \mathrm{C}_{m+1}^k$
2. $\frac{k}{n}\mathrm{C}_n^k = \mathrm{C}_{n-1}^{k-1}$, $\frac{k(k-1)}{n^2}\mathrm{C}_n^k = \frac{n-1}{n}\mathrm{C}_{n-2}^{k-2}$

$$\frac{k(k-1)}{n^2}\mathrm{C}_n^k = \frac{k(k-1)}{n^2}\frac{n!}{k!\,(n-k)!} = \frac{k-1}{n}\frac{(n-1)!}{(k-1)!\,(n-k)!} = \frac{k-1}{n}\mathrm{C}_{n-1}^{k-1} = \frac{n-1}{n}\frac{k-1}{n-1}\mathrm{C}_{n-1}^{k-1} = \frac{n-1}{n}\mathrm{C}_{n-2}^{k-2}$$

8 Calculus

8.1 Indefinite integrals

1. $\int \frac{1}{x}\,\mathrm{d}\,x = \ln|x| + C$. (Note that it is $|x|$ on the right hand side.)
2. $\int x^\alpha\,\mathrm{d}\,x = \begin{cases} \frac{1}{\alpha+1}x^{\alpha+1} + C, & \alpha \neq -1, \\ \ln|x| + C, & \alpha = -1. \end{cases}$
3. $\int \ln x\,\mathrm{d}\,x = x(\ln x - 1) + C$.
4. $\int \sin x\,\mathrm{d}\,x = -\cos x + C$; $\int \cos x\,\mathrm{d}\,x = \sin x + C$.
5. $\int \tan x\,\mathrm{d}\,x = -\ln|\cos x| + C$; $\int \cot x\,\mathrm{d}\,x = \ln|\sin x| + C$.
6. $\int \sec x\,\mathrm{d}\,x = \ln|\sec x + \tan x| + C$; $\int \csc x\,\mathrm{d}\,x = \ln|\csc x - \cot x| + C$.
7. $\int \sinh x\,\mathrm{d}\,x = \cosh x + C$; $\int \cosh x\,\mathrm{d}\,x = \sinh x + C$.
8. $\int \frac{\mathrm{d}\,x}{\sqrt{a^2 - x^2}} = \arcsin\frac{x}{a} + C$; $\int \frac{\mathrm{d}\,x}{\sqrt{x^2 \pm a^2}} = \ln|x + \sqrt{x^2 \pm a^2}| + C$.
9. $\int \frac{\mathrm{d}\,x}{x^2 + a^2} = \frac{1}{a}\arctan\frac{x}{a} + C$; $\int \frac{\mathrm{d}\,x}{x^2 - a^2} = \frac{1}{2a}\ln\left|\frac{x-a}{x+a}\right| + C$.
10. $\int \sqrt{a^2 - x^2}\,\mathrm{d}\,x = \frac{1}{2}x\sqrt{a^2 - x^2} + \frac{a^2}{2}\arcsin\frac{x}{a} + C$.
11. $\int \sqrt{x^2 \pm a^2} = \frac{1}{2}(x\sqrt{x^2 \pm a^2} \pm a^2 \ln|x + \sqrt{x^2 \pm a^2}|) + C$.

12.

$$\begin{aligned}\int \frac{(ax+b)\,\mathrm{d}\,x}{x^2+2\xi x+\eta^2} &= \frac{a}{2}\int \frac{\mathrm{d}(x^2+2\xi x+\eta^2)}{x^2+2\xi x+\eta^2} + (b-a\xi)\int \frac{\mathrm{d}\,x}{x^2+2\xi x+\eta^2} \\ &= \frac{a}{2}\ln|x^2+2\xi x+\eta^2| + (b-a\xi)\int \frac{\mathrm{d}\,x}{(x+\xi)^2+\eta^2-\xi^2} \\ &= \begin{cases} a\ln|x+\xi| - \frac{b-a\xi}{x+\xi} + C, & |\eta| = |\xi| \\ \frac{a}{2}\ln(x^2+2\xi x+\eta^2) + \frac{b-a\xi}{\sqrt{\eta^2-\xi^2}}\arctan\frac{x+\xi}{\sqrt{\eta^2-\xi^2}} + C, & |\eta| > |\xi| \\ \frac{a}{2}\ln|x^2+2\xi x+\eta^2| + \frac{b-a\xi}{2\sqrt{\xi^2-\eta^2}}\ln\left|\frac{x+\xi-\sqrt{\xi^2-\eta^2}}{x+\xi+\sqrt{\xi^2-\eta^2}}\right| + C, & |\eta| < |\xi|. \end{cases}\end{aligned}$$

13.

$$\begin{aligned}\int \frac{(ax+b)\,\mathrm{d}\,x}{\sqrt{x^2+2\xi x+\eta^2}} &= \frac{a}{2}\int \frac{\mathrm{d}(x^2+2\xi x+\eta^2)}{\sqrt{x^2+2\xi x+\eta^2}} + (b-a\xi)\int \frac{\mathrm{d}\,x}{\sqrt{x^2+2\xi x+\eta^2}} \\ &= a\sqrt{x^2+2\xi x+\eta^2} + (b-a\xi)\int \frac{\mathrm{d}\,x}{\sqrt{(x+\xi)^2+\eta^2-\xi^2}} \\ &= a\sqrt{x^2+2\xi x+\eta^2} + (b-a\xi)\ln|x+\xi+\sqrt{x^2+2\xi x+\eta^2}|\end{aligned}$$

where $x^2+2\xi x+\eta^2 > 0$ implies $4\xi^2-4\eta^2<0$, i.e., $\xi^2<\eta^2$.

14.

$$\begin{aligned}&\int (ax+b)\sqrt{x^2+2\xi x+\eta^2}\,\mathrm{d}\,x \\ &= \frac{a}{2}\int \sqrt{x^2+2\xi x+\eta^2}\,\mathrm{d}(x^2+2\xi x+\eta^2) + (b-a\xi)\int \sqrt{x^2+2\xi x+\eta^2}\,\mathrm{d}\,x \\ &= \frac{a}{2}\frac{(x^2+2\xi x+\eta^2)^{3/2}}{3/2} + (b-a\xi)\int \sqrt{(x+\xi)^2+\eta^2-\xi^2}\,\mathrm{d}\,x \\ &= \frac{a}{3}(x^2+2\xi x+\eta^2)^{3/2} \\ &\quad + \frac{1}{2}(b-a\xi)\left((x+\xi)\sqrt{x^2+2\xi x+\eta^2} + (\eta^2-\xi^2)\ln|(x+\xi)+\sqrt{x^2+2\xi x+\eta^2}|\right) + C\end{aligned}$$

15. Let $\tan\frac{x}{2}=t$, then $\sin x = 2\sin\frac{x}{2}\cos\frac{x}{2} = 2\frac{\sin\frac{x}{2}}{\cos\frac{x}{2}}\cos^2\frac{x}{2} = 2\tan\frac{x}{2}\cdot\frac{1}{\sec^2 x} = 2\tan\frac{x}{2}\cdot\frac{1}{1+\tan^2 x} = \frac{2t}{1+t^2}$, $\cos x = \sqrt{1-\sin^2 x} = \sqrt{1-\frac{4t^2}{(1+t^2)^2}} = \frac{1-t^2}{1+t^2}$, $\tan x = \frac{2t}{1-t^2}$, and $\mathrm{d}\,x = \mathrm{d}(2\arctan t) = \frac{2}{1+t^2}\,\mathrm{d}\,t$. Thus,

$$\int R(\sin x,\cos x)\,\mathrm{d}\,x = \int R\left(\frac{2t}{1+t^2},\frac{1-t^2}{1+t^2}\right)\frac{2}{1+t^2}\,\mathrm{d}\,t.$$

16. $\int \mathrm{e}^x\sin x\,\mathrm{d}\,x = \frac{\mathrm{e}^x}{2}(\sin x-\cos x)+C$; $\int \mathrm{e}^x\cos x\,\mathrm{d}\,x = \frac{\mathrm{e}^x}{2}(\sin x+\cos x)+C$

17. $\int x\,\mathrm{e}^x\sin x\,\mathrm{d}\,x = \frac{\mathrm{e}^x}{2}\,(x(\sin x-\cos x)+\cos x)+C$; $\int x\,\mathrm{e}^x\cos x\,\mathrm{d}\,x = \frac{\mathrm{e}^x}{2}\,(x(\sin x+\cos x)-\sin x)+C$

18. Provided $f(x)$ continuous on $[0,1]$, then

$$\int_0^{\frac{\pi}{2}} f(\cos x)\,\mathrm{d}\,x = \int_0^{\frac{\pi}{2}} f(\sin x)\,\mathrm{d}\,x.$$

Particularly, $\int_0^{\frac{\pi}{2}}\cos^2 x\,\mathrm{d}\,x = \int_0^{\frac{\pi}{2}}\sin^2 x\,\mathrm{d}\,x = \frac{1}{2}\int_0^{\pi}\sin^2 x\,\mathrm{d}\,x = \frac{1}{2}\int_0^{\pi}\frac{\sin^2+\cos^2}{2}x\,\mathrm{d}\,x = \frac{1}{2}\cdot\frac{\pi}{2}$.

Proof. Let $x = \frac{\pi}{2} - u$, then

$$\int_0^{\frac{\pi}{2}} f(\cos x)\,\mathrm{d}\,x = \int_{\frac{\pi}{2}}^0 f(\cos(\frac{\pi}{2} - u))\,\mathrm{d}(\frac{\pi}{2} - u) = -\int_{\frac{\pi}{2}}^0 f(\sin u)\,\mathrm{d}\,u = \int_0^{\frac{\pi}{2}} f(\sin u)\,\mathrm{d}\,u$$

□

19.

$$\int_0^{\pi} xf(\sin x)\,\mathrm{d}\,x = \frac{\pi}{2}\int_0^{\pi} f(\sin x)\,\mathrm{d}\,x$$

Proof. Let $x = \pi - u$,

$$\begin{aligned} I = \int_0^{\pi} xf(\sin x)\,\mathrm{d}\,x &= \int_{\pi}^0 (\pi - u)f(\sin(\pi - u))\,\mathrm{d}(\pi - u) \\ &= -\int_{\pi}^0 (\pi - u)f(\sin u)\,\mathrm{d}\,u \\ &= -\int_{\pi}^0 \pi f(\sin u)\,\mathrm{d}\,u + \int_{\pi}^0 uf(\sin u)\,\mathrm{d}\,u \\ &= \int_0^{\pi} \pi f(\sin u)\,\mathrm{d}\,u - \int_0^{\pi} uf(\sin u)\,\mathrm{d}\,u = \pi\int_0^{\pi} f(\sin u)\,\mathrm{d}\,u - I \end{aligned}$$

Thus,

$$I = \frac{\pi}{2}\int_0^{\pi} f(\sin u)\,\mathrm{d}\,u.$$

□

20. $\int (1 + x^{-2})\,\mathrm{d}\,x = x - x^{-1} + C$

21. Let $I_n = \int_0^{\frac{\pi}{2}} \sin^n x\,\mathrm{d}\,x$, then the following holds

$$I_n = (n-2)I_{n-2}.$$

Furthermore, since $I_0 = \int_0^{\frac{\pi}{2}} \mathrm{d}\,x = \pi/2$ and $I_0 = \int_0^{\frac{\pi}{2}} \sin x\,\mathrm{d}\,x = 1$, Furthermore, we have

$$\int_0^{\frac{\pi}{2}} \sin^n x\,\mathrm{d}\,x = \begin{cases} \frac{(n-1)\cdot(n-3)\cdot\cdots\cdot 3\cdot 1}{n\cdot(n-2)\cdot\cdots\cdot 4\cdot 2}\cdot\frac{\pi}{2}, & \text{if } n \text{ is even.} \\ \frac{(n-1)\cdot(n-3)\cdot\cdots\cdot 4\cdot 2}{n\cdot(n-2)\cdot\cdots\cdot 5\cdot 3}, & \text{if } n \text{ is odd.} \end{cases}$$

22. $\int_0^{\pi} \sin^n x\,\mathrm{d}\,x = 2\int_0^{\frac{\pi}{2}} \sin^n x\,\mathrm{d}\,x$.

Proof.

$$\int_0^{\pi} \sin^n x\,\mathrm{d}\,x = \int_0^{\pi/2} \sin^n x\,\mathrm{d}\,x + \int_{\pi/2}^{\pi} \sin^n x\,\mathrm{d}\,x$$

Letting $x = \pi - u$ in the second term of the RHS yields

$$\int_{\pi/2}^{\pi} \sin^n x\,\mathrm{d}\,x = \int_{\pi/2}^0 \sin^n(\pi - u)\,\mathrm{d}(\pi - u) = -\int_{\pi/2}^0 \sin^n u\,\mathrm{d}\,u = \int_0^{\pi/2} \sin^n u\,\mathrm{d}\,u.$$

Thus,

$$\int_0^{\pi} \sin^n x\,\mathrm{d}\,x = \int_0^{\pi/2} \sin^n x\,\mathrm{d}\,x + \int_{\pi/2}^{\pi} \sin^n u\,\mathrm{d}\,u = 2\int_0^{\pi/2} \sin^n x\,\mathrm{d}\,x.$$

□

8.2 Calculating gradients via Frobenius inner product

Given two complex number-valued $n \times m$ matrices $\mathbf{A}$ and $\mathbf{B}$, the Frobenius inner product is defined as[1]

$$\langle \mathbf{A}, \mathbf{B} \rangle_{\mathrm{F}} = \sum_{i,j} \overline{A_{ij}} B_{ij} = \mathrm{Tr}(\overline{\mathbf{A}^T}\mathbf{B}) \equiv \mathrm{Tr}(\mathbf{A}^\dagger \mathbf{B}) \tag{1}$$

where the overline denotes the complex conjugate, and † denotes the Hermitian conjugate. Sometimes we use the colon sign to denote the operation of Frobenius inner product. Following the definition, we have

$$\begin{aligned} A : BC &= BC : A \\ &= AC^T : B \\ &= B^T A : C \\ &= A^T : (BC)^T \\ &= \mathrm{Tr}(A^T BC) \end{aligned}$$

This operation is powerful in calculating gradients[2]. Here we take a good example from Section 6.5.6 of the well-known deep learning book[3], we use a matrix multiplication operation to create a variable $\mathbf{C} = \mathbf{AB}$. Suppose that the gradient of a scalar z with respect to $\mathbf{C}$ is given by $\mathbf{G}$. We can calculate the gradient of z w.r.t $\mathbf{A}$ as follows.

$$\begin{aligned} \mathrm{d}z &= \mathbf{G} : \mathrm{d}\mathbf{C} \\ &= \mathbf{G} : \mathrm{d}(\mathbf{AB}) \\ &= \mathbf{G} : (\mathrm{d}\mathbf{A}\mathbf{B} + \mathbf{A}\mathrm{d}\mathbf{B}) \\ &= \mathbf{G} : \mathrm{d}\mathbf{A}\mathbf{B} + \mathbf{G} : \mathbf{A}\mathrm{d}\mathbf{B} \\ &= \mathbf{G}\mathbf{B}^T : \mathrm{d}\mathbf{A} + \mathbf{G} : \mathbf{A} \cdot \mathbf{0} \\ &= \mathbf{G}\mathbf{B}^T : \mathrm{d}\mathbf{A} \end{aligned}$$

which shows that

$$\frac{\partial z}{\partial \mathbf{A}} = \mathbf{G}\mathbf{B}^T.$$

The intuition behind this is that if you want to have the total variation on z, you must multiply each small variation of $\mathbf{C}$ (i.e., $[\mathrm{d}\,C]_{ij}$) by the gradient in that direction (i.e. G_{ij}), and add them all up: $\sum_{i,j} G_{i,j}\,\mathrm{d}[C]_{ij}$ and that is the Frobenius inner product $\mathbf{G} : \mathrm{d}\mathbf{C}$, which is written as $\mathrm{Tr}(\mathbf{G}^T\,\mathrm{d}\,\mathbf{C})$.[4]

Another example[5] is to calculate the gradient of $\mathbf{x}^T(\mathbf{W}^2)^T\mathbf{W}^2\mathbf{x}$ w.r.t. $\mathbf{W}$. Let $A = W^2$, then it reduces to

$$\begin{aligned} \frac{\partial \mathbf{x}^T\mathbf{A}^T\mathbf{A}\mathbf{x}}{\partial \mathbf{A}} &= \mathbf{A}(\mathbf{x}\mathbf{x}^T + \mathbf{x}\mathbf{x}^T) \\ &= 2\mathbf{A}\mathbf{x}\mathbf{x}^T \end{aligned}$$

which follows from the formula (77) in Matrix Cookbook, specifically,

$$\frac{\partial \mathbf{b}^T\mathbf{X}^T\mathbf{X}\mathbf{c}}{\partial \mathbf{X}} = \mathbf{X}(\mathbf{b}\mathbf{c}^T + \mathbf{c}\mathbf{b}^T).$$

[1] https://handwiki.org/wiki/Frobenius_inner_product

[2] https://math.stackexchange.com/questions/1846339/why-does-the-gradient-of-matrix-product-ab-w-r-t-a-equal-bt

[3] https://github.com/janishar/mit-deep-learning-book-pdf

[4] https://math.stackexchange.com/questions/1846339/why-does-the-gradient-of-matrix-product-ab-w-r-t-a-equal-bt#comment5799146_1846803

[5] https://math.stackexchange.com/questions/4529245/how-to-calculate-the-gradient-of-mathbfxt-mathbfw2-mathbfw2t-mathb

Let $z = \mathbf{x}^T(\mathbf{W}^2)^T\mathbf{W}^2\mathbf{x}$, then we have

$$\begin{aligned}
\mathrm{d}z &= 2\mathbf{A}\mathbf{x}\mathbf{x}^T : \mathrm{d}\mathbf{A} \\
&= 2\mathbf{A}\mathbf{x}\mathbf{x}^T : \mathrm{d}\mathbf{W}^2 \\
&= 2\mathbf{A}\mathbf{x}\mathbf{x}^T : (\mathrm{d}\mathbf{W}\mathbf{W} + \mathbf{W}\mathrm{d}\mathbf{W}) \\
&= 2\mathbf{A}\mathbf{x}\mathbf{x}^T : \mathrm{d}\mathbf{W}\mathbf{W} + 2\mathbf{A}\mathbf{x}\mathbf{x}^T : \mathbf{W}\mathrm{d}\mathbf{W} \\
&= 2\mathbf{A}\mathbf{x}\mathbf{x}^T\mathbf{W}^T : \mathrm{d}\mathbf{W} + 2\mathbf{W}^T\mathbf{A}\mathbf{x}\mathbf{x}^T : \mathrm{d}\mathbf{W} \\
&= 2(\mathbf{A}\mathbf{x}\mathbf{x}^T\mathbf{W}^T + \mathbf{W}^T\mathbf{A}\mathbf{x}\mathbf{x}^T) : \mathrm{d}\mathbf{W}
\end{aligned}$$

which gives the solution

$$\frac{\partial \mathbf{x}^T(\mathbf{W}^2)^T\mathbf{W}^2\mathbf{x}}{\partial \mathbf{W}} = 2(\mathbf{A}\mathbf{x}\mathbf{x}^T\mathbf{W}^T + \mathbf{W}^T\mathbf{A}\mathbf{x}\mathbf{x}^T).$$

I tested this result with a 2×2 $\mathbf{W}$ using the auto-differentiation tool of PyTorch which gives an identical result. This implies that the above derivations are correct.

9 Projection operator

1. Projection on a hyperplane $C = \{x | a^T x = b, a \neq 0\}$

$$P_C(x) = x + \frac{b - a^T x}{\|a\|_2^2} a$$

2. Projection on an affine set $C = \{x | A^T x = b, A \in \mathbb{R}^{p \times n}, \mathrm{rank}(A) = p\}$

$$P_C(x) = x + A^T(AA^T)^{-1}(b - Ax)$$

   If $p \ll n$ or $AA^T = I$, then $P_C(x)$ is cheap.

3. Projection on a nonnegative orthant $C = \mathbb{R}^n_+$

$$P_C(x) = (x)^+ \quad \Leftrightarrow \quad [(x)^+]_i = \max\{x_i, 0\}$$

4. Projection onto a halfspace $C = \{x | a^T x \leq b, a \neq 0\}$

$$P_C(x) = \begin{cases} x + \frac{b - a^T x}{\|a\|_2^2} a, & \text{if } a^T x > b \\ x, & \text{otherwise.} \end{cases}$$

5. Projection onto a rectangular set $C = \{x \mid a \leq x \leq b\}$

6. Projection onto an ℓ_2 norm ball $C = \{x \mid \|x\|_2 \leq 1\}$

7. Projection onto an ℓ_1 norm ball $C = \{x \mid \|x\|_1 \leq 1\}$

8. Projection onto a second-order cone $C = \{(x, t) \mid \|x\|_2 \leq t, x \in \mathbb{R}^n, t \in \mathbb{R}_+\}$

   Proof. Given a point $(y, s) \in \mathbb{R}^n \times \mathbb{R}$, we want to find a point $(x, t) \in \mathbb{R}^n \times \mathbb{R}$ satisfying $\|x\|_2 \leq t$ and ensure that the distance between (y, s) and (x, t) is minimal. This is exactly what a projection

onto a second-order cone means. We can formulate this projection problem as an optimization problem as follows,

$$\min_x \ \frac{1}{2}\|x-y\|_2^2+\frac{1}{2}(t-s)^2, \ \text{subject to} \ \frac{1}{2}\|x\|_2^2\leq \frac{1}{2}t^2$$

The Lagrangian is

$$L(x,t,\lambda)=\frac{1}{2}\|x-y\|_2^2+\frac{1}{2}(t-s)^2+\lambda(\frac{1}{2}\|x\|_2^2-\frac{1}{2}t^2)$$

According to the KKT optimality conditions, we have

$$\frac{1}{2}\|x\|_2^2\leq \frac{1}{2}t^2 \Longleftrightarrow \|x\|_2^2\leq t^2 \qquad \text{(primal feasibility)} \quad (2)$$

$$\lambda \geq 0 \qquad \text{(dual feasibility)} \quad (3)$$

$$\lambda(\frac{1}{2}\|x\|_2^2-\frac{1}{2}t^2)=0 \qquad \text{(complementary slackness)} \quad (4)$$

$$\frac{\partial L(x,t,\lambda)}{\partial x}=x-y+\lambda x=0 \Longleftrightarrow (1+\lambda)x=y \qquad \text{(first-order optimality condition)} \quad (5)$$

$$\frac{\partial L(x,t,\lambda)}{\partial t}=t-s-\lambda t=0 \Longleftrightarrow (1-\lambda)t=s \qquad \text{(first-order optimality condition)} \quad (6)$$

In terms of the complementary slackness, we consider two cases separately.

- Case 1: If $\lambda=0$, we have $x=y$ and $0<t=s$ by the first-order optimality conditions. Combining with the primal feasibility, we get $\|y\|_2\leq s$.
- Case 2: If $\lambda>0$, by complementary slackness, we have $\|x\|_2= t>0$. Note that the case of $t=0$ is trivial in which $(\mathbf{0},0)$ is the only solution. By the first-order optimality conditions, we get

$$(1+\lambda)\|x\|_2=\|y\|_2 \Longrightarrow 1+\lambda=\frac{\|y\|_2}{\|x\|_2} \quad (7)$$

$$1-\lambda=\frac{s}{t} \quad (8)$$

(8)+(7),

$$2=\frac{\|y\|_2}{\|x\|_2}+\frac{s}{t}=\frac{\|y\|_2+s}{t} \Longrightarrow t=\frac{\|y\|_2+s}{2} \ (\|y\|_2>-s) \quad (9)$$

Since $\|x\|_2=t$ and (5), we have

$$x=\frac{\|y\|_2+s}{2}\cdot\frac{y}{\|y\|_2}$$

Thus,

$$(x,t)=\frac{\|y\|_2+s}{2}(\frac{y}{\|y\|_2},1) \quad (10)$$

From (8), we get

$$\begin{cases} 0<\lambda<1, & \text{if } s>0 \\ \lambda=1, & \text{if } s=0 \\ \lambda>1, & \text{if } s<0. \end{cases}$$

Now we consider the three subcases one by one. Combining (8) and (7),

$$1+\lambda=\frac{\|y\|_2}{t} \Longrightarrow 1+\lambda=\frac{\|y\|_2}{\frac{s}{1-\lambda}} \Longrightarrow \|y\|_2=\frac{1+\lambda}{1-\lambda}\cdot s=(\frac{2}{1-\lambda}-1)s=(\frac{2}{\lambda-1}+1)(-s)$$

Thus, when $s > 0$ and $0 < \lambda < 1$,

$$\|y\|_2 = (\frac{2}{1-\lambda} - 1)s \Longrightarrow \|y\|_2 > s.$$

If $s = 0$ and $\lambda = 1$, according to (5), $x = \frac{y}{2}$ which is consistent with (10).
When $s < 0$ and $\lambda > 1$,

$$\|y\|_2 = (\frac{2}{\lambda-1} + 1)(-s) \Longrightarrow \|y\|_2 > -s$$

If $\|y\|_2 \leq -s$, then it implies $\|y\|_2 + s \leq 0$ and $s \leq 0$ of which $s = 0$ has been analyzed. When $\|y\|_2 + s = 0$, by (9), $t = 0$, then $(\mathbf{0}, 0)$ is the projection. If $\|y\|_2 + s < 0$ in which $s < 0$, this is beyond the scope of the above two cases. Specifically, the first case requires $s > 0$ and the second case requires $\|y\|_2 + s > 0$. Now we the original formulation

$$\begin{aligned}
\min_{x,t\geq 0} \frac{1}{2}\|x-y\|_2^2 + \frac{1}{2}(t-s)^2 &= \min_{x\in\mathbb{R}^n, t\geq 0} \frac{1}{2}\|x\|_2^2 - x^T y + \frac{1}{2}\|y\|_2^2 + \frac{1}{2}t^2 - ts + \frac{1}{2}s^2 \\
&\geq \min_{x\in\mathbb{R}^n, t\geq 0} \frac{1}{2}\|x\|_2^2 - \|x\|_2\|y\|_2 + \frac{1}{2}\|y\|_2^2 + \frac{1}{2}t^2 - ts + \frac{1}{2}s^2 \\
&\geq \min_{x\in\mathbb{R}^n, t\geq 0} \frac{1}{2}\|x\|_2^2 - t\|y\|_2 + \frac{1}{2}\|y\|_2^2 + \frac{1}{2}t^2 - ts + \frac{1}{2}s^2 \\
&\geq \min_{x\in\mathbb{R}^n, t\geq 0} \frac{1}{2}\|x\|_2^2 + \frac{1}{2}t^2 - t(\|y\|_2 + s) + \frac{1}{2}\|y\|_2^2 + \frac{1}{2}s^2 \\
&\geq \frac{1}{2}\|y\|_2^2 + \frac{1}{2}s^2 \quad (\because \|y\|_2 + s < 0, t > 0)
\end{aligned}$$

where in the second line we used Cauchy-Schwarz inequality and the second inequality follows from the constraint $\|x\|_2 \leq t$. In the second last line only the first three terms contain x and t. Since $t > 0$ and $\|y\|_2 + s < 0$, the lower bound $(\frac{1}{2}\|y\|_2^2 + \frac{1}{2}s^2)$ can be achieved if and only if $(x, t) = (\mathbf{0}, 0)$.
In summary,

$$(x,t) = \begin{cases} (y, s), & \text{if } \|y\|_2 \leq s \\ \frac{\|y\|_2 + s}{2}(\frac{y}{\|y\|_2}, 1), & \text{if } \|y\|_2 > |s| \\ (\mathbf{0}, 0), & \text{if } \|y\|_2 \leq -s. \end{cases}$$

□

9. **Projection onto a positive semidefinite cone $C = \mathbf{S}_+^n$**

10 Geometry

10.1 Right Triangle Altitude Theorem

Part a: The measure of the altitude drawn from the vertex of the right angle of a right triangle to its hypotenuse is the geometric mean between the measures of the two segments of the hypotenuse. To put it in math,

$$AD^2 = BD \cdot CD$$

Proof. In Fig. 1, we have

$$AD = BD \cdot \tan \angle B$$

$$AD = CD \cdot \tan \angle C$$

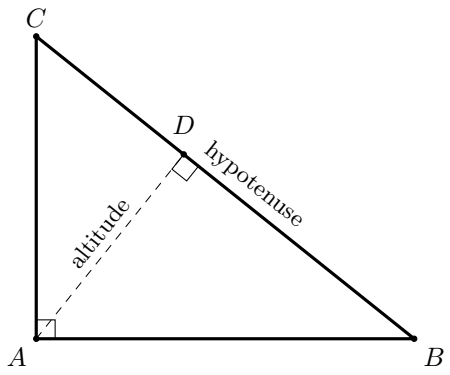

Figure 1: A right triangle

$$\tan \angle C = \cot \angle B$$

$$\begin{aligned} AD^2 &= BD \cdot \tan \angle B \cdot CD \cdot \tan \angle C \\ &= BD \cdot \tan \angle B \cdot CD \cdot \cot \angle B \\ &= BD \cdot CD \cdot \underbrace{\tan \angle B \cdot \cot \angle B}_{=1} \\ &= BD \cdot CD \end{aligned}$$

□

Part b: If the altitude is drawn to the hypotenuse of a right triangle, each leg of the right triangle is the geometric mean of the hypotenuse and the segment of the hypotenuse adjacent to the leg.

$$AC^2 = BC \cdot CD, \quad AD^2 = BC \cdot BD$$

Proof.

$$AC = BC \cdot \cos \angle C$$

$$AC = \frac{CD}{\cos \angle C}$$

$$AC^2 = AC \cdot AC = BC \cdot \cos \angle C \cdot \frac{CD}{\cos \angle C} = BC \cdot CD$$

$AD^2 = BC \cdot BD$ can be shown in the same way. □

10.2 A class of functions generalized from an inequality

Recall that

$$ax + \frac{b}{x} \geq 2\sqrt{ab}, \quad a, b, x > 0$$

$$ax + \frac{b}{x} \leq -2\sqrt{ab}, \quad a, b > 0 \text{ and } x < 0$$

The following functions can be induced by these two inequalities naturally

$$f(x) = ax^n + \frac{b}{x^n},$$

where $a, b > 0$ and $n \in \mathbf{Z}$. Since $x^n = \frac{1}{x^{-n}}$, we only consider the case when $n \in \mathbf{N}_+$. When n is odd (even, resp.), $f(x)$ is an odd (even, resp.) function. Hence, we only need to talk about the case of $x > 0$ for its monotonicity.

Specifically, when n is odd, f is decreasing on $(0, \sqrt[2n]{ab}]$ and $[\sqrt[2n]{ab}, 0)$, and increasing on $(-\infty, -\sqrt[2n]{ab}]$ and $[\sqrt[2n]{ab}, +\infty)$.

When n is even, f is decreasing on $(0, \sqrt[2n]{ab}]$ and $(-\infty, \sqrt[2n]{ab}]$, and increasing on $[-\sqrt[2n]{ab}, 0)$ and $[\sqrt[2n]{ab}, +\infty)$.

We give an example to make use of the above results to analyze the monotonicity of the following function,

$$\begin{aligned}
F(x) &= (x^2 + \frac{1}{x})^n + (x + \frac{1}{x^2})^n \\
&= C_n^0(x^2)^n + C_n^1(x^2)^{n-1}\frac{1}{x} + \cdots + C_n^{n-1}x^2(\frac{1}{x})^{n-1} \\
&\quad + C_n^n(\frac{1}{x})^n + C_n^0x^n + C_n^1x^{n-1}\frac{1}{x^2} + \cdots + C_n^{n-1}x(\frac{1}{x^2})^{n-1} + C_n^n(\frac{1}{x^2})^n \\
&= C_n^0x^{2n} + C_n^1x^{2n-3} + \cdots + C_n^{n-1}\frac{1}{x^{n-3}} + C_n^n\frac{1}{x^n} + C_n^0x^n + C_n^1x^{n-3} + \cdots + C_n^{n-1}\frac{1}{x^{2n-3}} + C_n^n\frac{1}{x^{2n}} \\
&= C_n^0(x^{2n} + \frac{1}{x^{2n}}) + C_n^1(x^{2n-3} + \frac{1}{x^{2n-3}}) + \cdots + C_n^{n-1}(x^{n-3} + \frac{1}{x^{n-3}}) + C_n^n(x^n + \frac{1}{x^n})
\end{aligned}$$

Thus, $F(x)$ is decreasing on $(0, 1]$ and increasing on $[1, +\infty)$.

11 The introduction of natural logarithm e

Show that the sequence $\{(1 + \frac{1}{n})^n\}$ is increasing and the sequence $\{(1 + \frac{1}{n})^{n+1}\}$ is decreasing, and they converge to the same limit.

Proof. Hint: AM-GM inequality and the fact that monotone bounded sequences converge. Refer to Example 2.4.6 in Jixiu Chen et al. Mathematical Analysis, third edition. This will be done later. □

We use e to denote this limit, namely,

$$\lim_{n\to\infty}(1 + \frac{1}{n})^n = \lim_{n\to\infty}(1 + \frac{1}{n})^{n+1} = e = 2.718281828459\cdots.$$

e is an irrational number. The logarithm in the base of e is called natural logarithm, usually denoted as $\ln x = \log_e x$.

12 Factorizing fractions

Let's see the following factorization,

$$\begin{aligned}
\frac{1}{n(n+1)(n+2)} &= \frac{1}{2}\frac{(n+2)-n}{n(n+1)(n+2)} \\
&= \frac{1}{2}\left[\frac{1}{n(n+1)} - \frac{1}{(n+1)(n+2)}\right] \\
&= \frac{1}{2}\left[\frac{1}{n} - \frac{1}{n+1} - \frac{1}{n+1} + \frac{1}{n+2}\right] \\
&= \frac{1}{2}\left[\frac{1}{n} - \frac{2}{n+1} + \frac{1}{n+2}\right]
\end{aligned}$$

This pattern can be generalized to

$$\begin{aligned}
\frac{1}{n(n+1)\cdots(n+m)} &= \frac{1}{m}\frac{(n+m)-n}{n(n+1)\cdots(n+m)} \\
&= \frac{1}{m}\left[\frac{1}{n(n+1)\cdots(n+m-1)} - \frac{1}{(n+1)(n+2)\cdots(n+m)}\right] \\
&= \frac{1}{m(m-1)}\left[\frac{1}{n(n+1)\cdots(n+m-2)} - \frac{2}{(n+1)(n+2)\cdots(n+m-1)}\right. \\
&\qquad\qquad \left. + \frac{1}{(n+2)(n+3)\cdots(n+m)}\right] \\
&= \frac{1}{m(m-1)(m-2)}\left[\frac{1}{n\cdots(n+m-3)} - \frac{3}{(n+1)\cdots(n+m-2)}\right. \\
&\qquad\qquad \left. + \frac{3}{(n+2)\cdots(n+m-1)} - \frac{1}{(n+3)\cdots(n+m)}\right] \\
&= \frac{1}{m\cdots(m-3)}\left[\frac{1}{n\cdots(n+m-4)} - \frac{4}{(n+1)\cdots(n+m-3)}\right. \\
&\quad \left. + \frac{6}{(n+2)\cdots(n+m-2)} - \frac{4}{(n+3)\cdots(n+m-1)} + \frac{1}{(n+4)\cdots(n+m)}\right]
\end{aligned}$$

13 A basic inequality

For any real number a, b, we always have $a(a-b) \geq b(a-b)$; if $b > 0$, we have $\frac{a^2}{b} \geq 2a - b$; if $ab^2 > 0$, $\frac{a}{b^2} \geq \frac{2}{b} - \frac{1}{a}$. These can be obtained by easy derivations.

$$(a-b)^2 \geq 0 \Leftrightarrow a(a-b) - b(a-b) \geq 0 \Leftrightarrow a(a-b) \geq b(a-b)$$

$$a(a-b) \geq b(a-b) \overset{b>0}{\Longleftrightarrow} \frac{a}{b}(a-b) \geq a-b \Longleftrightarrow \frac{a^2}{b} - a \geq a - b \Longleftrightarrow \frac{a^2}{b} \geq 2a - b$$

$$a(a-b) \geq b(a-b) \overset{ab^2>0}{\Longleftrightarrow} \frac{a-b}{b^2} \geq \frac{a-b}{ab} \Longleftrightarrow \frac{a}{b^2} - \frac{1}{b} \geq \frac{1}{b} - \frac{1}{a} \Longleftrightarrow \frac{a}{b^2} \geq \frac{2}{b} - \frac{1}{a}$$

14 Euler formula

Before we get to the famous Euler formula, let's introduce a problem. To calculate

$$I = 1 + \cos\theta + \cos 2\theta + \cdots + \cos n\theta$$

$$\begin{aligned}
I &= \frac{2\sin\frac{\theta}{2}(1 + \cos\theta + \cos 2\theta + \cdots + \cos n\theta)}{2\sin\frac{\theta}{2}} \\
&= \frac{2\sin\frac{\theta}{2} + 2\sin\frac{\theta}{2}\cos\theta + 2\sin\frac{\theta}{2}\cos 2\theta + \cdots + 2\sin\frac{\theta}{2}\cos n\theta}{2\sin\frac{\theta}{2}} \\
&= \frac{2\sin\frac{\theta}{2} + \sin(\frac{\theta}{2} - \theta) + \sin(\frac{\theta}{2} + \theta) + \sin(\frac{\theta}{2} - 2\theta) + \sin(\frac{\theta}{2} + 2\theta) + \cdots + \sin(\frac{\theta}{2} - n\theta) + \sin(\frac{\theta}{2} + n\theta)}{2\sin\frac{\theta}{2}} \\
&= \frac{2\sin\frac{\theta}{2} + \sin(\frac{-\theta}{2}) + \sin(\frac{3\theta}{2}) + \sin(\frac{-3\theta}{2}) + \sin(\frac{5\theta}{2}) + \cdots + \sin(\frac{-(n-1)\theta}{2}) + \sin(\frac{(2n+1)\theta}{2})}{2\sin\frac{\theta}{2}} \\
&= \frac{\sin\frac{\theta}{2} + \sin\frac{(2n+1)\theta}{2}}{2\sin\frac{\theta}{2}} = \frac{1}{2} + \frac{\sin\frac{(2n+1)\theta}{2}}{2\sin\frac{\theta}{2}}
\end{aligned}$$

The Euler formula is

$$e^{i\theta} = \cos\theta + i\sin\theta, \quad e^{-i\theta} = \cos\theta - i\sin\theta$$
$$e^{in\theta} = \cos n\theta + i\sin n\theta, \quad e^{-in\theta} = \cos n\theta - i\sin n\theta$$
$$\cos\theta = \frac{e^{i\theta}+e^{-i\theta}}{2}, \quad \sin\theta = \frac{e^{i\theta}-e^{-i\theta}}{2i}$$

$$\begin{aligned}
\sum_{j=1}^{n}\cos j\theta &= \mathrm{Re}(\sum_{j=1}^{n}(e^{i\theta})^j) \\
&= \mathrm{Re}(\frac{e^{i\theta}(1-e^{in\theta})}{1-e^{i\theta}}) = \mathrm{Re}(\frac{e^{-\frac{i\theta}{2}}e^{i\theta}(1-e^{in\theta})}{e^{-\frac{i\theta}{2}}(1-e^{i\theta})}) \\
&= \mathrm{Re}(\frac{e^{\frac{i\theta}{2}}(1-e^{in\theta})}{-2i\sin\frac{\theta}{2}}) = \mathrm{Re}(\frac{ie^{\frac{i\theta}{2}}-ie^{\frac{i(2n+1)\theta}{2}}}{2\sin\frac{\theta}{2}}) \\
&= \mathrm{Re}(\frac{i(\cos\frac{\theta}{2}+i\sin\frac{\theta}{2})-i(\cos\frac{(2n+1)\theta}{2}+i\sin\frac{(2n+1)\theta}{2})}{2\sin\frac{\theta}{2}}) \\
&= \mathrm{Re}(\frac{\sin\frac{(2n+1)\theta}{2}-\sin\frac{\theta}{2}+i(\cos\frac{\theta}{2}-\cos\frac{(2n+1)\theta}{2})}{2\sin\frac{\theta}{2}}) \\
&= \mathrm{Re}\left(-\frac{1}{2}+\frac{\sin\frac{(2n+1)\theta}{2}}{2\sin\frac{\theta}{2}}+i(\frac{\cos\frac{\theta}{2}-\cos\frac{(2n+1)\theta}{2}}{2\sin\frac{\theta}{2}})\right)
\end{aligned}$$

Thus,

$$1+\cos\theta+\cos 2\theta+\cdots+\cos n\theta = 1-\frac{1}{2}+\frac{\sin\frac{(2n+1)\theta}{2}}{2\sin\frac{\theta}{2}} = \frac{1}{2}+\frac{\sin\frac{(2n+1)\theta}{2}}{2\sin\frac{\theta}{2}}$$
$$\sin\theta+\sin 2\theta+\cdots+\sin n\theta = \frac{\tan\frac{\theta}{2}}{2}-\frac{\cos\frac{(2n+1)\theta}{2}}{2\sin\frac{\theta}{2}}$$

15 Some special functions

15.1 Gamma function

The gamma function is defined by

$$\Gamma(x) \equiv \int_0^\infty u^{x-1}e^{-u}\,\mathrm{d}\,u. \tag{11}$$

Using integration by parts, we can get the relation

$$\Gamma(x+1) = x\Gamma(x).$$

It is easy to get $\Gamma(1)=1$ and hence that

$$\Gamma(x+1) = x!\,.$$

Additionally, for any non-negative integer n, we have:

$$\Gamma(\frac{1}{2}+n) = \frac{(2n)!}{4^n n!}\sqrt{\pi}.$$

Particularly, $\Gamma(1/2)=\sqrt{\pi}$.

15.2 Digamma function (a.k.a. psi function)

The definition of the digamma function is

$$\psi(a) \equiv \frac{\mathrm{d}}{\mathrm{d}\,a}\Gamma(a) = \frac{\Gamma'(a)}{\Gamma(a)},$$

which is also known as the psi function.

16 Probability distributions

16.1 Beta distribution

The beta distribution is defined by

$$\mathrm{Beta}(\mu|a,b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\mu^{a-1}(1-\mu)^{b-1} \tag{12}$$

where $\Gamma(x)$ is the gamma function defined by (11), and the coefficient in (12) ensures that the beta distribution is normalized, so that

$$\int_0^1 \mathrm{Beta}(\mu|a,b)\,\mathrm{d}\,\mu = 1 \tag{13}$$

The mean and variance of the beta distribution are given by

$$\mathbb{E}[\mu] = \frac{a}{a+b} \tag{14}$$

$$\mathrm{var}[\mu] = \frac{ab}{(a+b)^2(a+b+1)} \tag{15}$$

The parameters a and b are often called hyperparameters because they control the distribution of the parameter μ.

16.2 Gaussian distribution

This part is largely taken from Section 1.2.4 of PRML book.

16.2.1 Univariate Gaussian

For the case of a single real-valued variable x, the Gaussian distribution is defined by

$$\mathcal{N}(x|\mu,\sigma^2) = \frac{1}{(2\pi\sigma^2)^{1/2}}\exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\} \tag{16}$$

which is governed by two parameters: μ, called the **mean**, and σ^2, called the **variance**. The square root of the variance, given by *sigma*, is called the **standard deviation**, and the reciprocal of the variance, written as $\beta = 1/\sigma^2$, is called the **precision**.

From the form of (16) we see that the Gaussian distribution satisfies

$$\mathcal{N}(x|\mu,\sigma^2) > 0. \tag{17}$$

Also it is straightforward to show that the Gaussian is normalized, so that

$$\int_{-\infty}^{+\infty} \mathcal{N}(x|\mu,\sigma^2)\,\mathrm{d}\,x = 1. \tag{18}$$

Thus (16) satisfies the two requirements for a valid probability density.

16.2.2 Proof: the Gaussian is normalized.

This proof is from Exercise 1.7 of PRML book. Consider this integral

$$I = \int_{-\infty}^{+\infty} \exp\left(-\frac{1}{2\sigma^2}x^2\right) \mathrm{d}\, x \tag{19}$$

which we can evaluate by first writing its square in the form

$$\begin{aligned}
I^2 =& \int_{-\infty}^{+\infty} \exp\left(-\frac{1}{2\sigma^2}x^2\right) \mathrm{d}\, x \cdot \int_{-\infty}^{+\infty} \exp\left(-\frac{1}{2\sigma^2}y^2\right) \mathrm{d}\, y \\
=& \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} \exp\left(-\frac{1}{2\sigma^2}x^2 - \frac{1}{2\sigma^2}y^2\right) \mathrm{d}\, x \,\mathrm{d}\, y \\
=& \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} \exp\left(-\frac{1}{2\sigma^2}(x^2+y^2)\right) \mathrm{d}\, x \,\mathrm{d}\, y \\
=& \int_{0}^{2\pi} \int_{0}^{+\infty} \exp\left(-\frac{1}{2\sigma^2}r^2\right) r \,\mathrm{d}\, r \,\mathrm{d}\, \theta \\
=& \int_{0}^{2\pi} \mathrm{d}\, \theta \int_{0}^{+\infty} \exp\left(-\frac{1}{2\sigma^2}r^2\right) r \,\mathrm{d}\, r \\
=& 2\pi \cdot \left(-\sigma^2 \exp\left(-\frac{1}{2\sigma^2}r^2\right)\Big|_0^{+\infty}\right) = 2\pi\sigma^2
\end{aligned}$$

where we make the transformation from Cartesian coordinates (x, y) to polar coordinates (r, θ) and then substitute $x^2 + y^2 = r^2$ in the third last line. Finally, we get

$$I = \int_{-\infty}^{+\infty} \exp\left(-\frac{1}{2\sigma^2}x^2\right) \mathrm{d}\, x = \sqrt{2\pi\sigma^2} \tag{20}$$

Thus,

$$\int_{-\infty}^{+\infty} \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right) \mathrm{d}\, x = \frac{1}{(2\pi\sigma^2)^{1/2}} \cdot \sqrt{2\pi\sigma^2} = 1 \tag{21}$$

16.3 Multivariate Gaussian

The multivariate gaussian distribution is defined by

$$\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{D/2}} \frac{1}{|\Sigma|^{1/2}} \exp\left\{\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right\} \tag{22}$$

where the D-dimensional vector $\boldsymbol{\mu}$ is called the mean, the $D \times D$ matrix $\boldsymbol{\Sigma}$ is called the covariance, and $|\Sigma|$ denotes the determinant of $\boldsymbol{\Sigma}$.

16.4 Gamma distribution

The gamma distribution is defined as

$$\mathrm{Gam}(\lambda|a, b) = \frac{1}{\Gamma(a)} b^a \lambda^{a-1} \exp(-b\lambda). \tag{23}$$

Here $\Gamma(a)$ is the gamma function that is defined by (11) and that ensures that (23) is correctly normalized.

The mean and variance of the gamma distribution are given by

$$\mathbb{E}[\lambda] = \frac{a}{b} \tag{24}$$

$$\text{var}[\lambda] = \frac{a}{b^2} \tag{25}$$

16.5 Student's t-distribution

16.5.1 The univariate case

Recall that the conjugate prior for the precision of a Gaussian is a gamma distribution. If we have a univariate Gaussian $\mathcal{N}(\mu, \tau)$ together with a Gamma prior $\text{Gam}(\tau|a, b)$ and we integrate out the precision, we obtain the marginal distribution of x in the form

$$\begin{aligned} p(x|\mu, a, b) &= \int_0^\infty \mathcal{N}(x|\mu, \tau^{-1})\text{Gam}(\tau|a, b)\,\mathrm{d}\,\tau \\ &= \int_0^\infty (\frac{\tau}{2\pi})^{1/2} \exp\left(-\frac{\tau}{2}(x-\mu)^2\right) \frac{1}{\Gamma(a)} b^a \tau^{a-1} \exp\left(-b\tau\right)\mathrm{d}\,\tau \\ &= \frac{1}{\Gamma(a)} b^a \int_0^\infty \tau^{a-1} \left(\frac{\tau}{2\pi}\right)^{1/2} \exp\left(-\tau\left(b + \frac{1}{2}(x-\mu)^2\right)\right)\mathrm{d}\,\tau \\ &= \frac{1}{\Gamma(a)} b^a \left(\frac{1}{2\pi}\right)^{1/2} \int_0^\infty \tau^{a-1/2} \exp\left(-\tau\left(b + \frac{1}{2}(x-\mu)^2\right)\right)\mathrm{d}\,\tau \end{aligned} \tag{26}$$

To fit the definition of the gamma function, namely (11), let $z = \tau\left(b + \frac{1}{2}(x-\mu)^2\right)$. Then we get

$$\mathrm{d}\,z = \left(b + \frac{1}{2}(x-\mu)^2\right)\mathrm{d}\,\tau \iff \mathrm{d}\,\tau = \left(b + \frac{1}{2}(x-\mu)^2\right)^{-1}\mathrm{d}\,z.$$

Thus,

$$\begin{aligned} p(x|\mu, a, b) &= \frac{b^a}{\Gamma(a)}\left(\frac{1}{2\pi}\right)^{1/2} \int_0^\infty \left(\left(b + \frac{1}{2}(x-\mu)^2\right)^{-1} z\right)^{a-1/2} \left(b + \frac{1}{2}(x-\mu)^2\right)^{-1} \exp\left(-z\right)\mathrm{d}\,z \\ &= \frac{b^a}{\Gamma(a)}\left(\frac{1}{2\pi}\right)^{1/2} \left(b + \frac{1}{2}(x-\mu)^2\right)^{-(a+1/2)} \int_0^\infty z^{a-1/2} \exp\left(-z\right)\mathrm{d}\,z \\ &= \frac{b^a}{\Gamma(a)}\left(\frac{1}{2\pi}\right)^{1/2} \left(b + \frac{1}{2}(x-\mu)^2\right)^{-(a+1/2)} \int_0^\infty z^{a+1/2-1} \exp\left(-z\right)\mathrm{d}\,z \\ &= \frac{b^a}{\Gamma(a)}\left(\frac{1}{2\pi}\right)^{1/2} \left(b + \frac{1}{2}(x-\mu)^2\right)^{-(a+1/2)} \Gamma(a + \frac{1}{2}) \end{aligned}$$

By convention we define $\nu = 2a$ and $\lambda = a/b$, in terms of which the distribution $p(x|\mu, a, b)$ takes the form

$$\begin{aligned} \text{St}(x|\mu, \lambda, \nu) &= \frac{\Gamma(\nu/2 + 1/2)}{\Gamma(\nu/2)} b^a \left(\frac{1}{2\pi}\right)^{1/2} \left(b(1 + \frac{1}{\nu}(x-\mu)^2)\right)^{-(a+1/2)} \\ &= \frac{\Gamma(\nu/2 + 1/2)}{\Gamma(\nu/2)} \left(\frac{1}{2\pi}\right)^{1/2} b^{-1/2} \left(1 + \frac{(x-\mu)^2}{\nu}\right)^{-\nu/2-1/2} \\ &= \frac{\Gamma(\nu/2 + 1/2)}{\Gamma(\nu/2)} \left(\frac{1}{2\pi}\right)^{1/2} (2\lambda/\nu)^{1/2} \left(1 + \frac{(x-\mu)^2}{\nu}\right)^{-\nu/2-1/2} \\ &= \frac{\Gamma(\nu/2 + 1/2)}{\Gamma(\nu/2)} \left(\frac{\lambda}{\pi\nu}\right)^{1/2} \left(1 + \frac{(x-\mu)^2}{\nu}\right)^{-\nu/2-1/2} \end{aligned} \tag{27}$$

If we substitute the alternative parameters $\nu = 2a, \lambda = a/b$, and $\eta = \tau b/a$, we see that the t-distribution can be written in the form

$$\begin{aligned}
\int_0^\infty \mathcal{N}(x|\mu, \tau^{-1})\text{Gam}(\tau|a,b)\,\mathrm{d}\,\tau &= \int_0^\infty \mathcal{N}(x|\mu,(a\eta/b)^{-1})\text{Gam}(a\eta/b|\nu/2,b)\,\mathrm{d}(a\eta/b) \\
&= \int_0^\infty \mathcal{N}(x|\mu,(\eta\lambda)^{-1})\text{Gam}(\eta\lambda|\nu/2,b)\frac{a}{b}\,\mathrm{d}\,\eta \\
&= \int_0^\infty \mathcal{N}(x|\mu,(\eta\lambda)^{-1})\frac{1}{\Gamma(\nu/2)}b^{\nu/2}(\eta\lambda)^{\nu/2-1}\exp\left(-b\eta\lambda\right)\frac{a}{b}\,\mathrm{d}\,\eta \\
&= \int_0^\infty \mathcal{N}(x|\mu,(\eta\lambda)^{-1})\frac{1}{\Gamma(\nu/2)}b^{\nu/2}(\eta\lambda)^{\nu/2-1}\exp\left(-a\eta\right)\lambda\,\mathrm{d}\,\eta \\
&= \int_0^\infty \mathcal{N}(x|\mu,(\eta\lambda)^{-1})\frac{1}{\Gamma(\nu/2)}(b\eta\lambda)^{\nu/2}\exp\left(-a\eta\right)\frac{1}{\eta}\,\mathrm{d}\,\eta \\
&= \int_0^\infty \mathcal{N}(x|\mu,(\eta\lambda)^{-1})\frac{1}{\Gamma(\nu/2)}(\nu/2)^{\nu/2}\eta^{\nu/2-1}\exp\left(-\frac{\nu}{2}\eta\right)\mathrm{d}\,\eta \\
\text{St}(x|\mu,\lambda,\nu) &= \int_0^\infty \mathcal{N}\left(x|\mu,(\eta\lambda)^{-1}\right)\text{Gam}(\eta|\nu/2,\nu/2)\,\mathrm{d}\,\eta \qquad (28)
\end{aligned}$$

16.5.2 The multivariate case

We can then generalize this to a multivariate Gaussian $\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Lambda}^{-1})$ to obtain the corresponding multivariate Student's t-distribution in the form

$$\text{St}(\mathbf{x}|\boldsymbol{\mu},\boldsymbol{\Lambda},\nu) = \int_0^\infty \mathcal{N}\left(\mathbf{x}|\boldsymbol{\mu},(\eta\Lambda)^{-1}\right)\text{Gam}(\eta|\nu/2,\nu/2)\,\mathrm{d}\,\eta \qquad (29)$$

Integrating out η yields

$$\text{St}(\mathbf{x}|\boldsymbol{\mu},\boldsymbol{\Lambda},\nu) = \frac{\Gamma(D/2+\nu/2)}{\Gamma(\nu/2)}\frac{|\Lambda|^{1/2}}{(\pi\nu)^{D/2}}\left[1+\frac{\Delta^2}{\nu}\right]^{-D/2-\nu/2} \qquad (30)$$

where D is the dimensionality of $\mathbf{x}$, and Δ^2 is the squared Mahalanobis distance defined by

$$\Delta^2 = (\mathbf{x}-\boldsymbol{\mu})^T\Lambda(\mathbf{x}-\boldsymbol{\mu}).$$

This is the multivariate form of Student's t-distribution and satisfies the following properties

$$\mathbb{E}[\mathbf{x}] = \boldsymbol{\mu}, \qquad \text{if} \quad \nu > 1 \qquad (31)$$

$$\text{cov}[\mathbf{x}] = \frac{\nu}{(\nu-2)}\boldsymbol{\Lambda}^{-1}, \qquad \text{if} \quad \nu > 2 \qquad (32)$$

$$\text{mode}[\mathbf{x}] = \boldsymbol{\mu} \qquad (33)$$

with corresponding results for the univariate case.

16.5.3 The proof for Student's t-distribution becoming a Gaussian when $\nu \to \infty$

As can be seen from (24) and (25), as $\nu \to +\infty$, the mean and variance of a gamma distribution will be 1 and 0, respectively. In that case, the gamma distribution degenerates to a Dirac delta function $\delta(\eta-1)$. Then the multivariate Student's t-distribution (29) will be in the form of

$$\begin{aligned}
\text{St}(\mathbf{x}|\boldsymbol{\mu},\boldsymbol{\Lambda},+\infty) &= \int_0^\infty \mathcal{N}\left(\mathbf{x}|\boldsymbol{\mu},(\eta\boldsymbol{\Lambda})^{-1}\right)\delta(\eta-1)\,\mathrm{d}\,\eta \\
&= \mathcal{N}(\mathbf{x}|\boldsymbol{\mu},\boldsymbol{\Lambda}^{-1})
\end{aligned} \qquad (34)$$

16.5.4 The proof for the mean, covariance, and mode of multivariate Student's t-distribution

To compute the mean of the multivariate Student's t-distribution, we make use of (29) as follows.

$$\begin{aligned}\mathbb{E}[\mathbf{x}] &= \int \mathbf{x} \cdot \mathrm{St}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Lambda}, \nu)\,\mathrm{d}\,\mathbf{x} \\ &= \int_0^\infty \left(\int \mathbf{x} \mathcal{N}\left(\mathbf{x}|\boldsymbol{\mu}, (\eta\boldsymbol{\Lambda})^{-1}\right) \mathrm{d}\,\mathbf{x} \right) \mathrm{Gam}(\eta|\nu/2, \nu/2)\,\mathrm{d}\,\eta \\ &= \boldsymbol{\mu} \cdot \int_0^\infty \mathrm{Gam}(\eta|\nu/2, \nu/2)\,\mathrm{d}\,\eta = \boldsymbol{\mu}.\end{aligned}$$

To calculate the variance of the multivariate Student's t-distribution, we follow the definition of variance and substitute the definition of gamma distribution into (29).

$$\begin{aligned}\mathrm{cov}[\mathbf{x}] &= \int (\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})^T \cdot \mathrm{St}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Lambda}, \nu)\,\mathrm{d}\,\mathbf{x} \\ &= \int (\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})^T \cdot \int_0^\infty \mathcal{N}\left(\mathbf{x}|\boldsymbol{\mu}, (\eta\boldsymbol{\Lambda})^{-1}\right) \mathrm{Gam}(\eta|\nu/2, \nu/2)\,\mathrm{d}\,\eta\,\mathrm{d}\,\mathbf{x} \\ &= \int_0^\infty \left(\int (\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})^T \mathcal{N}\left(\mathbf{x}|\boldsymbol{\mu}, (\eta\boldsymbol{\Lambda})^{-1}\right) \mathrm{d}\,\mathbf{x} \right) \mathrm{Gam}(\eta|\nu/2, \nu/2)\,\mathrm{d}\,\eta \\ &= (\eta\boldsymbol{\Lambda})^{-1} \int_0^\infty \mathrm{Gam}(\eta|\nu/2, \nu/2)\,\mathrm{d}\,\eta \\ &= \boldsymbol{\Lambda}^{-1} \int_0^\infty \frac{1}{\eta} \frac{1}{\Gamma(\nu/2)} (\nu/2)^{\nu/2} \eta^{\nu/2-1} \exp\left(-\nu/2 \cdot \eta\right) \mathrm{d}\,\eta \\ &= \boldsymbol{\Lambda}^{-1} \int_0^\infty \frac{1}{\Gamma(\nu/2-1+1)} (\nu/2)^{\nu/2-1+1} \eta^{\nu/2-1-1} \exp\left(-\nu/2 \cdot \eta\right) \mathrm{d}\,\eta \\ &= \boldsymbol{\Lambda}^{-1} \int_0^\infty \frac{\nu}{2} \cdot \frac{1}{\nu/2-1} \cdot \frac{1}{\Gamma(\nu/2-1)} (\nu/2)^{\nu/2-1} \eta^{\nu/2-1-1} \exp\left(-\nu/2 \cdot \eta\right) \mathrm{d}\,\eta \\ &= \frac{\nu}{\nu-2} \cdot \boldsymbol{\Lambda}^{-1} \int_0^\infty \mathrm{Gam}(\eta|\nu/2-1, \nu/2)\,\mathrm{d}\,\eta \\ &= \frac{\nu}{(\nu-2)} \boldsymbol{\Lambda}^{-1}, \quad \text{if } \nu > 2\end{aligned}$$

To get the mode of the multivariate Student's t-distribution, taking derivatives of (29) w.r.t. $\mathbf{x}$ and setting it equal to $\mathbf{0}$ gives $\mathbf{x} - \boldsymbol{\mu} = \mathbf{0}$, which shows $\mathrm{mode}[\mathbf{x}] = \mathbf{0}$.

16.6 Von Mises distribution

The von Mises distribution, or the circular normal, is defined as

$$p(\theta|\theta_0, m) = \frac{1}{2\pi I_0(m)} \exp\{m\cos(\theta - \theta_0)\} \tag{35}$$

Here the parameter θ_0 corresponds to the mean of the distribution, while m, which is known as the concentration parameter, is analogous to the precision for the Gaussian. The normalization coefficient is expressed in terms of $I_0(m)$, which is the zeroth-order modified Bessel function of the first kind and is defined by

$$I_0(m) = \frac{1}{2\pi} \int_0^{2\pi} \exp\{m\cos(\theta)\}\,\mathrm{d}\,\theta. \tag{36}$$

For large m, the distribution becomes approximately Gaussian.

17 The exponential family

The exponential family over $\mathbf{x}$, given parameters $\boldsymbol{\eta}$, is defined to be the set of distributions of the form

$$p(\mathbf{x}|\boldsymbol{\eta}) = h(\mathbf{x})g(\boldsymbol{\eta}) \exp\left\{\boldsymbol{\eta}^T \mathbf{u}(\mathbf{x})\right\} \tag{37}$$

where $\mathbf{x}$ may be scalar or vector, and may be discrete or continuous. Here $\boldsymbol{\eta}$ are called the **natural parameters** of the distribution, and $\mathbf{u}(\mathbf{x})$ is some function of $\mathbf{x}$. The function $g(\boldsymbol{\eta})$ can be interpreted as the coefficient that ensures that the distribution is normalized and therefore satisfies

$$g(\boldsymbol{\eta}) \int h(\mathbf{x}) \exp\left\{\boldsymbol{\eta}^T \mathbf{u}(\mathbf{x})\right\} \mathrm{d}\,\mathbf{x} = 1 \tag{38}$$

where the integration is replaced by summation if $\mathbf{x}$ is a discrete variable.

17.1 Bernoulli distribution

Consider the Bernoulli distribution

$$p(x|\mu) = \mathrm{Bern}(x|\mu) = \mu^x (1-\mu)^{1-x}. \tag{39}$$

Expressing the RHS as the exponential of the logarithm, we have

$$\begin{aligned} p(x|\mu) =& \exp\{x \ln \mu + (1-x)\ln(1-\mu)\} && (40)\\ =& \exp\left\{\ln(1-\mu) + x \ln \frac{\mu}{1-\mu}\right\} && (41)\\ =& (1-\mu)\exp\left\{\ln\left(\frac{\mu}{1-\mu}\right) x\right\}. && (42) \end{aligned}$$

Comparison with (37) allows us to identify

$$\eta = \ln\left(\frac{\mu}{1-\mu}\right) \tag{43}$$

which we can solve for μ to give

$$\begin{aligned} \exp(\eta) = \frac{\mu}{1-\mu} &\Leftrightarrow \exp(-\eta) = \frac{1-\mu}{\mu} = \frac{1}{\mu} - 1 \\ &\Leftrightarrow \frac{1}{\mu} = 1 + \exp(-\eta) \\ &\Leftrightarrow \mu = \frac{1}{1+\exp(-\eta)} = \sigma(\eta) \end{aligned}$$

where $\sigma(\eta)$ is called the **logistic sigmoid** function. Thus we can write the Bernoulli distribution in the form

$$\begin{aligned} p(x|\mu) =& \left(1 - \frac{1}{1+\exp(-\eta)}\right)\exp(\eta x) && (44)\\ =& \frac{\exp(-\eta)}{1+\exp(-\eta)}\exp(\eta x) && (45)\\ =& \frac{1}{1+\exp(\eta)}\exp(\eta x) && (46)\\ =& \sigma(-\eta)\exp(\eta x). && (47) \end{aligned}$$

Comparison with (37) shows that

$$u(x) = x \tag{48}$$
$$h(x) = 1 \tag{49}$$
$$g(\eta) = \sigma(-\eta). \tag{50}$$

17.2 Multinomial distribution

Consider the multinomial distribution that, for a single observation $\mathbf{x}$, takes the form

$$p(\mathbf{x}|\boldsymbol{\mu}) = \prod_{k=1}^{M} \mu_k^{x_k} = \exp\left\{\sum_{k=1}^{M} x_k \ln \mu_k\right\} \tag{51}$$

where $\mathbf{x} = (x_1, \cdots, x_M)^T$. Again, we can write this in the standard representation

$$p(\mathbf{x}|\boldsymbol{\eta}) = \exp(\boldsymbol{\eta}^T \mathbf{x})$$

where $\eta_k = \ln \mu_k$, and we have defined $\boldsymbol{\eta} = (\eta_1, \cdots, \eta_M)^T$. Again, comparing with (37) we have

$$\mathbf{u}(\mathbf{x}) = \mathbf{x} \tag{52}$$
$$h(\mathbf{x}) = 1 \tag{53}$$
$$g(\boldsymbol{\eta}) = 1. \tag{54}$$

Note that the parameters μ_k are subject to the constraint

$$\sum_{k=1}^{M} \mu_k = 1 \tag{55}$$

so that, given any $M-1$ of the parameters μ_k, the value of the remaining parameter is fixed. In some circumstances, it will be convenient to remove this constraint by expressing the distribution in terms of only $M-1$ parameters. The notes on this will be done in future.

17.3 Gaussian distribution

For the univariate Gaussian, we have

$$p(x|\mu, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\} \tag{56}$$
$$= \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left\{-\frac{1}{2\sigma^2}x^2 + \frac{\mu}{\sigma^2}x - \frac{1}{2\sigma^2}\mu^2\right\} \tag{57}$$
$$= \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left(-\frac{1}{2\sigma^2}\mu^2\right) \exp\left\{-\frac{1}{2\sigma^2}x^2 + \frac{\mu}{\sigma^2}x\right\} \tag{58}$$
$$= \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left(-\frac{1}{2\sigma^2}\mu^2\right) \exp\left\{\begin{pmatrix}\frac{\mu}{\sigma^2} & -\frac{1}{2\sigma^2}\end{pmatrix}\begin{pmatrix}x\\x^2\end{pmatrix}\right\} \tag{59}$$
$$= \frac{1}{(2\pi)^{1/2}}(\sigma^2)^{-1/2} \exp\left(-\frac{1}{2\sigma^2}\mu^2\right) \exp\left\{\begin{pmatrix}\frac{\mu}{\sigma^2} & -\frac{1}{2\sigma^2}\end{pmatrix}\begin{pmatrix}x\\x^2\end{pmatrix}\right\} \tag{60}$$

Let $\eta_1 = \frac{\mu}{\sigma^2}$ and $\eta_2 = -\frac{1}{2\sigma^2}$, and then $\sigma^2 = -1/(2\eta_2)$ and $\mu = -\eta_1/(2\eta_2)$. We get

$$p(x|\mu, \sigma^2) = \frac{1}{(2\pi)^{1/2}}(-1/(2\eta_2))^{-1/2} \exp\left(-\frac{1}{-2/(2\eta_2)}(-\eta_1/(2\eta_2))^2\right) \exp\left\{\boldsymbol{\eta}^T \mathbf{u}(x)\right\}$$
$$= (2\pi)^{-1/2}(-2\eta_2)^{1/2} \exp\left(\frac{\eta_1^2}{4\eta_2}\right) \exp\left\{\boldsymbol{\eta}^T \mathbf{u}(x)\right\}$$

which, after some simple rearrangement, can be cast in the standard exponential family form with

$$\boldsymbol{\eta} = \begin{pmatrix} \frac{\mu}{\sigma^2} \\ -\frac{1}{2\sigma^2} \end{pmatrix} \tag{61}$$

$$\mathbf{u}(x) = \begin{pmatrix} x \\ x^2 \end{pmatrix} \tag{62}$$

$$h(x) = (2\pi)^{-1/2} \tag{63}$$

$$g(\boldsymbol{\eta}) = (-2\eta_2)^{1/2} \exp\left(\frac{\eta_1^2}{4\eta_2}\right). \tag{64}$$

17.4 Multivariate Gaussian distribution

Recall the definition of the multivariate gaussian distribution, i.e. (22),

$$\begin{aligned}
\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) &= \frac{1}{(2\pi)^{D/2}} \frac{1}{|\boldsymbol{\Sigma}|^{1/2}} \exp\left\{\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right\} \\
&= \frac{1}{(2\pi)^{D/2}} \frac{1}{|\boldsymbol{\Sigma}|^{1/2}} \exp\left\{\frac{1}{2}\mathbf{x}^T\boldsymbol{\Sigma}^{-1}\mathbf{x} - \boldsymbol{\mu}^T\boldsymbol{\Sigma}^{-1}\mathbf{x} + \frac{1}{2}\boldsymbol{\mu}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}\right\} \\
&= \frac{1}{(2\pi)^{D/2}} \frac{1}{|\boldsymbol{\Sigma}|^{1/2}} \exp\left\{\frac{1}{2}\boldsymbol{\mu}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}\right\} \exp\left\{\frac{1}{2}\mathbf{x}^T\boldsymbol{\Sigma}^{-1}\mathbf{x} - \boldsymbol{\mu}^T\boldsymbol{\Sigma}^{-1}\mathbf{x}\right\}
\end{aligned}$$

Then

$$\boldsymbol{\eta}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \begin{pmatrix} \frac{1}{2}\,\mathrm{vec}(\boldsymbol{\Sigma}^{-1}) \\ -\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} \end{pmatrix} \tag{65}$$

$$\mathbf{u}(x) = \begin{pmatrix} \mathrm{vec}(\mathbf{x}\mathbf{x}^T) \\ \mathbf{x} \end{pmatrix} \tag{66}$$

$$h(x) = \frac{1}{(2\pi)^{D/2}} \tag{67}$$

$$g\left(\boldsymbol{\eta}(\boldsymbol{\mu}, \boldsymbol{\Sigma})\right) = \frac{1}{|\boldsymbol{\Sigma}|^{1/2}} \exp\left\{\frac{1}{2}\boldsymbol{\mu}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}\right\}. \tag{68}$$

where $\mathrm{vec}(\cdot)$ denotes the vectorization operator which concatenates all columns of a matrix into a column vector.

17.5 Beta distribution

Recall the definition of the beta distribution, i.e. (12),

$$\begin{aligned}
\mathrm{Beta}(x|a,b) &= \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} x^{a-1}(1-x)^{b-1} \\
&= \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \exp\{(a-1)\ln x + (b-1)\ln(1-x)\} \\
&= \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \exp\{a\ln x - \ln x + b\ln(1-x) - \ln(1-x)\} \\
&= \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \exp\{a\ln x + b\ln(1-x) - \ln[x(1-x)]\} \\
&= \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \frac{\exp\{a\ln x + b\ln(1-x)\}}{x(1-x)} \\
&= \frac{1}{x(1-x)} \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \exp\{a\ln x + b\ln(1-x)\}
\end{aligned}$$

Comparison with (37) allows us to identify

$$\begin{aligned}\boldsymbol{\eta} &= \begin{pmatrix}\eta_1\\ \eta_2\end{pmatrix} = \begin{pmatrix}a\\ b\end{pmatrix}\\ \mathbf{u}(x) &= \begin{pmatrix}\ln x\\ \ln(1-x)\end{pmatrix}\\ h(x) &= \frac{1}{x(1-x)}\\ g(\boldsymbol{\eta}) &= \frac{\Gamma(\eta_1+\eta_2)}{\Gamma(\eta_1)\Gamma(\eta_2)}.\end{aligned}$$

17.6 Gamma distribution

Recall the definition of the gamma distribution

$$\begin{aligned}\mathrm{Gam}(x|a,b) &= \frac{1}{\Gamma(a)}b^a x^{a-1}\exp(-bx)\\ &= \frac{1}{\Gamma(a)}x^{a-1}\exp\left\{\begin{pmatrix}a & b\end{pmatrix}\begin{pmatrix}0\\ -x\end{pmatrix}\right\}\end{aligned}$$

It is easy to identify that

$$\begin{aligned}\boldsymbol{\eta}(a,b) &= \begin{pmatrix}a\\ b\end{pmatrix}\\ \mathbf{u}(x) &= \begin{pmatrix}0\\ -x\end{pmatrix}\\ h(x) &= x^{a-1}\\ g(\boldsymbol{\eta}) &= \frac{1}{\Gamma(a)}.\end{aligned}$$

Another form is given by

$$\begin{aligned}\mathrm{Gam}(x|a,b) &= \frac{1}{\Gamma(a)}b^a x^{a-1}\exp(-bx)\\ &= \frac{1}{\Gamma(a)}\exp\{a\ln b\}x^{a-1}\exp(-bx)\\ &= \frac{1}{\Gamma(a)}x^{a-1}\exp(a\ln b - bx)\\ &= \frac{1}{\Gamma(a)}x^{a-1}\exp\left\{\begin{pmatrix}a\ln b & b\end{pmatrix}\begin{pmatrix}1\\ -x\end{pmatrix}\right\}\end{aligned}$$

Let $\eta_1 = a\ln b$ and $\eta_2 = b$, then $a = \eta_1/\ln\eta_2$. Thus, we have

$$\mathrm{Gam}(x|\eta_1,\eta_2) = \frac{1}{\Gamma(\eta_1/\ln\eta_2)}x^{\eta_1/\ln\eta_2-1}\exp\left\{\begin{pmatrix}\eta_1 & \eta_2\end{pmatrix}\begin{pmatrix}1\\ -x\end{pmatrix}\right\} \tag{69}$$

which gives

$$\begin{aligned}\boldsymbol{\eta} &= \begin{pmatrix}\eta_1\\ \eta_2\end{pmatrix} = \begin{pmatrix}a\ln b\\ b\end{pmatrix}\\ \mathbf{u}(x) &= \begin{pmatrix}1\\ -x\end{pmatrix}\\ h(x) &= x^{a-1}\\ g(\boldsymbol{\eta}) &= \frac{1}{\Gamma(\eta_1/\ln\eta_2)}.\end{aligned}$$

17.7 Von Mises distribution

Recall the definition of the von Mises distribution

$$\begin{aligned}p(\theta|\theta_0, m) &= \frac{1}{2\pi I_0(m)}\exp\{m\cos(\theta-\theta_0)\}\\ &= \frac{1}{2\pi I_0(m)}\exp\{m(\cos(\theta)\cos(\theta_0)-\sin(\theta)\sin(\theta_0))\}\\ &= \frac{1}{2\pi I_0(m)}\exp\left\{\begin{pmatrix}m\cos(\theta_0) & m\sin(\theta_0)\end{pmatrix}\begin{pmatrix}\cos(\theta)\\ \sin(\theta)\end{pmatrix}\right\}\end{aligned}$$

where $I_0(m)$ is given by (36). It is easy to identify that

$$\begin{aligned}\boldsymbol{\eta} &= \begin{pmatrix}\eta_1\\ \eta_2\end{pmatrix} = \begin{pmatrix}m\cos(\theta_0)\\ m\sin(\theta_0)\end{pmatrix}\\ \mathbf{u}(\theta) &= \begin{pmatrix}\cos(\theta)\\ \sin(\theta)\end{pmatrix}\\ h(x) &= 1\\ g(\boldsymbol{\eta}) &= \frac{1}{2\pi I_0(\sqrt{\eta_1^2+\eta_2^2})}.\end{aligned}$$

18 Maximum likelihood and sufficient statistics under the form of exponential family

18.1 The gradient and hessian of natural parameters

Taking the gradient of both sides of (37) w.r.t. $\boldsymbol{\eta}$, we have

$$\nabla g(\boldsymbol{\eta})\int h(\mathbf{x})\exp\{\boldsymbol{\eta}^T\mathbf{u}(\mathbf{x})\}\,\mathrm{d}\,\mathbf{x} + g(\boldsymbol{\eta})\int h(\mathbf{x})\exp\{\boldsymbol{\eta}^T\mathbf{u}(\mathbf{x})\}\mathbf{u}(\mathbf{x})\,\mathrm{d}\,\mathbf{x} = 0$$

Rearranging, and making use of (38) then gives

$$\begin{aligned}-\frac{1}{g(\boldsymbol{\eta})}\nabla g(\boldsymbol{\eta}) &= \frac{\int h(\mathbf{x})\exp\{\boldsymbol{\eta}^T\mathbf{u}(\mathbf{x})\}\mathbf{u}(\mathbf{x})\,\mathrm{d}\,\mathbf{x}}{\int h(\mathbf{x})\exp\{\boldsymbol{\eta}^T\mathbf{u}(\mathbf{x})\}\,\mathrm{d}\,\mathbf{x}} &(70)\\ &= g(\boldsymbol{\eta})\int h(\mathbf{x})\exp\{\boldsymbol{\eta}^T\mathbf{u}(\mathbf{x})\}\mathbf{u}(\mathbf{x})\,\mathrm{d}\,\mathbf{x} &(71)\\ &= \int g(\boldsymbol{\eta})h(\mathbf{x})\exp\{\boldsymbol{\eta}^T\mathbf{u}(\mathbf{x})\}\mathbf{u}(\mathbf{x})\,\mathrm{d}\,\mathbf{x} &(72)\\ -\nabla\ln g(\boldsymbol{\eta}) &= \mathbb{E}[\mathbf{u}(\mathbf{x})] &(73)\end{aligned}$$

Furthermore,

$$\begin{aligned}
-\nabla\nabla^T \ln g(\boldsymbol{\eta}) &= \nabla g(\boldsymbol{\eta}) \int h(\mathbf{x}) \exp\{\boldsymbol{\eta}^T \mathbf{u}(\mathbf{x})\} \mathbf{u}(\mathbf{x})^T \,\mathrm{d}\,\mathbf{x} + g(\boldsymbol{\eta}) \int h(\mathbf{x}) \exp\{\boldsymbol{\eta}^T \mathbf{u}(\mathbf{x})\} \mathbf{u}(\mathbf{x}) \mathbf{u}(\mathbf{x})^T \,\mathrm{d}\,\mathbf{x} \\
&= \frac{\nabla g(\boldsymbol{\eta})}{g(\boldsymbol{\eta})} \int g(\boldsymbol{\eta}) h(\mathbf{x}) \exp\{\boldsymbol{\eta}^T \mathbf{u}(\mathbf{x})\} \mathbf{u}(\mathbf{x})^T \,\mathrm{d}\,\mathbf{x} + \int g(\boldsymbol{\eta}) h(\mathbf{x}) \exp\{\boldsymbol{\eta}^T \mathbf{u}(\mathbf{x})\} \mathbf{u}(\mathbf{x}) \mathbf{u}(\mathbf{x})^T \,\mathrm{d}\,\mathbf{x} \\
&= \mathbb{E}[\mathbf{u}(\mathbf{x}) \mathbf{u}(\mathbf{x})^T] - \mathbb{E}[\mathbf{u}(\mathbf{x})] \mathbb{E}[\mathbf{u}(\mathbf{x})^T] \\
&= \mathrm{Cov}[\mathbf{u}(\mathbf{x})]
\end{aligned}$$

18.2 Maximum likelihood

Now consider a set of i.i.d. data denoted by $\mathbf{X} = \{\mathbf{x}_1, \cdots, \mathbf{x}_n\}$, for which the likelihood function is given by

$$p(\mathbf{X}|\boldsymbol{\eta}) = \left(\prod_{n=1}^{N} h(\mathbf{x}_n) \right) g(\boldsymbol{\eta})^N \exp\left\{ \boldsymbol{\eta}^T \sum_{n=1}^{N} \mathbf{u}(\mathbf{x}_n) \right\} \tag{74}$$

Setting the gradient of $\ln p(\mathbf{X}|\boldsymbol{\eta})$ w.r.t. $\boldsymbol{\eta}$ to zero, we get

$$\begin{aligned}
\nabla \ln p(\mathbf{X}|\boldsymbol{\eta}) &= N \frac{\nabla g(\boldsymbol{\eta})}{g(\boldsymbol{\eta})} + \sum_{n=1}^{N} \mathbf{u}(\mathbf{x}_n) = 0 \\
-N \frac{\nabla g(\boldsymbol{\eta})}{g(\boldsymbol{\eta})} &= \sum_{n=1}^{N} \mathbf{u}(\mathbf{x}_n) \\
-\frac{\nabla g(\boldsymbol{\eta})}{g(\boldsymbol{\eta})} &= \frac{1}{N} \sum_{n=1}^{N} \mathbf{u}(\mathbf{x}_n) \\
-\nabla \ln g(\boldsymbol{\eta}_{\mathrm{ML}}) &= \frac{1}{N} \sum_{n=1}^{N} \mathbf{u}(\mathbf{x}_n)
\end{aligned}$$

where $\boldsymbol{\eta}_{\mathrm{ML}}$ is the desired maximum likelihood estimator. This result can be used to obtain $\boldsymbol{\eta}_{\mathrm{ML}}$.

18.3 Sufficient statistic

We see that the solution for the maximum likelihood estimator depends on the data only through $\sum_n \mathbf{u}(\mathbf{x}_n)$, which is therefore called the **sufficient statistic** of the exponential family distribution. As a result, we do not need to store the entire data set itself but only the value of the sufficient statistic. For the Gaussian $\mathbf{u}(x) = (x, x^2)^T$, we should keep both the sum of $\{x_n\}$ and the sum of $\{x_n^2\}$.

19 Conjugate priors

In general, for a given probability distribution $p(\mathbf{x}|\boldsymbol{\eta})$, we can seek a prior $p(\boldsymbol{\eta})$ that is conjugate to the likelihood function, so that the posterior distribution has the same functional form as the prior. For example, the conjugate prior of the Bernoulli distribution is the beta distribution. For the Gaussian, the conjugate prior for the mean is a Gaussian, and the conjugate prior for the precision is a gamma distribution and the Wishart distribution for the precision matrix of a multivariate Gaussian.

For any member of the exponential family, there exists a conjugate prior that can be written in the form

$$p(\boldsymbol{\eta}|\boldsymbol{\chi}, \nu) = f(\boldsymbol{\chi}, \nu) g(\boldsymbol{\eta})^{\nu} \exp\{\nu \boldsymbol{\eta}^T \boldsymbol{\chi}\} \tag{75}$$

Table 1: Conjugate priors for discrete distributions

Likelihood	Model parameter	Conjugate prior distribution	Prior hyperparameters	Posterior hyperparameters
Bernoulli				
book	✓	✓		
report	✓	✓	✓	

where $f(\boldsymbol{\chi}, \nu)$ is a normalization coefficient, and $g(\boldsymbol{\eta})$ is the same function as appears in the standard form of the exponential family. To see that this is indeed conjugate, let us multiply the prior (75) by the likelihood function (74) to obtain the posterior distribution, up to a normalization coefficient, in the form

$$p(\boldsymbol{\eta}|\mathbf{X}, \boldsymbol{\chi}, \nu) \propto g(\boldsymbol{\eta})^{\nu+N} \exp\left\{\boldsymbol{\eta}^T \left(\sum_{n=1}^{N} \mathbf{u}(\mathbf{x}_n) + \nu\boldsymbol{\chi}\right)\right\} \tag{76}$$

This again takes the same functional form as the prior (75), confirming conjugacy. Furthermore, we see that the parameter ν can be interpreted as an effective number of pseudo-observations in the prior, each of which has a value for the sufficient statistic $\mathbf{u}(\mathbf{x})$ given by $\boldsymbol{\chi}$.