RESEARCH NOTES IN STATISTICAL MACHINE LEARNING

JINGHUA YAO

Abstract. The probability foundations for statistical machine learning is of fundamental importance. It is our opinion that as the trend of automation of machine learnings develops, the probability or more primarily the mathematical background behind the machine learning algorithms will become more and more important for people who want to use machine learning technique correctly and effectively. We collect the probability distributions used in machine learning study and give the detailed derivations for some related quantities such as expectation, variance and characteristic functions. Meanwhile, we present the expectation maximization (EM) algorithm in integral form, supply the hidden Markov model (HMM) with a specific example to demonstrate the computation of probabilities using forward or backward algorithm, and illustrate the predictions in HMM using Viterbi algorithm. Also, we collect some typical computations and ideas in using machine learning algorithms including the artificial neural network. We give some concentration inequalities with proofs, and record some useful ideas of the author in this evolving notes. We emphasize the geometrical intuition and methods from calculus of variations during the writing.

Contents

Date: Updated February, 2018; © Jinghua Yao.

2010 *Mathematics Subject Classification.* Mathematics, Statistics, Physics, and Engineering.

Key words and phrases. Probability, Statistics, Statistical Machine Learning, EM, HMM, Entropy, Inequality, Gamma function, Beta function, Stirling formula.

1. Statistics Historical Notes

1. 1890—1900, Francis Galton and Karl Pearson.

Galton: the concepts of standard deviation, correlation, regression analysis and the

Pearson: Pearson product-moment correlation coefficient, the method of moments for the fitting of distributions to samples and the Pearson distribution Ronald Fishe: null hypothesis.

2. 1910—1920, William Gosset, Ronald Fisher

Fisher: the concepts of sufficiency, ancillary statistics, Fisher's linear discriminator and Fisher information

3. 1930s, Egon Pearson and Jerzy Neyman

the concepts of "Type II" error, power of a test and confidence intervals.

4. Today

Statistical methods are applied in all fields that involve decision making. The use of modern computers has expedited large-scale statistical computations, and has also made possible new methods that are impractical to perform manually.

2. Maxwell-Boltzmann distribution

The distribution is given by

$$f(v) = \left(\frac{m}{2\pi kT}\right)^{3/2} 4\pi v^2 \exp(-\frac{mv^2}{2kT})$$

for $v \in [0, +\infty)$. Using variance of normal distribution, it is easy to verify that

$$\int_0^\infty f(v)\, dv = 1.$$

Let $\alpha = \frac{m}{2kT}$. We have

$$f(v;\alpha) = \left(\frac{\alpha}{\pi}\right)^{3/2} 4\pi v^2 \exp(-\alpha v^2).$$

$f(v)$ satisfies the following ODE

$$\begin{cases} kTvf'(v) + f(v)(mv^2 - 2kT) = 0, \\ f(1) = (2/\pi)^{1/2} \exp(-m/(2kT))(m/(kT))^{3/2}. \end{cases} \tag{2.1}$$

In unitless form, it is as follows

$$\begin{cases} a^2xf'(x) + (x^2 - 2a^2)f(x) = 0, \\ f(1) = \frac{1}{a^3}\left(\frac{2}{\pi}\right)^{1/2} e^{-\frac{1}{2a^2}} \end{cases} \tag{2.2}$$

3. Probability distributions

3.1. **Bernoulli distribution.** (a) Distribution. Consider toss a coin and denote the result by $X \in \{1 = \text{head}, 0 = \text{tail}\}$ where $p(X = 1|\mu) = \mu$. Consequently, $p(X = 0|\mu) = 1 - \mu$. The distribution can be written as

$$X \sim \text{Bernoulli}(x|\mu) = \mu^x(1-x)^{1-x}, \quad x \in \{0, 1\}.$$

It is easy to verify that $E[X] = \mu$ and $Var[X] = \mu - \mu^2$.

(b) Maximum likelihood estimate. Assume that $X \sim \text{Bernoulli}(x|\mu)$ and we independently sampled N data points $D = \{x_1, \cdots, x_N\}$. We aim to estimate the parameter μ. For this, we consider the log maximum likelihood function

$$\begin{aligned}\ln p(D|\mu) &= \ln(\Pi_i \text{Bernoulli}(x_i|\mu)) = \sum_i \ln \text{Bernoulli}(x_i|\mu) \\ &= \sum_i \{x_i \ln \mu + (1 - x_i) \ln(1-\mu)\}.\end{aligned} \tag{3.1}$$

Letting

$$\frac{d \ln p(D|\mu)}{d\mu} = \frac{1}{\mu}\sum_j x_i - \frac{1}{1-\mu}\sum_i (1 - x_i) = 0,$$

we get

$$\mu_{ML} = \frac{\sum_i x_i}{N}.$$

3.2. **Binomial distribution.** (a) Distribution. Consider N times independent Bernoulli trails $X_1, \cdots, X_N$ and consider the number of heads X obtained, i.e., $X = X_1 + \cdots + X_N$. The probability $p(X = m)$ for $m = 0, 1, \cdots, N$ is

$$P(X = m) = \binom{N}{m} \mu^m (1-\mu)^{N-m},$$

which is the Binomial distribution, denoted by $X \sim \text{Binomial}(\mu, N)$.

(b) $E[X]$ and $Var[X]$. We can easily show that

$$E[X] = N\mu, \quad Var[X] = N(\mu - \mu^2)$$

by noticing the independence of X_1, X_2, $\cdots$, X_N. Next, we give a direct verification. For $E[X]$, we have

$$\begin{aligned}E[X] &= \sum_{m=0}^{N} m\frac{N!}{(N-m)!m!}\mu^m(1-\mu)^{N-m}\\ &= N\mu \sum_{m-1=0}^{m-1=N-1} \frac{(N-1)!}{[(N-1)-(m-1)]!(m-1)!}\mu^{m-1}(1-\mu)^{(N-1)-(m-1)}\\ &= N\mu\sum_{j=0}^{N-1}\frac{(N-1)!}{(N-1-j)!j!}\mu^j(1-\mu)^{N-1-j}\\ &= N\mu(\mu+1-\mu)^{N-1}\\ &= N\mu.\end{aligned} \tag{3.2}$$

For $Var[X]$, it is sufficient to compute $E[X^2]$ and use the relation $Var[X]=E[X^2]-E[X]^2$.

$$\begin{aligned}E[X^2] &= \sum_{m=0}^{N} m^2\frac{N!}{(N-m)!m!}\mu^m(1-\mu)^{N-m}\\ &= N\mu\sum_{m-1=0}^{m-1=N-1} m\frac{(N-1)!}{[(N-1)-(m-1)]!(m-1)!}\mu^{m-1}(1-\mu)^{(N-1)-(m-1)}\\ &= N\mu\sum_{m-1=0}^{m-1=N-1}(m-1+1)\frac{(N-1)!}{[(N-1)-(m-1)]!(m-1)!}\mu^{m-1}(1-\mu)^{(N-1)-(m-1)}\\ &= N\mu\Big\{\sum_{j=0}^{N-1} j\frac{(N-1)!}{(N-1-j)!j!}\mu^j(1-\mu)^{N-1-j}+\sum_{j=0}^{N-1}\frac{(N-1)!}{(N-1-j)!j!}\mu^j(1-\mu)^{N-1-j}\Big\}\\ &= N\mu\big\{E[\text{Binomial}(\mu,N-1)]+1\big\}\\ &= N\mu((N-1)\mu+1),\end{aligned} \tag{3.3}$$

from which we get

$$Var[X]=E[X^2]-E[X]^2=N\mu((N-1)\mu+1)-(N\mu)^2=N\mu(1-\mu).$$

3.3. **The Beta distribution.** (a) Preliminary on the Beta function $B(x,y)$ and Gamma function $\Gamma(x)$. The Beta function, also known as Euler integral of the first kind, is defined as

$$B(x,y)=\int_0^1 t^{x-1}(1-t)^{y-1}\,dt, \quad \mathrm{Re}(x)>0, \mathrm{Re}(y)>0.$$

It is easy to see that $B(x,y)=B(y,x)$ and $B(1,1)=0$. The function $\Gamma(x)$ is defined as

$$\Gamma(x)=\int_0^\infty u^{x-1}e^{-u}du, \quad \mathrm{Re}(x)>0.$$

It is easy to see that $\Gamma(1)=1$ and $\Gamma(x+1)=x\Gamma(x)$, hence $\Gamma(n)=(n-1)!$. Some other special values are

$$\Gamma(2)=1, \quad \Gamma(1/2)=\sqrt{\pi}, \quad \Gamma(3/2)=\sqrt{\pi}/2.$$

In particular, if n is a positive integer, we have $\Gamma(n+1)=n!$. Therefore, the Γ-function can be regarded as a generalization of factorial to the complex variable case. For the factorial, we know that Stirling's formula which says

$$\text{(Stirling's formula)}\quad \ln(n!)=n\ln n-n+O(\ln n),\quad n\to+\infty. \tag{3.4}$$

More precisely, we have

$$n!\sim\sqrt{2\pi n}\Big(\frac{n}{e}\Big)^n,\quad n\to+\infty. \tag{3.5}$$

For the Gamma function, there is a Stirling's formula which reads

$$\text{(Stirling's formula)}\quad \Gamma(z+1)\sim\sqrt{2\pi z}\Big(\frac{z}{e}\Big)^z,\quad z\to\infty, |\arg(z+1)|\leqslant\pi-\varepsilon. \tag{3.6}$$

An equivalent form is

$$\text{(Stirling's formula)}\quad \Gamma(z)=\sqrt{\frac{2\pi}{z}}\Big(\frac{z}{e}\Big)^z\Big(1+O(1/2)\Big),\quad z\to\infty, |\arg(z)|\leqslant\pi-\varepsilon. \tag{3.7}$$

The following approximation is now obvious

$$\lim_{n\to\infty}\frac{\Gamma(n+\alpha)}{\Gamma(n)n^\alpha}=1, \tag{3.8}$$

where $\alpha\in\mathbb{C}^1$.

Proposition 3.1.

$$B(x,y)=\frac{\Gamma(x)\Gamma(y)}{\Gamma(x+y)}.$$

Proof. Consider $\Gamma(x)\Gamma(y)$:

$$\begin{aligned}\Gamma(x)\Gamma(y)&=\int_0^\infty e^{-u}u^{x-1}dx\int_0^\infty e^{-v}v^{y-1}dy\\&=\int_{u=0}^\infty\int_{v=0}^\infty e^{-(u+v)}u^{x-1}v^{y-1}\,dudv\end{aligned} \tag{3.9}$$

Using the change of variables

$$u=zt\geqslant 0,\quad v=z(1-t)\geqslant 0,$$

i.e.,

$$z=u+v,\quad t=\frac{u}{u+v},$$

we have

$$\frac{\partial(u,v)}{\partial(z,t)}=z,$$

and

$$\begin{aligned}\Gamma(x)\Gamma(y)&=\int_{u=0}^\infty\int_{v=0}^\infty e^{-(u+v)}u^{x-1}v^{y-1}\,dudv\\&=\int_{z=0}^\infty\int_{t=0}^1 e^{-z}z^{x+y-1}t^{x-1}(1-t)^{y-1}\,dzdt\\&=\int_{z=0}^\infty e^{-z}z^{x+y-1}\,dz\int_{t=0}^1 t^{x-1}(1-t)^{y-1}\,dt\\&=\Gamma(x+y)B(x,y).\end{aligned} \tag{3.10}$$

□

Definition 3.2. (Multivariate Beta function) The multivariate Beta function is defined as

$$B(a_1, a_2, \cdots, a_n) := \frac{\prod_j \Gamma(a_j)}{\Gamma(\sum_j a_j)}.$$

(b) The Beta distribution. Let $\mu \in [0, 1]$ and $X \sim \text{Beta}(\mu|a, b)$ for $a > 0$ and $b > 0$ is the following pdf on the interval $[0, 1]$:

$$\text{Beta}(\mu|a, b) := \frac{1}{B(a, b)}\mu^{a-1}(1-\mu)^{b-1} = \frac{\Gamma(x+y)}{\Gamma(x)\Gamma(y)}\mu^{a-1}(1-\mu)^{b-1}. \tag{3.11}$$

Proposition 3.3. *Let $X \sim Beta(\mu|a, b)$. Then*

$$E[X] = \frac{a}{a+b}, \quad Var[X] = \frac{ab}{(a+b)^2(a+b+1)}.$$

Proof. We use the relation between Gamma and Beta functions to give the proof.

$$\begin{aligned} E[X] &= \int \mu \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\mu^{a-1}(1-\mu)^{b-1}\,d\mu \\ &= \int \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\mu^{a+1-1}(1-\mu)^{b-1}\,d\mu \\ &= \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}B(a+1, b) \\ &= \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\frac{\Gamma(a+1)\Gamma(b)}{\Gamma(a+1+b)} \\ &= \frac{\Gamma(a+b)}{\Gamma(a+b+1)}\frac{\Gamma(a+1)}{\Gamma(a)} \\ &= \frac{a}{a+b}. \end{aligned} \tag{3.12}$$

Similarly, we could get

$$\begin{aligned} E[X^2] &= \int \mu^2 \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\mu^{a-1}(1-\mu)^{b-1}\,d\mu \\ &= \int \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\mu^{a+2-1}(1-\mu)^{b-1}\,d\mu \\ &= \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}B(a+2, b) \\ &= \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\frac{\Gamma(a+2)\Gamma(b)}{\Gamma(a+2+b)} \\ &= \frac{\Gamma(a+b)}{\Gamma(a+b+2)}\frac{\Gamma(a+2)}{\Gamma(a)} \\ &= \frac{(a+1)a}{(a+b+1)(a+b)}. \end{aligned} \tag{3.13}$$

Therefore, we get

$$Var[X] = E[X^2] - E[X]^2 = \frac{(a+1)a}{(a+b+1)(a+b)} - \Big(\frac{a}{a+b}\Big)^2 = \frac{ab}{(a+b+1)(a+b)^2}.$$

□

Definition 3.4. (Conjugate distributions) Let $X \sim p(x|\theta)$ and consider its maximum likelihood function $L(\theta) = \prod_i p(x_i|\theta)$ with respect to N independent samples $D = \{x_1, \cdots, x_N\}$. If we choose a prior for the parameter $p(\theta) \propto L(\theta)$. Then the posterior distribution $p(\theta|D) \propto p(\theta)L(\theta)$. If $p(\theta)L(\theta)$ and $p(\theta)$ have the same functional form, we say that $p(\theta)$ is the conjugate prior of the maximum likelihood of $X \sim p(x|\theta)$.

In other words, for a given probability distribution $p(\vec{x}|\vec{\mu})$, we can seek a prior $p(\vec{\mu})$ that is conjugate to the likelihood function, so that the posterior distribution $p(\vec{\mu}|D)$ (computed using Bayes' Theorem) after observing $D = \{\mathbf{x}_1, \cdots, \mathbf{x}_N\}$ has the same functional form as the prior.

Proposition 3.5. *The beta distribution is a conjugate prior for the Binomial distribution.*

Proposition 3.6. *The Dirichlet distribution is a conjugate prior for the multinomial distribution.*

Proposition 3.7. *For the Gaussian $X \sim N(x|\mu, \sigma^2)$ with σ^2 being fixed, we consider the inference of μ. Then the conjugate prior distribution of μ is Gaussian; If we fixed μ, consider the inference of the precision $\lambda := 1/\sigma^2$, then the conjugate prior for λ is the Gamma distribution; If we vary both μ, and λ, then the conjugate prior for (μ, λ) is the Gaussian-Gamma distribution. For multivariate Gaussian $N(\mathbf{x}|\mu, \mathbf{\Lambda}^{-1})$, if both μ and $\mathbf{\Lambda}$ are to be inferred, then the conjugate prior for $(\mu, \mathbf{\Lambda})$ is the Gaussian-Wishart distribution.*

3.4. **Multinomial Variables.** (a) Consider the set D of K dimensional binary vectors $\mathbf{x} = (x_1, x_2, \cdots, x_k)^T$ where $x_k \in \{0, 1\}$ for all k and $\sum_k x_k = 1$, i.e.,

$$D = \Big\{\mathbf{x} = (x_1, \cdots, x_k, \cdots, x_K)^T; x_k \in \{0, 1\}, \sum_k x_k = 1\Big\}.$$

Assume $p(x_k = 1) = \mu_k \in [0, 1]$ and $\sum_k \mu_k = 1$, and denote

$$\mu = (\mu_1, \cdots, \mu_K)^T.$$

For any $\mathbf{x} \in D$, we have

$$p(\mathbf{x}|\mu) = \prod_k \mu_k^{x_k} := \mu^{\mathbf{x}}.$$

We can see easily that

$$\sum_{\mathbf{x}} p(\mathbf{x}|\mu) = 1, \quad E[\mathbf{x}|\mu] = \sum_{\mathbf{x}} \mathbf{x} p(\mathbf{x}|\mu)\, dx = \mu.$$

(b) Maximum likelihood function. Assume we have N independent samples $\mathcal{D} = \{\mathbf{x}_1, \cdots \mathbf{x}_N\}$, then the maximum likelihood function is

$$p(\mathcal{D}|\mu) = \prod_n \prod_k \mu_k^{x_{nk}} = \prod_k \mu_k^{\sum_n x_{nk}} = \prod_k \mu_k^{n_k}, \tag{3.14}$$

where n_k is the number of observations with $x_k = 1$. We can estimate the parameter μ with Lagrangian multiplier method by considering $\ln p(\mathcal{D}|\mu) + \lambda(\sum_k \mu_k - 1)$ to get

$$\mu_k^{ML} = m_k/N, \quad k = 1, 2, \cdots, K.$$

(c) Multinomial distribution of $m_1, \cdots, m_K$. From (3.14), and choosing proper normalization constant, we get the multinomial distribution with parameters μ and N as below

$$\text{Multinomial}(m_1, m_2, \cdots, m_K; \mu, N) = \binom{N}{m_1, m_2, \cdots, m_K} \prod_{k=1}^{K} \mu_k^{m_k}, \tag{3.15}$$

where $\sum_k m_k = N$ and

$$\binom{N}{m_1, m_2, \cdots, m_K} = \frac{N!}{m_1! m_2! \cdots m_K!}.$$

(d) Multinomial theory. From (c) above, we know that

$$\sum_{m_1, \cdots, m_K} \text{Multinomial}(m_1, m_2, \cdots, m_K; \mu, N) = (\mu_1 + \mu_2 + \cdots + \mu_K)^N = 1.$$

Multiplying the above equality by M^N for any nonzero number M, we have

$$(M\mu_1 + \cdots + M\mu_K)^N = \sum_{m_1 + \cdots + m_K = N} \binom{N}{m_1, m_2, \cdots, m_K} \prod_{k=1} (M\mu_k)^{m_k}.$$

Denote $Mm_k = a_k$, we have

$$(a_1 + \cdots + a_K)^N = \sum_{m_1 + \cdots + m_K = N} \binom{N}{m_1, m_2, \cdots, m_K} \prod_{k=1}^{K} a_k^{m_k}.$$

The above derivation restricts all a_k are of the same sign. Actually, if we regard these a_k as symbols, then formally, the above equality holds for any a_k. Therefore, it holds for any $\mathbf{a} = (a_1, a_2, \cdots, a_K) \in \mathbb{R}^K$. Therefore, we have

Theorem 3.8. *(Multinomial theorem) Let* $\mathbf{a} = (a_1, a_2, \cdots, a_K) \in \mathbb{R}^K$ *and* N *be a positive integer. Then*

$$(a_1 + a_2 + \cdots + a_K)^N = \sum_{m_1 + \cdots + m_K = N} \binom{N}{m_1, m_2, \cdots, m_K} \prod_{k=1}^{K} a_k^{m_k}.$$

3.5. **The Dirichlet distribution.** Let μ be as in the multinomial distribution. We consider the conjugate prior for μ to get the Dirichlet distribution

$$\text{Dirichlet}(\mu|\alpha) = \frac{1}{B(\alpha_1, \cdots, \alpha_K)} \prod_{k=1}^{K} \mu_k^{\alpha_k - 1} = \frac{\Gamma(\sum_k \alpha_k)}{\Gamma(\alpha_1) \cdots \Gamma(\alpha_K)} \prod_{k=1}^{K} \mu_k^{\alpha_k - 1}, \tag{3.16}$$

where α_k can be interpreted as

$$\alpha_k = \{\mathbf{x}; x_k = 1\}$$

for the $\sum_k \alpha_k$ samples in multinomial distribution.

3.6. **The Gamma distribution.** We say that a continuous random variable $X \in (0, +\infty]$ obeys gamma distribution with parameters $a > 0$ and $b > 0$, denoted by $X \sim \text{Gamma}(x|a, b)$, iff

$$P[X \in (x, x + dx)] = \int_x^{x+dx} \frac{1}{\Gamma(a)} b^a \lambda^{a-1} \exp(-b\lambda)\, d\lambda, \tag{3.17}$$

i.e., X has pdf

$$\mathrm{Gamma}(x|a,b)=\frac{1}{\Gamma(a)}b^a x^{a-1}\exp(-bx).$$

It is easy to compute that

$$E[X]=a/b,\quad \mathrm{Var}(X)=a/b^2.$$

Indeed, we have

$$\begin{aligned}
E[X]&=\int_0^{+\infty}\frac{1}{\Gamma(a)}b^a x^{a-1}\exp(-bx)x\,dx\\
&=\int_0^{+\infty}\frac{1}{\Gamma(a)}(bx)^a\exp(-bx)\frac{d(bx)}{b}\\
&=\int_0^{+\infty}\frac{1}{b\Gamma(a)}u^{a+1}\exp(-u)\frac{du}{u}\ (bx:=u)\\
&=\frac{1}{b\Gamma(a)}\Gamma(a+1)\\
&=\frac{a\Gamma(a)}{b\Gamma(a)}\\
&=\frac{a}{b},
\end{aligned}\tag{3.18}$$

and

$$\begin{aligned}
E[X^2]&=\int_0^{+\infty}\frac{1}{\Gamma(a)}b^a x^{a-1}\exp(-bx)x^2\,dx\\
&=\int_0^{+\infty}\frac{1}{b^2\Gamma(a)}(bx)^a\exp(-bx)(bx)^2\frac{d(bx)}{bx}\\
&=\int_0^{+\infty}\frac{1}{b^2\Gamma(a)}u^{a+2}\exp(-u)\frac{du}{u}\ (bx:=u)\\
&=\frac{1}{b^2\Gamma(a)}\Gamma(a+2)\\
&=\frac{(a+1)a\Gamma(a)}{b^2\Gamma(a)}\\
&=\frac{(a+1)a}{b^2}.
\end{aligned}\tag{3.19}$$

When $a=1$ in the Gamma distribution, X obeys the exponential distribution

$$X\sim b\exp(-bx),\quad b>0.$$

Typically we write the parameter b as λ. Therefore,

$$X\sim \mathrm{Exp}(\lambda)=\lambda\exp(-\lambda x),\lambda>0.$$

Obviously,

$$E[\mathrm{Exp}(\lambda)]=1/\lambda,\quad \mathrm{Var}[\mathrm{Exp}(\lambda)]=1/\lambda^2.$$

3.7. **Gaussian-Gamma distribution.** Let (μ, λ) be a random vector where $\mu \in \mathbb{R}^1$ and $\lambda > 0$. We say (μ, λ) obeys the Gaussian-Gamma distribution with parameters μ_0, β, a, b iff

$$p(\mu, \lambda|\mu_0, \beta, a, b) = N(\mu|\mu_0, 1/(\beta\lambda))\mathrm{Gamma}(\lambda|a, b).$$

It is easy to check that

$$\begin{aligned}
&\int_{\mu=0}^{+\infty}\int_{\lambda=0}^{+\infty} N(\mu|\mu_0, 1/(\beta\lambda))\mathrm{Gamma}(\lambda|a, b)\, d\lambda d\mu \\
=& \int_{\mu=0}^{+\infty}\int_{\lambda=0}^{+\infty} \frac{1}{\sqrt{2\pi(\beta\lambda)^{-1}}} \exp(-(\mu-\mu_0)^2/2(\beta\mu)^{-1}) \frac{1}{\Gamma(a)} b^a \lambda^{a-1} \exp(-b\lambda)\, d\lambda d\mu \\
=& \int_{\lambda=0}^{+\infty} \Big\{ \int_{\mu=0}^{+\infty} \frac{1}{\sqrt{2\pi(\beta\lambda)^{-1}}} \exp(-(\mu-\mu_0)^2/2(\beta\mu)^{-1})\, d\mu \Big\} \frac{1}{\Gamma(a)} b^a \lambda^{a-1} \exp(-b\lambda)\, d\lambda \\
=& \int_{\lambda=0}^{+\infty} \frac{1}{\Gamma(a)} b^a \lambda^{a-1} \exp(-b\lambda)\, d\lambda \\
=& 1.
\end{aligned} \tag{3.20}$$

3.8. **The multivariate Gaussian.** Let $X \in \mathbb{R}^D$. We say X obeys multivariate Gaussian distribution iff X has the following pdf:

$$N(\mathbf{x}|\mu, \Sigma) = \frac{1}{(2\pi)^{D/2}|\Sigma|^{1/2}} \exp\Big\{ -\frac{1}{2}(\mathbf{x}-\mu)^T \Sigma^{-1} (\mathbf{x}-\mu) \Big\}, \tag{3.21}$$

with

$$E[X] = \mu, \quad \mathrm{Var}[X] = E[(X-\mu)(X-\mu)^T] = \Sigma.$$

The matrix $\Lambda := \Sigma^{-1}$ is called the *precision.*

3.9. **The Student's t-distribution.** As the conjugate prior for the precision in the univariate normal distribution is the Gamma distribution. Assume that we have a univariate Gaussian $N(x|\mu, \tau^{-1})$ where $\tau > 0$ is the precision. and that τ has a prior $\mathrm{Gamma}(\tau|a, b)$.

Then the marginal distribution of x is

$$\begin{aligned}
p(x|\mu,a,b) &= \int_{\tau=0}^{+\infty} N(x|\mu,\tau^{-1})\mathrm{Gamma}(\tau|a,b)\,d\tau \\
&= \int_{\tau=0}^{+\infty} \frac{1}{\sqrt{2\pi\tau^{-1}}}\exp\Big\{-\frac{1}{2}(x-\mu)\tau(x-\mu)\Big\}\frac{1}{\Gamma(a)}b^a\tau^{a-1}\exp(-b\tau)\,d\tau \\
&= \int_{\tau=0}^{+\infty} \frac{\tau^{1/2}}{\sqrt{2\pi}}\exp\Big\{-\frac{1}{2}(x-\mu)^2\tau\Big\}\frac{1}{\Gamma(a)}b^a\tau^{a-1}\exp(-b\tau)\,d\tau \\
&= \int_{\tau=0}^{+\infty} \frac{\tau^{1/2}}{\sqrt{2\pi}}\exp\Big\{-\frac{1}{2}(x-\mu)^2\tau\Big\}\frac{1}{\Gamma(a)}b^a\tau^{a}\exp(-b\tau)\,\frac{d\tau}{\tau} \\
&= \frac{b^a}{\Gamma(a)\sqrt{2\pi}}\int_{\tau=0}^{+\infty}\exp\Big\{-\tau[b+(x-\mu)^2/2]\Big\}\tau^{a+1/2}\,d\tau \\
&= \frac{b^a}{\Gamma(a)\sqrt{2\pi}}[b+(x-\mu)^2/2]^{-a-1/2}\int_{z=0}^{+\infty}\exp(-z)z^{a+1/2}\,\frac{dz}{z} \quad (\tau[b+(x-\mu)^2/2] := z) \\
&= \frac{b^a}{\Gamma(a)\sqrt{2\pi}}[b+(x-\mu)^2/2]^{-a-1/2}\Gamma(a+1/2)
\end{aligned} \tag{3.22}$$

Letting $\nu = 2a$ and $\lambda = a/b$, we have

$$\begin{aligned}
p(x|\mu,a,b) &= \frac{b^a}{\Gamma(a)\sqrt{2\pi}}[b+(x-\mu)^2/2]^{-a-1/2}\Gamma(a+1/2) \\
&= \frac{\Gamma(\nu/2+1/2)}{\Gamma(\nu/2)}\frac{1}{(2\pi)^{1/2}}\Big(\frac{\nu}{2\lambda}\Big)^{\nu/2}[\nu/(2\lambda)+(x-\mu)^2/2]^{-\nu/2-1/2} \\
&= \frac{\Gamma(\nu/2+1/2)}{\Gamma(\nu/2)}\Big[\frac{1}{(2\pi)^{1/2}}\Big(\frac{\nu}{2\lambda}\Big)^{-1/2}\Big]\Big[\Big(\frac{\nu}{2\lambda}\Big)^{\nu/2+1/2}[\nu/(2\lambda)+(x-\mu)^2/2]^{-\nu/2-1/2}\Big] \\
&= \frac{\Gamma(\nu/2+1/2)}{\Gamma(\nu/2)}\Big[\frac{\lambda}{\pi\nu}\Big]^{1/2}\Big[1+\frac{\lambda(x-\mu)^2}{\nu}\Big]^{-\nu/2-1/2}.
\end{aligned} \tag{3.23}$$

Definition 3.9. A real-valued random variable X is said to obey the Student's t-distribution iff it has pdf

$$\mathrm{Student}(x|\mu,\lambda,\nu) = \frac{\Gamma(\nu/2+1/2)}{\Gamma(\nu/2)}\Big[\frac{\lambda}{\pi\nu}\Big]^{1/2}\Big[1+\frac{\lambda(x-\mu)^2}{\nu}\Big]^{-\nu/2-1/2}, \tag{3.24}$$

where λ is called precision, ν is called the degrees of freedom, and $E[X] = \mu$.

When $\nu = 1$, the Student's t-distribution is called Cauchy distribution with the pdf becoming

$$\mathrm{Cauchy}(x|\mu,\lambda) = \frac{\lambda^{1/2}}{\pi}\frac{1}{1+\lambda(x-\mu)^2}. \tag{3.25}$$

In particular, we have

$$\mathrm{Cauchy}(x|0,1) = \frac{1}{\pi}\frac{1}{1+x^2}. \tag{3.26}$$

It is easy to verify that

$$E[\mathrm{Cauchy}(x|\mu,\lambda)] = \mu, \quad E[\mathrm{Cauchy}^2(x|\mu,\lambda)] = +\infty.$$

When $\nu \to +\infty$, we have Student$(x|\mu, \lambda, \nu) \to N(x|\mu, \lambda^{-1})$. This is an easy consequence of Stirling's formula and $\lim_{\delta\to 0}(1+\delta)^{1/\delta} = e$:

$$\begin{aligned} & \frac{\Gamma(\nu/2+1/2)}{\Gamma(\nu/2)}\Big[\frac{\lambda}{\pi\nu}\Big]^{1/2}\Big[1+\frac{\lambda(x-\mu)^2}{\nu}\Big]^{-\nu/2-1/2} \\ \sim\ & (\nu/2)^{1/2}\Big[\frac{\lambda}{\pi\nu}\Big]^{1/2}\Big\{\Big[1+\frac{\lambda(x-\mu)^2}{\nu}\Big]^{\nu/(\lambda(x-\mu)^2)}\Big\}^{-\frac{\lambda}{2}(x-\mu)^2}\Big[1+\frac{\lambda(x-\mu)^2}{\nu}\Big]^{-1/2} \\ \sim\ & \Big[\frac{\lambda}{2\pi}\Big]^{1/2}\exp\Big\{-\frac{\lambda}{2}(x-\mu)^2\Big\} \\ \sim\ & N(x|\mu,\lambda^{-1}). \end{aligned} \tag{3.27}$$

Proposition 3.10. *Let X be a random variable with pdf $f(x)$. Then $Y = aX + b$ has pdf $\frac{1}{|a|}f(\frac{y-b}{a})$ for $a \neq 0$.*

Proof. First, consider the case $a > 0$.

$$P[Y \leqslant y] = P[X \leqslant (y-b)/a] = \int_{-\infty}^{(y-b)/a} f(x)\,dx = \int_{-\infty}^{y}\frac{1}{a}f((y-b)/a)\,dy.$$

For the case $a < 0$, we have

$$\begin{aligned} P[Y \leqslant y] = P[X \geqslant (y-b)/a] &= \int_{(y-b)/a}^{+\infty} f(x)\,dx \\ &= \int_{y}^{-\infty}\frac{1}{a}f((y-b)/a)\,dy = \int_{-\infty}^{y}\frac{1}{-a}f((y-b)/a)\,dy. \end{aligned}$$

□

Remark 3.11. The general form of the above proposition is the change of measures formula. It is the idea that is important here. Consider X with pdf $p(x)$, and we want to get the pdf of $Y = f(X)$. We do the following for $y \in \text{Range}(Y)$:

$$P[Y \leqslant y] = P[f(X) \leqslant y] = \int_{\{x; f(x)\leqslant y\}} p(x)\,dx.$$

then we use

$$\text{pdf}_Y(y) = \frac{d}{dy}P[Y \leqslant y] = \frac{d}{dy}\int_{\{x; f(x)\leqslant y\}} p(x)\,dx.$$

In particular, if f is monotonically increasing, we have

$$\text{pdf}_Y(y) = \frac{d}{dy}\int_{-\infty}^{f^{-1}(y)} p(x)\,dx = p(f^{-1}(y))\frac{df^{-1}(y)}{dy}.$$

In (3.22), we use the following change of variables

$$\nu = 2a, \quad \lambda = a/b, \quad \eta = \tau b/a.$$

As $\tau \sim \text{Gamma}(\tau|a,b)$, then η has pdf $\frac{1}{b/a}\text{Gamma}(a\eta/b|a,b)$ which is

$$\begin{aligned}
&\frac{a}{b}\frac{1}{\Gamma(a)}b^a(a\eta/b)^{a-1}\exp\{-b\frac{a\eta}{b}\}\\
=&\frac{1}{\Gamma(a)}a^a\eta^{a-1}\exp\{-a\eta\}\\
=&\text{Gamma}(\eta/|a,a)=\text{Gamma}(\eta/|\nu/2,\nu/2).
\end{aligned}\tag{3.28}$$

Now, we can write (3.22) as follows, which is useful to generalize Student's t-distribution to multi-dimensional case

$$\text{Student}(x|\mu,\lambda,\nu)=\int_0^{+\infty}N(x|\mu,(\eta\lambda)^{-1})\text{Gamma}(\eta/|\nu/2,\nu/2)\,d\eta.\tag{3.29}$$

Now in (3.29), we use multivariate normal $N(\mathbf{x}|\mu,\mathbf{\Lambda})$ to get D-dimensional Student's t-distribution

$$\begin{aligned}
\text{Student}(\mathbf{x}|\mu,\mathbf{\Lambda},\nu)&=\int_0^{+\infty}N(\mathbf{x}|\mu,(\eta\mathbf{\Lambda})^{-1})\text{Gamma}(\eta/|\nu/2,\nu/2)\,d\eta\\
&=\int_{\eta=0}^{+\infty}\frac{\eta^{D/2}|\Lambda|^{1/2}}{(2\pi)^{D/2}}\exp\{-\frac{\eta\Delta^2}{2}\}\frac{1}{\Gamma(\nu/2)}\eta^{\nu/2-1}\exp\{-\nu\eta/2\}\,d\eta\\
&=\int_{\eta=0}^{+\infty}\frac{\eta^{D/2}|\Lambda|^{1/2}}{(2\pi)^{D/2}}\exp\{-\frac{\eta}{2}[\Delta^2+\nu]\}\eta^{\nu/2-1}\,d\eta\;\;((\mathbf{x}-\mu)^T\Lambda(\mathbf{x}-\mu):=\Delta^2)\\
&=\int_{z=0}^{+\infty}\frac{(\nu/2)^{\nu/2}}{\Gamma(\nu/2)}\frac{|\Lambda|^{1/2}}{(2\pi)^{D/2}}\exp\{-z\}[2z(\Delta^2+\nu)^{-1}]^{D/2+\nu/2-1}[\Delta^2+\nu]^{-1}z\,dz\\
&\quad(\frac{\eta}{2}(\Delta^2+\nu):=z)\\
&=\frac{\Gamma(D/2+\nu/2)}{\Gamma(\nu/2)}\frac{|\Lambda|^{1/2}}{(\pi\nu)^{D/2}}\Big[1+\Delta^2/\nu\Big]^{-D/2-\nu/2}.
\end{aligned}\tag{3.30}$$

If $X\sim\text{Student}(\mathbf{x}|\mu,\mathbf{\Lambda},\nu)$, we can also verify that

$$E[X]=\mu,\quad \nu>1,$$

$$\text{Var}[X]=\frac{\nu}{(\nu-2)}|\Lambda|^{-1},\quad \nu>2,$$

and

$$\text{mode}[X]=\mu.$$

3.10. **The χ^2 distribution.**

(a) Let us first state a simple proposition.

Proposition 3.12. *If $X\sim N(0,1)$, then X^2 has pdf*

$$f(x)=\begin{cases}0, & x\leqslant 0,\\ \frac{1}{\sqrt{2\pi}}x^{-1/2}\exp\{-x/2\}, & x>0.\end{cases}\tag{3.31}$$

Proof. It is sufficient to consider the probability $P[X^2 \leqslant x]$ for $x > 0$. We have

$$\begin{aligned} P[X^2 \leqslant x] = P[-\sqrt{x} \leqslant X \leqslant \sqrt{x}] &= \int_{-\sqrt{x}}^{\sqrt{x}} \frac{1}{\sqrt{2\pi}} \exp\{-t^2/2\}\, dt \\ &= 2\int_0^{\sqrt{x}} \frac{1}{\sqrt{2\pi}} \exp\{-t^2/2\}\, dt \quad (t^2 := u) \\ &= \int_0^u \frac{1}{\sqrt{2\pi}} u^{-1/2} \exp\{-u/2\}\, du. \end{aligned} \tag{3.32}$$

□

Remark 3.13. We can do, as in Remark 3.11, the following proof which is more general

$$\begin{aligned} &\frac{d}{dx}\int_{-\sqrt{x}}^{\sqrt{x}} \frac{1}{\sqrt{2\pi}} \exp\{-t^2/2\}\, dt \\ =\ &\frac{1}{\sqrt{2\pi}} \exp\{-t^2/2\}\Big|_{t=\sqrt{x}} \frac{d\sqrt{x}}{dx} - \frac{1}{\sqrt{2\pi}} \exp\{-t^2/2\}\Big|_{t=-\sqrt{x}} \frac{d(-\sqrt{x})}{dx} \\ =\ &\frac{1}{\sqrt{2\pi}} \exp\{-x/2\} x^{-1/2}. \end{aligned} \tag{3.33}$$

Definition 3.14. Let $X_1, \cdots, X_n$ be i.i.d standard normal distributions. Then

$$\chi_n^2 := X_1^2 + \cdots + X_n^2$$

has pdf

$$\frac{1}{2^{n/2}\Gamma(n/2)} x^{n/2-1} \exp\{-x/2\}, \tag{3.34}$$

which is called the χ^2-distribution with n degrees of freedom.

Next, we show that $X_1^2 + \cdots + X_n^2$ has pdf (3.34). We know each X_i^2 has density given by (3.32). We will use convolution to give an inductive proof.

If $n = 2$, we know the density of $X_1^2 + X_2^2$ is given by $f * f$ as below

$$\begin{aligned} f_2(x) := f * f(x) &= \int_0^x \frac{1}{\sqrt{2\pi}} (x-y)^{-1/2} \exp\{-(x-y)/2\} \frac{1}{\sqrt{2\pi}} y^{-1/2} \exp\{-y/2\}\, dy \\ &= \Big[\frac{1}{\sqrt{2\pi}}\Big]^2 \exp\{-x/2\} \int_0^x (x-y)^{-1/2} y^{-1/2}\, dy \quad (y := tx) \\ &= \Big[\frac{1}{\sqrt{2\pi}}\Big]^2 \exp\{-x/2\} \int_0^1 (1-t)^{-1/2} t^{-1/2}\, dt \quad (y := tx) \\ &= \Big[\frac{1}{\sqrt{2\pi}}\Big]^2 \exp\{-x/2\} B(1/2, 1/2) \\ &= \Big[\frac{1}{\sqrt{2\pi}}\Big]^2 \exp\{-x/2\} \frac{\Gamma(1/2)\Gamma(1/2)}{\Gamma(1)} \\ &= \frac{1}{2^{2/2}\Gamma(2/2)} x^{2/2-1} \exp\{-x/2\}. \end{aligned} \tag{3.35}$$

Assume the conclusion holds for n. Let us consider the $n+1$ case. We have

$$\begin{aligned} f_{n+1} := f_n * f(x) &= \int_0^x \frac{1}{\sqrt{2\pi}}(x-y)^{-1/2}\exp\{-(x-y)/2\}\frac{1}{2^{n/2}\Gamma(n/2)}y^{n/2-1}\exp\{-y/2\} \\ &= \frac{1}{\sqrt{2\pi}2^{n/2}\Gamma(n/2)}x^{n/2-1/2}\exp\{-x/2\}\int_0^1 (1-t)^{-1/2}t^{n/2-1}\,dt \\ &= \frac{1}{\sqrt{2\pi}2^{n/2}\Gamma(n/2)}x^{n/2-1/2}\exp\{-x/2\}B(1/2,n/2) \\ &= \frac{1}{\sqrt{2\pi}2^{n/2}\Gamma(n/2)}x^{n/2-1/2}\exp\{-x/2\}\Gamma(1/2)\Gamma(n/2)/\Gamma(n/2+1/2) \\ &= \frac{1}{2^{(n+1)/2}\Gamma((n+1)/2)}x^{(n+1)/2-1}\exp\{-x/2\}. \end{aligned} \tag{3.36}$$

We can also verify that

$$E[\chi^2(n)] = n, \quad \mathrm{Var}[\chi^2(n)] = 2n. \tag{3.37}$$

First, in view of $E[X_i^2]=1$, it is easy to see that

$$E[\chi^2(n)] = E[X_1^2] + \cdots E[X_n^2] = n.$$

Now, in view of independence of the X_i, hence independence of X_i^2, we have

$$\mathrm{Var}[\chi^2(n)] = n\mathrm{Var}[X^2] = n(E[X^4]-(E[X^2])^2) = n(E[X^4]-1) = 2n,$$

where we have used the fact $E[X^4]=3$. This fact can be verified using characteristic function $E[e^{itX}] = e^{-t^2/2}$ for a standard normal distribution X. Hence $E[X^4] = \frac{d^4}{dt^4}\Big|_{t=0} e^{-t^2/2} = 3$.

3.11. **Periodic random variables: the von Mises distribution.** Consider the random variable Θ which is periodic with period 2π. Assume Θ has pdf $p(\theta)$ and so $p(\theta)$ satisfies

$$p(\theta) \geqslant 0, \quad \int_0^{2\pi} p(\theta)\,d\theta = 1, \quad p(\theta+2\pi) = p(\theta).$$

To derive the von Mises distribution, we consider the two dimensional Gaussian with mean $\mu = (\mu_1,\mu_2)^T$ and covariance $\Sigma = \sigma^2 I_2$. Then the pdf is

$$p(x_1,x_2) = \frac{1}{2\pi\sigma^2}\exp\Big\{-\frac{(x_1-\mu_1)^2+(x_2-\mu_2)^2}{2\sigma^2}\Big\}. \tag{3.38}$$

The contours of $p(x_1,x_2) =$ constant are circles

$$-\frac{(x_1-\mu_1)^2+(x_2-\mu_2)^2}{2\sigma^2} = \text{constant}.$$

Next, we shall condition on the unit circle $x_1^2+x_2^2=1$. Use polar coordinates

$$x_1 = r\cos\theta, \quad x_2 = r\sin\theta; \quad \mu_1 = r_0\cos\theta_0, \quad \mu_2 = r_0\sin\theta_0,$$

and consider the exponent with $r = 1$:

$$\begin{aligned}
&-\frac{(x_1-\mu_1)^2+(x_2-\mu_2)^2}{2\sigma^2}\\
&=\frac{-1}{2\sigma^2}\{(r\cos\theta-r_0\cos\theta_0)^2+(r\sin\theta-r_0\sin\theta_0)^2\}\\
&=\frac{-1}{2\sigma^2}\{1+r_0^2-2r_0\cos\theta\cos\theta_0-2r_0\sin\theta\sin\theta_0\}\\
&=\frac{r_0}{\sigma^2}\cos(\theta-\theta_0)+\text{constant independent of }\theta\\
&=m\cos(\theta-\theta_0)+\text{constant independent of }\theta\quad(m:=\frac{r_0}{\sigma^2}>0).
\end{aligned}\tag{3.39}$$

Definition 3.15. (von Mises, or circular normal) Θ obeys von Mises distribution if it has pdf

$$p(\theta|\theta_0,m)=\frac{1}{2\pi I_0(m)}\exp\{m\cos(\theta-\theta_0)\},\tag{3.40}$$

where $E[\Theta]=m>0$ and m is the concentration parameter playing the role precision $1/\sigma^2$ and

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

When $m=0$, the von Mises distribution reduces to uniform distribution on $[0,2\pi]$:

$$p(\theta|\theta_0,0)=\frac{1}{2\pi}I_{[0,2\pi]}(\theta).\tag{3.41}$$

When $m\to+\infty$, the von Mises distribution turns to be $N(\theta|\theta_0,1/m)$.

3.12. The exponential family of pdfs.

Definition 3.16. We say that the random variable X belongs to the exponential family if it has the pdf of the form

$$p(\mathbf{x}|\eta)=h(\mathbf{x})g(\eta)\exp\{\eta^Tu(\mathbf{x})\},\tag{3.42}$$

where $\mathbf{x}$ can be continuous or discrete, η is called the natural parameters of the distribution, and $u(\mathbf{x})$ is some function of $\mathbf{x}$. The function $g(\eta)$ satisfies the normalization requirement:

$$g(\eta)\int h(\mathbf{x})\exp\{\eta^Tu(\mathbf{x})\}\,d\mathbf{x}=1.\tag{3.43}$$

The distributions we have seen above are all members of the exponential family. Let us look at several examples.

Bernoulli distribution. Let $X\in\{0,1\}$ and $\mu\in[0,1]$ and X has the pdf

$$p(x|\mu)=\mu^x(1-\mu)^{1-x}.$$

We can write the above pdf as follows

$$\begin{aligned}\mu^x(1-\mu)^{1-x} &= \exp\Big\{x\ln\mu + (1-x)\ln(1-\mu)\Big\}\\ &= (1-\mu)\exp\Big\{x\ln\frac{\mu}{1-\mu}\Big\} \quad [\text{If } \ln\frac{\mu}{1-\mu} := \eta, \text{then } \mu = \frac{1}{1+\exp\{-\eta\}} := \sigma(\eta)]\\ &= (1-\sigma(\eta))\exp\{\eta x\}\\ &= \sigma(-\eta)\exp\{\eta x\}.\end{aligned} \tag{3.44}$$

Therefore, the Bernoulli distribution takes the form

$$p(x|\eta) = \sigma(-\eta)\exp\{\eta x\},$$

which belongs to the exponential family with

$$u(x) = x, \quad h(x) \equiv 1, \quad g(\eta) = \sigma(-\eta).$$

Multinomial distribution[1]**.** The multinomial distribution pdf

$$p(\mathbf{x}|\mu) = \prod_{k=1}^{M} \mu_k^{x_k} = \exp\Big\{\sum_{k=1}^{M} x_k \ln\mu_k\Big\} \tag{3.45}$$

where $\mu_k \in [0,1]$ and $\sum_k \mu_k = 1$, $\mathbf{x}$ is binary vectors with $\sum_k x_k = 1$. Due to the constraint $\sum_k \mu_k = 1$, we assume $\ln\mu_k = \eta_k - S$, i.e.,

$$\mu_k = \exp\{\eta_k\}/\exp\{S\}, \quad k = 1, \cdots, K \quad \text{(softmax functions)}, \tag{3.46}$$

and we have $\exp\{S\} = \sum_{k=1}^{K}\exp\{\eta_k\}$. Now, the multinomial pdf can be written as

$$p(\mathbf{x}|\vec{\eta}) = \exp\Big\{\sum_{k=1}^{K} x_k(\eta_k - S)\Big\} = \exp\Big\{\mathbf{x}\cdot\vec{\eta} - S\Big\} = \frac{\exp\{\mathbf{x}\cdot\vec{\eta}\}}{\exp\{S\}}, \tag{3.47}$$

which belongs to the exponential family with

$$\mathbf{x} \longleftrightarrow \mathbf{x}, \quad u(\mathbf{x}) = \mathbf{x}, \quad g(\vec{\eta}) = \frac{1}{\sum_{k=1}^{K}\exp\{\eta_k\}}.$$

Normal distribution. Consider the pdf $N(x|\mu,\sigma^2)$:

$$\begin{aligned} N(x|\mu,\sigma^2) &= \frac{1}{(2\pi\sigma^2)^{1/2}}\exp\Big\{-\frac{1}{2\sigma^2}(x-\mu)^2\Big\}\\ &= \frac{1}{(2\pi\sigma^2)^{1/2}}\exp\Big\{-\frac{1}{2\sigma^2}x^2 + \frac{\mu}{\sigma^2}x - \frac{1}{2\sigma^2}\mu^2\Big\}\\ &= h(x)g(\vec{\eta})\exp\{\eta^T u(x)\}\end{aligned} \tag{3.48}$$

where $h(x) \equiv \frac{1}{(2\pi)^{1/2}}$, $u(x) = (x, x^2)^T$, $\vec{\eta} = (\mu/\sigma^2, -1/2\sigma^2)^T$ and $g(\vec{\eta}) = (-2\eta_2)^{1/2}\exp\{\eta_1^2/4\eta_2\}$. Therefore, the normal distribution belongs to the exponential family.

[1] See also Section 6.

3.13. **Poisson Distribution.** Let X be a discrete random variable taking nonnegative integer values. X is the Possion distribution with parameter $\lambda > 0$ iff

$$P[X = k] = e^{-\lambda}\lambda^k/k!, \quad k = 0, 1, 2, \cdots$$

It is easy to verify $\sum_k P[X = k] = 1$, and

$$E[X] = \lambda, \quad \mathrm{Var}[X^2] = \lambda.$$

Actually, we can compute the characteristic function $\varphi(t) := E[e^{itX}]$ of X:

$$\begin{aligned} E[e^{itX}] &= \sum_k e^{itk} P[X = k] = \sum_k e^{itk} e^{-\lambda}\lambda^k/k! \\ &= e^{-\lambda}\sum_k \frac{(\lambda e^{it})^k}{k!} \\ &= \exp\big\{\lambda(e^{it} - 1)\big\}. \end{aligned}$$

Proposition 3.17. *(Sum of independent Possion random variables) Let $X \sim Poisson(\lambda)$ and $Y \sim Possion(\mu)$ for some $\lambda > 0$ and $\mu > 0$ be two independent Poisson random variables. Then $X + Y$ is still Poisson: $X + Y \sim Poisson(\lambda + \mu)$.*

Proof. As X and Y are independent, the pmf of $X + Y$ is the convolution of those of X and Y. Notice also that X and Y both take nonnegative integer values. We have for any nonnegative integer s that

$$\begin{aligned} P[X + Y = s] &= \sum_{k=0}^{s} P[X = s - k]P[Y = k] \\ &= \sum_{k=0}^{s} e^{-\lambda}\frac{\lambda^{s-k}}{(s-k)!}e^{-\mu}\frac{\mu^k}{k!} \\ &= \frac{e^{-(\lambda+\mu)}}{s!}\sum_{k=0}^{s}\frac{s!}{(s-k)!k!}\lambda^{s-k}\mu^k \\ &= e^{-(\lambda+\mu)}\frac{(\lambda+\mu)^s}{s!} \quad \text{(Binomial Theorem)}. \end{aligned} \tag{3.49}$$

□

Proposition 3.18. *(Stirling's formula, Problem 27.18, Page 370 in Billingsley) Let $S_n := X_1 + X_2 + \cdots + X_n$, where X_n are independent and each has Poisson distribution with parameter 1. Define $X^- := \max\{-X, 0\} \geqslant 0$ for a random variable. Then, with respect to n,*

(a) $E\Big[\Big(\frac{S_n - n}{\sqrt{n}}\Big)^-\Big] = e^{-n}\sum_{k=0}^{n}\Big(\frac{n-k}{\sqrt{n}}\Big)\frac{n^k}{k!} = \frac{n^{n+1/2}e^{-n}}{n!}$.

(b) $\Big(\frac{S_n - n}{\sqrt{n}}\Big)^- \Longrightarrow N^-$.

Consequently, the following hold

$$E\Big[\Big(\frac{S_n - n}{\sqrt{n}}\Big)^-\Big] \to E[N^-] = \frac{1}{\sqrt{2\pi}}$$

and

$$n! \sim \sqrt{2\pi n}\Big(\frac{n}{e}\Big)^n.$$

Let us give some digestion about N^-. $N^- = \max\{-N, 0\}$. When $N \geqslant 0$, we have $N^- = 0$; when $N < 0$, we have $N^- = -N > 0$. Therefore,

$$P[N^- = 0] = P[N \geqslant 0] = 1/2,$$

and

$$P[N^- > x] = P[-N > x] = P[N < -x] = \int_{-\infty}^{-x} \frac{1}{\sqrt{2\pi}} e^{-t^2/2}\,dt.$$

Therefore, the distribution of N^- is a mixture of pmf and pdf

$$\begin{cases} P[N^- = 0] = 1/2, \\ P[N^- \leqslant x] = 1/2 + \int_0^x \frac{1}{\sqrt{2\pi}} e^{-t^2/2}\,dt. \end{cases} \tag{3.50}$$

For a nonnegative random variable X, its expectation $E[X]$ can be computed using

$$E[X] = \int_0^{+\infty} P[X > t]\,dt.$$

Therefore, for N^-, we have

$$\begin{aligned} E[N^-] &= \int_0^{+\infty} P[N^- > x]\,dx = \int_0^{+\infty}\int_{-\infty}^{-x} \frac{1}{\sqrt{2\pi}} e^{-t^2/2}\,dt\,dx \\ &= \int_{-\infty}^{0} \Big\{ \int_0^{-t} \frac{1}{\sqrt{2\pi}} e^{-t^2/2}\,dx \Big\}\,dt \\ &= \int_{-\infty}^{0} \frac{1}{\sqrt{2\pi}} e^{-t^2/2}(-t)\,dt \\ &= \frac{1}{\sqrt{2\pi}}. \end{aligned} \tag{3.51}$$

We can use exactly the same idea to compute $E[N^+]$ which is defined as $N^+ = \max\{N, 0\}$. When $N \leqslant 0$, when have $N^+ = 0$; when $N > 0$, we have $N^+ > 0$. Therefore,

$$P[N^+ = 0] = P[N \leqslant 0] = 1/2,$$

and

$$P[N^+ > x] = P[N > x] = \int_x^{+\infty} \frac{1}{\sqrt{2\pi}} e^{-t^2/2}\,dt.$$

Therefore, the distribution of N^+ is a mixture of pmf and pdf

$$\begin{cases} P[N^+ = 0] = 1/2, \\ P[N^+ \leqslant x] = 1/2 + \int_0^x \frac{1}{\sqrt{2\pi}} e^{-t^2/2}\,dt. \end{cases} \tag{3.52}$$

Therefore, for N^+, we have

$$\begin{aligned} E[N^+] &= \int_0^{+\infty} P[N^+ > x]\,dx = \int_0^{+\infty}\int_x^{+\infty} \frac{1}{\sqrt{2\pi}} e^{-t^2/2}\,dt\,dx \\ &= \int_0^{+\infty} \Big\{ \int_0^t \frac{1}{\sqrt{2\pi}} e^{-t^2/2}\,dx \Big\}\,dt \\ &= \int_0^{+\infty} \frac{1}{\sqrt{2\pi}} e^{-t^2/2} t\,dt \\ &= \frac{1}{\sqrt{2\pi}}. \end{aligned} \tag{3.53}$$

In fact, comparing (3.51) and (3.52), we notice that they have the same distribution. This fact is due to the symmetry of the normal distribution N.

We can also see $E[|N|] = E[N^+ + N^-] = 2E[N^+] = 2E[N^-] = 2/\sqrt{2\pi} = \sqrt{2/\pi}$.

We first prove (a). The first equality in (a) is nothing but the definition of $E\Big[\Big(\frac{S_n - n}{\sqrt{n}}\Big)^-\Big]$. We only need to show

$$\sum_{k=0}^{n} \Big(\frac{n-k}{\sqrt{n}}\Big)\frac{n^k}{k!} = \frac{n^{n+1/2}}{n!}.$$

This is very easy by splitting the left hand side:

$$\Big(\frac{n-k}{\sqrt{n}}\Big)\frac{n^k}{k!} = \frac{n^{k+1/2}}{k!} - \frac{n^{k-1/2}}{(k-1)!}.$$

Due to cancellation, we have

$$\sum_{k=0}^{n} \Big(\frac{n-k}{\sqrt{n}}\Big)\frac{n^k}{k!} = \frac{n^{k+1/2}}{k!}\Big|_{k=n} = \frac{n^{n+1/2}}{n!}.$$

Now, we prove (b). Notice that $E[X_i] = 1$ and $\mathrm{Var}(X_i) = 1$. By Central Limit Theorem, we know

$$\frac{S_n - n}{\sqrt{n}} \Longrightarrow N(0,1).$$

Define the function $\phi(x)$ which is 0 for $x \geqslant 0$ and $-x$ for $x < 0$ (in fact, $\phi(x) = -\mathrm{ReLu}(-x)$). Obviously, $\phi(x)$ is a continuous function. Then, we have due to continuity of $\phi(x)$ and the definition of convergence in distribution that

$$\phi(\frac{S_n - n}{\sqrt{n}}) \Longrightarrow \phi(N(0,1)).$$

Then we observe that $\phi(X) = X^-$ for a random variable. Therefore, we have

$$\Big(\frac{S_n - n}{\sqrt{n}}\Big)^- \Longrightarrow N^-,$$

and (b) is proved.

Now, due to (b), we immediately have

$$E\Big[\Big(\frac{S_n - n}{\sqrt{n}}\Big)^-\Big] \to E[N^-] = \frac{1}{\sqrt{2\pi}}.$$

In view of (a), we conclude that

$$\frac{n^{n+1/2}e^{-n}}{n!} \to 1/\sqrt{2\pi}.$$

By now, we complete the proof.

4. Density estimation

1.**The concept of density estimation.**

Given a set of n data samples $\mathbf{x}_1, ..., \mathbf{x}_n$, we can estimate the density function $p(\mathbf{x})$, so that we can output $p(\mathbf{x})$ for any new sample x. This is called density estimation.

2. **Some nonparametric methods**.

In a broad sense, all probabilistic models are some ways of modeling the probabilistic distributions. Some nonparametric methods are:

- Histogram method
- Kernel method
- K nearest neighborhood method

We briefly discuss the kernel density method here. To estimate a pdf $p(x)$, we consider a small domain $\mathcal{R}$ in $\mathbb{R}^D$, the probability is $P = \int_{\mathcal{R}} p(x)\,dx$. When $|\mathcal{R}| := V$ is small, we may assume $P \approx p(x)V$. Assume we have N data points and each has probability P lying inside $\mathcal{R}$. So the number K of points inside $\mathcal{R}$ among the N data points is Binomial$(K|N,P)$. Binomial$(K|N,P)$ is peaked around $K \approx NP$. Therefore, we have

$$p(x) \approx P/V \approx K/NV. \tag{4.1}$$

Below, we shall take $\mathcal{R}$ to be a unit cube:

$$k(u) = \begin{cases} 1, & \text{if } |u_i| \leqslant 1/2, \forall\ i = 1, 2, \cdots D; \\ 0, & \text{otherwise.} \end{cases} \tag{4.2}$$

$k(\cdot)$ is the so-called *Parzen window.* The scaled function $k((x-x_n)/h)$ can be interpreted as the indicator function of unit cube centered at x_n or x with side length h. Then let $K = \sum_{i=1}^N k((x-x_n)/h)$ be the number of data x_n that lie insider a cube centered at x with side length 1. Therefore, the (4.1) can be written as

$$p(x) = \frac{1}{N}\frac{1}{h^D}\sum_{i=1}^N k((x-x_n)/h) = \frac{1}{N}\sum_{i=1}^N \frac{1}{h^D}k(\frac{x-x_n}{h}). \tag{4.3}$$

For inference, we can regard $\sum_{i=1}^N k((x-x_n)/h)$ as a count of the number of windows of x_n which contain x due to the symmetry of distance. Here $\frac{1}{h^D}k(\frac{x-x_n}{h}) := k_h(x-x_n)$ is the L^1 scaling of $k(\cdot)$.

Some thoughts. The kernel density methods suffer from the problem of discontinuity, for example, it may be due to the bins' boundaries in histogram method, or due to the choice of kernel functions in kernel method. The idea of approximation of identity may be used to smooth out the discontinuity. The main consideration is that we shall keep the normalization property of pdfs. Actually , this is guaranteed during the process. The process is as follows. Let $0 \leqslant \rho(x) \in C_c^\infty(\mathbb{R}^D)$ be a mollifier with $\int \rho(x)\, dx = 1$ and support $\bar{B}(0,1)$, and $\rho_\varepsilon(x) = \rho(x/\varepsilon)/\varepsilon^D$ is the L^1 scaling for a small positive parameter ε. Then given any kernel $k(x)$, we consider the convolution $k * \rho_\varepsilon(x) := \int k(x-y)\rho_\varepsilon(y)\, dy$. Then for any $\varepsilon > 0$, $k * \rho_\varepsilon$ is a valid kernel function for density estimation and we have $k * \rho_\varepsilon \geqslant 0$,

$$\text{support}(k * \rho_\varepsilon) \subset \{x; \text{dist}(x, \text{support}(k(\cdot))) \leqslant \varepsilon\},$$

and

$$|k * \rho_\varepsilon|_{L^1} = |k|_{L^1}.$$

A reference for the kernel density estimation. https://arxiv.org/abs/1704.03924.

5. The Robbins-Monro algorithm

(a) The problem. Consider two random variables θ and z governed by the joint distribution $p(\theta, z)$. The conditional expectation of z given θ $E[z|\theta]$ is a function of θ:

$$f(\theta) := E[z|\theta] = \int zp(z|\theta)\, dz. \tag{5.1}$$

We call $E[z|\theta]$ a *regression function*. The problem is to find the root θ^* of $f(\theta)$.

(b) The Robbins-Monro algorithm. We assume that $E[(z-f)|\theta] < \infty$ and that f is increasing near θ^*. The iteration step is given below:

$$\theta^{(N)} = \theta^{(N-1)} + a_{N-1} z(\theta^{(N-1)}), \tag{5.2}$$

where the sequence $\{a_N\}$ of positive numbers satisfies:

$$\lim_{N\to\infty} a_N = 0, \quad \sum_{N=1}^{\infty} a_N = \infty, \quad \sum_{N=1}^{\infty} a_N^2 < \infty.$$

See Robbins and Monro (1951), Blum (1965) for mathematical details.

6. The softmax functions and cross entropy

Assume P, Q are two probability distributions on the same σ-field $(\Omega, \mathcal{F})$. Then the cross entropy $H(P,Q)$ is defined as

$$H(P,Q) = E_P[-\log Q]. \tag{6.1}$$

If both P and Q are absolutely continuous with an Borel measure, for example, the m-dimensional Lebesgue measure dx

$$\frac{dP}{dx} = p(x), \quad \frac{dQ}{dx} = q(x), \tag{6.2}$$

then

$$H(P,Q) = \int_{\mathbb{R}^m} -p(x) \log q(x)\, dx \tag{6.3}$$

It is easy to verify that

$$H(P,Q) = H(P) + D_{KL}(P||Q) \tag{6.4}$$

where

$$H(P) = E_P[-\log P] = -\int p(x)\log p(x)\,dx$$

and

$$D_{KL}(P||Q) = \int p(x)\log\frac{p(x)}{q(x)}\,dx.$$

The softmax functions for K-class classification are

$$y_i = \frac{\exp(z_i)}{\sum_{k=1}^{K}\exp(z_k)}, \quad 1 \leqslant i \leqslant K, \tag{6.5}$$

which form a probability distribution.

Similar to that of a Sigmoid function, a good property of the softmax functions is

$$\frac{\partial y_i}{\partial z_i} = y_i(1-y_i). \tag{6.6}$$

More generally, we have

$$\frac{\partial y_j}{\partial z_i} = y_j(\delta_{ji} - y_i). \tag{6.7}$$

Assume the predicted probabilities for the K classes are $t_1, t_2, \cdots, t_K$, the the cross entropy cost function is

$$C = -\sum_{j=1}^{K} t_j \log y_j. \tag{6.8}$$

We also have

$$\begin{aligned}
\partial_{z_i} C &= \sum_j \partial_{y_j} C \partial_{z_i} y_j \\
&= \sum_j (-t_j/y_j) y_j (\delta_{ji} - y_i) \\
&= \sum_j -t_j(\delta_{ji} - y_i) \\
&= y_i - t_i.
\end{aligned}$$

7. Bayesian interpretation of weight constraint (a.k.a weight decay) in neural network

(1) Assume the interpretation of the network is $y_c = f(\text{input}_c; W)$ and

$$p(t_c|y_c) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp(-(t_c - y_c)^2/2\sigma^2).$$

Then, we see

$$-\log p(t_c|y_c) = \frac{(t_c - y_c)^2}{2\sigma^2} + \text{constant}. \tag{7.1}$$

Therefore, maximizing the log probability is equivalent to minimizing the squared distance for Gaussian prior.

(2) (Bayesian Theorem) From $P(W|D)P(D) = P(D,W) = P(D|W)P(W)$, we see

$$P(W|D) = \frac{P(D|W)P(W)}{P(D)}, \tag{7.2}$$

where $P(D)$ can be regarded as a normalization of the numerator

$$P(D) = \int_W P(D|W)P(W).$$

(3) Taking -log in (7.2), we have

$$\text{cost} := -\log P(W|D) = -\log P(D|W) - \log P(W) + \log P(D) \tag{7.3}$$

where $\log P(D)$ is independent of W and can be regarded as a constant in the optimization process.

Assume that $P(w_i) = \frac{1}{\sqrt{2\pi\sigma_{w_i}^2}} \exp(-w_i^2/2\sigma^2)$ and again

$$P(t_c|y_c) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp(-(t_c - y_c)^2/2\sigma^2).$$

Then, minimization in (7.3) is equivalent to the minimization

$$\text{cost} = \frac{1}{2\sigma_D^2} \sum_c (y_c - t_c)^2 + \frac{1}{2\sigma_w^2} \sum_i w_i^2 + \text{constant}, \tag{7.4}$$

and further equivalent to

$$2\sigma_D^2 \text{cost} = \sum_c (y_c - t_c)^2 + \frac{\sigma_D^2}{\sigma_w^2} \sum_i w_i^2 + \text{constant}. \tag{7.5}$$

8. Heuristics about "Ensemble average improves learning"

(1) Minimizing squared error function. Assume the N predictors' predictions for the ground truth are y_i, $1 \leqslant i \leqslant N$. Then there average would be

$$\bar{y} = \langle y_i \rangle_i = \frac{1}{N} \sum_{i=}^{N} y_i.$$

We examine the following equality

$$\begin{aligned} \langle (t - y_i)^2 \rangle_i &= \langle ((t - \bar{y}) + (\bar{y} - y_i))^2 \rangle_i \\ &= \langle (t - \bar{y})^2 \rangle_i + 2\langle (t - \bar{y})(\bar{y} - y_i) \rangle_i + \langle (\bar{y} - y_i)^2 \rangle_i \\ &= (t - \bar{y})^2 + \langle (\bar{y} - y_i)^2 \rangle_i + 2(t - \bar{y})\langle (\bar{y} - y_i) \rangle_i \\ &= (t - \bar{y})^2 + \langle (\bar{y} - y_i)^2 \rangle_i. \end{aligned}$$

From the above expression, we can conclude that

$$\langle(\bar{y}-y_i)^2\rangle_i \leqslant \langle(t-y_i)^2\rangle_i.$$

Actually, the above idea is equivalent to the fact that

$$\arg\min_c E[(X-c)^2] = E[X].$$

(2) Maximizing log probability. Assume N predictors predicts a class label with probability p_i for $1 \leqslant i \leqslant N$. Then due to concavity of log and Jensen's inequality, we have

$$\log \bar{p} = \log((p_1+p_2+\cdots+p_N)/N) \geqslant (\log p1 + \log p_2 + \cdots + \log p_N)/N. \tag{8.1}$$

9. Concavity of entropy

Proposition 9.1. *Let* $\mathbf{P} = \{p_1, p_2, \cdots, p_n\}$ *be a discrete probability distribution. Then the function* $H(\mathbf{P}) = -\sum_{i=1}^n p_i \log p_i$ *is concave.* $H(\mathbf{P})$ *attains its maximum at the uniform distribution* $\mathbf{P} = \{1/n, \cdots, 1/n\}$.

To show the above proposition, we first verify that $-D^2H(\mathbf{P})$ is positive definite. An easy computation shows that

$$\partial_{p_i}(-H) = 1 + \log p_i, \quad \partial^2_{p_ip_i}(-H) = \frac{1}{pi}\delta_{ij}.$$

Therefore, we know

$$-D^2H(\mathbf{P}) = \mathrm{diag}\{1/p_1, \cdots, 1/p_n\} > 0.$$

From the above computation, we know that the function $f(p) = -p\log p$ for $p \in I = (0,1)$ is concave. Note also f have two zeros 0, 1, and attains maximum at $p = 1/e$.

Second, the maximum of $H(\mathbf{P})$ can be computed using Lagrange Multiplier. Consider the function $F(\mathbf{P}, \lambda)$:

$$F(\mathbf{P}, \lambda) = H(\mathbf{P}) + \lambda(\sum_{i=1}^n p_i - 1).$$

It is easy to show that

$$\begin{cases} \partial_{p_i}F = -(1+\log p_i) + \lambda, \quad i = 1, 2, \cdots, n, \\ \partial_\lambda F = \sum_{i=1}^n p_i - 1. \end{cases} \tag{9.1}$$

Letting $\partial_{p_i}F = 0$ and $\partial_\lambda F = 0$, we have

$$p_i = \exp(\lambda - 1) = 1/n, \quad i = 1, 2, \cdots n.$$

Definition 9.2. Let X be a random variable on the probability space $(\Omega, \mathcal{F}, P)$ with probability density function $f(x)$ with respect to the Borel measure dx. Then the entropy $H(X) := H(f)$ is defined as

$$H(f) = -\int f(x) \log f(x)\, dx.$$

H is a functional on $(\Omega, \mathcal{F}, P)$. We could compute its first and second variations as follows. Let ϕ be a density function such that $\int \phi(x)\, dx = 0$ and $f + t\phi$ is still probability density function for $|t| < \varepsilon$. Then

$$\langle DH(f), \phi \rangle = \frac{d}{dt} H(f + t\phi)\Big|_{t=0} = -\int (1 + \log f)\phi\, dx,$$

$$\langle D^2 H(f)\phi, \phi \rangle = \frac{d^2}{dt^2} H(f + t\phi)\Big|_{t=0} = -\int \phi^2 / f\, dx.$$

Example 9.3. Let X obey uniform distribution on a interval $I = (a, b)$, i.e, $X \sim \text{Uniform}(a, b)$. The probability density function (pdf) for X is $p(x) = \frac{1}{b-a}$ for all $x \in (a, b)$ and $p(x) = 0$ for other x. Then the entropy of X is

$$H(X) = -\int_x f(x) \ln f(x)\, dx = -\int_a^b \frac{1}{b-a} \ln \frac{1}{b-a}\, dx = \ln(b-a). \tag{9.2}$$

From the above result, we see that the large the length $|b-a|$ is, the large the entropy $H(X)$ is for the uniform distribution. In particular, if $|b-a| = 1$, then $H(X) = 0$.

Example 9.4. The entropy of $f \sim \mathcal{N}(\mu, \sigma^2) \sim \frac{1}{\sqrt{2\pi\sigma^2}} \exp(-(x-\mu)^2/2\sigma^2)$.

$$\begin{aligned}
-H(f) &= \int \frac{1}{\sqrt{2\pi\sigma^2}} \exp(-(x-\mu)^2/2\sigma^2)(\log \frac{1}{\sqrt{2\pi\sigma^2}} - (x-\mu)^2/2\sigma^2) \\
&= \log \frac{1}{\sqrt{2\pi\sigma^2}} - \frac{1}{2\sigma^2} Var(\mathcal{N}(\mu, \sigma^2)) \\
&= \log \frac{1}{\sqrt{2\pi\sigma^2}} - 1/2 \\
&= -\frac{1}{2} \log(2\pi e \sigma^2).
\end{aligned}$$

Therefore, $H(f) = \frac{1}{2} \log(2\pi e \sigma^2)$

Proposition 9.5. *Let $X \sim f, Y \sim g$ be two independent random variables with joint probability density function $p(x, y) = f(x)g(y)$. Then $H(p) = H(f) + H(g)$.*

Proof.

$$\begin{aligned}
-H(p) &= \int p(x,y)\log p(x,y)\,dxdy \\
&= \int f(x)g(y)(\log f(x) + \log g(y))\,dxdy \\
&= \int f(x)g(y)\log f(x)\,dxdy + \int f(x)g(y)\log g(y)\,dxdy \\
&= \int f(x)\log f(x)\,dx \int g(y)\,dy + \int g(y)\log g(y)\,du \int f(x)\,dx \\
&= -H(f) - H(g).
\end{aligned}$$

□

Example 9.6. Let $X \sim \mathcal{N}(\mu_1, \sigma_1^2)\cdots\mathcal{N}(\mu_n, \sigma_n^2)$. Then from the above example about the entropy for one dimensional Gaussian, we know

$$H(X) = \sum_{i=1}^{n} \frac{1}{2}\log(2\pi e\sigma_i^2) = \frac{1}{2}\log\left((2\pi e)^n \sigma_1^2\cdots\sigma_n^2\right).$$

Proposition 9.7. *Consider probability density functions f belonging to the set*

$$\mathcal{A} = \{f; -\infty < E[f] = \mu < \infty, -\infty < E[f^2] = \sigma^2 < \infty\}.$$

Then

$$\arg\max\{f \in \mathcal{A}; H(f)\} = N(\mu, \sigma^2).$$

Proof. Directly use of variational methods with constraints. □

Definition 9.8. (Conditional entropy of Y given X, continuous version) Let $(X,Y) \sim p(x,y)$ and $p(Y|X) \sim p(y|x)$. Then the conditional entropy of Y given X is defined as

$$\begin{aligned}
H(Y|X) &= -\int p(x,y)\ln p(y|x)\,dydx \\
&= \int\int [-p(y|x)\ln p(y|x)]\,dyp(x)\,dx \\
&= \int H(Y|X=x)p(x)\,dx \\
&= E_{x\sim p(x)}[H(Y|X=x)],
\end{aligned} \tag{9.3}$$

where $H(Y|X=x) := \int[-p(y|x)\ln p(y|x)]\,dy$.

Remark 9.9. It is easy to see that

$$H(Y|X) = H(X,Y) - H(X).$$

Similarly, we see

$$H(X|Y) = H(X,Y) - H(Y).$$

Here $H(X,Y) = -\int p(x,y)\ln p(x,y)\,dxdy$.

Definition 9.10. (Conditional entropy of Y given X, discrete version) Assume (X,Y) obey the joint distribution

$$p(X=x_i, Y=y_j) = p_{ij}, \quad i = 1,2,\cdots,n; j = 1,2,\cdots,m,$$

and the marginal distribution

$$p(X = x_i) = p_i, \quad i = 1, 2, \cdots, n.$$

Then the conditional entropy $H(Y|X)$ is defined as

$$\begin{aligned} H(Y|X) &:= -\sum_{i,j} p_{ij} \ln p(Y = y_j|X = x_i) \\ &= \sum_j p_i \big(\sum_j [-p(Y = y_j|X = x_i) \ln p(Y = y_j|X = x_i)] \big) \\ &= \sum_i p_i H(Y|X = x_i), \end{aligned} \tag{9.4}$$

where $H(Y|X = x_i) = -\sum_j [p(Y = y_j|X = x_i) \ln p(Y = y_j|X = x_i)]$.

Example 9.11. (Empirical entropy and empirical relative entropy) Let D be the training dataset and $|D|$ be the number of instances in D. Assume there are K classes with class labels $C_1, \cdots, C_K$, and $|C_k|$ is the number of instances in D with class label C_k for $1 \leqslant k \leqslant K$. Therefore, $|D| = \sum_k |C_k|$. Assume that a specific feature A for instances of D has n levels, denoted by $\{a_1, \cdots, c_n\}$. According to feature A, the dataset is divided into n subsets $D_1, \cdots, D_n$, in other words, $D_i = \{x \in D; A(x) = a_i\}$ for $1 \leqslant i \leqslant n$, and $|D| = \sum_i |D_i|$. Assume among the subset D_i, the set of instances of D belonging to class C_k is D_{ik}, i.e., $D_{ik} = D_i \cap C_k$, and $|D_{ik}|$ is the number of instances. Then the entropy $H(D)$ of D with respect to the classification $C_1, \cdots, C_K$ is

$$H(D) = -\sum_k \frac{|C_k|}{|D|} \ln \frac{|C_k|}{|D|}.$$

The conditional entropy of $H(D|A)$ given A with respect to the classification is

$$H(D|A) = \sum_i \frac{|D_i|}{|D|} H(D_i) = -\sum_i \frac{|D_i|}{|D|} \sum_k \frac{|D_{ik}|}{|D_i|} \ln \frac{|D_{ik}|}{|D_i|}.$$

From here we can define the *information gain* of D due to A by

$$g(D, A) = H(D) - H(D|A)$$

and the *relative information gain* by

$$g_{relative}(D, A) = g(D, A)/H(D) = \frac{H(D) - H(D|A)}{H(D)} = 1 - H(D|A)/H(D).$$

Definition 9.12. (Kullback-Leibner divergence) Assume $p(x)$ and $q(x)$ are two pdfs. The Kullback-Leibner divergence $D_{KL}(p||q)$ is defined as

$$D_{KL}(p||q) = -\int p(x) \ln \frac{q(x)}{p(x)} \, dx.$$

Definition 9.13. (Cross-entropy) As above, the cross entropy $H_{cross}(p, q)$ is defined as

$$H_{cross}(p, q) = -\int p(x) \ln q(x) \, dx.$$

Remark 9.14. It is easy to see from the above two definitions that

$$D_{KL}(p||q) = H_{cross}(p, q) - H(p).$$

Meanwhile, by Jensen's inequality, we see that $D_{KL}(p||q)$ is always nonnegative, and is zero if and only if $p(x) = q(x)$ in law.

Definition 9.15. (Mutual information) Assume $(X,Y) \sim p(x,y)$ and $X \sim p(x)$ and $Y \sim p(y)$. Then the mutual information of X, Y is defined as

$$I(X,Y) = D_{KL}(p(x,y)||p(x)p(y)) = -\int p(x,y) \ln \frac{p(x)p(y)}{p(x,y)}\, dxdy.$$

Remark 9.16. When X, Y are independent, we have $p(x,y) = p(x)p(y)$, which implies $I(X,Y) = 0$, as desired. We can easily see

$$I(X,Y) = H(X) + H(Y) - H(X,Y) = H(X) - H(X|Y) = H(Y) - H(Y|X).$$

10. Principle Component Analysis (PCA)

Let $\mathbf{x}_i \in \mathbb{R}^m$ $(1 \leqslant i \leqslant n)$ be n points and $\mathbf{X} = \begin{bmatrix} \mathbf{x}_1^T \\ \mathbf{x}_2^T \\ \vdots \\ \mathbf{x}_n^T \end{bmatrix} \in \mathbb{R}^{n\times m}$ be the dataset. We assume for convenience $E[\mathbf{X}^{(j)}] = \vec{0} \in \mathbb{R}^n$ for each column of $\mathbf{X}$ for $1 \leqslant j \leqslant m$.

Definition 10.1. The PCA of $\mathbf{X}$ is the eigenvalue decomposition of the covariance matrix $\mathbf{X}^T\mathbf{X} \in \mathbb{R}^{m\times m}$.

Assume the eigenvalues in descending order are given by $\sigma_1^2 \geqslant \sigma_2^2 \geqslant \cdots \geqslant \sigma_m^2$, and the corresponding unit eigenvectors are $w_1, w_2, \cdots, w_m$.

Definition 10.2. The w_is are called *principle components*; the matrix $W = [w_1, w_2, \cdots, w_m] \in \mathbb{R}^{m\times m}$ is called *loadings*; $T = \mathbf{X}W \in \mathbb{R}^{n\times}$ is called *scores*.

We have

$$\mathbf{X}^T\mathbf{X} = W\text{diag}\{\sigma_1^2, \cdots, \sigma_m^2\}W^{-1} \tag{10.1}$$

Definition 10.3. (Dimension reduction) $W_r := [w_1, \cdots, w_r] \in \mathbb{R}^{m\times r}$ and $T_r := \mathbf{X}W_r \in \mathbb{R}^{n\times r}$ is called the projected data.

Reconstruction and reconstruction error. The reconstruction from projected data towards the original data is $\mathbf{X}_{\text{recovered}} = T_r W_d^T$, which equals $\mathbf{X}W_dW_d^T$. The construction error Error $= \|\mathbf{X}_{\text{recovered}} - \mathbf{X}\|$ is thus $\|\mathbf{X}(W_dW_d^T - \mathbf{I}_n)\|$ in a suitable norm.

Explained variance. Denote by $\sigma^2 = \sum_{i=1}^m \sigma_i^2$ the total variance of $\mathbf{X}$. The ratio of explained variance, denoted by $\eta^2(r)$, is defined by $\eta^2(r) = \frac{\sum_{i=1}^r \sigma_i^2}{\sigma^2}$.

Definition 10.4. The singular value decomposition (SVD) of $\mathbf{X}$ is defined as $\mathbf{X} = U\Sigma V^*$ where $U \in \mathbb{C}^{n\times n}$, $V \in \mathbb{C}^{m\times m}$ are unitary matrices, and $\Sigma \in \mathbb{C}^{n\times m}$ comprise the singular values $\{\sigma_1, \sigma_2, \cdots, \sigma_m\}$ of X is the main diagonal.

By SVD, we know that

$$X^*X = (U\Sigma V^*)^*(U\Sigma V^*) = V(\Sigma^*\Sigma)V^* = V\text{diag}\{\sigma_1^2, \cdots, \sigma_m^2\}V^*. \tag{10.2}$$

Comparing (10.1) and (10.2), we can identify, up to a sign, that

$$W = V.$$

Meanwhile,

$$T = XW = U\Sigma V^* W = U\Sigma V^* V = U\Sigma,$$

and

$$T_r = U_r \Sigma_r$$

with U_r being the $(n \times r)$-block of U and Σ_r the principle $(r \times r)$-block of Σ.

11. EM Algorithm

EM Algorithm, the expectation maximization algorithm, is a general method for finding maximum likelihood solutions for probabilistic model having latent variables (Dempster et al. 1977; McLachlan and Krishnan, 1997; Neal and Hinton, 1999).

Here we aim to demonstrate the EM algorithm (the $\mathcal{L}$-function maximization-maximization, the F-function maximization-maximization) in integral form.

Consider a probabilist model. Assume the observables $X_i \in \mathbb{R}^D$ and hidden variables $Z_i \in \mathbb{R}^K$ for $i = 1, 2, \cdots, N$. We shall denote $\mathbf{X} = (X_1, \cdots, X_N)^T \in \mathbb{R}^{N \times D}$ the observable dataset and $\mathbf{Z} = (Z_1, \cdots, Z_N)^T \in \mathbb{R}^{N \times K}$ the latent dataset. We assume that the complete data is $\{X, Z\}$ obey the joint distribution $p(\mathbf{X}, \mathbf{Z}|\theta)$ where θ is the model parameter. Our goal is to maximize to maximum likelihood function

$$p(\mathbf{X}|\theta) = \int_{\mathbf{Z}} p(\mathbf{X}, \mathbf{Z}|\theta). \tag{11.1}$$

Proposition 11.1. *(The decomposition of log maximum likelihood function) Assume the latent variable $\mathbf{Z}$ obeys the distribution $q(\mathbf{Z})$. Then for any choice of $q(\mathbf{Z})$, the following decomposition holds*

$$\ln p(\mathbf{X}|\theta) = \mathcal{L}(q, \theta) + KL(q||p(\mathbf{Z}|\mathbf{X}, \theta)) \tag{11.2}$$

where

$$\begin{aligned} \mathcal{L}(q, \theta) &= \int_{\mathbf{Z}} q(\mathbf{Z}) \ln \frac{p(\mathbf{X}, \mathbf{Z}|\theta)}{q(\mathbf{Z})} \\ &= \int_{\mathbf{Z}} q(\mathbf{Z}) \ln p(\mathbf{X}, \mathbf{Z}|\theta) + H(q). \end{aligned} \tag{11.3}$$

and

$$KL(q||p(\mathbf{Z}|\mathbf{X}, \theta)) = -\int_{\mathbf{Z}} q(\mathbf{Z}) \ln \frac{p(\mathbf{Z}|\mathbf{X}, \theta)}{q(\mathbf{Z})}. \tag{11.4}$$

Proof.

$$\begin{aligned}\mathcal{L}(q,\theta)+\mathrm{KL}(q||p(\mathbf{Z}|\mathbf{X},\theta)) &= \int_{\mathbf{Z}} q(\mathbf{Z})\Big(\ln\frac{p(\mathbf{X},\mathbf{Z}|\theta)}{q(\mathbf{Z})}-\ln\frac{p(\mathbf{Z}|\mathbf{X},\theta)}{q(\mathbf{Z})}\Big)\\ &= \int_{\mathbf{Z}} q(\mathbf{Z})\ln\frac{p(\mathbf{X},\mathbf{Z}|\theta)}{p(\mathbf{Z}|\mathbf{X},\theta)}\\ &= \int_{\mathbf{Z}} q(\mathbf{Z})\ln p(\mathbf{X}|\theta)\\ &= \ln p(\mathbf{X}|\theta).\end{aligned} \tag{11.5}$$

□

Remark 11.2. As $\mathrm{KL}(q||p(\mathbf{Z}|\mathbf{X},\theta))\geqslant 0$, we know that the function $\mathcal{L}(q,\theta)$ is a lower bound of the $\ln p(\mathbf{X}|\theta)$. In some books (e.g., Li Hang's), $\mathcal{L}(q,\theta)$ is called the F-function. Note also that Theorem 9.1 in Li Hang's book is a direct consequence of the above decomposition.

As we know, $\mathcal{L}(q,\theta)$ as a functional of $q(\mathbf{Z})$ is a linear perturbation of the entropy $H(q)$, hence, $\mathcal{L}(q,\theta)$ is a concave functional of $q(\mathbf{Z})$. We have the following proposition.

Proposition 11.3. *The maximum of $\mathcal{L}(q,\theta)$ over all admissible probability distributions $q(\mathbf{Z})$ is attained at the posterior distribution $\tilde{q}(\mathbf{Z})=p(\mathbf{Z}|\mathbf{X},\theta)$, and*

$$\mathcal{L}(\tilde{q},\theta)=\int_{\mathbf{Z}} p(\mathbf{Z}|\mathbf{X},\theta)\ln p(\mathbf{X},\mathbf{Z}|\theta)+H(p(\mathbf{Z}|\mathbf{X},\theta)).$$

When $q(\mathbf{Z})=\tilde{q}(\mathbf{Z})$, we have $\mathcal{L}(\tilde{q},\theta)=\ln p(\mathbf{X}|\theta)$.

Proof. (Lagrangian Multiplier method) Consider the first variation of $\mathcal{L}(q,\theta)$ with respect to $q(\mathbf{Z})$ under the constraint $\int_{\mathbf{Z}} q(\mathbf{Z})\,dZ=1$, i.e., consider the functional $f(q,\lambda)$:

$$f(q,\lambda):=\mathcal{L}(q,\theta)+\lambda(\int_{\mathbf{Z}} q(\mathbf{Z})\,dZ-1).$$

It is easy to obtain that

$$\langle D_q f,\phi\rangle=\int_{\mathbf{Z}}(\ln p(\mathbf{X},\mathbf{Z}|\theta)-\ln q(\mathbf{Z})-1+\lambda)\phi\,dZ \tag{11.6}$$

and $D_\lambda f=\int_{\mathbf{Z}} q(\mathbf{Z})-1$ where $\phi(\mathbf{Z})$ is such that $\int_{\mathbf{Z}}\phi(\mathbf{Z})\,dZ=0$.

Letting$\langle D_q f,\phi\rangle=0$, we have

$$\ln p(\mathbf{X},\mathbf{Z}|\theta)-\ln q(\mathbf{Z})-1+\lambda=C$$

for some constant C. Therefore, we see

$$p(\mathbf{X},\mathbf{Z}|\theta)=\exp(C+1-\lambda)q(\mathbf{Z}). \tag{11.7}$$

Integrating over Z in (11.7) and using $D_\lambda f=0$, we obtain

$$\int_{\mathbf{Z}} p(\mathbf{X},\mathbf{Z}|\theta)\,dZ=\exp(C+1-\lambda).$$

Then from (11.7), we obtain

$$\begin{aligned} q(\mathbf{Z}) &= p(\mathbf{X},\mathbf{Z}|\theta)/\exp(C+1-\lambda) = p(\mathbf{X},\mathbf{Z}|\theta)/\int_{\mathbf{Z}} p(\mathbf{X},\mathbf{Z}|\theta)\,dZ \\ &= p(\mathbf{X},\mathbf{Z}|\theta)/p(\mathbf{X}|\theta) \\ &= p(\mathbf{Z}|\mathbf{X},\theta). \end{aligned} \tag{11.8}$$

□

Definition 11.4. (Q-function) $Q(\theta,\theta^{(i)}) := \int_{\mathbf{Z}} p(\mathbf{Z}|\mathbf{X},\theta^{(i)}) \ln p(\mathbf{X},\mathbf{Z}|\theta)$ is known as the $\mathcal{Q}$-function for the EM algorithm.

Now, we are ready to state the EM algorithm based on Propositions 11.1 and 11.3. The algorithm can be understood as the maximization-maximization of the $\mathcal{L}$-function.

(**Maximization-maximization of the $\mathcal{L}$-function**.) Suppose the current parameter is $\theta^{(i)}$. (1) E-step. We maximize $\max_q \mathcal{L}(q,\theta^{(i)})$ to find $\arg\max_q \mathcal{L}(q,\theta^{(i)}) = p(\mathbf{Z}|\mathbf{X},\theta^{(i)}) := \tilde{q}$. Then we compute $\mathcal{L}(q,\theta^{(i)}) = E_{p(\mathbf{Z}|\mathbf{X},\theta^{(i)})}[\ln p(\mathbf{X},\mathbf{Z}|\theta)] + H(p(\mathbf{Z}|\mathbf{X},\theta^{(i)}))$ in order to maximize $\max_\theta \mathcal{L}(\tilde{q},\theta^{(i)})$. (2) M-step. We consider the problem $\max_\theta \mathcal{L}(\tilde{q},\theta^{(i)})$. As $H(p(\mathbf{Z}|\mathbf{X},\theta^{(i)}))$ is constant for the problem, we see

$$\arg\max_\theta \mathcal{L}(\tilde{q},\theta^{(i)}) = \arg\max_\theta E_{p(\mathbf{Z}|\mathbf{X},\theta^{(i)})}[\ln p(\mathbf{X},\mathbf{Z}|\theta)].$$

Note that $E_{p(\mathbf{Z}|\mathbf{X},\theta^{(i)})}[\ln p(\mathbf{X},\mathbf{Z}|\theta)]$ is exactly the Q-function. We iterate through the above two-stage procedure to approach the optimal parameter θ.

Using Q-function, we state the EM algorithm as follows

(**EM algorithm**.) Given observable data $\mathbf{X}$, assume $\mathbf{Z}$ is the latent data and $\{\mathbf{X},\mathbf{Z}\}$ obeys the joint distribution $p(\mathbf{X},\mathbf{Z}|\theta)$ and conditional distribution $p(\mathbf{Z}|\mathbf{X},\theta)$. We aim to find the model parameter θ. (1) Initialize $\theta^{(0)}$ and begin to iterate; (2) E-step: let $\theta^{(i)}$ be the parameter value for iteration i. In iteration $i+1$, compute the Q-function:

$$Q(\theta,\theta^{(i)}) = \int_{\mathbf{Z}} p(\mathbf{Z}|\mathbf{X},\theta^{(i)}) \ln p(\mathbf{X},\mathbf{Z}|\theta).$$

(3) M-step: compute $\theta^{(i+1)} = \arg\max\{\theta; Q(\theta,\theta^{(i)})\}$. (4) Repeat (2) and (3) until convergence or break at a threshold value.

(**Derivation of Q-function, another formulation of Proposition 11.1**.)

We aim to maximize $\ln p(\mathbf{X}|\theta) := L(\theta)$, i.e., the log maximum likelihood of the observable data with respect to θ via iteration. Assume the current parameter value is $\theta^{(i)}$. Consider the difference $L(\theta) - L(\theta^{(i)})$. We hope to find $\theta^{(i+1)}$ such that the difference is greater than zero.

First, we have

$$L(\theta) = \ln p(\mathbf{X}|\theta) = \ln \int_{\mathbf{Z}} p(\mathbf{Y},\mathbf{Z}|\theta)\,d\mathbf{Z} = \ln\Big(\int_{\mathbf{Z}} p(\mathbf{X}|\mathbf{Z},\theta)p(\mathbf{Z}|\theta)\,d\mathbf{Z}\Big). \tag{11.9}$$

Therefore,

$$\begin{aligned}
L(\theta) - L(\theta^{(i)}) &= \ln\Big(\int_{\mathbf{Z}} p(\mathbf{X}|\mathbf{Z},\theta)p(\mathbf{Z}|\theta)\,d\mathbf{Z}\Big) - \ln p(\mathbf{X}|\theta^{(i)}) \\
&= \ln\Big(\int_{\mathbf{Z}} \frac{p(\mathbf{X}|\mathbf{Z},\theta)p(\mathbf{Z}|\theta)}{p(\mathbf{Z}|\mathbf{X},\theta^{(i)})} p(\mathbf{Z}|\mathbf{X},\theta^{(i)})\,d\mathbf{Z}\Big) - \ln p(\mathbf{X}|\theta^{(i)}) \\
&\geqslant \int_{\mathbf{Z}} \ln\Big(\frac{p(\mathbf{X}|\mathbf{Z},\theta)p(\mathbf{Z}|\theta)}{p(\mathbf{Z}|\mathbf{X},\theta^{(i)})}\Big) p(\mathbf{Z}|\mathbf{X},\theta^{(i)})\,d\mathbf{Z}\Big) - \ln p(\mathbf{X}|\theta^{(i)}) \\
&= \int_{\mathbf{Z}} \ln\Big(\frac{p(\mathbf{X}|\mathbf{Z},\theta)p(\mathbf{Z}|\theta)}{p(\mathbf{Z}|\mathbf{X},\theta^{(i)})p(\mathbf{X}|\theta^{(i)})}\Big) p(\mathbf{Z}|\mathbf{X},\theta^{(i)})\,d\mathbf{Z}.
\end{aligned} \tag{11.10}$$

Denote by

$$B(\theta,\theta^{(i)}) = L(\theta^{(i)}) + \int_{\mathbf{Z}} \ln\Big(\frac{p(\mathbf{X}|\mathbf{Z},\theta)p(\mathbf{Z}|\theta)}{p(\mathbf{Z}|\mathbf{X},\theta^{(i)})p(\mathbf{X}|\theta^{(i)})}\Big) p(\mathbf{Z}|\mathbf{X},\theta^{(i)})\,d\mathbf{Z}.$$

We see that $L(\theta) \geqslant B(\theta,\theta^{(i)})$ and $L(\theta^{(i)}) = B(\theta^{(i)},\theta^{(i)})$. Therefore, in order to make $L(\theta)$ increase, it is sufficient to make $B(\theta,\theta^{(i)})$ increase.

Now, we consider the problem $\arg\max\{\theta; B(\theta,\theta^{(i)})\}$. Ignoring the irrelevant terms for the maximization, we have

$$\arg\max\{\theta; B(\theta,\theta^{(i)})\} = \arg\max\{\theta; \int_{\mathbf{Z}} \ln\big(p(\mathbf{X}|\mathbf{Z},\theta)p(\mathbf{Z}|\theta)\big) p(\mathbf{Z}|\mathbf{X},\theta^{(i)})\,d\mathbf{Z}\}.$$

The expression in the right hand side is nothing but the Q-function:

$$\int_{\mathbf{Z}} \ln\big(p(\mathbf{X}|\mathbf{Z},\theta)p(\mathbf{Z}|\theta)\big) p(\mathbf{Z}|\mathbf{X},\theta^{(i)})\,d\mathbf{Z} = Q(\theta,\theta^{(i)}).$$

(**Monotonicity of the maximum likelihood function sequence** $p(\mathbf{X}|\theta^{(i)})$.).

We aim to show that $p(\mathbf{X}|\theta^{(i+1)}) \geqslant p(\mathbf{X}|\theta^{(i)})$ for the EM iteration sequence. It is sufficient to show $\ln p(\mathbf{X}|\theta^{(i+1)}) \geqslant \ln p(\mathbf{X}|\theta^{(i)})$. By Bayes' Theorem, we know

$$\ln p(\mathbf{X}|\theta) = \frac{p(\mathbf{X},\mathbf{Z}|\theta)}{p(\mathbf{Z}|\mathbf{X},\theta)}.$$

Recall that

$$Q(\theta,\theta^{(i)}) = \int_{\mathbf{Z}} \ln\big(p(\mathbf{X},\mathbf{Z}|\theta)\big) p(\mathbf{Z}|\mathbf{X},\theta^{(i)})\,d\mathbf{Z}\}$$

and define

$$H(\theta,\theta^{(i)}) := \int_{\mathbf{Z}} \ln\big(p(\mathbf{Z}|\mathbf{X},\theta)\big) p(\mathbf{Z}|\mathbf{X},\theta^{(i)})\,d\mathbf{Z}\}.$$

It is easy to see

$$\ln p(\mathbf{X}|\theta) = Q(\theta,\theta^{(i)}) - H(\theta,\theta^{(i)}).$$

Therefore, we have

$$\begin{aligned}
\ln p(\mathbf{X}|\theta^{(i+1)}) - \ln p(\mathbf{X}|\theta^{(i)}) &= [Q(\theta^{(i+1)},\theta^{(i)}) - Q(\theta^{(i+1)},\theta^{(i)})] - [Q(\theta^{(i+1)},\theta^{(i)}) - Q(\theta^{(i+1)},\theta^{(i)})] \\
&= [Q(\theta^{(i+1)},\theta^{(i)}) - Q(\theta,\theta^{(i)})] + \mathrm{KL}(p(\mathbf{Z}|\mathbf{X},\theta^{(i)})||p(\mathbf{Z}|\mathbf{X},\theta^{(i+1)}))
\end{aligned} \tag{11.11}$$

Both of the two terms in (11.11) are nonnegative, we conclude that $\ln p(\mathbf{X}|\theta^{(i+1)}) \geqslant \ln p(\mathbf{X}|\theta^{(i)})$.

Remark 11.5. For an example of application of EM on Gaussian mixture model, see [Li]; for more applications, see [Bishop].

12. Hidden Markov Model

1. The HMM concept.

Definition 12.1. (Hidden Markov Model, a.k.a HMM) An HMM is a probabilistic model for time sequence. A *state sequence* is generated randomly by a hidden Markov chain, then the state sequence generates an observable random sequence, called *observation sequence.* The indices of the sequences are called *time sequence.*

An HMM is determined by initial probability distribution π, probability transition matrix A and observation matrix B. Assume the state set is given by $Q = \{q_1, q_2, \cdots, q_N\}$, the observation set is $V = \{v_1, v_2, \cdots, v_M\}$. Then $\pi \in \mathbb{R}^N$, $A \in \mathbb{R}^{N\times N}$ and $B \in \mathbb{R}^{N\times M}$.

Assume $I = (i_1, i_2, \cdots, i_T)$ (where $i_t \in Q$) is a state sequence of length $|I| = T$ and O is the corresponding observation sequence $O = (o_1, o_2, \cdots, o_T)$ with $o_i \in V$ for $i = 1, 2, \cdots, T$. We shall define A, B and π as follows:

$$A = [a_{ij}]_{N\times N}, \qquad a_{ij} := P[i_{t+1} = q_j | i_t = q_i],$$

$$B = [b_{jk}]_{N\times K}, \qquad b_j(k) := P[o_t = v_k | i_t = q_j],$$

and

$$\pi = (\pi_1, \pi_2, \cdots, \pi_N), \qquad \pi_i := P[i_1 = q_i].$$

Denote the HMM by $\lambda := (A, B, \pi)$ and make the following Markovian assumptions:

(1) Homogeneous Markov property.

$$P[i_t | i_{t-1}, o_{t-1}, \cdots, i_1, o_1] = P[i_t | i_{t-1}], \quad 1 \leqslant t \leqslant T;$$

(2) Observation Independence.

$$P[o_t | i_T, o_T, i_{T-1}, o_{T-1}, \cdots, i_{t+1}, o_{t+1}, i_t, i_{t-1}, o_{t-1}, \cdots, i_1, o_1] = P[o_t | i_t], \quad 1 \leqslant t \leqslant T.$$

The above assumptions can be summarized in terms of the joint distribution $P[o_1, i_1, \cdots, o_T, i_T]$ as follows for any $T \geqslant 1$:

$$P[o_1, i_1, \cdots, o_T, i_T] = P[i_1]P[o_1|i_1]\prod_{i=2}^{T} P[i_t|i_{t-1}]P[o_t|i_t].$$

The three questions we are interested in are:

- Compute probabilities. For example, given $\lambda := (A, B, \pi)$ and $O = (o_1, o_2, \cdots, o_T)$, compute $P[O|\lambda]$. We may use *forward- or backward-algorithm.*

- Learning. Given $O = (o_1, o_2, \cdots, o_T)$, estimate the parameters A, B, π such that the posterior probability $P[O|\lambda]$ is maximal. We can use the Baum-Welch algorithm, which is the application of EM algorithm in HMM.
- Inference/prediction/Decoding. Given $\lambda := (A, B, \pi)$ and $O = (o_1, o_2, \cdots, o_T)$, find the corresponding state sequence $I = (i_1, i_2, \cdots, i_T)$ such that $P[I|O]$ is maximal. We can use the Viterbi algorithm, which is a dynamic programming method for finding the optimal path.

2. Some probabilities.

Given $\lambda := (A, B, \pi)$, $O = (o_1, o_2, \cdots, o_T)$ and $I = (i_1, i_2, \cdots, i_T)$. The probability for the state sequence I is

$$P[I|\lambda] = \pi_{i_1} a_{i_1 i_2} a_{i_2 i_3} \cdots a_{i_{T-1} i_T}. \tag{12.1}$$

For a given $I = (i_1, i_2, \cdots, i_T)$, the probability of observing $O = (o_1, o_2, \cdots, o_T)$ is

$$P[O|I, \lambda] = b_{i_1}(o_1) b_{i_2}(o_2) \cdots b_{i_T}(o_T). \tag{12.2}$$

Therefore, O and I happen simultaneously is

$$\begin{aligned} P[O, I|\lambda] &= P[O|I, \lambda] P[I|\lambda] = [b_{i_1}(o_1) b_{i_2}(o_2) \cdots b_{i_T}(o_T)][\pi_{i_1} a_{i_1 i_2} a_{i_2 i_3} \cdots a_{i_{T-1} i_T}] \\ &= [\pi_{i_1} b_{i_1}(o_1)][a_{i_1 i_2} b_{i_2}(o_2)] \cdots [a_{i_{T-1} i_T} b_{i_T}(o_T)]. \end{aligned} \tag{12.3}$$

Marginalizing with respect to I, we get

$$P[O|\lambda] = \sum_I P[O, I|\lambda] = \sum_{i_1, i_2, \cdots, \mathrm{\imath}_T} [\pi_{i_1} b_{i_1}(o_1)][a_{i_1 i_2} b_{i_2}(o_2)] \cdots [a_{i_{T-1} i_T} b_{i_T}(o_T)], \tag{12.4}$$

where

$$\sum_{i_1, i_2, \cdots, \mathrm{\imath}_T} = \sum_{i_1 \in Q} \sum_{i_2 \in Q} \cdots \sum_{i_T \in Q}.$$

Define the *forward probability* $\alpha_t(i) := \alpha_t(q_i)$ as

$$\alpha_t(q_i) := P[o_1, o_2, \cdots, o_t, i_t = q_i|\lambda],$$

i.e., the probability that the observation is $o_1, \cdots, o_t$ up to time t and the state at time t is q_i.

Proposition 12.2. *(Forward algorithm)*

$$\alpha_{t+1}(i) = \sum_{j=1}^N \alpha_t(j) a_{ji} b_i(o_{t+1}), \quad P[O|\lambda] = \sum_{i=1}^N \alpha_T(i). \tag{12.5}$$

Here in $\alpha_{t+1}(i)$, the subindex corresponds to time sequence and i in the parentheses corresponds to state q_i.

Proof. The second equality follows from definition. For the first equality, we have

$$\begin{aligned}
\alpha_{t+1}(i) &:= P[o_1, \cdots, o_t, o_{t+1}, i_{t+1} = q_i] \\
&= P[o_{t+1}|o_1, o_2, \cdots, o_t, i_{t+1} = q_i] \times P[o_1, o_2, \cdots, o_t, i_{t+1} = q_i] \\
&= P[o_{t+1}|i_{t+1} = q_i] \times \sum_j P[o_1, o_2, \cdots, o_t, i_t = q_j, i_{t+1} = q_i] \quad \text{(Observation independence)} \\
&= b_i(o_{t+1}) \sum_j P[i_{t+1} = q_i|o_1, o_2, \cdots, o_t, i_t = q_j] P[o_1, \cdots, o_t, i_t = q_j] \\
&= b_i(o_{t+1}) \sum_j P[i_{t+1} = q_i|i_t = q_j] P[o_1, \cdots, o_t, i_t = q_j] \quad \text{(Homogeneous Markov property)} \\
&= b_i(o_{t+1}) \sum_j a_{q_j q_i} P[o_1, \cdots, o_t, i_t = q_j] \\
&= b_i(o_{t+1}) \sum_j a_{ji} \alpha_t(j).
\end{aligned} \tag{12.6}$$

□

Define the *backward probability* $\beta_t(i) := \beta_t(q_i)$ as

$$\beta_t(q_i) := P[o_{t+1}, o_{t+2}, \cdots, o_T|i_t = q_i, \lambda].$$

Proposition 12.3. *(Backward algorithm) Set* $\beta_T(i) = 1, \forall\, i \in \{1, 2, \cdots, N\}$*. For any* $t \in \{T-1, T-2, \cdots, 2, 1\}$*, there hold*

$$\beta_t(i) = \sum_{j=1}^N a_{ij} b_j(o_{t+1}) \beta_{t+1}(j); \quad P[O|\lambda] = \sum_{i=1}^N \pi_i b_i(o_1) \beta_1(i). \tag{12.7}$$

Using both the forward and the backward probabilities, we have

$$P[O|\lambda] = \sum_{i=1}^N \sum_{j=1}^N \alpha_t(i) a_{ij} b_j(o_{t+1}) \beta_{t+1}(j), \forall\ t \in \{1, 2, \cdots, T-1\}. \tag{12.8}$$

When $t = 0, 1$, the above equation can be modified to be Proposition 12.3; when $t = T-1, T$, it can be modified to be Proposition 12.2.

Given λ and O, define the probability that at time t, the state is q_i is $\gamma_t(i) := P[i_t = q_i|O, \lambda]$. Using conditional probability, we know

$$\gamma_t(i) = P[i_t = q_i, O|\lambda]/P[O|\lambda].$$

In view of the definition of forward and backward probability, we see

$$\alpha_t(i)\beta_t(i) = P[i_t = q_i, O|\lambda]. \tag{12.9}$$

Therefore, we have

$$\gamma_t(i) = \frac{\alpha_t(i)\beta_t(i)}{P[O|\lambda]} = \frac{\alpha_t(i)\beta_t(i)}{\sum_{j=1}^N \alpha_t(j)\beta_t(j)}. \tag{12.10}$$

Given λ and O, define the probability that at time t, the state is q_i and at time $t+1$, the state is q_j, is $\xi_t(i,j) := P[i_t = q_i, i_{t+1} = q_j | O, \lambda]$. Similarly, we have

$$\begin{aligned}\xi_t(i,j) := P[i_t = q_i, i_{t+1} = q_j|O,\lambda] &= \frac{P[i_t = q_i, i_{t+1} = q_j, O|\lambda]}{P[O|\lambda]} \\ &= \frac{P[i_t = q_i, i_{t+1} = q_j, O|\lambda]}{\sum_{1\leqslant i\leqslant N}\sum_{1\leqslant j\leqslant N} P[i_t = q_i, i_{t+1} = q_j, O|\lambda]} \\ &= \frac{\alpha_t(i)a_{ij}b_j(o_{t+1})\beta_{t+1}(j)}{\sum_{i,j}\alpha_t(i)a_{ij}b_j(o_{t+1})\beta_{t+1}(j)}.\end{aligned} \tag{12.11}$$

From the definitions of $\gamma_t(i)$ and $\xi_t(i,j)$, we have

$$E[q_i \text{ occurs}|O,\lambda] = \sum_{t=1}^{T}\gamma_t(i), \tag{12.12}$$

$$E[q_i \text{ transmites}|O,\lambda] = \sum_{t=1}^{T-1}\gamma_t(i), \tag{12.13}$$

and

$$E[q_i \text{ transmites to } q_j|O,\lambda] = \sum_{t=1}^{T-1}\xi_t(i,j). \tag{12.14}$$

3. Viterbi algorithm.

Viterbi method is a way of finding optimal path $I^* = (i_1^*, i_2^*, \cdots, i_T^*)$ by dynamic programming. A useful property about the optimal path is used.

Proposition 12.4. *(A property of optimal path) If I^* is an optimal path in predicting I for given λ and O, then with $i_1^*, \cdots, i_t^*$ fixed, the path $i_t^*, \cdots, i_T^*$ is also optimal among the paths starting at i_t^* and ending at i_T^*.*

Proof. An easy contradiction argument. □

Define the largest probability among the paths $(i_1, i_2, \cdots, i_t)$ with $i_t = q_i$ by

$$\delta_t(i) := \max_{i_1,i_2,\cdots,i_{t-1}} P[i_t = i, i_{t-1}, \cdots, i_1, o_t, \cdots, o_1|\lambda], i \in \{1,2,\cdots,N\}.$$

We have the recursion relation

$$\begin{aligned}\delta_{t+1}(i) &= \max_{i_1,i_2,\cdots,i_{t-1},i_t} P[i_{t+1} = i, i_t, i_{t-1}, \cdots, i_1, o_t, \cdots, o_1|\lambda] \\ &= \max_{1\leqslant j\leqslant N}\{\delta_t(j)a_{ji}\}b_i(o_{t+1}).\end{aligned} \tag{12.15}$$

Define the $(t-1)^{\text{th}}$ index of the path $(i_1, \cdots, i_{t-1}, i)$ which has maximal probability among paths with $i_t = q_i$ as $\Psi_t(i) := \arg\max_{1\leqslant j\leqslant N}\{\delta_{t-1}(j)a_{ji}\}$.

The **Viterbi algorithm** is as follows:

Given input $\lambda = (A, B, \pi)$ and observation sequence $O = (o_1, o_2, \cdots, o_T)$, find the optimal path $I^* = (i_1^*, i_2^*, \cdots, i_T^*)$ that maximize $P(I|O)$.

(1) Initialization. $\delta_1(i) = \pi_i b_i(o_1)$ for $i = 1, 2, \cdots, N$; $\Psi_1(i) = 0$ for all i.

(2) Recursion. For $t = 2, 3, \cdots, T$, recursively compute

$$\delta_t(i) = \max_j \{\delta_{t-1}(j) a_{ji}\} b_i(o_t), \quad i = 1, 2, \cdots, N,$$

$$\Psi_t(i) = \arg\max_j \{\delta_{t-1}(j) a_{ji}\}, \quad i = 1, 2, \cdots, N.$$

(3) Stop. $P^* = \max_i \delta_T(i)$, $i_T^* = \arg\max_i \{\delta_T(i)\}$.

(4) Find the optimal path. For $t = T-1, T-2, 2, 1$, $i_t^* = \Psi_{t+1}(i^*_(t+1))$.

4. An example of HMM.

Let $\lambda = (A, B, \pi)$ with $\pi = (0.2, 0.4, 0.4)^T$ and

$$A = \begin{bmatrix} 0.5 & 0.2 & 0.3 \\ 0.3 & 0.5 & 0.2 \\ 0.2 & 0.3 & 0.5 \end{bmatrix}, \quad B = \begin{bmatrix} 0.5 & 0.5 \\ 0.4 & 0.6 \\ 0.7 & 0.3 \end{bmatrix}$$

corresponding to the states q_1, q_2, q_3 and observations v_1, v_2. In other words, $|Q| = N = 3$ and $|V| = M = 2$. (a) Assume $O = (v_1, v_2, v_1)$, compute $P[O|\lambda]$. (b) Assume $O = (v_1, v_2, v_1)$, find the optimal state sequence $I^* = (i_1^*, i_2^*, i_3^*)$.

(a) Here we use forward algorithm to compute $P[O|\lambda]$.

(a1) Initialization $\alpha_1(i) := \pi_i b_i(o_1) = \pi_i b_i(v_1)$.

$$\alpha_1(1) = \pi_1 b_1(v_1) = 0.2 \times 0.5 = 0.1,$$

$$\alpha_1(2) = \pi_2 b_2(v_1) = 0.4 \times 0.4 = 0.16,$$

$$\alpha_1(3) = \pi_3 b_3(v_1) = 0.4 \times 0.7 = 0.28.$$

(a2) Recursion $\alpha_t(i) = \sum_j \alpha_{t-1}(j) a_{ji} b_i(o_t)$.

$$\alpha_2(1) = [\sum_j \alpha_1(j) a_{j1}] b_1(v_2) = (0.1 \times 0.5 + 0.16 \times 0.3 + 0.28 \times 0.2) \times 0.5 = 0.077,$$

$$\alpha_2(2) = [\sum_j \alpha_1(j) a_{j2}] b_2(v_2) = (0.1 \times 0.2 + 0.16 \times 0.5 + 0.28 \times 0.3) \times 0.6 = 0.1104,$$

$$\alpha_2(3) = [\sum_j \alpha_1(j) a_{j3}] b_3(v_2) = (0.1 \times 0.3 + 0.16 \times 0.2 + 0.28 \times 0.5) \times 0.3 = 0.0606.$$

$$\alpha_3(1) = \sum_j \alpha_2(j) a_{j1} b_1(v_1) = (0.077 \times 0.5 + 0.1104 \times 0.3 + 0.0606 \times 0.2) \times 0.5 = 0.04187,$$

$$\alpha_3(2) = \sum_j \alpha_2(j) a_{j2} b_2(v_1) = (0.077 \times 0.2 + 0.1104 \times 0.5 + 0.0606 \times 0.3) \times 0.4 = 0.03551,$$

$$\alpha_3(3) = \sum_j \alpha_2(j) a_{j3} b_3(v_1) = (0.077 \times 0.3 + 0.1104 \times 0.2 + 0.0606 \times 0.5) \times 0.7 = 0.05284.$$

(a3) Stop. $P[O|\lambda] = \alpha_3(1) + \alpha_3(2) + \alpha_3(3) = 0.04187 + 0.03551 + 0.05284 = 0.13022.$

(b) We use Viterbi algorithm to find the optimal path.

(b1) Initialization $\delta_1(i) = \pi_i b_i(v_1)$.

$$\begin{aligned} \delta_1(1) &= \pi_1 b_1(v_1) = 0.2 \times 0.5 = 0.1, \\ \delta_1(2) &= \pi_2 b_2(v_1) = 0.4 \times 0.4 = 0.16, \\ \delta_1(3) &= \pi_3 b_3(v_1) = 0.4 \times 0.7 = 0.28. \end{aligned}$$

We have $\Psi_1(i) = 0$ for $i = 1, 2, 3$.

(b2) Recursion $\delta_2(i) = \max_j[\delta_1(j)a_{ji}]b_i(v_2)$.

First, compute $\delta_2(1) = \max_j[\delta_1(j)a_{j1}]b_1(v_2)$. We have

$$\begin{aligned} \delta_1(1)a_{11} &= 0.1 \times 0.5 = 0.05, \\ \delta_1(2)a_{21} &= 0.16 \times 0.3 = 0.048, \\ \delta_1(3)a_{31} &= 0.28 \times 0.2 = 0.056, \end{aligned}$$

Therefore, $\delta_2(1) = 0.056 \times 0.5 = 0.028$. Also, $\Psi_2(1) = 3$.

Second, $\delta_2(2) = \max_j[\delta_1(j)a_{j2}]b_2(v_2)$. We have

$$\begin{aligned} \delta_1(1)a_{12} &= 0.1 \times 0.2 = 0.02, \\ \delta_1(2)a_{22} &= 0.16 \times 0.5 = 0.08, \\ \delta_1(3)a_{32} &= 0.28 \times 0.3 = 0.084, \end{aligned}$$

Therefore, $\delta_2(2) = 0.084 \times 0.6 = 0.0504$. Also, $\Psi_2(2) = 3$.

Third, $\delta_2(3) = \max_j[\delta_1(j)a_{j3}]b_3(v_2)$. We have

$$\begin{aligned} \delta_1(1)a_{13} &= 0.1 \times 0.3 = 0.03, \\ \delta_1(2)a_{23} &= 0.16 \times 0.2 = 0.032, \\ \delta_1(3)a_{33} &= 0.28 \times 0.5 = 0.14, \end{aligned}$$

Therefore, $\delta_2(2) = 0.14 \times 0.3 = 0.042$. Also, $\Psi_2(3) = 3$.

Now, we do another recursion $\delta_3(i) = \max_j[\delta_1(j)a_{ji}]b_i(v_1)$.

First, compute $\delta_3(1) = \max_j[\delta_2(j)a_{j1}]b_1(v_1)$. We have

$$\begin{aligned} \delta_2(1)a_{11} &= 0.028 \times 0.5 = 0.014, \\ \delta_2(2)a_{21} &= 0.0504 \times 0.3 = 0.01512, \\ \delta_2(3)a_{31} &= 0.042 \times 0.2 = 0.0082, \end{aligned}$$

Therefore, $\delta_2(1) = 0.01512 \times 0.5 = 0.00756$. Also, $\Psi_3(1) = 2$.

Second, compute $\delta_3(2) = \max_j[\delta_2(j)a_{j2}]b_2(v_1)$. We have

$$\begin{aligned} \delta_2(1)a_{12} &= 0.028 \times 0.2 = 0.0056, \\ \delta_2(2)a_{22} &= 0.0504 \times 0.5 = 0.0252, \\ \delta_2(3)a_{32} &= 0.042 \times 0.3 = 0.0126, \end{aligned}$$

Therefore, $\delta_2(1) = 0.0252 \times 0.4 = 0.01008$. Also, $\Psi_3(2) = 2$.

Third, compute $\delta_3(3) = \max_j[\delta_2(j)a_{j3}]b_3(v_1)$. We have

$$\delta_2(1)a_{13} = 0.028 \times 0.3 = 0.0084,$$
$$\delta_2(2)a_{23} = 0.0504 \times 0.2 = 0.01008,$$
$$\delta_2(3)a_{33} = 0.042 \times 0.5 = 0.021,$$

Therefore, $\delta_2(1) = 0.021 \times 0.7 = 0.0147$. Also, $\Psi_3(3) = 3$.

(b3) Stop. $P^* = \max_i \delta_3(i) = \max\{0.00756, 0.01008, 0.0147\} = 0.0147$. And also, we have $i_3^* = \arg\max_i \delta_3(i) = 3$, $i_2^* = \Psi_3(i_3^*) = \Psi_3(3) = 3$, and $i_1^* = \Psi_2(i_2^*) = \Psi_2(3) = 3$.

(b4) Find the optimal path. $I = (3, 3, 3)$.

13. Appendix I: Linear algebra

1. **Basic Matrix Identities**

Proposition 13.1. *$(AB)^T = B^T A^T$.*

Proof. $[(AB)^T]_{ij} = [AB]_{ji} = A_{jk}B_{ki} = [B^T]_{ik}[A^T]_{kj} = [B^T A^T]_{ij}$. □

Proposition 13.2.

$$(P^{-1} + B^T R^{-1} B)^{-1} B^T R^{-1} = P B^T (B P B^T + R)^{-1}. \tag{13.1}$$

Proof. Multiplying the equality by $(BPB^T + R)$ on the right, and $(P^{-1} + B^T R^{-1} B)$ on the left, we obtain

$$B^T R^{-1}(BPB^T + R) = (P^{-1} + B^T R^{-1} B) P B^T.$$

It is easy to see that both sides equal to $B^T + B^T R^{-1} B P B^T$. □

In (13.1), let $P = Id$ and $B^T R = A$, we have

Proposition 13.3.

$$(I + AB)^{-1} A = A(I + BA)^{-1}. \tag{13.2}$$

Proposition 13.4. *(Woodbury identity)*

$$(A + UCV)^{-1} = A^{-1} - A^{-1}U(C^{-1} + VA^{-1}U)^{-1}VA^{-1}. \tag{13.3}$$

Proof.

$$\begin{aligned}
&(A + UCV)[A^{-1} - A^{-1}U(C^{-1} + VA^{-1}U)^{-1}VA^{-1} \\
&= I - U(C^{-1} + VA^{-1}U)^{-1}VA^{-1} + UCVA^{-1} - UCVA^{-1}U(C^{-1} + VA^{-1}U)^{-1}VA^{-1} \\
&= [I + UCVA^{-1}] - [U(C^{-1} + VA^{-1}U)^{-1}VA^{-1} + UCVA^{-1}U(C^{-1} + VA^{-1}U)^{-1}VA^{-1}] \\
&= [I + UCVA^{-1}] - [U + UCVA^{-1}U](C^{-1} + VA^{-1}U)^{-1}VA^{-1} \\
&= [I + UCVA^{-1}] - UC(C^{-1} + VA^{-1}U)(C^{-1} + VA^{-1}U)^{-1}VA^{-1} \\
&= I + UCVA^{-1} - UCVA^{-1} \\
&= I.
\end{aligned} \tag{13.4}$$

□

Proposition 13.5. *(Inverse of partitioned matrix)*

$$\begin{bmatrix} A, & B \\ C, & D \end{bmatrix}^{-1} = \begin{bmatrix} M, & -MBD^{-1} \\ -D^{-1}CM, & D^{-1} - D^{-1}CMBD^{-1} \end{bmatrix}, \tag{13.5}$$

where $M := (A - BD^{-1}C)^{-1}$ is the Schur complement.

2. **Trace and determinant**

Proposition 13.6. *Tr(AB) = Tr(BA), i.e., Tr is cyclic.*

Proof.

$$\mathrm{Tr}(AB) = A_{ij}B_{ji} = B_{ji}A_{ij} = \mathrm{Tr}(BA). \tag{13.6}$$

□

The *determinant* $|A|$ of a square matrix A is defined by

$$|A| = \sum_{\sigma \in S_n} \mathrm{sign}(\sigma) A_{1\sigma(1)} A_{2\sigma(2)} \cdots A_{n\sigma(n)}.$$

We have $|AB| = |A||B|$ and $|A^{-1}| = 1/|A|$. Assume A, B are matrices of size $N \times M$, then

$$|I_N + AB^T| = |I_M = A^T B|.$$

In particular, if $\mathbf{a}, \mathbf{b}$ are both N dimensional column vectors, then we have

$$|I_N + \mathbf{a}\mathbf{b}^T| = |I_1 + \mathbf{a}^T\mathbf{b}| = 1 + \mathbf{a}^T\mathbf{b}.$$

3. **Spectrum decomposition**

Let A be a $M \times M$ matrix, and with right eigenvalues and eigenvectors given by

$$\lambda_1, \lambda_2, \cdots, \lambda_M \in \mathbb{C}^1; u_1, u_2, \cdots, u_M \in \mathbb{C}^M,$$

i.e.,

$$Au_j = \lambda_j u_j, \quad j = 1, 2, \cdots, M. \tag{13.7}$$

We can assume that $\|u_j\|_2 = 1$ for all j.

We also assume $v_1, v_2, \cdots, v_M \in \mathbb{C}^{1\times M}$ are the left eigenvectors of A associating with $\mu_1, \mu_2, \cdots, \mu_K \in \mathbb{C}^1$, i.e.,

$$v_j A = \mu_j v_j.$$

Taking transpose of the above equation, we get

$$A^T v_j^T = \mu_j v_j^T,$$

which shows that $\mu_1, \mu_2, \cdots, \mu_M \in \mathbb{C}^1$ are eigenvalues of A^T. Since $\det(\lambda - A) = \det(\lambda - A^T)$, we easily see

$$\{\mu_1, \mu_2, \cdots, \mu_M\} = \{\lambda_1, \lambda_2, \cdots, \lambda_M\}.$$

Assume the M eigenvectors form a basis of $\mathbb{C}^M$ and $U = [u_1, \cdots, u_M]$, we can write (13.7) in matrix form as

$$AU = U\text{Diag}\{\lambda_1, \cdots, \lambda_M\} := U\Lambda,$$

which gives $A = U\Lambda U^{-1}$.

Let A^* be the conjugate transpose of A. A is Hermitian iff $A^* = A$. If A is real matrix, then that A is Hermitian means A is symmetric.

Proposition 13.7. *For a Hermitian matrix, its eigenvalues are real, and eigenvectors associated to different eigenvectors are orthogonal.*

Proof. Let λ, u be an eigenpair for A, i.e., $Au = \lambda u$. Then we have

$$\lambda^* \|u\|^2 = \langle u, \lambda u\rangle = \langle u, Au\rangle = \langle u, A^* u\rangle = \langle Au, u\rangle = \lambda \|u\|^2, \tag{13.8}$$

which gives $(\lambda^* - \lambda)\|u\|^2 = 0$, implying $\lambda^* = \lambda$, i.e., λ is real. Consider $\langle Au_i, u_j\rangle$:

$$\lambda_i \langle u_i, u_j\rangle = \langle Au_i, u_j\rangle = \langle u_i, Au_j\rangle = \lambda_j \langle u_i, u_j\rangle, \tag{13.9}$$

which gives $(\lambda_i - \lambda_j)\langle u_i, u_j\rangle = 0$, and hence $\langle u_i, u_j\rangle = 0$ when $\lambda_i - \lambda_j \ne 0$. □

From the above proposition and Grad-Schmidt orthogonalization, we know that

Proposition 13.8. *For a Hemitian matrix A, the eigenvectors $u_1, \cdots, u_M$ can be made mutually orthogonal: $\langle u_i, u_j\rangle = u_i^* u_j = \delta_{ij}$, or $U^* U = I$*

From $U^* U = I$, we see $UU^* = I$ by definition of matrix inverse, therefore, for a Hermition matrix A, the corresponding U has the property that both and row vectors and the column vectors of U form a orthogonal basis of $\mathbb{C}^M$.

Definition 13.9. (Unitary matrix, normal matrix) If $U^* U = UU^* = I$, we say U is a unitary matrix. If $U^* U = UU^*$, we say U is a normal matrix.

From definition, it is easy to see that $\det(U) = \det(U^*) = 1$. Meanwhile, unitary matrix induces unitary transformation which preserves distance and angle. Indeed,

$$\langle Uv, Uw\rangle = \langle U^* Uv, w\rangle = \langle v, w\rangle.$$

Proposition 13.10. *(Spectral decomposition) Let A be Hermitian and $U = [u_1, \cdots, u_M]$ be the unitary matrix consisting of the eigenvectors of A. Then $A = U\Lambda U*$ and $A^{-1} = U\Lambda^{-1}U*$. These two relations can be written as*

$$A = \sum_i \lambda_i u_i u_i^T, \quad A^{-1} = \sum_i \lambda_i^{-1} u_i u_i^T.$$

The diagonal matrix Λ is a representation of the linear transform A under transformed basis. We can give a direct proof of the spectral decomposition. Let us take $A = \sum_i \lambda_i u_i u_i^T$ as an example. We regard both the left hand side and right hand side as linear transformations. To show they are actually the same, we just need to verify that their action on a or any basis is the same. We can use the basis $\{u_i\}_{i=1}^M$. For any u_j, we have

$$Au_j = \lambda_j u_j, \quad \sum_i \lambda_i u_i u_i^T u_j = \sum_i \lambda_i u_i \delta_{ij} = \lambda_j u_j.$$

Comparing the above two equalities, we verified the conclusion.

4. **Matrix derivatives**

Matrix is a representation of linear transformations between finite dimensional spaces. More precisely, a linear operator from $\mathbb{R}^m$ into $\mathbb{R}^n$ can be represented by a $n\times m$ matrix. Because in order to map a vector $u \in \mathbb{R}^m$ to result in a vector $v \in \mathbb{R}^n$, we must have A be a $n\times m$ matrix, which is easily see by considering the expression $Au = v$. In general, a linear transformation between two Banach spaces X, Y can be represented by a linear operator $L : X \to Y$.

The above viewpoint is the key to understand derivatives. The idea of derivative is local linearization. Any derivative is a linear operator. More precisely, $f : X \to Y$ for any Banach space X, Y (in other words, both can be infinitely dimensional), $Df(x)$ is a linear operator from X to Y. In particular, if $X = \mathbb{R}^m$ and $Y = \mathbb{R}^n$, then $Df(x)$ is a linear operator from $\mathbb{R}^m$ into $\mathbb{R}^n$, hence a $n \times m$ matrix.

About the gradient. Let $f : \mathbb{R}^n \to \mathbb{R}^1$ with $y = f(x)$. The derivative of $\frac{\partial y}{\partial x} \in \mathbb{R}^{1\times n}$, in other words, a n-dimensional row vector. This is different from gradient $\nabla_x f(x)$ which is typically defined as a column vector. Therefore, $\nabla_x f(x) = \left[\frac{\partial y}{\partial x}\right]^T$.

Now, we can define the derivative rules. Let x be a scaler, $\mathbf{f}, \mathbf{x} \in \mathbb{R}^n$ be a vector. Then

$$(D_x\mathbf{f})_i = D_x f_i,$$
$$(D_{\mathbf{x}}\mathbf{f})_{ij} = D_{x_j} f_i.$$

The above rule extends naturally to any tensors.

Next, we collect some computation propositions. Let $\mathbf{a}, \mathbf{x}$ be two vectors, then

$$D_{\mathbf{x}}(\mathbf{a}\cdot\mathbf{x}) = \mathbf{a}^T.$$

Let A, B be two matrices which can be multiplied, then

$$D_x(AB) = (D_xA)B + AD_xB.$$

Let A be an invertible matrix, we have $A^{-1}A = I_n$. Taking x-derivative in the equation gives us

$$\partial_x(A^{-1})A + A^{-1}\partial_x A = 0_n.$$

From the above equation, we get

$$\partial_x(A^{-1}) = -A^{-1}(\partial_x A)A^{-1}.$$

The derivative of trace. Now we compute $D_A\mathrm{Tr}(AB)$ for two matrices A, B. First, consider $\partial_{A_{ij}}\mathrm{Tr}(AB)$, and we have

$$\partial_{A_{ij}}\mathrm{Tr}(AB) = \partial_{A_{ij}}(A_{ij}B_{ji}) = B_{ji} = (B^T)_{ij},$$

therefore, we have

$$\partial_A\mathrm{Tr}(AB) = B^T.$$

Similarly, we have

$$\partial_A\mathrm{Tr}(A^TB) = B,$$

$$\partial_A \mathrm{Tr}(A) = \partial_A \mathrm{Tr}(A^T) = I,$$
$$\partial_A \mathrm{Tr}(ABA^T) = A(B + B^T).$$

Let F be a square matrix and $|F|$ be the determinant, then we have

$$d|F| = \mathrm{Tr}(F^\sharp)dF = |F|\mathrm{Tr}(F^{-1})dF,$$

which $F^\sharp$ is the cofactor matrix of F whose (i, j)-element is given by the F_{ji}'s signed cofactor, i.e, $F^\sharp/|F| = F^{-1}$. In particular, we have $d \ln |F| = \mathrm{Tr}(F^{-1}dF$.

The derivative of inverse matrix F^{-1} is given by

$$d(F^{-1}) = -F^{-1}(dF)F^{-1}.$$

5. **Cholesky decompoistion and normal distribution**

Theorem 13.11. *A is Hermitian positive-definite matrix iff $\mathbf{A} = \mathbf{L}\mathbf{L}^*$ uniquely for some lower diagonal matrix $\mathbf{L}$ with real and positive diagonal matrix. The decomposition is called Cholesky decomposition.*

If the matrix $\mathbf{A}$ is Hermitian and positive semi-definite, then it still has a decomposition of the form $\mathbf{A} = \mathbf{L}\mathbf{L}^*$ if the diagonal entries of $\mathbf{L}$ are allowed to be zero. When $\mathbf{A}$ has real entries, $\mathbf{L}$ has real entries as well, and the factorization may be written $\mathbf{A} = \mathbf{L}\mathbf{L}^T$. The Cholesky decomposition is unique when $\mathbf{A}$ is positive definite; there is only one lower triangular matrix $\mathbf{L}$ with strictly positive diagonal entries such that $\mathbf{A} = \mathbf{L}\mathbf{L}^*$. However, the decomposition need not be unique when $\mathbf{A}$ is positive semidefinite. The converse holds trivially: if $\mathbf{A}$ can be written as $\mathbf{L}\mathbf{L}^*$ for some invertible $\mathbf{L}$, lower triangular, then $\mathbf{A}$ is Hermitian and positive definite.

It is well known that linear transformation of a Gaussian random vector is Gaussian. The mean and covariance matrix determine a Gaussian vector. Therefore, we can construct Gaussian vectors with specified mean μ and real covariance matrix Σ which we assume is positive semi-definite. Assume $X_1, X_2, \cdots, X_n$ are n independent standard univariate Gaussian.Then $X = (X_1, \cdots, X_n)^T$ is a Gaussian with mean $\vec{0}$ and covariance matrix I_n. Now assume $\Sigma = LL^T$ for some lower diagonal matrix L with nonnegative diagonal entries. Then $Y := LX + \mu$ is Gaussian with mean μ and covariance matrix Σ. Indeed, we have

$$E[Y] = E[LX + \mu] = LE[X] + \mu = \mu,$$

and

$$\mathrm{cov}(Y) = E[LX(LX)^T] = LE[XX^T]L^T = LL^T = \Sigma.$$

14. Appendix II: Inequalities

Lemma 14.1. *(Hoeffding) Let X be any real-valued random variable with $E[X] = 0$ and $P[a \leqslant X \leqslant b] = 1$. Then for any $\lambda \in \mathbb{R}$, there holds*

$$E[\exp\{\lambda X\}] \leqslant \exp\{\lambda^2(b - a)^2/8\}. \tag{14.1}$$

Proof. By assumption, if one of a and b is zero, so is the other. Then $X = 0$ almost surely, and (14.1) becomes $E[e^0] \leqslant e^0 = 1$, which is obviously true. Now consider the case $a < 0 < b$. As the function $x \to e^{sx}$ is convex, we consider the convex combination

$$x = \frac{b-x}{b-a}a + \frac{x-a}{b-a}b,$$

and use Jensen's inequality to get

$$e^{sx} \leqslant \frac{b-x}{b-a}e^a + \frac{x-a}{b-a}e^b, \quad \forall x \in [a, b].$$

Taking expectation in the above inequality, we have

$$\begin{aligned}
E[e^{sX}] &\leqslant \frac{b-E[X]}{b-a}e^a + \frac{E[X]-a}{b-a}e^b \qquad (E[X]=0)\\
&= \frac{b}{b-a}e^a + \frac{-a}{b-a}e^b\\
&= \frac{-a}{b-a}e^{sa}(-b/a + e^{s(b-a)})\\
&= \frac{-a}{b-a}e^{sa}(-\frac{b-a+a}{a} + e^{s(b-a)})\\
&= \frac{-a}{b-a}e^{sa}(-\frac{b-a}{a} - 1 + e^{s(b-a)}) \qquad (-\frac{a}{b-a} := \theta > 0, -s\theta(b-a) = sa)\\
&= e^{-s\theta(b-a)}(1-\theta+\theta e^{s(b-a)}) \qquad (s(b-a) := u)\\
&= (1-\theta+\theta e^u)e^{-\theta u}\\
&= \exp\{-\theta u + \ln(1-\theta+\theta e^u)\} \qquad (-\theta u + \ln(1-\theta+\theta e^u) := \phi(u))\\
&= e^{\phi(u)}.
\end{aligned} \tag{14.2}$$

We shall find a bound of $\phi(u)$. First $\phi : \mathbb{R} \to \mathbb{R}$ is well-defined. To see this, we need to verify that $1-\theta+\theta e^u > 0$ is always true. Indeed, we have

$$1-\theta+\theta e^u = \theta(1/\theta - 1 + e^u) = \theta(-(b-a)/a - 1 + e^u) = \theta(-b/a + e^u) > 0.$$

Now, by Taylor's formula, we know there exist v between 0 and u such that

$$\phi(u) = \phi(0) + \phi'(0)u + \frac{1}{2}\phi''(v)u^2.$$

By direct computations, we know

$$\phi(0) = 0, \quad \phi'(0) = -\theta + \frac{\theta e^u}{1-\theta+\theta e^u}\Big|_{u=0} = 0, \tag{14.3}$$

and

$$\begin{aligned}\phi''(v) &= \frac{\theta e^v(1-\theta+\theta e^v)-\theta e^u\theta e^u}{(1-\theta+\theta e^v)^2}\\ &= \frac{\theta e^v(1-\theta)}{(1-\theta+\theta e^v)^2}\\ &= \frac{\theta e^v}{(1-\theta+\theta e^v)^2}\frac{1-\theta}{(1-\theta+\theta e^v)^2}\\ &= \frac{\theta e^v}{1-\theta+\theta e^v}(1-\frac{\theta e^v}{1-\theta+\theta e^v})\\ &= t(1-t) \qquad (\frac{\theta e^v}{1-\theta+\theta e^v}:=t>0)\\ &\leqslant [(t+(1-t))/2]^2\\ &= 1/4.\end{aligned}$$

Then from (14.3), we conclude

$$\phi(u) \leqslant u^2/8 = s^2(b-a)^2/8.$$

Therefore, for the moment generating function $E[e^{sX}]$, we have

$$E[e^{sX}] \leqslant \exp\{s^2(b-a)^2/8\}.$$

□

Proposition 14.2. *(Hoeffding's inequality, 1963) Let $X_1,\cdots,X_n$ be independent r.v. bounded by the intervals $[a_i,b_i]$ for $1\leqslant i\leqslant n$. Define $S_n := X_1+\cdots+X_n$. Then, for $t\geqslant 0$, there hold*

$$P[S_n-E[S_n]\geqslant t] \leqslant \exp\{-2t^2/\sum_i(b_i-a_i)^2\},$$
$$P[|S_n-E[S_n]|\geqslant t] \leqslant 2\exp\{-2t^2/\sum_i(b_i-a_i)^2\}.$$

Writing in the form of average $\bar{X}:=S_n/n$, the above inequalities are

$$P[\bar{X}-E[\bar{X}]\geqslant t] \leqslant \exp\{-2n^2t^2/\sum_i(b_i-a_i)^2\},$$
$$P[|\bar{X}-E[\bar{X}]|\geqslant t] \leqslant 2\exp\{-2n^2t^2/\sum_i(b_i-a_i)^2\}.$$

Proof. When $t=0$, the inequalities hold trivially. In the following, assume $t>0$. Then for any $s>0$, by Markov's inequality and independence of X_is, we have

$$\begin{aligned}P[S_n-E[S_n]\geqslant t] &= P[e^{s(S_n-E[S_n])}\geqslant e^{st}]\\ &\leqslant e^{-st}E[e^{s(S_n-E[S_n])}]\\ &\leqslant e^{-st}\prod_i E[e^{s(X_i-E[X_i])}] \qquad \text{(Hoeffding's Lemma)}\\ &\leqslant e^{-st}\prod_i e^{\frac{s^2(b_i-a_i)^2}{8}}\\ &= \exp\{-st+\sum_i\frac{s^2(b_i-a_i)^2}{8}\}.\end{aligned}$$

The quadratic function $f(s) := -st + \sum_i \frac{s^2(b_i-a_i)^2}{8}$ attains its minimum $-2t^2/\sum_i(b_i-a_i)^2$ at $s = 4t/\sum_i(b_i-a_i)^2$. Therefore, we conclude

$$P[S_n - E[S_n] \geqslant t] \leqslant \exp\{-2t^2/\sum_i(b_i-a_i)^2\}.$$

The above proof also gives us

$$P[S_n - E[S_n] \leqslant -t] \leqslant \exp\{-2t^2/\sum_i(b_i-a_i)^2\},$$

by replacing S_n by $-S_n$ and $X_i s$ by $-X_i s$.

Together, we have

$$P[|S_n - E[S_n]| \geqslant t] \leqslant 2\exp\{-2t^2/\sum_i(b_i-a_i)^2\}.$$

□

Proposition 14.3. *(Sample size estimate for confidence interval) Let the random variable $X \in [a,b]$. To acquire $(1-\alpha)$-confidence interval $E[\bar{X}] \pm t$, one needs at least $[(b-a)^2\ln(2/\alpha)/2t^2]+1$ samples.*

Proof. Let $X_i, \cdots, X_n$ be n samples. Then we have by Hoeffding's inequality

$$E[|\bar{X} - E[\bar{X}]| \geqslant t] \leqslant 2e^{-2n^2t^2/n(b-a)^2} = 2e^{-2nt^2/(b-a)^2},$$

which implies

$$E[|\bar{X} - E[\bar{X}]| < t] > 1 - 2e^{-2nt^2/(b-a)^2}.$$

To acquire a $(1-\alpha)$-confidence interval $E[\bar{X}] \pm t$, we need $1 - 2e^{-2nt^2/(b-a)^2} \geqslant 1-\alpha$, i.e., $2e^{-2nt^2/(b-a)^2} \leqslant \alpha$, from which we easily see that $n \geqslant -(b-a)^2\ln(\alpha/2)/2t^2$. □

An example. Consider i.i.d Bernoulli random variables $X_1, X_2, \cdots, X_n$ where each X_i a Bernoulli trial of tossing a coin with $P[\text{head}] = P[X_i = 1] = p$ and $P[\text{tail}] = p[X_i = 0] = 1-p$. Then $S_n := X_1 + \cdots X_n$ is the number of heads for n independent tosses of the same coin. We know $E[S_n] = \sum_{i=1}^n E[X_i] = np$. Then for any $\varepsilon > 0$, we apply the Hoeffding inequality to get

$$P[|S_n - np| \geqslant n\varepsilon] \leqslant 2\exp\{-2(n\varepsilon)^2/\sum_{i=1}^n(1-0)^2\} = 2\exp\{-2n\varepsilon^2\}.$$

Therefore, we have

$$P[|S_n - np| < n\varepsilon] \geqslant 1 - 2\exp\{-2n\varepsilon^2\}.$$

Letting $\varepsilon = \sqrt{\frac{\ln n}{n}}$, we have

$$P[|S_n - np| < \sqrt{n\ln n}] \geqslant 1 - 2/n^2,$$

or in the average form

$$P\Big[|S_n/n - p| < \sqrt{\frac{\ln n}{n}}\Big] \geqslant 1 - 2/n^2.$$

Lemma 14.4. *(Gibbs' inequality) Let $P = \{p_1, \cdots, p_n\}$ and $Q = \{q_1, \cdots, q_n\}$ are two probability distributions. Then there always holds*

$$-\sum_{i=1}^{n} p_i \log_2 p_i \leqslant -\sum_{i=1}^{n} p_i \log_2 q_i, \tag{14.4}$$

with equality iff $p_i = q_i$ for all i. In other words, entropy is always less than or equal to the cross entropy.

Proof. (A rough proof) We just need to show that

$$-\sum_{i=1}^{n} p_i \log_2 q_i - (-\sum_{i=1}^{n} p_i \log_2 p_i) = \sum_{i=1}^{n} p_i \log_2 \frac{p_i}{q_i} \geqslant 0.$$

The above inequality is a consequence of the Jensen's inequality in view of the convexity of $-\log_2$:

$$\sum_{i=1}^{n} p_i \log_2 \frac{p_i}{q_i} = \sum_i p_i(-\log_2)(q_i/p_i) \geqslant (-\log_2)(\sum_i p_i \frac{q_i}{p_i}) = 0.$$

When $q_i = p_i$ for all i, Jensen's inequality takes equality. □

The above proof also shows that the Kullback-Leibler divergence $D_{KL}(p||q)$ is always non-negative:

$$D_{KL}(p||q) = \sum_{i=1}^{n} p_i \log_2(p_i/q_i) \geqslant 0.$$

Proof. (A more illuminating proof) As $\log_2 a = \ln a/\ln 2$, we could use ln to give a proof. We have the following inequality:

$$\ln x \leqslant x - 1, \text{ for all } x > 0; \text{with} \text{``} = \text{''} \text{ when } x = 1.$$

Denote $I = \{i; p_i \neq 0\}$. Then,

$$-\sum_{i=i}^{n} p_i \ln(q_i/p_i) = -\sum_{i \in I} p_i \ln(q_i/p_i) - \sum_{i \in I^c} p_i \ln(q_i/p_i).$$

The second term $-\sum_{i \in I^c} p_i \ln(q_i/p_i) = 0$ and for the first term, we have

$$-\sum_{i \in I} p_i \ln(q_i/p_i) \geqslant -\sum_{i \in I} p_i(q_i/p_i - 1) = -\sum_{i \in I} q_i + \sum_{i \in I} p_i = 1 - \sum_{i \in I} q_i \geqslant 0.$$

Therefore, we have proved the inequality. When $q_i/p_i = 1$ for $i \in I$, we have equality. In the meantime, as P and Q are probability distributions, we must have $p_i = q_i = 0$ for $i \in I^c$. Therefore, when $q_i = p_i$ for all $1 \leqslant 1i \leqslant n$, we have equality in the entropy and cross-entropy inequality (14.4). □

Lemma 14.5. *(Log-sum inequality) Let $a_1, \cdots, a_n$ and $b_1, \cdots, b_n$ are two sequences of non-negative numbers and $a := \sum_{i=1}^n a_i$ and $b := \sum_{i=1}^n b_i$. Then,*

$$\sum_{i=1}^{n} a_i \log(a_i/b_i) \geqslant a \log(a/b) \tag{14.5}$$

with equality iff a_i/b_i are equal for all i, i.e., $a_i = cb_i$ for some common positive constant c.

Proof. We use the convexity of the function $f(x) = x \log x$ defined for $x \geqslant 0$. Due to change of base formula for logarithm, we can regard the base as e without loss of generality. In fact, it is easy to see

$$f'(x) = \ln x + 1; \quad f''(x) = 1/x \geqslant 0.$$

Using f, the left hand side of (14.6) can be written as

$$\sum_{i=1}^{n} b_i(a_i/b_i)\log(a_i/b_i) = \sum_{i=1}^{n} b_i f(a_i/b_i) = b\sum_{i=1}^{n}(b_i/b)f(a_i/b_i).$$

Now, we can use Jensen's inequality to get

$$b\sum_{i=1}^{n}(b_i/b)f(a_i/b_i) \geqslant b\sum_i f(\sum_i \frac{b_i}{b}\frac{a_i}{b_i}) = bf(a/b) = a\log\frac{a}{b}.$$

In the above step, we have inequality iff a_i/b_i are equal for all $1 \leqslant i \leqslant n$. □

Remark 14.6. The log-sum inequality still holds for $n = \infty$ as long as $\sum_i^n a_n < \infty$ and $\sum_i^n b_n < \infty$. The log-sum inequality can also be generalized to arbitrary g such that $f(x) := xg(x)$ is convex on $x \geqslant 0$.

Lemma 14.7. *(Generalized log-sum inequality-the g-sum inequality) Let $a_1, \cdots, a_n, \cdots$ and $b_1, \cdots, b_n, \cdots$ are two sequences of nonnegative numbers and $a := \sum_{i=1}^{\infty} a_i < \infty$ and $b := \sum_{i=1}^{\infty} b_i < \infty$. Then, for any function $g(x)$ defined on $x \geqslant 0$ such that $f(x) := xg(x)$ is well-defined on $x \geqslant 0$ and is convex, there holds*

$$\sum_{i=1}^{\infty} a_i g(a_i/b_i) \geqslant ag(a/b) \tag{14.6}$$

with equality iff a_i/b_i are equal for all i, i.e., $a_i = cb_i$ for some common positive constant c.

Next we discuss the *Chernoff bound* which relates the tail probability of a random variable with moment generating functions. Consider the tail probability $P[|X| \geqslant a]$ for a random variable X with a positive number a, then we have the Markov inequality

$$aP[|X| \geqslant a] = \int_{[|X|\geqslant a]} a\,dP \leqslant \int_{[|X|\geqslant a]} |X|\,dP \leqslant \int_{\Omega} |X|\,dP = E[|X|] \Longrightarrow P[|X| \geqslant a] \leqslant E[|X|]/a.$$

Now let $X = X_1 + \cdots + X_n$. By Markov's inequality, we have for any $t > 0$, that

$$P[X \geqslant a] = P[e^{tX} \geqslant e^{ta}] \leqslant e^{-ta}E[e^{tX}] = e^{-ta}E[\prod_{i=1}^{n} e^{tX_i}],$$

and

$$P[X \leqslant a] = P[-X \geqslant -a] = P[e^{-tX} \geqslant e^{-ta}] \leqslant e^{ta}E[e^{-tX}] = e^{ta}E[\prod_{i=1}^{n} e^{-tX_i}].$$

Lemma 14.8. *(Idea of Chernoff bound) Let X is the sum of n independent random variables $X_1, \cdots, X_n$. Then for any positive numbers a and t, we have the following bounds*

$$P[X \geqslant a] \leqslant \min\Big\{\inf_{t>0}\{e^{-ta}\prod_{i=1}^{n} E[e^{tX_i}]\}, 1\Big\},$$

$$P[X \leqslant a] \leqslant \min\Big\{\inf_{t>0}\{e^{ta}\prod_{i=1}^{n} E[e^{-tX_i}]\}, 1\Big\}.$$

Theorem 14.9. *(Chernoff bounds) Let $X = X_1 + \cdots + X_n$, where $X_i = 1$ with probability p_i and $X_i = 0$ with probability $1 - p_i$, and all X_i are independent. Let $\mu = E[X] = \sum_{i=1}^n p_i$. Then*

(i) (Upper tail)
$$P[X \geqslant (1+\delta)\mu] \leqslant e^{-\frac{\delta^2}{2+\delta}\mu}, \ \forall\ \delta > 0;$$

(ii) (Lower tail)
$$P[X \leqslant (1-\delta)\mu] \leqslant e^{-\mu\delta^2/2} \ \forall\ 0 \leqslant \delta \leqslant 1;$$

(iii) (Full tail)
$$P[|X - \mu| \geqslant \delta\mu] \leqslant 2e^{-\mu\delta^2/3} \ \forall\ 0 \leqslant \delta \leqslant 1.$$

Proof. We first compute the moment generating function for each X_i with $t \in \mathbb{R}^1$:
$$E[e^{tX_i}] = e^{t\times 1}p_i + e^{t\times 0}(1-p_i) = 1 + p_i(e^t - 1).$$
By the elementary inequality $1 + x \leqslant e^x$ for all $x \in \mathbb{R}^1$, we notice that
$$E[e^{tX_i}] \leqslant \exp\{p_i(e^t-1)\}.$$
As X is the sum of independent random variables, we know that
$$E[e^{tX}] = \prod_{i=1}^n E[e^{tX_i}] \leqslant \prod_i \exp\{p_i(e^t-1)\} = \exp\{\sum_i p_i(e^t-1)\} = \exp\{(e^t-1)\mu\}.$$
(i) By Lemma 14.8, we have for the upper tail
$$P[X \geqslant (1+\delta)\mu] \leqslant \inf_{t>0}\{e^{-(1+\delta)\mu t}\exp\{(e^t-1)\mu\}\} = \inf_{t>0}\{\exp\{\mu[e^t - 1 - (1+\delta)t]\}\}.$$
Next, we compute the inf. For this, it is sufficient to minimize $g(t) := e^t - 1 - (1+\delta)t$. It is easy to see that $g(0) = 0$ and $g(+\infty) = +\infty$, and it has a global minimizer at t^* such that $g'(t^*) = e^{t^*} - (1+\delta) = 0$, i.e., $e^{t^*} = 1 + \delta$. Putting this information back to g, we have
$$g(t^*) = \delta - (1+\delta)\ln(1+\delta).$$
Now, we obtain an upper bound for the upper tail:
$$P[X \geqslant (1+\delta)\mu] \leqslant \exp\{\mu g(t^*)\} = \exp\{\mu[\delta - (1+\delta)\ln(1+\delta)]\}.$$
To obtain the desired upper bound, we try to bound
$$s(\delta) := \delta - (1+\delta)\ln(1+\delta).$$
We use another elementary inequality
$$\ln(1+x) \geqslant \frac{x}{1+x/2}, \quad \forall x > 0. \tag{14.7}$$
Therefore, for any $\delta > 0$, we have
$$s(\delta) \leqslant \delta - (1+\delta)\frac{\delta}{1+\delta/2} = -\frac{\delta^2}{2+\delta}.$$
Therefore, we have
$$P[X \geqslant (1+\delta)\mu] \leqslant \exp\{\mu s(\delta)\} \leqslant \exp\{-\mu\delta^2/(\delta+2)\}.$$

(ii) By Lemma 14.8, we have for the lower tail

$$P[X \leqslant (1-\delta)\mu] \leqslant \inf_{t>0}\{e^{(1-\delta)\mu t}\exp\{(e^{-t}-1)\mu\}\} = \inf_{t>0}\{\exp\{\mu[e^{-t}-1+(1-\delta)t]\}\}.$$

To control the lower tail, we consider to control the function

$$h(t) := e^{-t} - 1 + (1-\delta)t.$$

It is easy to see that $h(0) = 0$, $h(+\infty) = +\infty$ and $h'(0) = [-e^{-t} + (1-\delta)]\Big|_{t=0} = -\delta < 0$. Therefore, $h(t)$ admits a global minimizer on $[0, +\infty)$ at t^* such that $h'(t^*) = -e^{-t^*}+(1-\delta) = 0$. Putting this information back to h, we have

$$h(t^*) = -\delta - (1-\delta)\ln(1-\delta).$$

Now, we obtained an upper bound for the lower tail

$$P[X \leqslant (1-\delta)\mu] \leqslant \exp\{\mu h(t^*)\} = \exp\{\mu[-\delta - (1-\delta)\ln(1-\delta)]\}.$$

To obtain the desired bound, we use the following bound holds

$$-\delta - (1-\delta)\ln(1-\delta) \leqslant -\frac{\delta^2}{2}, \quad \forall\ \delta \in [0,1]. \tag{14.8}$$

Therefore, we have

$$P[X \geqslant (1-\delta)\mu] \leqslant \exp\{-\mu\delta^2/2\}, \quad \forall\ \delta \in (0,1).$$

(iii) For the full tail, we have for all $\delta \in (0,1)$,

$$P[|X-\mu| \geqslant \delta\mu] = P[X \leqslant (1+\delta)\mu] + P[X \leqslant (1-\delta)\mu] \leqslant e^{-\frac{\delta^2}{2+\delta}\mu} + e^{-\mu\delta^2/2} \leqslant 2\exp\{-\mu\delta^2/3\}.$$

□

To make the above complete, we show the validity of (14.7) and (14.8). To show (14.7), we define the auxiliary function $f_1(x) := \ln(1+x) - \frac{x}{1+x/2}$. We notice that $f_1(0) = 0$ and

$$f_1'(x) = \frac{1}{1+x} - \frac{(1+x/2) - x \times 1/2}{(1+x/2)^2} = \frac{(x/2)^2}{(1+x)(1+x/2)^2} > 0, \quad \forall\, x > 0.$$

Therefore, (14.7) holds. To show (14.8), it is sufficient to show for $\delta \in (0,1)$ (at the two end points, the desired inequality holds),

$$f_2(\delta) := \ln(1-\delta) + \frac{\delta - \delta^2/2}{1-\delta}.$$

It is easy to observe that $f_2(0) = 0$ and to compute that

$$f_2'(\delta) = \frac{\delta^2/2}{(1-\delta)^2} \geqslant 0.$$

Therefore, (14.8) is valid.

The generalization of the Chernoff bound can lead to Hoeffding's inequality. Now, we consider an example of n independent coin tosses. Assume the coin is fair, i.e, for each X_i, $P[\text{head}] = 1/2$. Consider $S_n := X_1 + \cdots + X_n$, the number of heads in the n tosses. Then the weak law of large number says

$$P[|S_n/n - 1/2| \geqslant \varepsilon] \to 0,\ n \to +\infty.$$

The above convergence in (probability) measure does not specify the rate of convergence. This rate can be specified using either Chebyshev or Chernoff. Let's first use Chebyshev:

$$\begin{aligned} P[|S_n/n-1/2|\geqslant\varepsilon] &= P[|S_n-n/2|\geqslant n\varepsilon]\leqslant(n\varepsilon)^{-2}\mathrm{Var}(S_n)=(n\varepsilon)^{-2}\sum_i\mathrm{Var}(X_i)\\ &=(n\varepsilon)^{-2}n/4=\frac{1}{4\varepsilon^2n}.\end{aligned}\tag{14.9}$$

If $\varepsilon=1/4$, we have

$$P[|S_n/n-1/2|\geqslant 1/4]\leqslant\frac{4}{n}.$$

Now, we use Chernoff bound:

$$P[|S_n/n-1/2|\geqslant\delta/2]=P[|S_n-n/2|\geqslant\delta n/2]\leqslant 2\exp\{-(n/2)\delta^2/3\}=2\exp\{-n\delta^2/6\}.\tag{14.10}$$

If $\delta=1/2$, we have

$$P[|S_n/n-1/2|\geqslant 1/4]\leqslant 2\exp\{-n/24\}.\tag{14.11}$$

If we consider a smaller δ: let $\delta=\sqrt{\frac{6k\ln n}{n}}$ for any positive number k, we have

$$P[|S_n/n-1/2|\geqslant\frac{1}{2}\sqrt{\frac{6k\ln n}{n}}]\leqslant 2/n^k.\tag{14.12}$$

15. Appendix III: Elementary things

Proposition 15.1. $1^2+2^2+\cdots+n^2=\frac{1}{6}n(n+1)(2n+1)$.

Proof. Denote $f(n):=\sum_{i=1}^n i^2$ and use induction. For $n=1$, $f(1)=1=\frac{1}{6}\times1\times2\times3$. Now assume the desired formula holds for n, and we show it also holds for $n+1$ as follows:

$$\begin{aligned} f(n+1)=f(n)+(n+1)^2&=\frac{1}{6}n(n+1)(2n+1)+(n+1)^2\\ &=\frac{1}{6}(n+1)[n(2n+1)+6(n+1)]=\frac{1}{6}(n+1)(n+1+1)[2(n+1)+1].\end{aligned}$$

Therefore, the desired formula holds true for all n. □

Proposition 15.2. $1^3+2^3+\cdots+n^3=\frac{1}{4}n^2(n+1)^2$.

Proof. Denote $f(n):=\sum_{i=1}^n i^3$ and use induction. For $n=1$, $f(1)=1=\frac{1}{4}\times1\times2^2$. Now assume the desired formula holds for n, and we show it also holds for $n+1$ as follows:

$$\begin{aligned} f(n+1)=f(n)+(n+1)^3&=\frac{1}{4}n^2(n+1)^2+(n+1)^3\\ &=\frac{1}{4}(n+1)^2(n^2+4(n+1))=\frac{1}{4}(n+1)^2(n+1+1)^2.\end{aligned}$$

Therefore, the desired formula holds true for all n. □

Some elementary inequalities. Consider the function $f(x) = e^x$. It is convex, and the tangent line passing through (0, 1) is $y - 1 = e^x\Big|_{x=0}(x - 0) = x$, i.e., $y = x + 1$. Therefore, we have

$$1 + x \leqslant e^x, \quad \forall\, x \in \mathbb{R}^1. \tag{15.1}$$

Replacing x by $-x$, we have

$$1 - x \leqslant e^{-x}, \quad \forall\, x \in \mathbb{R}^1. \tag{15.2}$$

When $1 + x \geqslant 0$, we can also apply monotone functions to (15.1) to obtain

$$(1 + x)^\alpha \leqslant \exp\{\alpha x\}, \quad \forall\, \alpha > 0, \tag{15.3}$$

and

$$(1 + x)^{-\alpha} \geqslant \exp\{-\alpha x\}, \quad \forall\, \alpha > 0, \tag{15.4}$$

Consider the function $f(x) = \ln(1 + x)$ for $x > -1$. This function is concave and passes the point $(0, 0)$. The corresponding tangent line is $y = x$. Therefore, we have

$$\ln(1 + x) \leqslant x, \quad \forall\, x > -1. \tag{15.5}$$

Replacing $x > 1$ by $-x > -1$, we have

$$\ln(1 - x) \leqslant -x, \quad \forall\, x < 1. \tag{15.6}$$

Using convexity and concavity, and local analysis, Taylor expansion, we could generate more inequalities. For example, consider the $f(x) = x \ln x$ for $x > 0$. This function is globally convex, and passes through the point $(1, 0)$. The corresponding tangent line passing through this point is $y = x - 1$. Therefore, we have

$$x - 1 \leqslant x \ln x, \quad \forall\, x > 0.$$

Now, we use the definition of convexity or concavity to derive some inequalities. For a convex function f, we have the Jensen ineqaulity

$$f(\alpha x_1 + (1 - \alpha)x_2) \leqslant \alpha f(x_1) + (1 - \alpha)f(x_2),$$

or more generally, assume P is a probability distribution on the domain of f and X is random variable with values in domain of f, then the Jensen inequality can be written as

$$f(E[X]) \leqslant E[f(X)].$$

Now consider the power function $f(x) = x^\alpha$ for $x \geqslant 0$ if $\alpha > 0$ and for $x > 0$ if α is allowed to be negative. We shall consider the case $\alpha > 0$, as otherwise, we could consider the reciprocal. We assume $\alpha \neq 1, 2$, as these two cases are either linear or well-known for us. The power function is convex globally on its domain when $\alpha > 1$. Hence, we have by Jensen's inequality for positive numbers a, b,

$$(\frac{a + b}{2})^\alpha \leqslant \frac{1}{2}a^\alpha + \frac{1}{2}b^\alpha, \quad \forall \alpha > 1.$$

i.e.,

$$(a + b)^\alpha \leqslant 2^{\alpha-1}(a^\alpha + b^\alpha), \quad \forall \alpha > 1.$$

By concavity, we have

$$(a + b)^\alpha \geqslant 2^{\alpha-1}(a^\alpha + b^\alpha), \quad \forall\, 0 < \alpha < 1. \tag{15.7}$$

We also have for $0 < \alpha < 1$ that

$$(a + b)^\alpha \leqslant a^\alpha + b^\alpha. \tag{15.8}$$

To show (15.8), it sufficient to show that $(1+x)^\alpha \leqslant 1+x^\alpha$ for $x \geqslant 0$. This is an elementary exercise: Let $f(x) = 1 + x^\alpha - (1+x)^\alpha$. We check that $f(0) = 0$ and

$$f'(x) = \alpha\Big(\frac{1}{x^{1-\alpha}} - \frac{1}{(1+x)^{1-\alpha}}\Big) > 0, \quad \forall\, x > 0.$$

From (15.7) and (15.8), we have

$$a^\alpha + b^\alpha \geqslant (a+b)^\alpha \geqslant 2^{\alpha-1}(a^\alpha + b^\alpha), \quad \forall\, 0 < \alpha < 1. \tag{15.9}$$

The first “ $\geqslant$ ” is convenient to be generalized:

$$a_1^\alpha + \cdots a_n^\alpha \geqslant (a_1 + \cdots + a_n)^\alpha, \quad \forall\, 0 < \alpha < 1. \tag{15.10}$$

References

[Billingsley] P. Billingsley, Probability and Measure, 3rd edition, Wiley, 1995, +593 pages.
[Bishop] C. Bishop, Patten Recognition and Machine Learning, Springer, 2006, +738 pages.
[Li] Hang Li, Statistical Machine Learning, Tsinghua University Press, 2012, +235 pages.
[Zhou] Zhihua Zhou, Machine Learning, Tsinghua University Press, 1st edition of January 2016, +425 pages.

E-mail address: `jinghuayao@gmail.com`