Lecture 2: Mathematical Preliminaries for ML

Shukai Gong

Not all required mathematical preliminaries are included in the sections above. It's just a reminder of some of my rusty math knowledge.

1 Norm

1. Vector Norms: A norm is a function $\|\cdot\| : \mathbb{C}^N \mapsto \mathbb{R}$ satisfying

1. $\|\boldsymbol{x}\| \geq 0$ and $\|\boldsymbol{x}\| = 0$ iff $\boldsymbol{x} = \boldsymbol{0}$
2. $\|\alpha \boldsymbol{x}\| = |\alpha|\|\boldsymbol{x}\|$ for any $\alpha \in \mathbb{C}$
3. $\|\boldsymbol{x} + \boldsymbol{y}\| \leq \|\boldsymbol{x}\| + \|\boldsymbol{y}\|$ (triangle inequality)

Commonly-used Vector Norms:

- ℓ_0-norm: $\|\boldsymbol{x}\|_0 = \sum_{i=1}^{N} \mathbb{I}(x_i \neq 0)$(number of non-zero elements)
- ℓ_1-norm: $\|\boldsymbol{x}\|_1 = \sum_{i=1}^{N} |x_i|$, ℓ_2-norm: $\|\boldsymbol{x}\|_2 = (\sum_{i=1}^{N} |x_i|^2)^{\frac{1}{2}}$, ℓ_p-norm: $\|\boldsymbol{x}\|_p = (\sum_{i=1}^{N} |x_i|^p)^{\frac{1}{p}}$
- ℓ_∞-norm: $\|\boldsymbol{x}\|_\infty = \max_{i=1,\cdots,N} |x_i|$
- Weighted norm: $\|\boldsymbol{x}\|_{\boldsymbol{W}} = \|\boldsymbol{W}\boldsymbol{x}\|$
- Frobenius norm: $\|\boldsymbol{X}\|_F = \sqrt{\sum_{i=1}^{M} \sum_{j=1}^{N} |x_{ij}|^2}$

2. Induced Matrix Norms: $\|\boldsymbol{A}\|_p = \max_{\boldsymbol{x} \neq \boldsymbol{0}} \frac{\|\boldsymbol{A}\boldsymbol{x}\|}{\|\boldsymbol{x}\|} = \max_{\|\boldsymbol{x}\|=1} \|\boldsymbol{A}\boldsymbol{x}\|$

- $\|\boldsymbol{A}\|_1 = \max_i \|\boldsymbol{a}_i\|_1$, where $\boldsymbol{a}_i$ is the i-th column of $\boldsymbol{A}$
- $\|\boldsymbol{A}\|_2 = \sqrt{\rho(\boldsymbol{A}^H \boldsymbol{A})}$, where $\rho(\cdot)$ is the spectral radius.
- $\|\boldsymbol{A}\|_\infty = \max_j \|\boldsymbol{A}_j\|_1$, where $\boldsymbol{A}_j$ is the j-th row of $\boldsymbol{A}$

Some properties of induced matrix norms:

- Consistency: $\|\boldsymbol{A}\boldsymbol{x}\| \leq \|\boldsymbol{A}\|\|\boldsymbol{x}\|$, $\|\boldsymbol{A}\boldsymbol{B}\| \leq \|\boldsymbol{A}\|\|\boldsymbol{B}\|$
- If $\boldsymbol{\Lambda}$ is a diagonal matrix, then $\|\boldsymbol{\Lambda}\|_p = \max_i |d_{ii}|$
- When $\boldsymbol{A} = \boldsymbol{a}$ is a vector, $\|\boldsymbol{A}\|_2 = \|\boldsymbol{a}\|_2$
- If $\boldsymbol{A} = \boldsymbol{u}\boldsymbol{v}^H$ is a rank-1 matrix, then $\|\boldsymbol{A}\|_2 = \|\boldsymbol{u}\|_2\|\boldsymbol{v}\|_2$

2 Algebra

1. Vector products: given $\boldsymbol{a}, \boldsymbol{b} \in \mathbb{R}^N$

1. **Inner product:** $\boldsymbol{a}^\top \boldsymbol{b} = \boldsymbol{a} \cdot \boldsymbol{b} \, \langle \boldsymbol{a}, \boldsymbol{b} \rangle = \sum_{i=1}^{N} a_i b_i$

2. **Outer product:** $\boldsymbol{a}\boldsymbol{b}^\top = \boldsymbol{a} \otimes \boldsymbol{b} = [a_i b_j] \in \mathbb{R}^{N \times N}$

2. Matrix products: given $\boldsymbol{A}, \boldsymbol{B} \in \mathbb{R}^{M \times N}$

1. **Inner product:** $\langle \boldsymbol{A}, \boldsymbol{B} \rangle = \mathrm{tr}(\boldsymbol{A}^\top \boldsymbol{B}) = \sum_{i=1}^{M} \sum_{j=1}^{N} a_{ij} b_{ij}$

2. **Outer product:** $\boldsymbol{A} \otimes \boldsymbol{B} = [a_{ij} b_{kl}] \in \mathbb{R}^{MN \times MN}$

3. **Kronecker product:** $\boldsymbol{A} \otimes \boldsymbol{B} = [a_{ij} b_{kl}] \in \mathbb{R}^{M^2 \times N^2}$

3. Orthognal vector and matrix:

- **Orthognal vector:**
  - $\boldsymbol{a}, \boldsymbol{b} \in \mathbb{R}^N$ are orthogonal if $\boldsymbol{a}^\top \boldsymbol{b} = 0$ and $\|\boldsymbol{a}\|_2 \neq 0, \|\boldsymbol{b}\|_2 \neq 0$
  - $\boldsymbol{a}, \boldsymbol{b} \in \mathbb{C}^N$ are orthogonal if $\boldsymbol{a}^H \boldsymbol{b} = \sum_{i=1}^{N} \overline{a_i} b_i = 0$ and $\|\boldsymbol{a}\|_2 \neq 0, \|\boldsymbol{b}\|_2 \neq 0$
  - H represents Hermitian transpose for matrix (or vector) $\boldsymbol{A} \in \mathbb{C}^{M \times N}$
- **Orthonormal vector:** $\boldsymbol{a}, \boldsymbol{b}$ are orthonormal if $\boldsymbol{a}^\top \boldsymbol{b} = 0$ and $\|\boldsymbol{a}\|_2 = \|\boldsymbol{b}\|_2 = 1$
- **Orthognal matrix:** $\boldsymbol{A}$ is orthogonal if $\boldsymbol{A}^\top \boldsymbol{A} = \boldsymbol{I}$, i.e. $\boldsymbol{A}^\top = \boldsymbol{A}^{-1}$
  - Norm preserving: $\|\boldsymbol{A}\boldsymbol{x}\|_2 = \|\boldsymbol{x}\|_2$ if $\boldsymbol{A}$ is orthogonal

4. Inverse Matrix: A matrix which has an inverse is called **non-singular**, otherwise **singular**. $\boldsymbol{A}^{-1}$ exists $\iff \det(\boldsymbol{A}) \neq 0$

- **Ill-conditioned matrix:** $\boldsymbol{A}$ is close to being singular, that $\kappa(\boldsymbol{A}) = \|\boldsymbol{A}\| \|\boldsymbol{A}^{-1}\|$ is large.
- **Pseudo-inverse:** $\boldsymbol{A}^+ = (\boldsymbol{A}^\top \boldsymbol{A})^{-1} \boldsymbol{A}^\top$, where $\boldsymbol{A}$ could be non-square.
  - It can be shown that $\boldsymbol{A}^+ \boldsymbol{A} = \boldsymbol{I}$ provided that $\boldsymbol{A}^\top \boldsymbol{A}$ is non-singular.

5. Linear Systems: $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}$, $\boldsymbol{A} \in \mathbb{R}^{M \times N}$, $\boldsymbol{x} \in \mathbb{R}^N$, $\boldsymbol{b} \in \mathbb{R}^M$

- $\boldsymbol{A}$ is a **linear mapping**: $\boldsymbol{x} \mapsto \boldsymbol{A}\boldsymbol{x}$
- $\boldsymbol{b}$ is a **linear combination** of columns of $\boldsymbol{A}$: $\boldsymbol{b} = \sum_{i=1}^{N} x_i \boldsymbol{a}_i$

There are 3 key tasks of linear systems

$$\underbrace{\boldsymbol{A}}_{\text{System}} \underbrace{\boldsymbol{x}}_{\text{Input}} = \underbrace{\boldsymbol{b}}_{\text{Output}}$$

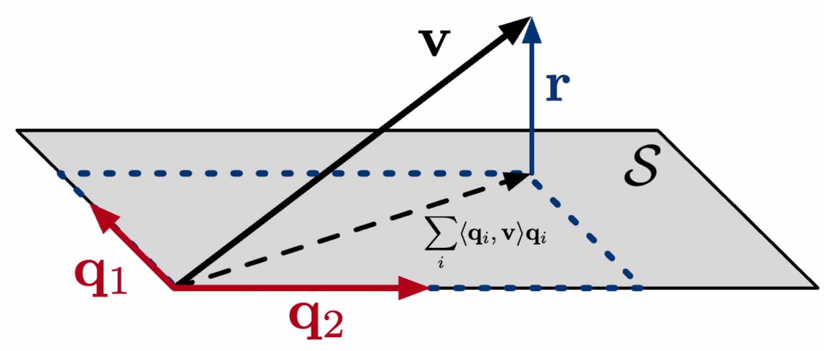
- **Inverse Problem:** Given $\boldsymbol{A}$ and $\boldsymbol{b}$, solve $\boldsymbol{x}$ or $\min_{\boldsymbol{x}} d(\boldsymbol{A}\boldsymbol{x}, \boldsymbol{b})$
- **Modelling:** Given sets of $\boldsymbol{b}$'s(Denote $\boldsymbol{B}$) and $\boldsymbol{x}$'s(Denote $\boldsymbol{X}$), solve $\boldsymbol{A}$ or $\min_{\boldsymbol{A}} d(\boldsymbol{A}\boldsymbol{X}, \boldsymbol{B})$
- **Factorization:** Given $\boldsymbol{B}$, solve the decompostion/factorization $\boldsymbol{B} = \boldsymbol{A}\boldsymbol{X}$ or $\min_{\boldsymbol{A},\boldsymbol{X}} d(\boldsymbol{A}\boldsymbol{X}, \boldsymbol{B})$

6. **Range and Null Space:**

- $\text{Range}(\boldsymbol{A}) = \text{Im}(\boldsymbol{A}) = \{\boldsymbol{Ax} : \boldsymbol{x} \in \mathbb{R}^N\}$, i.e. the column space of $\boldsymbol{A}$
- $\text{Null}(\boldsymbol{A}) = \{\boldsymbol{x} : \boldsymbol{Ax} = \boldsymbol{0}\}$

7. **Components of a vector:** Let $\mathcal{S} = \text{span}\{\boldsymbol{q_1}, \cdots, \boldsymbol{q_N}\}$ where $\boldsymbol{q_1}, \cdots, \boldsymbol{q_N}$ is an orthonormal set in $\mathbb{R}^m$. Then for any $\boldsymbol{v} \in \mathbb{R}^m$, we have **the decompostion of $\boldsymbol{v}$** as shown below

$$\boldsymbol{v} = \sum_{i=1}^{N} \underbrace{\langle \boldsymbol{v}, \boldsymbol{q_i} \rangle \boldsymbol{q_i}}_{\in \mathcal{S}} + \underbrace{r}_{\in \mathcal{S}^\perp}$$



- **The residual of $\boldsymbol{v} \in \mathbb{R}^m$** w.r.t. the set $\boldsymbol{q_1}, \cdots, \boldsymbol{q_N}$: $\boldsymbol{r} = \boldsymbol{v} - \sum_{i=1}^{N} \langle \boldsymbol{v}, \boldsymbol{q_i} \rangle \boldsymbol{q_i}$

2.1 Eigenvalue Decomposition(EVD)

Eigenvalue Decomposition/Spectrum Decomposition

If $\boldsymbol{A} \in \mathbb{R}^{n \times n}$ is **symmetric**, denote the orthonormal eigenvectors of $\boldsymbol{A}$ as $\boldsymbol{Q} = [\boldsymbol{q_1}, \cdots, \boldsymbol{q_n}]$, and the corresponding eigenvalues as $\boldsymbol{\Lambda} = \text{diag}(\lambda_1, \cdots, \lambda_n)$, we can write

$$\boldsymbol{A} = \boldsymbol{Q \Lambda Q}^\top$$

The steps of EVD are as follows

1. **Find the eigenvalues:** By solving the characteristic equation $\det(\boldsymbol{A} - \lambda \boldsymbol{I}) = 0$, we get $\lambda_1, \cdots, \lambda_s$ where the sum of all algebraic multiplicity $m(\lambda_i)$ is n.
2. **Find the eigenvectors:** For each λ_i, solve the equation $(\boldsymbol{A} - \lambda_i \boldsymbol{I})\boldsymbol{q_i} = \boldsymbol{0}$, we get $\boldsymbol{v_{11}}, \cdots, \boldsymbol{v_{1n_1}}, \cdots, \boldsymbol{v_{s1}}, \cdots, \boldsymbol{v_{sn_s}}$ where the sum of all algebraic multiplicity $m(\lambda_i)$ is n.
3. **Orthogonalize the eigenvectors:** Use Gram-Schmidt process to orthogonalize $\boldsymbol{v_{11}}, \cdots, \boldsymbol{v_{1n_1}}, \cdots, \boldsymbol{v_{s1}}, \cdots, \boldsymbol{v_{sn_s}}$ into $\boldsymbol{q_{11}}, \cdots, \boldsymbol{q_{1n_1}}, \cdots, \boldsymbol{v_{q1}}, \cdots, \boldsymbol{v_{qn_s}}$
4. **Form the matrix $\boldsymbol{Q}$:** $\boldsymbol{Q} = [\boldsymbol{q_{11}}, \cdots, \boldsymbol{q_{1n_1}}, \cdots, \boldsymbol{v_{q1}}, \cdots, \boldsymbol{v_{qn_s}}]$, we have $\boldsymbol{Q}^\top \boldsymbol{AQ} = \text{diag}(\lambda_1, \cdots, \lambda_n) = \boldsymbol{\Lambda} \Rightarrow \boldsymbol{A} = \boldsymbol{Q\Lambda Q}^\top$

2.2 Singular Value Decomposition(SVD)

Singular Value Decomposition(SVD)

For any $\boldsymbol{A} \in \mathbb{R}^{m\times n}$, we can write

$$\boldsymbol{A} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top = \sum_{i=1}^{r} \sigma_i \boldsymbol{u}_i \boldsymbol{v}_i^\top$$

where $\boldsymbol{U} \in \mathbb{R}^{m\times m}$ and $\boldsymbol{V} \in \mathbb{R}^{n\times n}$ are orthogonal matrices, and $\boldsymbol{\Sigma} \in \mathbb{R}^{m\times n}$ is a diagonal matrix with the **singular values of $\boldsymbol{A}$** σ_i on its diagonal, $\boldsymbol{u_i}, \boldsymbol{v_i}$ are the i-th columns of $\boldsymbol{U}$ and $\boldsymbol{V}$. Only the first $r = \text{rank}(\boldsymbol{A})$ singular values are non-zero and by convention, they are ordered in non-increasing order: $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r > \sigma_{r+1} = \cdots = \sigma_{\min\{m,n\}} = 0$.

Observe that the SVD factors provide eigendecomposition for $\boldsymbol{A}^\top\boldsymbol{A}$ and $\boldsymbol{A}\boldsymbol{A}^\top$:

$$\begin{aligned} \boldsymbol{A}^\top\boldsymbol{A} &= (\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top)^\top(\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top) = \boldsymbol{V}\boldsymbol{\Sigma}^\top\boldsymbol{U}^\top\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top = \boldsymbol{V}(\boldsymbol{\Sigma}^\top\boldsymbol{\Sigma})\boldsymbol{V}^\top = \boldsymbol{V}\boldsymbol{\Lambda}_1\boldsymbol{V}^\top \\ \boldsymbol{A}\boldsymbol{A}^\top &= (\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top)(\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top)^\top = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top\boldsymbol{V}\boldsymbol{\Sigma}^\top\boldsymbol{U}^\top = \boldsymbol{U}(\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top)\boldsymbol{U}^\top \equiv \boldsymbol{U}\boldsymbol{\Lambda}_2\boldsymbol{U}^\top \end{aligned}$$

It follows immediately that the columns of $\boldsymbol{V}$ are eigenvectors of $\boldsymbol{A}^\top\boldsymbol{A}$ and the columns of $\boldsymbol{U}$ are eigenvectors of $\boldsymbol{A}\boldsymbol{A}^\top$.

The non-zero singular values of $\boldsymbol{A}$ are the square roots of the non-zero eigenvalues of $\boldsymbol{A}^\top\boldsymbol{A}$ and $\boldsymbol{A}\boldsymbol{A}^\top$.

The steps of SVD are as follows

1. **Find the orthogonal eigenvectors of $(\boldsymbol{A}^\top\boldsymbol{A})_{n\times n}$:** $\boldsymbol{V} = [\boldsymbol{v_1}, \cdots, \boldsymbol{v_n}]$, where $\boldsymbol{v_i}$ is the i-th orthogonal eigenvector of $\boldsymbol{A}^\top\boldsymbol{A}$.
2. **Find the orthogonal eigenvectors of $(\boldsymbol{A}\boldsymbol{A}^\top)_{m\times m}$:** $\boldsymbol{U} = [\boldsymbol{u_1}, \cdots, \boldsymbol{u_m}]$, where $\boldsymbol{u_i}$ is the i-th orthogonal eigenvector of $\boldsymbol{A}\boldsymbol{A}^\top$.
3. **Form the singular value matrix $\boldsymbol{\Sigma}$:** $\boldsymbol{A} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top \Rightarrow A\boldsymbol{V} = \boldsymbol{U}\boldsymbol{\Sigma} \Rightarrow \boldsymbol{A}\boldsymbol{v_i} = \sigma_i\boldsymbol{u_i} \Rightarrow \sigma_i = \sqrt{\frac{\|\boldsymbol{A}\boldsymbol{v_i}\|_2}{\|\boldsymbol{v_i}\|_2}}$, and then we can form $\boldsymbol{\Sigma} = \text{diag}(\sigma_1, \cdots, \sigma_r, 0, \cdots, 0)$

3 Matrix Calculus

1. Gradient:

- **Matrix Gradient:** Suppose that $f : \mathbb{R}^{m\times n} \mapsto \mathbb{R}$, then the gradient of f w.r.t. $\boldsymbol{A} \in \mathbb{R}^{m\times n}$ is defined as

$$\nabla_{\boldsymbol{A}} f(\boldsymbol{A}) = \begin{bmatrix} \frac{\partial f(\boldsymbol{A})}{\partial a_{11}} & \cdots & \frac{\partial f(\boldsymbol{A})}{\partial a_{1n}} \\ \vdots & \ddots & \vdots \\ \frac{\partial f(\boldsymbol{A})}{\partial a_{m1}} & \cdots & \frac{\partial f(\boldsymbol{A})}{\partial a_{mn}} \end{bmatrix} \in \mathbb{R}^{m\times n}$$

- **Vector Gradient:** Suppose that $f : \mathbb{R}^{n} \mapsto \mathbb{R}$, then the gradient of f w.r.t. $\boldsymbol{x} \in \mathbb{R}^{n}$ is defined as

$$\nabla_{\boldsymbol{x}} f(\boldsymbol{x}) = \begin{bmatrix} \frac{\partial f(\boldsymbol{x})}{\partial x_1} \\ \cdots \\ \frac{\partial f(\boldsymbol{x})}{\partial x_n} \end{bmatrix} \in \mathbb{R}^{n\times 1}$$

- Suppose $f : \mathbb{R}^n \mapsto \mathbb{R}^m$, then the gradient of f w.r.t. $\boldsymbol{x} \in \mathbb{R}^n$ is defined as

$$\nabla_{\boldsymbol{x}} f(\boldsymbol{x}) = \begin{bmatrix} \frac{\partial f_1(\boldsymbol{x})}{\partial x_1} & \cdots & \frac{\partial f_1(\boldsymbol{x})}{\partial x_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial f_m(\boldsymbol{x})}{\partial x_1} & \cdots & \frac{\partial f_m(\boldsymbol{x})}{\partial x_n} \end{bmatrix} = \begin{bmatrix} \nabla_{\boldsymbol{x}} f_1(\boldsymbol{x})^\top \\ \vdots \\ \nabla_{\boldsymbol{x}} f_m(\boldsymbol{x})^\top \end{bmatrix} = \boldsymbol{J}_f^\top(\boldsymbol{x}) \in \mathbb{R}^{m \times n}$$

2. Basic Facts about Matrix Derivatives:

1. $\frac{\partial \boldsymbol{x}^\top \boldsymbol{a}}{\partial \boldsymbol{x}} = \frac{\partial \boldsymbol{a}^\top \boldsymbol{x}}{\partial \boldsymbol{x}} = \boldsymbol{a}$
2. $\frac{\partial \boldsymbol{x}^\top \boldsymbol{A} \boldsymbol{x}}{\partial \boldsymbol{x}} = (\boldsymbol{A} + \boldsymbol{A}^\top)\boldsymbol{x}$
3. $\frac{\partial \boldsymbol{a}^\top \boldsymbol{X} \boldsymbol{b}}{\partial \boldsymbol{X}} = \boldsymbol{a}\boldsymbol{b}^\top$
4. $\frac{\partial \boldsymbol{a}^\top \boldsymbol{X}^\top \boldsymbol{b}}{\partial \boldsymbol{X}} = \boldsymbol{b}\boldsymbol{a}^\top$
5. $\frac{\partial \boldsymbol{a}^\top \boldsymbol{X} \boldsymbol{a}}{\partial \boldsymbol{X}} = \frac{\partial \boldsymbol{a}^\top \boldsymbol{X}^\top \boldsymbol{a}}{\partial \boldsymbol{X}} = \boldsymbol{a}\boldsymbol{a}^\top$

3. Hessian Matrix: Suppose that $f : \mathbb{R}^n \mapsto \mathbb{R}$, then the Hessian matrix of f w.r.t. $\boldsymbol{x} \in \mathbb{R}^n$ is defined as

$$\nabla^2_{\boldsymbol{x}} f(\boldsymbol{x}) = \begin{bmatrix} \frac{\partial^2 f(\boldsymbol{x})}{\partial x_1^2} & \cdots & \frac{\partial^2 f(\boldsymbol{x})}{\partial x_1 \partial x_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial^2 f(\boldsymbol{x})}{\partial x_n \partial x_1} & \cdots & \frac{\partial^2 f(\boldsymbol{x})}{\partial x_n^2} \end{bmatrix} \in \mathbb{R}^{n \times n}$$

4 Probability and Statistics

A typical ML scenario $(\mathcal{X}, d_{\mathcal{X}}, \mu_{\mathcal{X}})$ requires us to estimate $\mu_{\mathcal{X}}$ via a model $\hat{p}_\theta$ based on data $\boldsymbol{X} = \{\boldsymbol{x}_i\}_{i=1}^N \subset \mathcal{X}$.

4.1 Law of Large Numbers and Central Limit Theorem

1. LLN explains why ML requires lots of data: For $\boldsymbol{X} = \{\boldsymbol{x}_i\}_{i=1}^N \subset \mathcal{X}$

- WLLN: $\overline{X}_n \xrightarrow{P} \mu$
- SLLN: $\overline{X}_n \xrightarrow{a.s.} \mu$
- **Variance reduction:** $\text{Var}(\overline{X}_n) = \frac{\sigma^2}{n} \to 0$ as $n \to \infty$

2. CLT provides ML with Gaussian Distribution:

$$\frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}} \xrightarrow{d} \mathcal{N}(0, 1)$$

4.2 Method of Moments(MoM)

Suppose that we have a set of i.i.d. data $\boldsymbol{X} = \{\boldsymbol{x}_i\}_{i=1}^N \subset \mathcal{X}$ drawn from a distribution $P(\boldsymbol{X}|\boldsymbol{\theta})$ with l parameters $\boldsymbol{\theta} = (\theta_1, \cdots, \theta_l)$.

1. Compute the first l moments as functions of $\boldsymbol{\theta}$

$$\begin{cases} \mu_1 = E[X] = \int_{\mathcal{X}} \boldsymbol{x} P(\boldsymbol{x}|\boldsymbol{\theta})\mathrm{d}\boldsymbol{x} = g_1(\theta_1, \cdots, \theta_n) \\ \mu_2 = E[X^2] = \int_{\mathcal{X}} \boldsymbol{x}^2 P(\boldsymbol{x}|\boldsymbol{\theta})\mathrm{d}\boldsymbol{x} = g_2(\theta_1, \cdots, \theta_n) \\ \cdots \\ \mu_l = E[X^l] = \int_{\mathcal{X}} \boldsymbol{x}^l P(\boldsymbol{x}|\boldsymbol{\theta})\mathrm{d}\boldsymbol{x} = g_l(\theta_1, \cdots, \theta_n) \end{cases}$$

2. Algebraically invert the linear system of l equations to solve for $\theta_1, \cdots, \theta_l$ as functions of $\mu_1, \cdots, \mu_l$

$$\begin{cases} \theta_1 = h_1(\mu_1, \cdots, \mu_l) \\ \theta_2 = h_2(\mu_1, \cdots, \mu_l) \\ \cdots \\ \theta_l = h_l(\mu_1, \cdots, \mu_l) \end{cases} \tag{*}$$

3. Insert **sample moments** $\hat{\mu}_k = \frac{1}{N}\sum_{i=1}^N \boldsymbol{x}_i^k$ into (*) to obtain **MoM estimators** $\hat{\theta}_1, \cdots, \hat{\theta}_l$

$$\begin{cases} \hat{\theta}_1 = h_1(\hat{\mu}_1, \cdots, \hat{\mu}_l) \\ \hat{\theta}_2 = h_2(\hat{\mu}_1, \cdots, \hat{\mu}_l) \\ \cdots \\ \hat{\theta}_l = h_l(\hat{\mu}_1, \cdots, \hat{\mu}_l) \end{cases}$$

Drawbacks of MoM

1. High computation load
2. Lack of extensibility
3. MoM estimators may not exist as the linear system of $\boldsymbol{\mu} = \boldsymbol{G}\boldsymbol{\theta}$ may has no solution⇒Can only be used to estimate simple distributions with few parameters

4.3 Maximum Likelihood Estimation

Suppose that we have a set of i.i.d. data $\boldsymbol{X} = \{\boldsymbol{x}_i\}_{i=1}^N \subset \mathcal{X}$, we assume that the samples are sampled from a distribution $P(\boldsymbol{x}|\theta)$(a model with parameter θ).

Principle: Assume a **deterministic model** and learn the model via maximum likelihood estimation (MLE)

$$\hat{\theta}_{MLE} = \arg\max_\theta \prod_{i=1}^N \log P(\boldsymbol{X}|\theta) = \arg\max_\theta \sum_{i=1}^N \log P(\boldsymbol{x}_i|\theta)$$

- **Pros:** more efficient in general, avoid the design of prior.
- **Cons:** non-robust to sparse data, not easy to quantify the uncertainty of the estimation (doable, but not efficient).

4.4 Bayesian Estimation

Principle: Assume a **probabilistic model** and the model θ yields a prior distribution.

According to the Bayes' theorem, we have

$$\underbrace{P(\theta|\boldsymbol{X})}_{\text{Posterior}} = \frac{P(\boldsymbol{X}|\theta)P(\theta)}{P(\boldsymbol{X})} \propto \underbrace{P(\boldsymbol{X}|\theta)}_{\text{Likelihood } \theta} \underbrace{P(\theta)}_{\text{prior } \theta}$$

- The influence of prior decays with the increase of the number of samples.

MAP estimation: $\hat{\theta}_{MAP} = \arg\max_{\theta} P(\theta|\boldsymbol{X}) = \arg\max_{\theta} P(\boldsymbol{X}|\theta)P(\theta)$

- **Pros:** prior makes it (relatively) robust to sparse data, quantify the uncertainty of the estimation (obtain the distribution of θ)
- **Cons:** require sophisticated design of prior, time-consuming in general.