# Bayesian Learning

1. What is a probability distribution?
2. What is a random variable?
3. Different types of probability distributions
   - Bernoulli, Binomial, Poisson
   - Normal / Gaussian, Beta,

For any probability distribution
- What is its domain? $\mathcal{X}$
- What is its pmf or pdf.
- What are the parameters.

$$\mathbb{E}[X] = \sum_{\forall x \in \mathcal{X}} x P(X = x)$$

or

$$\int_{x} x\, p(x)\, dx$$

$$\mathbb{E}[f(x)] = \sum_{\forall x \in \mathcal{X}} f(x) P(X = x)$$

or

$$\int_{x} f(x)\, p(x)\, dx$$

Why deal with a probability distribution?

Poisson Distribution

$$Y = \{0, 1, 2, \dots\}$$

$$P(X = k) = \frac{\lambda^{k} e^{-\lambda}}{k!}$$

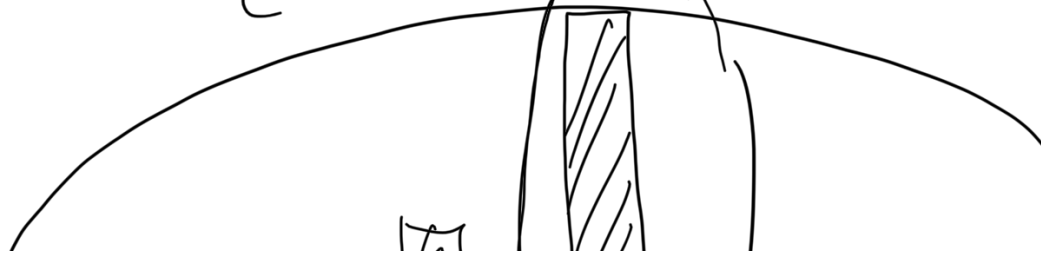
$$\lambda = 4$$

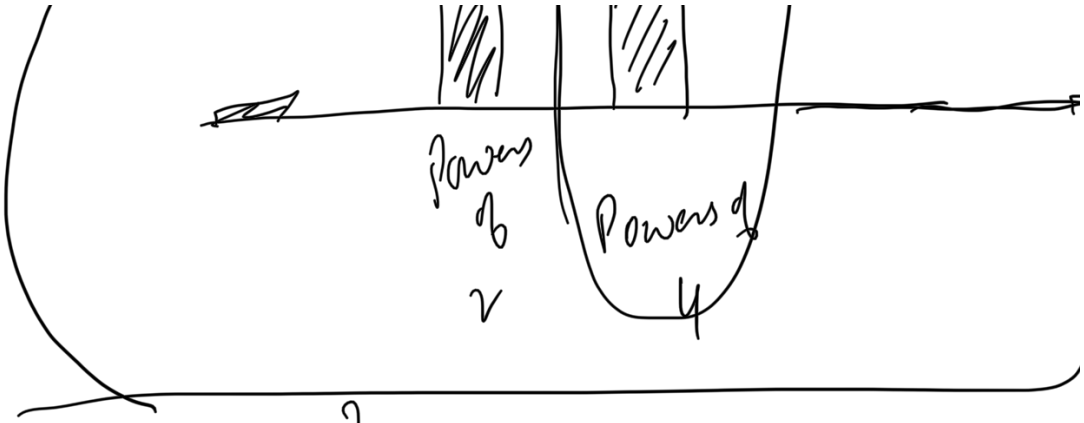
$$E[X] = \lambda$$

---

What are the parameters of the distribution?

$$\{1, 4, 16, 64\}$$






Why?

all numbers between 1 & 100

---

Concept : all numbers between 1 & 100

1, 4, 16, 64

Concept = all powers of 4

---

choose concept with highest likelihood

---

$D = \{1, 4, 16, 64\}$

$h =$ All powers of 4

---

$P(1|h) = \frac{1}{4}$

$P(4|h) = \frac{1}{4}$

$P(16|h) = \frac{1}{4}$

$P(64|h) = \frac{1}{4}$

$L(D|h) = P(x_1|h)\, P(x_2|h) \cdots P(x_N|h)$

likelihood

$= \prod_{x \in D} P(x|h)$

$$\log(L(D|h)) = \sum_{x \in D} \log P(x|h)$$

$L(D|h) = \frac{1}{4} * \frac{1}{4} * \frac{1}{4} * \frac{1}{4}$

Let $h$ = all even numbers.

$D = \{1, 4, 16, 64\}$

$P(1|h) = \frac{1}{50}$

$L(D|h) = \left(\frac{1}{50}\right)^4$

Let $h$ = multiples of 10

P(H|D) –

Calculate the likelihood of D

① given every h.
Pick h with highest likelihood.
Maximum likelihood Estimation
(MLE)

② Use some prior information
and pick h with highest prior.
Max a-Priori estimation

③ Combine likelihood and the prior.
Maximum a-Posteriori Estimation
(MAP)

$P(h)$ → Prior prob that h is the hypothesis.

$P(D|h)$ → likelihood

$$P(h|D) = \frac{P(D|h)\,P(h)}{\sum_{h' \in H} P(D|h')\,P(h')}$$

↑ Posterior

↑ Normalization

normalization constant

$$\sum_{h' \in H} p(h' \mid D) = 1$$

---

$h$ – is the right hypothesis

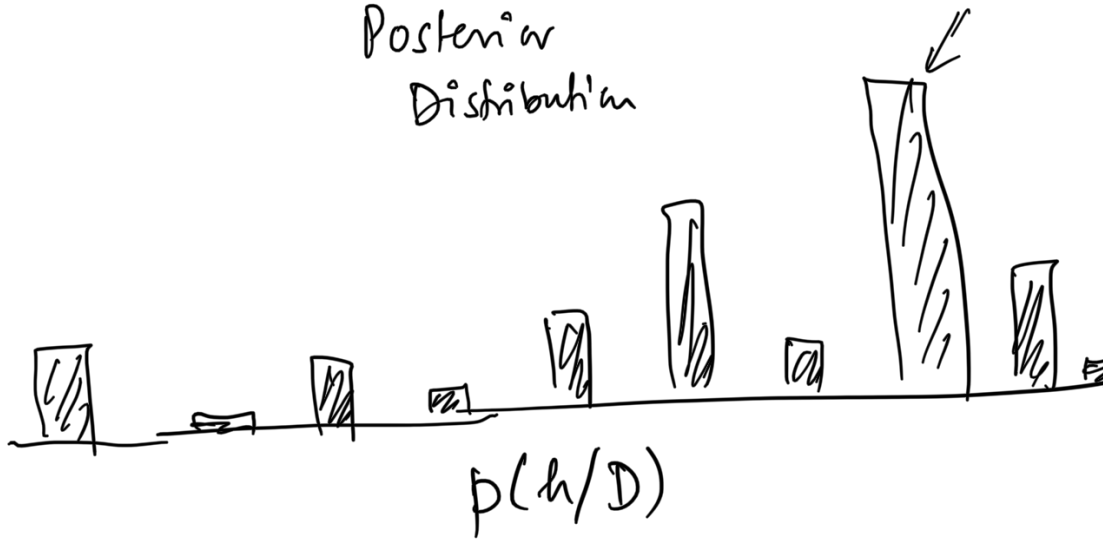

$h_{prior}$ → All even numbers

$$P(x^* = 12 \mid D) = P(x^* = 12 \mid h_{prior})$$

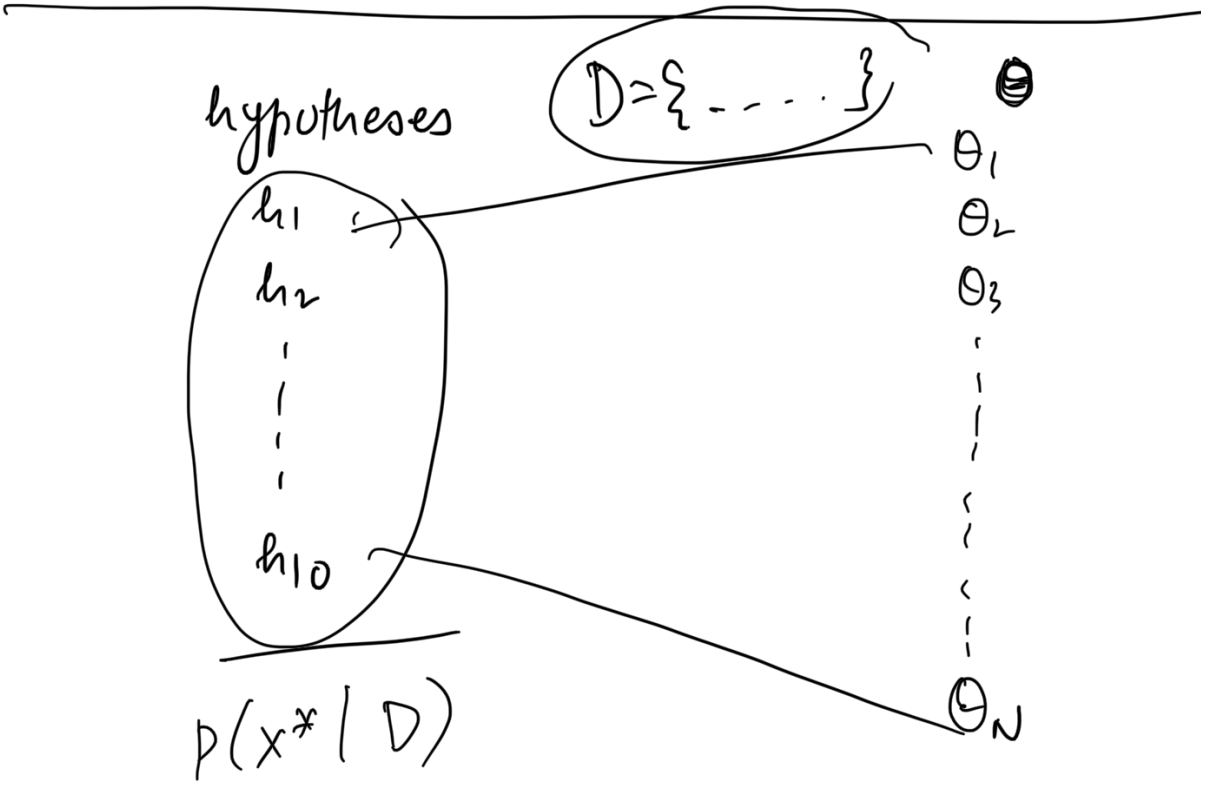
$$=$$

---

Bayesian Averaging




$$P(x^* = 12 \mid D) = \sum_{h \in H} p(x^* = 12 \mid h) p($$




## Bernoulli Distribution

$D = \{ H, H, H, T, T, H, T, H \}$

$N$

$\theta \rightarrow$ learn $\theta$ given $D$

$\frac{5}{8}$

$P(x^* = H | D)$

$0 \leq \theta \leq 1$ → infinite possible values.

---

## Likelihood of D given $\theta$.

likelihood: $\prod_{x \in D} P(x = x \mid \theta)$

$$P(x = H \mid \theta) = \theta$$
$$P(x = T \mid \theta) = 1 - \theta$$

Let D have $N_1$ heads & $N_0$ tails

$$N_1 + N_0 = N$$

$$L(D \mid \theta) = \prod P(x = x \mid \theta)$$

$$= \theta^{N_1} (1-\theta)^{N_0}$$

$$\hat{\theta}_{MLE} = \underset{\theta}{\operatorname{argmax}}\ L(D \mid \theta)$$

$$= \underset{\theta}{\operatorname{argmax}}\ \theta^{N_1} (1-\theta)^{N_0}$$

$$\hat{\theta}_{MLE} = \frac{N_1}{N_0 + N_1} = \frac{N_1}{N}$$

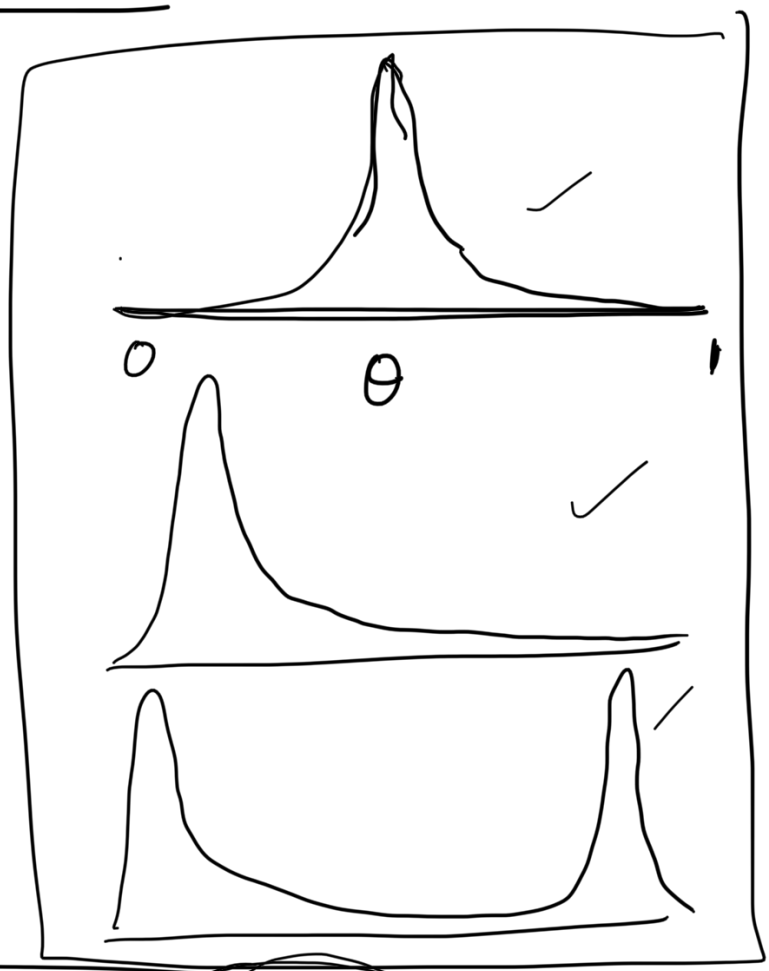
$$P(x^* = H \mid D) = P(x^* = H \mid \hat{\theta}_{MLE})$$

$$\hat{\theta}_{MLE} = \frac{N_1}{N}$$

## Incorporating Prior

**Beta**

**a, b**




$$\text{Posterior} = \frac{\text{likelihood} \times \text{Prior}}{Z}$$

$$L(D|\theta) = \theta^{N_1}(1-\theta)^{N_0}$$

$$b(\theta) \propto \theta^{a-1}(1-\theta)^{b-1}$$

$$p(\theta|D) = \frac{\theta^{N_1}(1-\theta)^{N_0}\,\theta^{a-1}(1-\theta)^{b-1}}{\int_0^1 \theta'^{N_1}(1-\theta')^{N_0}\,\theta'^{a-1}(1-\theta')^{b-1}\,d\theta'}$$

$$= \frac{\theta^{N_1+a-1}(1-\theta)^{N_0+b-1}}{\int_0^1 \theta'^{N_1+a-1}(1-\theta')^{N_0+b-1}\,d\theta'}$$

$p(\theta|D)$ is also a Beta distribution with parameters are $N_1+a$, $N_0+b$

---

If prior & likelihood pdf are conjugate pairs

Then the posterior will have the same form as the prior.

---

Now we know that the posterior

$$p(\theta|D) = \text{Beta}(\theta|a+N_1, b+N_0)$$

$$\hat{\theta}_{MAP} = \frac{a+N_1-1}{a+b+N-2} \qquad N = N_0+N_1$$

---

$$\hat{\theta}_{MAP} = \frac{a-1}{}$$

$\theta_{Prior} = \frac{}{a+b-2}$

$\hat{\theta}_{MLE} = \frac{N_1}{N_0+N_1}$

$\hat{\theta}_{MAP} = \frac{a+N_1-1}{a+b+N-2}$ $\quad N = N_0+N_1$

---

$P(x^* = H \mid D) =$ → $\hat{\theta}_{Prior}$

→ $\hat{\theta}_{MLE}$

→ $\hat{\theta}_{MAP}$

→ $\int_0^1 P(x^*=H \mid \theta)\, p(\theta \mid D)\, d\theta$

$= \frac{a+N_1}{a+b+N}$ $\quad N = N_0+N_1$

---

## Black Swan Paradox

$D = \{T, T, T\}$

$N_1 = 0 \qquad N_0 = 3 \qquad N_1+N_0 = 3$

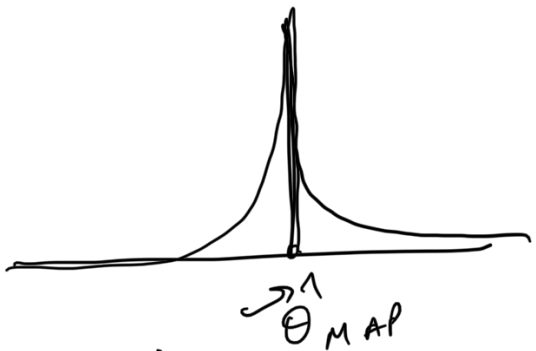
$\hat{\theta}_{MLE} = 0$

$\boxed{P(x^* = H \mid \hat{\theta}_{MLE}) = 0}$

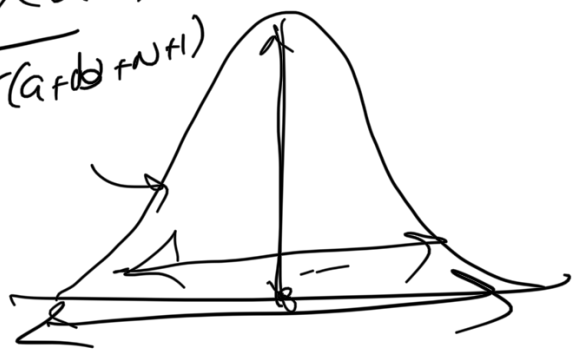
$$P(x^* = H \mid \hat{\theta}_{MAP}) = \frac{a + N_1}{a + b + N - 2}$$

$a = 2 \quad b = 10$

$$P(x^* = H \mid \hat{\theta}_{MAP}) = \frac{2 + 0 - 1}{2 + 10 + 3 - 2} = \frac{1}{13}$$




$$Var[\theta \mid D] = \frac{(a + N_1)(b + N_0)}{(a + b + N)^2 (a + b + N + 1)}$$



---

MVN

$\mu$ $D \times 1$ vector

$\Sigma$ $D \times D$ matrix

$$D = \begin{matrix} x_1 \\ x_2 \\ x_2 \end{matrix} \longrightarrow D \times 1 \text{ vector}$$

$x_N$

---

$$pdf(x \mid \mu, \Sigma) = \frac{1}{(2\pi)^{D/2}|\Sigma|^{1/2}} \exp\left[-\frac{1}{2}(x-\mu)^T \Sigma^{-1}\right.$$

$$L(D \mid \mu, \Sigma) = \prod_{i=1}^{N} pdf(x_i \mid \mu, \Sigma)$$

$$LL(D \mid \mu, \Sigma) = \sum_{i=1}^{N} \log pdf(x_i \mid \mu, \Sigma)$$

$$\hat{\mu}_{MLE} = \frac{1}{N}\sum_{i=1}^{N} x_i \triangleq \bar{x}$$

↳ sample mean

$$\hat{\Sigma}_{MLE} = \frac{1}{N}\sum_{i=1}^{N} (x_i - \bar{x})(x_i - \bar{x})^T$$

---

- Choosing a Gaussian prior on $\mu$ will give a Gaussian posterior for $\mu$

($\Sigma$-Wishart) distribution

① MAP for $\theta$

② MLE for univariate Gaussian

③ MLE for multivariate Gaussian.

---

$X \sim \{H, T\}$

$\theta \quad 0 \le \theta \le 1$

$p(X = H) = \theta$

$p(X = T) = 1 - \theta$

---

Prior on $\theta \rightarrow$ Beta distribution

$\theta \quad 0 \le \theta \le 1$

$a, b$

$pdf(\theta) = Beta(\theta | a, b)$

$$= \frac{1}{Beta(a, b)} \theta^{a-1} (1-\theta)^{b-1}$$

$$\mathbb{E}[\theta] = \int \theta \, p(\theta) \, d\theta = \int_0^1 \theta \frac{1}{Beta(a,b)} \theta^{a-1} (1-\theta)^{b-1}$$

$$= \frac{a}{a+b}$$

What $\theta$ gives the max pdf

Mode

$$\arg\max_{\theta} \; p(\theta)$$

$$\arg\max_{\theta} \; \theta^{a-1}(1-\theta)^{b-1}$$

$$\frac{d}{d\theta}\,\theta^{a-1}(1-\theta)^{b-1} = (a-1)\theta^{a-2}(1-\theta)^{b-1} + \theta^{a-1}(b-1)(1-\theta)^{b-2}(-$$

$$= (a-1)\theta^{a-2}(1-\theta)^{b-1} - (b-1)\theta^{a-1}(1-\theta)^{b-2}$$

Setting to 0 & solving for $\theta$

$$(a-1)\theta^{a-2}(1-\theta)^{b-1} = (b-1)\theta^{a-1}(1-\theta)$$

$$(a-1)(1-\theta) = (b-1)\theta$$

$$\theta[b-1+a-1] = a-1$$

$$\hat{\theta} = \frac{a-1}{a+b-2}$$

Bruin

---

Posterior for $\theta$: $p(\theta|D)$

$$= \text{Beta}(\theta | a+N_1, b+N_0)$$

$N_1$ – # heads in D

$N_0$ – # tails in D

$$\hat{\theta}_{MAP} = \frac{a+N_1-1}{a+N_1+b+N_0-2} = \frac{a+N_1-1}{a+b+N-2}$$

$N = N_1 + N_0$

## Univariate Gaussian Distribution

$X \qquad -\infty \le x \le \infty$

$$pdf(x) = \mathcal{N}(x \mid \mu, \sigma^2)$$

$$= \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{1}{2\sigma^2}(x-\mu)^2\right]$$

$$D = \{x_1, x_2, x_3 \cdots x_N\}$$

i.i.d independent & identically distributed

$$L(D) = \prod_{i=1}^{N} p(x_i)$$

$$= \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{1}{2\sigma^2}(x_i-\mu)^2\right]$$

$$ll(D) = \log L(D)$$

$$= \sum_{i=1}^{N} \log \left[ \qquad \right]$$

$$= \sum_{i=1}^{N} \left[-\frac{1}{2}\log(2\pi) - \log\sigma - \frac{1}{2\sigma^2}(x_i-\mu)^2\right]$$

$$= -\frac{N}{2}\log(2\pi) - N\log\sigma - \frac{1}{2\sigma^2}\sum_{i=1}^{N}(x_i-\mu)^2$$

$$\frac{\partial}{\partial\mu} ll(D) = -\frac{1}{2\sigma^2}\sum_{i=1}^{N} 2(x_i-\mu)(-1)$$

$$= \frac{1}{\sigma^2}\sum_{i=1}^{N}(x_i-\mu)$$

$$= \frac{1}{\sigma^2}\sum_{i=1}^{N} x_i - \frac{1}{\sigma^2} N\mu$$

Set to 0 and solve for $\mu$

$$\hat{\mu}_{MLE} = \frac{1}{N}\sum_{i=1}^{N} x_i$$

→ Sample mean or average

---

$$\frac{\partial}{\partial\sigma} ll(D) = -\frac{N}{\sigma} - \frac{1}{2\sigma^3}\cdot 2\sum_{i=1}^{N}(x_i-\mu)^2$$

Set to 0 and solve for $\sigma$

$$\hat{\sigma}^2_{MLE} = \frac{1}{N}\sum_{i=1}^{N}(x_i-\mu)^2$$

→ Sample variance

---

$$D = \{x_1, x_2, x_3 \cdots x_N\}$$

$$x_i \in \mathbb{R}^d$$

$$\text{pdf}(x_i) = \mathcal{N}(x_i \mid \mu, \Sigma)$$

($\mu$: $d \times 1$, $\Sigma$: $d \times d$)

$$= \frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}} \exp\left\{-\frac{1}{2}\left[(x_i-\mu)^T \Sigma^{-1}(x_i-\mu)\right]\right\}$$

$$ll(D) = -\frac{Nd}{2}\log(2\pi) - \frac{N}{2}\log|\Sigma| - \frac{1}{2}\sum_{i=1}^{N}(x_i-\mu)^T \Sigma^{-1}(x_i-\mu)$$

---

$$\frac{\partial}{\partial \mu} ll(D) = -\frac{1}{2}\sum_{i=1}^{N} \frac{\partial}{\partial \mu}\left[(x_i-\mu)^T \Sigma^{-1}(x_i-\mu)\right]$$

$$\frac{d}{da} a^T A a = (A + A^T)a$$

$$= -\frac{1}{2}\sum_{i=1}^{N} \left(\Sigma^{-1} + \Sigma^{-T}\right)(x_i-\mu)(-1)$$

$$\frac{\partial}{\partial \mu} ll(D) = 0$$

$$\sum_{i=1}^{N}(x_i-\mu) = 0$$

$$\hat{\mu}_{MLE} = \frac{1}{N}\sum_{i=1}^{N} x_i$$

$$ll(D) = -\frac{N}{2}\log|\Sigma| - \frac{1}{2}\sum (x_i-\mu)^T \Sigma^{-1}(x_i-\mu)$$

$$\Lambda = \Sigma^{-1}$$

$$\boxed{|\Lambda| = \frac{1}{|\Sigma|}} \quad \text{— cookbook}$$

$$ll(D) = \frac{N}{2}\log|\Lambda| - \frac{1}{2}\sum (x_i-\mu)^T \Lambda (x_i-\mu)$$

"Trace Trick"

$$a^T A a = tr(a a^T A)$$

$(1\times d)\ (d\times d)\ (d\times 1)$ — $(d\times 1)\ (1\times 1)\ (1\times d)\ (d\times d)$

$$tr(B) = \sum_{i=1}^{d} B_{ii}$$

$$ll(D) = \frac{N}{2}\log|\Lambda| - \frac{1}{2}\sum_{i=1}^{N} tr\left[(x_i-\mu)(x_i-\mu)^T \Lambda\right]$$

$$= \frac{N}{2}\log|\Lambda| - \frac{1}{2} tr\left(\left[\sum_{i=1}^{N}(x_i-\mu)(x_i-\mu)^T\right]\Lambda\right)$$

$$\sum_{i=1}^{N}(x_i-\mu)(x_i-\mu)^T = \text{scatter matrix} = S_\mu$$

$$ll(D) = \frac{N}{2}\log|\Lambda| - \frac{1}{2} tr(S_\mu \Lambda)$$

$$\frac{d\,\ell(D)}{d\,\Lambda} = \frac{N}{2}\frac{d}{d\Lambda}\log|\Lambda| - \frac{1}{2}\frac{d}{d\Lambda}\mathrm{tr}(S_\mu \Lambda)$$

$$\frac{d}{dA}\log|A| = (A^{-1})^T$$

$$\frac{d}{dA}\mathrm{tr}(BA) = B$$

$$\frac{d}{d\Lambda}\ell(D) = \frac{N}{2}(\Lambda^{-1})^T - \frac{1}{2}S_\mu$$

$$\Lambda^{-1} = \Sigma \qquad \Sigma^T = \Sigma$$

$$\frac{d}{d\Lambda}\ell(D) = \frac{N}{2}\Sigma - \frac{1}{2}S_\mu$$

Set it to 0, solve for $\Sigma$

$$\hat{\Sigma}_{MLE} = \frac{1}{N}S_\mu = \frac{1}{N}\sum_{i=1}^{N}(x_i-\mu)(x_i-\mu)^T$$

↑ sample covariance matrix