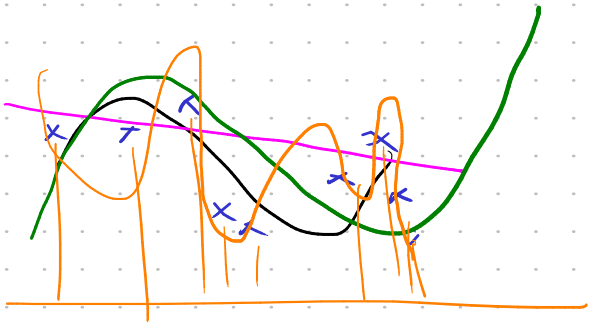
① Bayesian model selection



D $\quad M_1, M_2, \dots, M_K$

$p(D|\bar\theta_1)\ \bar\theta_2, \dots, \bar\theta_K$

$$p(M_k|D) \propto \underbrace{p(M_k)}_{\text{prior}} \underbrace{p(D|M_k)}_{\text{marginal likelihood}} \qquad M_k \to p(D|\bar\theta_k)$$

likelihood

$$p(\bar\theta_k|D, M_k) = \frac{\overbrace{p(\bar\theta_k|M_k)}^{\text{prior}} \cdot p(D|\bar\theta_k, M_k)}{p(D|M_k)}$$

evidence probability обоснованность

$$p(D|M_k) = \int p(\bar\theta_k|M_k)\, p(D|\bar\theta_k, M_k)\, d\bar\theta_k$$

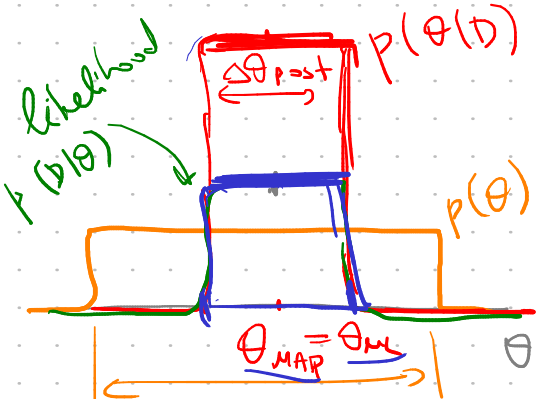
Пример: coin toss $\quad n = 20$

- M_0: $\theta = \frac{1}{2}$
- M_1: θ – парам., $p(\theta) \simeq$ Unif $=$ Beta$(1,1)$

$D = hhttth\ \dots$

$$\begin{cases} p(D|M_0) = 2^{-n} \\ p(D|M_1) = \int \theta^k (1-\theta)^{n-k}\, d\theta = \text{Beta}(k+1, n-k+1) \end{cases}$$

② Иллюстративный пример



$$p(\theta|D) = \frac{1}{\Delta\theta_{post}} \qquad p(\theta) = \frac{1}{\Delta\theta_{pri}}$$

$\bar\theta \in \mathbb{R}^d$

$$p(D) = \int p(D|\bar\theta)\, p(\bar\theta)\, d\theta = \int p(D|\theta_{ML}) \cdot \left(\frac{1}{\Delta\theta_{pri}}\right)^d d\theta$$

$$\left(\Delta\theta_{post}\right)^d$$

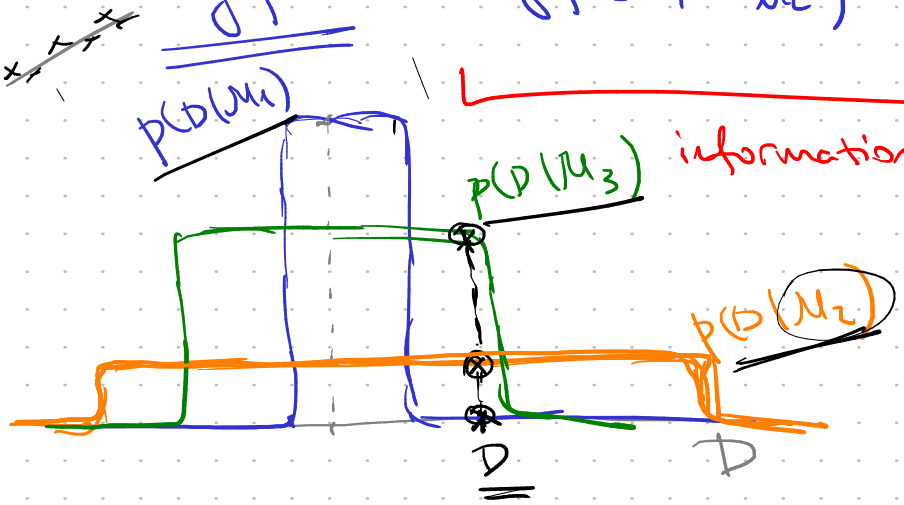
$$= p(D|\theta_{ML}) \left(\frac{\Delta\theta_{post}}{\Delta\theta_{pri}}\right)^d$$

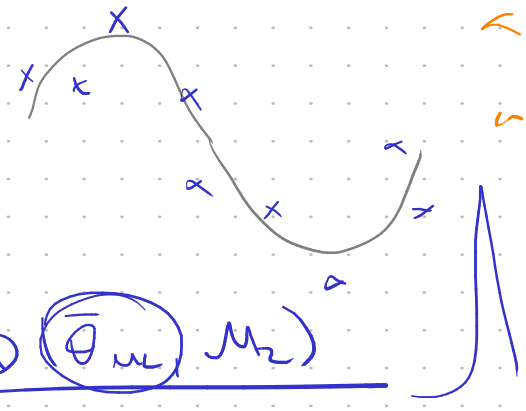
$$\log p(D) = \log p(D|\theta_{ML}) - \log \frac{\Delta\theta_{prior}}{\Delta\theta_{post}}$$

max log-likelihood − penalty

$$\log p(D) = \log p(D|\bar{\theta}_{ML}) - d \log \frac{\Delta\theta_{prior}}{\Delta\theta_{post}}$$

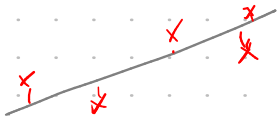
information criterion



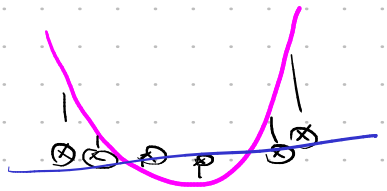


$p(D|\bar{\theta}_{ML}, M_2)$

(3) Sanity check $\quad D \sim M_{true}$



$p(D|M_{true}) \geq p(D|M) \quad \forall M$?



$$\forall M \; \mathbb{E}_{D\sim M_{true}}\left[p(D|M_{true}) \geq p(D|M)\right] \geq 0$$

$$= KL\left(p(D|M_{true}) \,\|\, p(D|M)\right)$$

$$\mathbb{E}_{D\sim M_{true}}\left[\log \frac{p(D|M_{true})}{p(D|M)}\right] = \int p(D|M_{true}) \log \frac{p(D|M_{true})}{p(D|M)} dD \geq 0$$

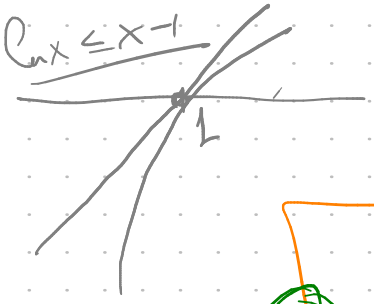
$$KL(p\|q) = \int p(\bar{x}) \log \frac{p(\bar{x})}{q(\bar{x})} d\bar{x} = \mathbb{E}_{p(\bar{x})}\left[\log \frac{p(\bar{x})}{q(\bar{x})}\right]$$

$$= \int p(\bar{x}) \log p(\bar{x}) d\bar{x} - \int p(\bar{x}) \log q(\bar{x}) d\bar{x} = H(p) - H(p,q)$$

cross-entropy

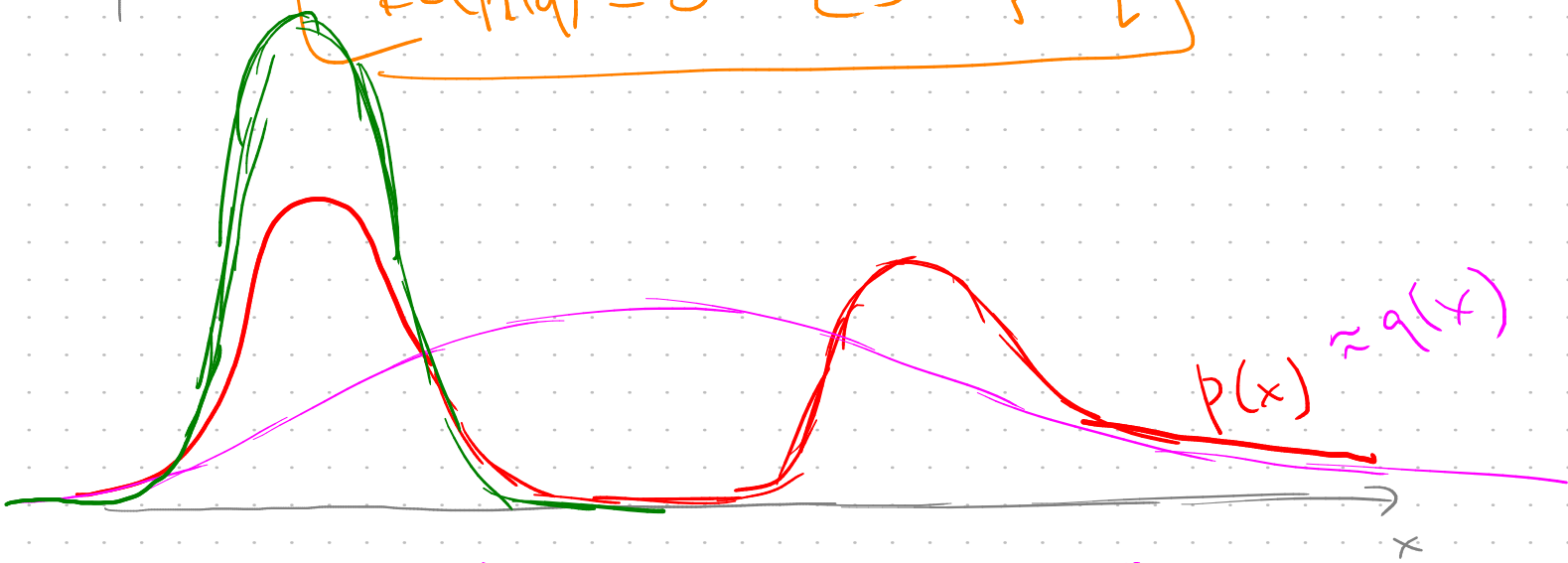
$KL(p\|q) \geq 0$ $\quad$ $(KL(p\|q) \neq KL(q\|p))$

$$\int p \log \frac{p}{q}\, dx = -\int p \log \frac{q}{p}\, dx \geq -\int p \left(\frac{q}{p} - 1\right) dx$$



$$= -\int q\, dx + \int p\, dx = -1 + 1 = 0$$

$KL(p\|q) = 0 \iff p \overset{a.e.}{=} q$



$KL(p\|q) \to \min$ $\qquad$ $KL(p\|q) = \int p \log \frac{p}{q}\, dx$

$q \neq 0 \quad p > 0$

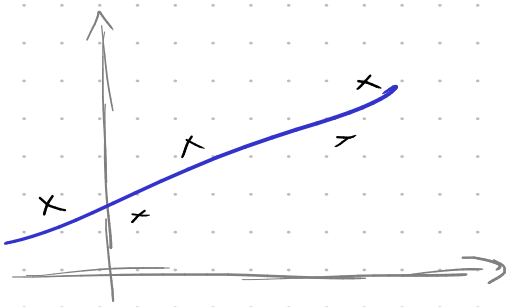
$KL(q\|p) \to \min$

$$KL(q\|p) = \int q \log \frac{q}{p}\, dx$$

$p \neq 0,\ q \neq 0$

(4) Empirical Bayes

hyperparameters



$\bar{w} \in \mathbb{R}^d$

$p(y \mid \bar{x}, \bar{w}) = \mathcal{N}(y \mid \bar{w}^T \bar{x}, \sigma^2)$

$p(\bar{w}) = \mathcal{N}(\bar{w} \mid \bar{0}, \sigma_0^2 I)$

$$\log p(\bar{w} \mid \bar{y}, X) =$$

$$= \text{Const} - \underbrace{\frac{1}{2\sigma^2}}_{=\beta} \cdot \sum_{n=1}^{N} (y_n - \bar{w}^T \bar{x}_n)^2 - \underbrace{\frac{1}{2\sigma_0^2}}_{=\alpha} \bar{w}^T \bar{w}$$

$$\log p(\bar{w}|y, X, \alpha, \beta) = \text{const} - \frac{\beta}{2}\sum_n \left(y_n - \bar{w}^T \bar{x}_n\right)^2 - \frac{\alpha}{2}\bar{w}^T\bar{w}$$

$$p(\bar{w}|D, \alpha, \beta) = \frac{p(\bar{w}|\alpha)\, p(D|\bar{w}, \beta)}{p(D|\alpha, \beta)}$$

MLE - II

marginal likelihood

$\alpha, \beta \to \max$

$$\log p(\bar{y}|X, \alpha, \beta) = \log \int p(\bar{w}|\alpha)\, p(\bar{y}|X, \bar{w}, \beta)\, d\bar{w}$$

$$= \log \int \left(\frac{\alpha}{2\pi}\right)^{d/2} \left(\frac{\beta}{2\pi}\right)^{N/2} e^{-\frac{\beta}{2}(\bar{y}-X\bar{w})^T(\bar{y}-X\bar{w}) - \frac{\alpha}{2}\bar{w}^T\bar{w}} d\bar{w} =$$

$$= E(\bar{w})$$

$$E(\bar{w}) = -\frac{\alpha}{2}\bar{w}^T\bar{w} - \frac{\beta}{2}\left(\bar{y}^T\bar{y} - 2\bar{w}^T X^T\bar{y} + \bar{w}^T X^T X\bar{w}\right)$$

$$= -\frac{1}{2}(\bar{w}-\bar{\mu})^T A(\bar{w}-\bar{\mu}) - \frac{\beta}{2}\bar{y}^T\bar{y} + \frac{1}{2}\bar{\mu}^T A\bar{\mu}$$

$$A = \beta X^T X + \alpha \cdot I$$

$$A\bar{\mu} = \beta \cdot X^T\bar{y}, \quad \bar{\mu} = \beta \cdot A^{-1} X^T\bar{y}$$

$$E(\bar{\mu}) = -\frac{\alpha}{2}\bar{\mu}^T\bar{\mu} - \frac{\beta}{2}(\bar{y} - X\bar{\mu})^T(\bar{y} - X\cdot\bar{\mu})$$

$$E(\bar{w}) = E(\bar{\mu}) - \frac{1}{2}(\bar{w}-\bar{\mu})^T A(\bar{w}-\bar{\mu})$$

$$= \log \int \left(\frac{\alpha}{2\pi}\right)^{d/2}\left(\frac{\beta}{2\pi}\right)^{N/2} e^{E(\bar{\mu}) - \frac{1}{2}(\bar{w}-\bar{\mu})^T A(\bar{w}-\bar{\mu})} d\bar{w} =$$

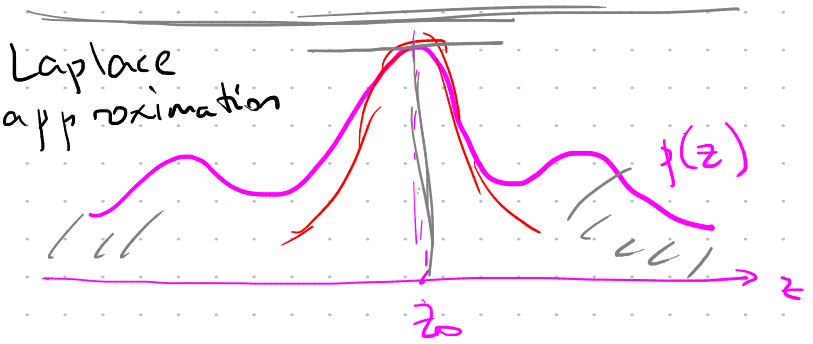
$$= \log(\cdot)(\sim) e^{E(\bar{\mu})} \cdot \underbrace{\int e^{-\frac{1}{2}(\bar{w}-\bar{\mu})^T A (\bar{w}-\bar{\mu})} d\bar{w}}_{\frac{1}{\sqrt{\det A}} \cdot (2\pi)^{d/2}} =$$

$$= \frac{d}{2}\log\alpha + \frac{N}{2}\log\beta - \frac{1}{2}\log\det A - \frac{\alpha}{2}\bar{\mu}^T\bar{\mu} - \frac{\beta}{2}(\bar{y}-X\bar{\mu})^T(\bar{y}-X\bar{\mu}) - \cancel{\frac{2d+N}{2}\log 2\pi} \longrightarrow \max_{\alpha,\beta}$$

(5) Bayesian information criterion (BIC)

$$p(D) = \int p(\bar{\theta}) \overbrace{p(D|\bar{\theta})}^{\text{Lapl. appr.}} d\bar{\theta} = \int \underline{p(\bar{\theta})\, e^{l(\bar{\theta})}}\, d\bar{\theta},$$

$$l(\bar{\theta}) = \log p(D|\bar{\theta}) = \sum_{n=1}^{N} \log p(d_n|\bar{\theta})$$



$$\log p(z) \approx \log p(z_0) + \frac{1}{2}\underbrace{\left.\frac{\partial^2 \log p(z)}{\partial z^2}\right|_{z_0}}_{-A} \cdot (z-z_0)^2$$

$$\log p(z) \approx \log p(z_0) - \frac{A}{2}(z-z_0)^2$$

$$p(z) \approx p(z_0) \cdot \underline{e^{-\frac{A}{2}(z-z_0)^2}}$$

$$p(\bar{z}) \approx p(\bar{z}_0) \cdot e^{-\frac{1}{2}(\bar{z}-\bar{z}_0)^T A (\bar{z}-\bar{z}_0)}$$

$$l(\bar\theta) \approx l(\bar\theta_{ML}) - \frac{N}{2}(\bar\theta - \bar\theta_{ML})^T J(\bar\theta_{ML})(\bar\theta - \bar\theta_{ML}),$$ где

$$J(\bar\theta_{ML}) = -\frac{1}{N}\cdot\left(\left.\frac{\partial^2 \log p(D|\bar\theta)}{\partial\theta_i \partial\theta_j}\right|_{\bar\theta=\bar\theta_{ML}}\right)$$

$$p(\bar\theta) \approx p(\bar\theta_{ML}) + (\bar\theta - \bar\theta_{ML})^T \nabla_{\bar\theta} p(\bar\theta)\Big|_{\bar\theta=\bar\theta_{ML}}$$

$$p(D) \approx \int \left(p(\bar\theta_{ML}) + (\bar\theta - \bar\theta_{ML})^T \nabla_\theta p(\bar\theta_{ML})\right) e^{l(\bar\theta_{ML}) - \frac{N}{2}(\bar\theta - \bar\theta_{ML})^T J(\bar\theta_{ML})(\bar\theta - \bar\theta_{ML})} d\bar\theta$$

$$\int (\bar\theta - \bar\theta_{ML})^T \nabla_{\bar\theta} \cdot e^{-\frac{N}{2}(\bar\theta - \bar\theta_{ML})^T J (\bar\theta - \bar\theta_{ML})} d\bar\theta$$

$$\int (\bar\theta - \bar\mu)\cdot \mathcal{N}\left(\bar\theta \,|\, \bar\mu, \tfrac{1}{N} J^{-1}\right) d\bar\theta = 0$$

$$= e^{l(\bar\theta_{ML})}\, p(\bar\theta_{ML}) \cdot \int e^{-\frac{N}{2}(\bar\theta - \bar\theta_{ML})^T J(\bar\theta_{ML})(\bar\theta - \bar\theta_{ML})} d\bar\theta$$

$$= e^{l(\bar\theta_{ML})}\, p(\bar\theta_{ML}) \cdot (2\pi)^{d/2} \cdot \det\left(N \cdot J(\bar\theta_{ML})\right)^{-\frac{1}{2}} =$$

$$= \text{—} \cdot N^{-\frac{d}{2}} \cdot (\det J)^{-1/2}$$

Occam's factor

$$\log p(D) \approx \log p(D|\bar\theta_{ML}) + \left(\log p(\bar\theta_{ML}) + \frac{d}{2}\log 2\pi - \frac{d}{2}\log N - \frac{1}{2}\log\det J(\bar\theta_{ML})\right)$$

$$BIC(M) = -2\log p(D|\bar{\theta}_{ML}) + d\log N$$

$d = \dim\bar{\theta}$, $N = |D|$

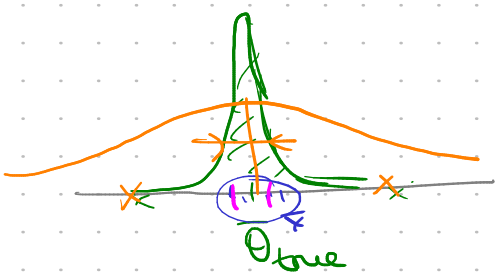
(6) Fisher information matrix

$p(D|\bar{\theta}) \xrightarrow{\bar{\theta}\to\max} \bar{\theta}_{ML}$

$p_{model}(\bar{x}) = p(\bar{x}|\bar{\theta}) \rightsquigarrow p_{data}(\bar{x})$

Assumption: $p_{data}(\bar{x}) = p(\bar{x}|\bar{\theta}_{true})$

D $\bar{x}$ ~ в среднем как хорошо $\bar{x}$ про $\bar{\theta}_{true}$?



$x \sim p_{data}(x) = p(x|\theta_{true})$

$$\mathbb{E}_x\left[\frac{\partial \log p(x|\theta)}{\partial\theta}\right] = \int \frac{\frac{\partial p(x|\theta)}{\partial\theta}}{p(x|\theta)} \cdot p(x|\theta)\,d\theta$$

$$= \frac{\partial}{\partial\theta}\int p(x|\theta)\,d\theta = \frac{\partial}{\partial\theta}1 = 0$$

Fisher information:

$$I(\theta) = \mathbb{E}_{x\sim p(x|\theta)}\left[\left(\frac{\partial}{\partial\theta}\log p(x|\theta)\right)^2\right] = \int\left(\frac{\partial}{\partial\theta}\log p(x|\theta)\right)^2 p(x|\theta)\,dx$$

$$\frac{\partial^2 \log p}{\partial\theta^2} = \frac{\partial}{\partial\theta}\left(\frac{\partial p/\partial\theta}{p}\right) = \frac{\frac{\partial^2 p}{\partial\theta^2}\cdot p - \left(\frac{\partial p}{\partial\theta}\right)^2}{p^2} = \frac{\frac{\partial^2 p}{\partial\theta^2}}{p} - \frac{\left(\frac{\partial p}{\partial\theta}\right)^2}{p^2} =$$

$$\mathbb{E}_p\left[\frac{\partial^2 p/\partial\theta^2}{p}\right] = \int \frac{\partial^2 p/\partial\theta^2}{p}\cdot p\,d\theta = \frac{\partial^2}{\partial\theta^2}\int p\,d\theta = 0$$

$$= -\frac{(\partial p/\partial\theta)^2}{p^2} = -\left(\frac{\partial \log p}{\partial\theta}\right)^2$$

$$I(\theta) = -\,E_{p(x|\theta)}\left[\frac{\partial^2 \log p(\bar{x}|\bar{\theta})}{\partial\theta^2}\right]$$

$$I(\bar{\theta})_{ij} = E_{p(\bar{x}|\bar{\theta})}\left[\frac{\partial \log p(\bar{x}|\bar{\theta})}{\partial\theta_i}\cdot\frac{\partial \log p(\bar{x}|\bar{\theta})}{\partial\theta_j}\right]$$

$$\| \text{ecm } p_{data}(\bar{x}) = p(\bar{x}|\bar{\theta}_{true})$$

$$J(\bar{\theta})_{ij} = -E_{p(\bar{x}|\bar{\theta})}\left[\frac{\partial^2 \log p(\bar{x}|\bar{\theta})}{\partial\theta_i \partial\theta_j}\right]$$

(7) Ugen AIC

$X = \{\bar{x}_1, \dots, \bar{x}_N\} \sim p_{data}(\bar{x})$

$p(\bar{x}|\bar{\theta})$, $\bar{\theta} \in \mathbb{R}^d$ ....

$$\bar{\theta}_{ML} = \arg\max_{\bar{\theta}} p(X|\bar{\theta})$$

2. $p(\bar{x}|\bar{\theta}_{ML})$

min $\bar{\theta}$

$$KL(p_{data} \| p(\bar{x}|\bar{\theta})) = E_{\dots}[\log \dots] =$$

$$= E_{p_{data}}[\log p_{data}(\bar{x})] - E_{p_{data}}[\log p(\bar{x}|\bar{\theta})]$$

$\bar{\theta} \to \max$

$$E_{p_{data}}[\log p(\bar{x}|\bar{\theta})] \approx \frac{1}{N}\sum_{n=1}^{N} \log p(\bar{x}_n|\bar{\theta})$$

$\bar{\theta} \to \max$

$$\mathbb{E}_{p_{data}}\left[\log p(\bar{x}|\bar{\theta}_{ML})\right] \approx \frac{1}{N}\sum_n \log p(\bar{x}_n|\bar{\theta}_{ML})$$

AIC

$\log p(D|\bar{\theta}_{ML})$

$$p_{data}(\bar{x}) = p(\bar{x}|\bar{\theta}_0)$$

$$\bar{\theta}_0 = \arg\max_{\bar{\theta}} \mathbb{E}_{p_{data}}\left[\log p(\bar{x}|\bar{\theta})\right]$$

- $\bar{\theta}_{ML}(X) \xrightarrow{N\to\infty} \bar{\theta}_0$
- $\sqrt{N}\left(\bar{\theta}_{ML} - \bar{\theta}_0\right) \to \mathcal{N}\left(\bar{0}, I(\bar{\theta}_0)^{-1}\right)$

$$p_{data}(\bar{x}) \neq p(\bar{x}|\bar{\theta}_0)$$

$$\bar{\theta}_{ML}(X) \to \bar{\theta}_0, \quad N\to\infty$$

$$\sqrt{N}\left(\bar{\theta}_{ML} - \bar{\theta}_0\right) \to \mathcal{N}\left(\bar{0}, J^{-1}(\bar{\theta}_0) I(\bar{\theta}_0) J^{-1}(\bar{\theta}_0)\right)$$

bias

$$b(p_{data}) = \mathbb{E}_{D\sim p_{data}}\left[\log p(D|\bar{\theta}_{ML}) - N\cdot \mathbb{E}_{p_{data}}\left[\log p(\bar{x}|\bar{\theta}_{ML})\right]\right] =$$

$$= \mathbb{E}_D\left[\log p(D|\bar{\theta}_{ML}(D)) - \log p(D|\bar{\theta}_0)\right] + B_1$$

$$+ \mathbb{E}_D\left[\log p(D|\bar{\theta}_0) - N\cdot \mathbb{E}_{p_{data}}\left[\log p(\bar{x}|\bar{\theta}_0)\right]\right] B_2$$

$$+ \mathbb{E}_D\left[N\cdot \mathbb{E}_{p_{data}}\left[\log p(\bar{x}|\bar{\theta}_0)\right] - N\cdot \mathbb{E}_{p_{data}}\left[\log p(\bar{x}|\bar{\theta}_{ML})\right]\right] B_3$$

$$AIC = 2\log p(D|\bar{\theta}_{ML}) - 2\,b(p_{data}) \quad 2d$$

$$BIC = 2\log p(D|\bar{\theta}_{ML}) - \dots \quad d\log N$$