Lecture 9

Bayesian Curve Fitting, Second Lesson on "Learning from Data"

Bao-Jun Cai, 4/29/2026

Introduction to Algorithms for Data Science and Physics IMP@Fudan, 2026

Topics of this lecture:

- high-dimensional Gaussian $\Delta^2 = (\mathbf{x} - \vec{\mu})^\top \vec{\Sigma}^{-1} (\mathbf{x} - \vec{\mu})$
- correlation among random variables $\rho_{xy} = \mathrm{cov}[x, y]/\sigma[x]\sigma[y]$
- Bayesian inference, maximum likelihood
- biased estimate for variance $\mathrm{E}[\widehat{\sigma}^2_{\mathrm{ML}}] = (1 - m^{-1})\sigma^2$
- variance reduction from data generation
- model selection, Bayes factor

$$\overbrace{p(\mathbf{w}|\{\mathrm{data}\})}^{\text{posterior for } \mathbf{w}} = \frac{\overbrace{p(\{\mathrm{data}\}|\mathbf{w})}^{\text{likelihood for } \mathbf{w}} \overbrace{p(\mathbf{w})}^{\text{prior}}}{\underbrace{p(\{\mathrm{data}\})}_{\text{evidence}}}$$

$$\sigma^2_m(\mathbf{x}) = \overbrace{1/\beta}^{\text{noise of data sample}} + \underbrace{\vec{\phi}^\top(\mathbf{x}) \mathbf{S}_m \vec{\phi}(\mathbf{x})}_{\text{uncertainty of learning parameter}}$$

Review: first lesson on "learning from data"

Lecture 5

First Lesson on "Learning from Data", Parameter Estimate

Bao-Jun Cai, 4/1/2026

Introduction to Algorithms for Data Science and Physics IMP@Fudan, 2026

Topics of this lecture:

- parameter estimate for the learning model $f_{\vec{\theta}}(x) = ax + b$
- parabolic loss function and its optimization $J(\vec{\theta}) = 2^{-1} \sum_{i=1}^{m} [f_{\vec{\theta}}(x^{(i)}) - y^{(i)}]^2$
- goodness of learning model: decomposition of bias, variance
- penalty term, belief in data, avoidance of singularity $J \to J + \lambda g$
- normal equation $\vec{\Phi}^\top \vec{\Phi} \vec{w} = \vec{\Phi}^\top \vec{y}$
- stochastic gradient descent

1

2

Some basic concepts:

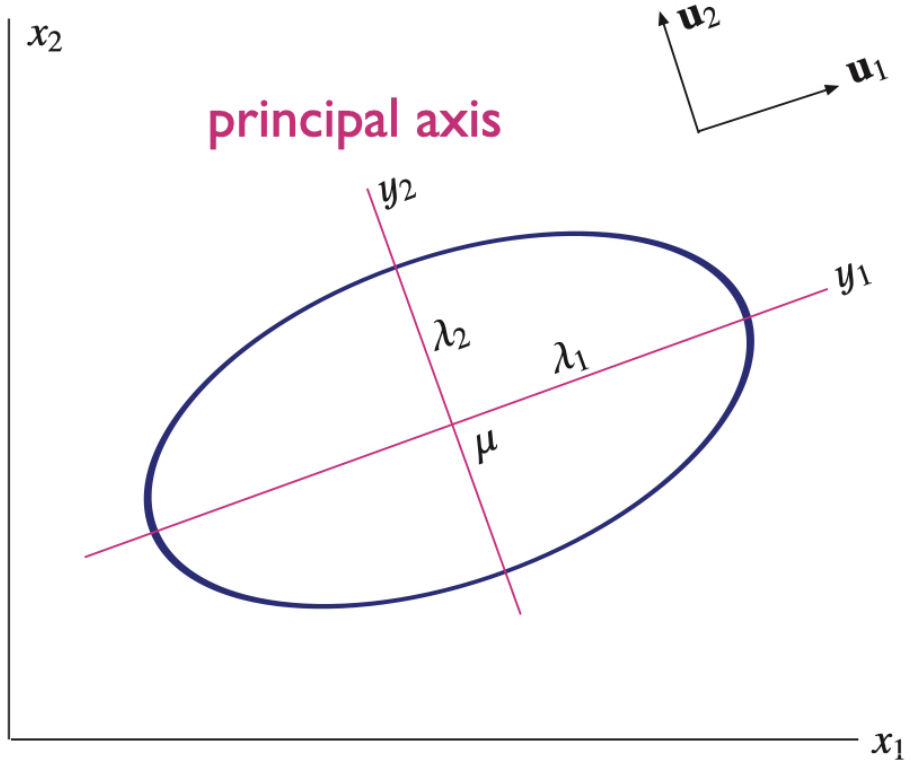
- learning model
- model parameter
- error/loss function
- bias+variance decomp.
- normal equation
- gradient descent
- regularization
- ....

2D Gaussian

$\mathbf{x} \in \mathbf{R}^d$

1D: $e^{-(x-\mu)^2/2\sigma^2}$

$$p(\mathbf{x}) = \mathcal{N}(\mathbf{x}|\vec{\mu}, \vec{\Sigma}) = \mathcal{N}(\vec{\mu}, \vec{\Sigma}) = \frac{1}{(2\pi)^{d/2}} \frac{1}{(\det \vec{\Sigma})^{1/2}} \exp\left[-\frac{1}{2}(\mathbf{x} - \vec{\mu})^\top \vec{\Sigma}^{-1}(\mathbf{x} - \vec{\mu})\right]$$



Mahalanobis distance:

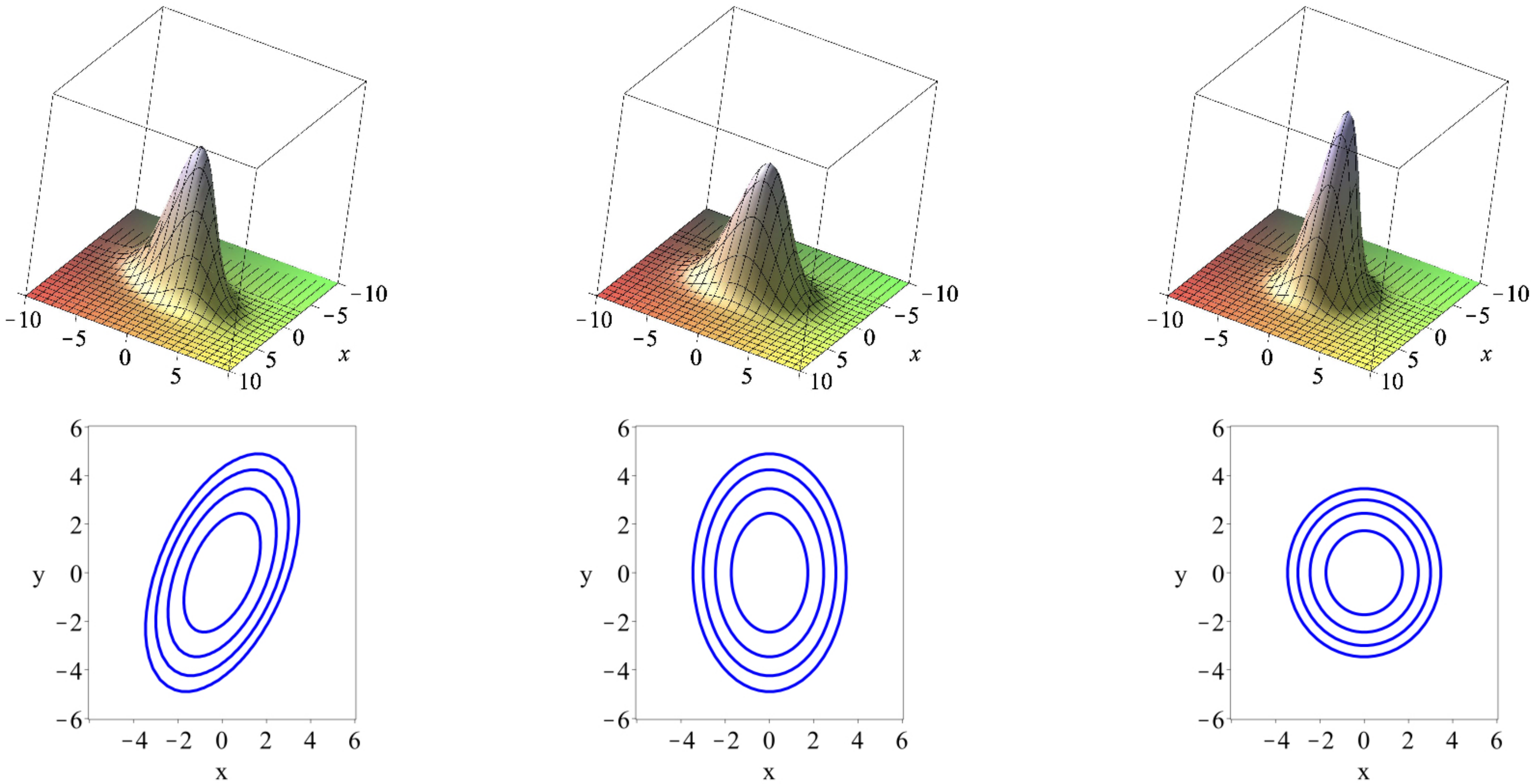
$$\Delta^2 = (\mathbf{x} - \vec{\mu})^\top \vec{\Sigma}^{-1}(\mathbf{x} - \vec{\mu})$$

$$\vec{\Sigma}\mathbf{u}_i = \lambda_i \mathbf{u}_i,\ \mathbf{u}_i^\top \mathbf{u}_j = \delta_{ij}$$

$$\vec{\Sigma} = \sum_{i=1}^{d} \lambda_i \mathbf{u}_i \mathbf{u}_i^\top,\ \vec{\Sigma}^{-1} = \sum_{i=1}^{d} \lambda_i^{-1} \mathbf{u}_i \mathbf{u}_i^\top$$

$$\Delta^2 = \sum_{i=1}^{d} y_i^2/\lambda_i,\ y_i = \mathbf{u}_i^\top (\mathbf{x} - \vec{\mu})$$

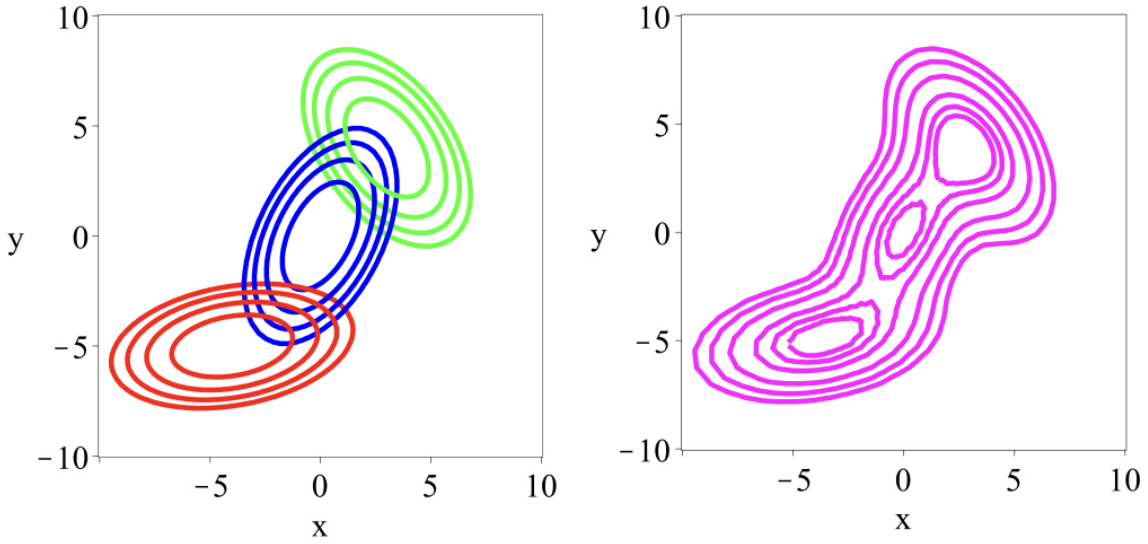
Examples: 2D Gaussian

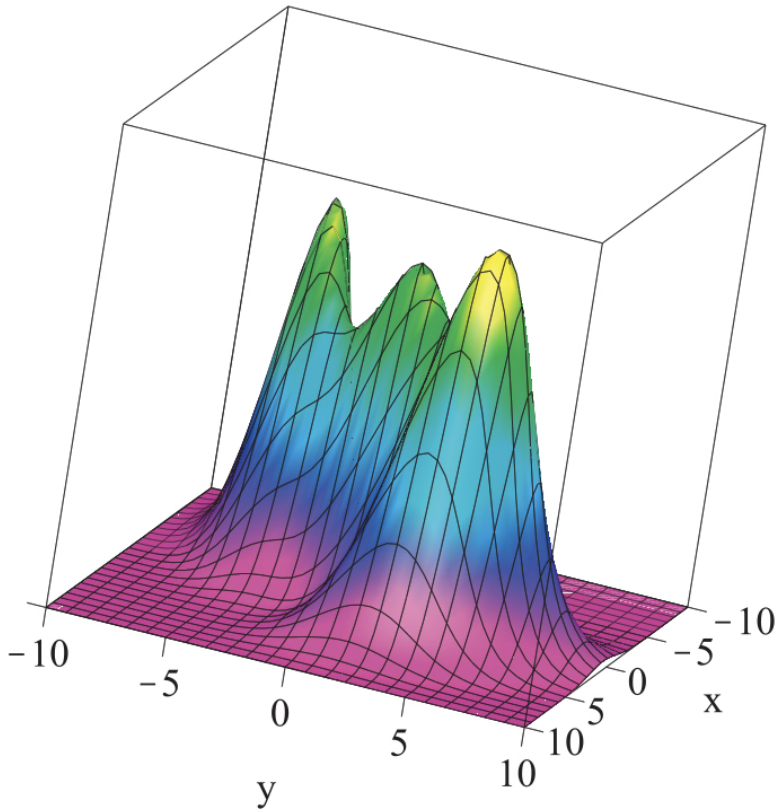


Importance of Gaussian: mixing

$$p(\mathbf{x}) = \sum_i \alpha_i \mathcal{N}_i(\vec{\mu}_i, \vec{\Sigma}_i)$$

$$\Delta^2_{\text{eff}} = -2 \ln \left[\sum_{j=1}^{3} \exp\left(-\frac{1}{2} (x - \vec{\mu}_j)^\top \vec{\Sigma}_j^{-1} (x - \vec{\mu}_j) \right) \right]$$





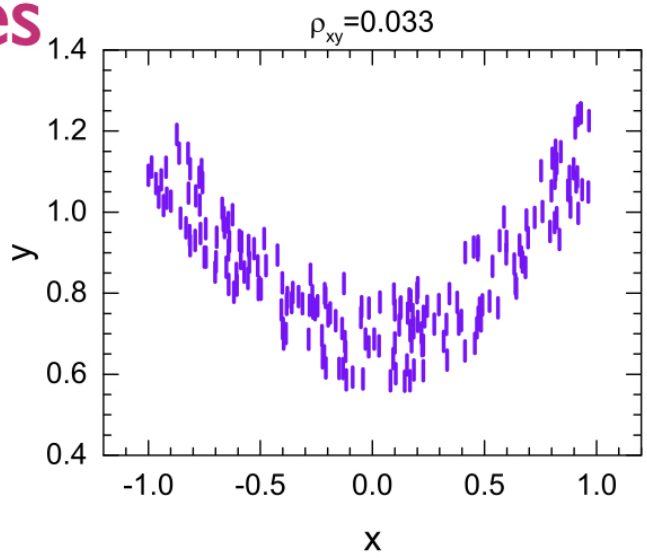
$$\vec{\Sigma}_1 = \begin{pmatrix} 3 & 2 \\ 2 & 6 \end{pmatrix}, \vec{\Sigma}_2 = \begin{pmatrix} 18/5 & -2 \\ -2 & 5 \end{pmatrix}, \vec{\Sigma}_3 = \begin{pmatrix} 8 & 1 \\ 1 & 2 \end{pmatrix}$$

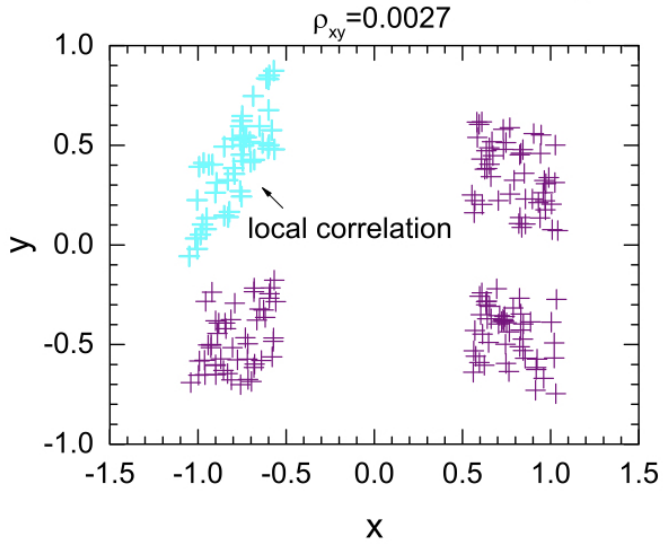
Correlations between random variables

$$\overbrace{\mathrm{cov}[x,y]}^{\text{correlation of fluctuations}} = \mathrm{E}\left[\delta x \delta y\right] = \mathrm{E}\left[\left(x - \mathrm{E}[x]\right)\left(y - \mathrm{E}[y]\right)\right] = \mathrm{E}[xy] - \mathrm{E}[x]\mathrm{E}[y]$$

$$\vec{\Sigma} = \mathrm{E}\left[\left\{\begin{pmatrix} x \\ y \end{pmatrix} - \mathrm{E}\begin{pmatrix} x \\ y \end{pmatrix}\right\}\left\{\begin{pmatrix} x \\ y \end{pmatrix} - \mathrm{E}\begin{pmatrix} x \\ y \end{pmatrix}\right\}^{\top}\right] = \begin{pmatrix} \mathrm{var}[x] & \mathrm{cov}[x,y] \\ \mathrm{cov}[x,y] & \mathrm{var}[y] \end{pmatrix}$$

correlation density: $\rho_{xy} = \dfrac{\mathrm{cov}[x,y]}{\sigma[x]\sigma[y]} = \dfrac{\mathrm{cov}[x,y]}{\mathrm{var}^{1/2}[x]\mathrm{var}^{1/2}[y]}$





Ex.: Write the expression for the correlation matrix in 3D.

Decomposition of Gaussian**

$\vec{\Lambda}_{kk'}, k, k' = a, b$

$\mathcal{N}(\mathbf{x}|\vec{\mu}, \vec{\Sigma})$, $\mathbf{x} = (\mathbf{x}_a, \mathbf{x}_b)^\top$, $\vec{\mu} = (\vec{\mu}_a, \vec{\mu}_b)^\top$, $\vec{\Sigma} = \begin{pmatrix} \vec{\Sigma}_{aa} & \vec{\Sigma}_{ab} \\ \vec{\Sigma}_{ba} & \vec{\Sigma}_{bb} \end{pmatrix}$, precision matrix: $\vec{\Lambda} = \vec{\Sigma}^{-1}$

(1) conditional probability $p(\mathbf{x}_a|\mathbf{x}_b) \sim \mathcal{N}(\vec{\mu}_{a|b}, \vec{\Sigma}_{a|b})$

$$\vec{\mu}_{a|b} = \vec{\mu}_a + \vec{\Sigma}_{ab}\vec{\Sigma}_{bb}^{-1}(\mathbf{x}_b - \vec{\mu}_b) = \vec{\mu}_a + \vec{\Lambda}_{aa}^{-1}\vec{\Lambda}_{ab}(\mathbf{x}_b - \vec{\mu}_b), \quad \vec{\Sigma}_{a|b} = \vec{\Sigma}_{aa} - \vec{\Sigma}_{ab}\vec{\Sigma}_{bb}^{-1}\vec{\Sigma}_{ba} = \vec{\Lambda}_{aa}^{-1}$$

(2) edge distribution $p(\mathbf{x}_a) = \int p(\mathbf{x}_a, \mathbf{x}_b) d\mathbf{x}_b \sim \mathcal{N}(\mathbf{x}_a|\vec{\mu}_a, \vec{\Sigma}_{aa})$

(3) Bayes' theorem:

known: $p(\mathbf{x}) = \mathcal{N}(\mathbf{x}|\vec{\mu}, \vec{\Lambda}^{-1})$, $p(\mathbf{y}|\mathbf{x}) = \mathcal{N}(\mathbf{y}|\mathbf{A}\mathbf{x} + \mathbf{b}, \mathbf{L}^{-1})$

$\rightarrow p(\mathbf{y}) \sim \mathcal{N}(\mathbf{A}\vec{\mu} + \mathbf{b}|\mathbf{L}^{-1} + \mathbf{A}\vec{\Lambda}^{-1}\mathbf{A}^{\mathrm{T}})$

Ex.: show the relations:

$$\vec{\Lambda}_{aa} = (\vec{\Sigma}_{aa} - \vec{\Sigma}_{ab}\vec{\Sigma}_{bb}^{-1}\vec{\Sigma}_{ba})^{-1}$$

$$\vec{\Lambda}_{ab} = -(\vec{\Sigma}_{aa} - \vec{\Sigma}_{ab}\vec{\Sigma}_{bb}^{-1}\vec{\Sigma}_{ba})^{-1}\vec{\Sigma}_{ab}\vec{\Sigma}_{bb}^{-1}$$

$$p(\mathbf{x}|\mathbf{y}) \sim \mathcal{N}\left(\overbrace{\left(\vec{\Lambda} + \mathbf{A}^\top\mathbf{L}\mathbf{A}\right)^{-1}\left[\mathbf{A}^\top\mathbf{L}(\mathbf{y} - \mathbf{b}) + \vec{\Lambda}\vec{\mu}\right]}^{\mathrm{E}[\mathbf{x}|\mathbf{y}]} \middle| \overbrace{\left(\vec{\Lambda} + \mathbf{A}^\top\mathbf{L}\mathbf{A}\right)^{-1}}^{\mathrm{cov}[\mathbf{x}|\mathbf{y}]} \right)$$

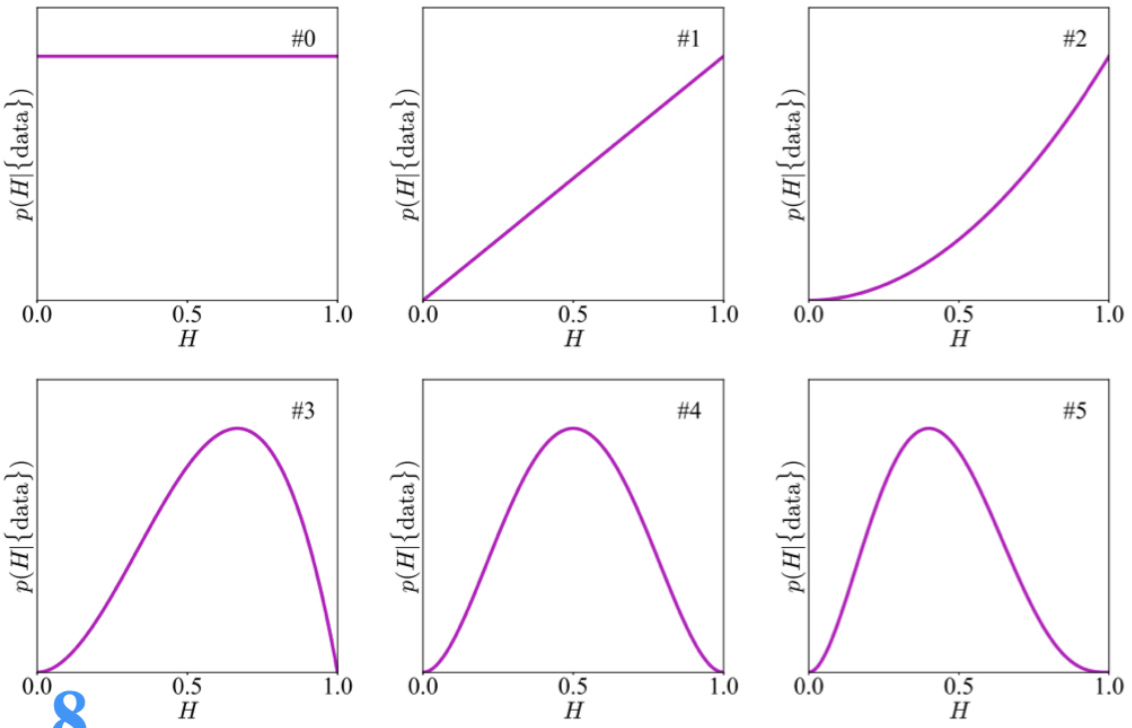
Bayesian formula revisited

example: $f_{\mathbf{w}}(x) = \sum_{j=0}^{n} w_j x^j$

$$\overbrace{p(\mathbf{w}|\{\text{data}\})}^{\text{posterior for } \mathbf{w}} = \frac{\overbrace{p(\{\text{data}\}|\mathbf{w})}^{\text{likelihood for } \mathbf{w}}\overbrace{p(\mathbf{w})}^{\text{prior}}}{\underbrace{p(\{\text{data}\})}_{\text{evidence}}}$$

$$p(A|B) = \frac{p(B|A)p(A)}{p(B)}$$

$$p(\{\text{data}\}) = \int dw p(\{\text{data}\}|\mathbf{w})p(\mathbf{w})$$



$p(H|\{\text{data}\}) \sim p(\{\text{data}\}|H) \sim H^r(1-H)^{n-r}$

(under a uniform prior)

$$\frac{(n+1)!}{r!(n-r)!}$$

$$p(H_0|\text{data}) = H_0 \Big/ \int_0^1 dH p(H|\text{data})$$

$p(5/6|\text{one head}) = 5/3 < p(5/6|\text{two heads}) = 25/12$

Maximum likelihood (ML)

problem: estimate μ, σ^2

Ex.: show that $\mathrm{E}[x^{(i)}x^{(j)}] = \mu^2 + \delta_{ij}\sigma^2$

$$p(\mathbf{X}|\mu,\sigma^2) = \prod_{i=1}^{m} \mathcal{N}(x^{(i)}|\mu,\sigma^2),\ \mathbf{X} = \left(x^{(1)}, x^{(2)}, \cdots, x^{(m)}\right)^\top$$

$$\ln p(\mathbf{X}|\mu,\sigma^2) = -\frac{1}{2\sigma^2}\sum_{i=1}^{m}(x^{(i)}-\mu)^2 - \frac{1}{2}m\ln\sigma^2 \overbrace{-\frac{1}{2}m\ln(2\pi)}^{\text{const.}}$$

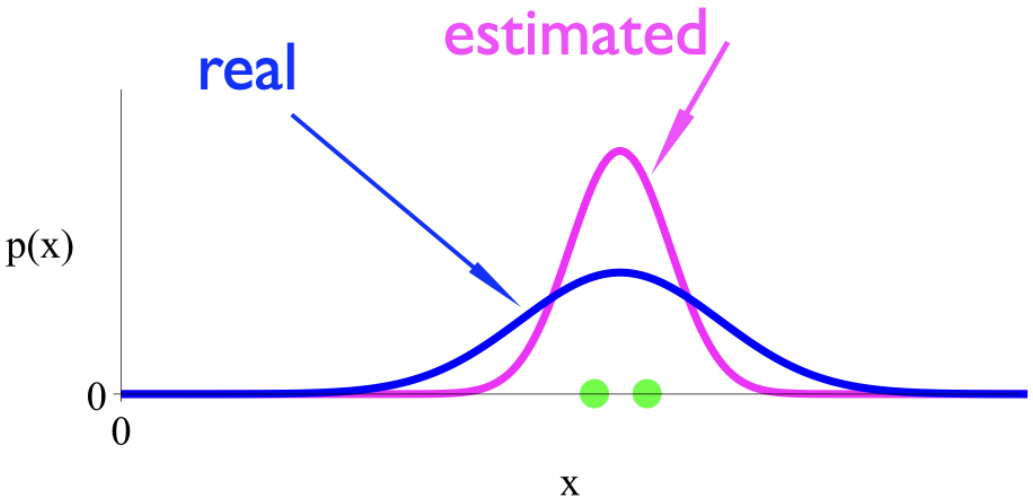
This is the least-squares!

$$\widehat{\mu}_{\mathrm{ML}} = \frac{1}{m}\sum_{i=1}^{m} x^{(i)},\ \widehat{\sigma}^2_{\mathrm{ML}} = \frac{1}{m}\sum_{i=1}^{m}\left(x^{(i)} - \widehat{\mu}_{\mathrm{ML}}\right)^2$$

Ex.: Derive them! Does the order of derivative matter?

Maximum likelihood estimate is biased for sigma^2

$$\mathrm{E}[\widehat{\mu}_{\mathrm{ML}}] = \mu,\ \mathrm{E}[\widehat{\sigma}^2_{\mathrm{ML}}] = \mathrm{E}\left[\frac{1}{m}\sum_{j=1}^{m}\left(x^{(j)} - \frac{1}{m}\sum_{i=1}^{m}x^{(i)}\right)^2\right] = \frac{m-1}{m}\sigma^2$$



$$\widetilde{\sigma}^2 = \frac{m}{m-1}\widehat{\sigma}^2_{\mathrm{ML}} = \frac{1}{m-1}\sum_{i=1}^{m}\left(x^{(i)} - \widehat{\mu}_{\mathrm{ML}}\right)^2$$

$$\begin{aligned}
\langle\widehat{\sigma}^2_{\mathrm{ML}}\rangle &= \left\langle\frac{1}{m}\sum_{j=1}^{m}\left(x^{(j)} - \frac{1}{m}\sum_{i=1}^{m}x^{(i)}\right)^2\right\rangle \\
&= \frac{1}{m}\sum_{j=1}^{m}\left\langle x^{(i),2} - \frac{2}{m}x^{(j)}\sum_{i=1}^{m}x^{(i)} + \frac{1}{m^2}\sum_{i,i'=1}^{m}x^{(i)}x^{(i')}\right\rangle \\
&= \frac{1}{m}\sum_{j=1}^{m}\left[\left\langle x^{(i),2}\right\rangle - \frac{2}{m}\sum_{i=1}^{m}\left\langle x^{(i)}x^{(j)}\right\rangle + \frac{1}{m^2}\sum_{i,i'=1}^{m}\left\langle x^{(i)}x^{(i')}\right\rangle\right] \\
&= \frac{1}{m}\sum_{j=1}^{m}\left[\mu^2 + \sigma^2 - \frac{2}{m}\left(m\mu^2 + \sigma^2\right) + \frac{1}{m^2}\sum_{i=1}^{m}\left(m\mu^2 + \sigma^2\right)\right] \\
&= \frac{1}{m}\sum_{j=1}^{m}\left(1 - \frac{1}{m}\right)\sigma^2 = \left(1 - \frac{1}{m}\right)\sigma^2
\end{aligned}$$

Maximum a posterior (MAP)

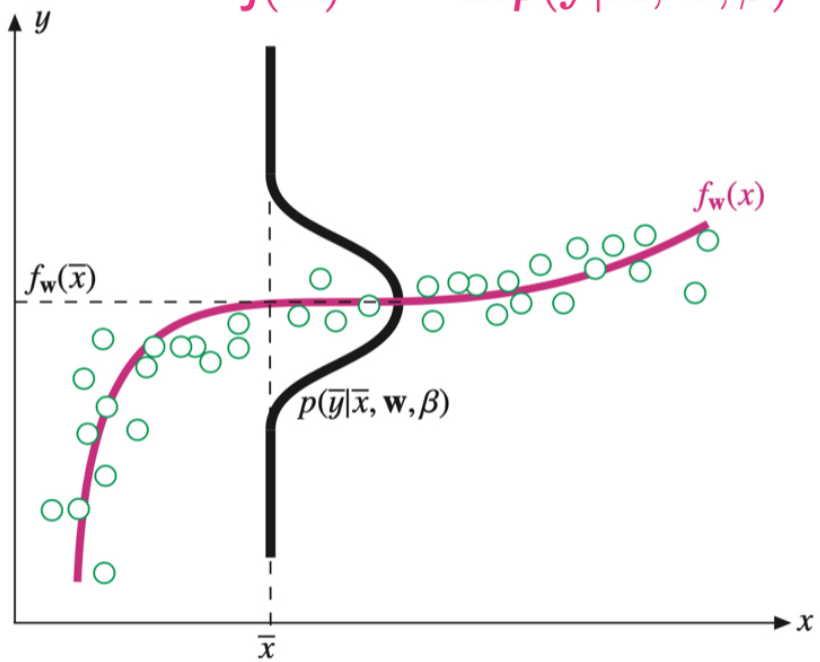
output when input arrives

$$p(\bar{y}|\bar{x}, \mathbf{w}, \beta) = \mathcal{N}(\bar{y}|f_{\mathbf{w}}(\bar{x}), \beta^{-1})$$

likelihood for $\mathbf{w}$ and β: $p(\mathbf{y}|\mathbf{X}, \mathbf{w}, \beta) = \prod_{i=1}^{m} \mathcal{N}(y^{(i)}|f_{\mathbf{w}}(x^{(i)}), \beta^{-1})$

least-squares (LS)

$$J(\mathbf{w}) \sim -\ln p(\mathbf{y}|\mathbf{X}, \mathbf{w}, \beta) \sim \frac{1}{2}\beta \sum_{i=1}^{m} \left[f_{\mathbf{w}}(x^{(i)}) - y^{(i)}\right]^2 - \frac{1}{2} m \ln \beta + \frac{1}{2} m \ln(2\pi)$$



$$p(\mathbf{w}|\alpha) = \mathcal{N}(\mathbf{w}|\mathbf{0}, \alpha^{-1}\mathbf{I}) = \left(\frac{\alpha}{2\pi}\right)^{(n+1)/2} \exp\left(-\frac{\alpha}{2}\mathbf{w}^{\top}\mathbf{w}\right)$$

posterior: $p(\mathbf{w}|\mathbf{X}, \mathbf{y}, \alpha, \beta) \sim p(\mathbf{y}|\mathbf{X}, \mathbf{w}, \beta) p(\mathbf{w}|\alpha)$

MAP: $\frac{\beta}{2}\left[\sum_{i=1}^{m} \left[f_{\mathbf{w}}(x^{(i)}) - y^{(i)}\right]^2 + \frac{\alpha}{\beta}\mathbf{w}^{\top}\mathbf{w}\right]$

Prediction for an output

$$p(\bar{y}|\bar{x}, \mathbf{X}, \mathbf{y}, \alpha, \beta) = \int \overbrace{p(\bar{y}|\bar{x}, \mathbf{w}, \beta)}^{p(\bar{y}|\bar{x},\mathbf{w})} \underbrace{p(\mathbf{w}|\mathbf{X}, \mathbf{y}, \alpha, \beta)}_{p(\mathbf{w}|\mathbf{X},\mathbf{y})} d\mathbf{w}$$

$$\mathbf{S}^{-1} = \alpha \vec{\mathbf{I}} + \beta \sum_{i=1}^{m} \vec{\phi}(\mathbf{x}^{(i)}) \vec{\phi}^{\top}(\mathbf{x}^{(i)})$$

$$m(\bar{x}) = \beta \vec{\phi}^{\top}(\bar{x}) \mathbf{S} \sum_{i=1}^{m} \vec{\phi}(\mathbf{x}^{(i)}) y^{(i)}, \; s^2(\bar{x}) = \beta^{-1} + \vec{\phi}^{\top}(\bar{x}) \mathbf{S} \vec{\phi}(\bar{x})$$

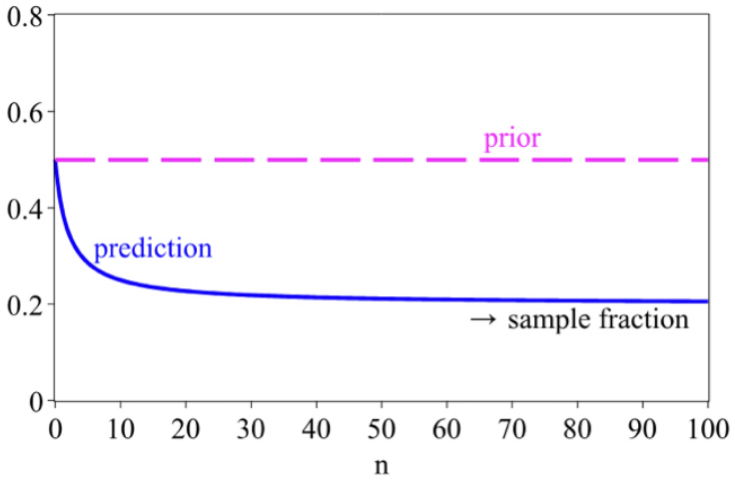
$$\phi_j(x) = x^j$$

Example (coin-drawing)

$$p(H|\{\text{data}\}) = \int_0^1 d\theta \overbrace{p(H|\theta)}^{\text{binomial: } \theta} p(\theta|\{\text{data}\}) = \int_0^1 d\theta \theta p(\theta|\{\text{data}\})$$

$$= \mathrm{E}[\theta|\{\text{data}\}] = \frac{n_H + 1}{n + 2}$$

θ: the success rate; n_H: the number of successful trials



Bayesian linear curve fitting

$$p(\mathbf{y}|\mathbf{X}, \mathbf{w}, \beta) = \prod_{i=1}^{m} \mathcal{N}(y^{(i)}|f_{\mathbf{w}}(\mathbf{x}^{(i)}), \beta^{-1}) = \prod_{i=1}^{m} \mathcal{N}(y^{(i)}|\overbrace{\mathbf{w}^\top \vec{\phi}(\mathbf{x}^{(i)})}^{\text{polynomial}}, \beta^{-1})$$

$$p(\mathbf{w}) \sim \mathcal{N}(\mathbf{m}_0, \mathbf{S}_0), \quad p(\mathbf{w}|\mathbf{y}) \sim \mathcal{N}(\mathbf{m}_m, \mathbf{S}_m)$$

$$\mathbf{m}_m = \mathbf{S}_m \left(\mathbf{S}_0^{-1}\mathbf{m}_0 + \beta\vec{\Phi}^\top \mathbf{y}\right), \; \mathbf{S}_m^{-1} = \mathbf{S}_0^{-1} + \beta\vec{\Phi}^\top\vec{\Phi}$$

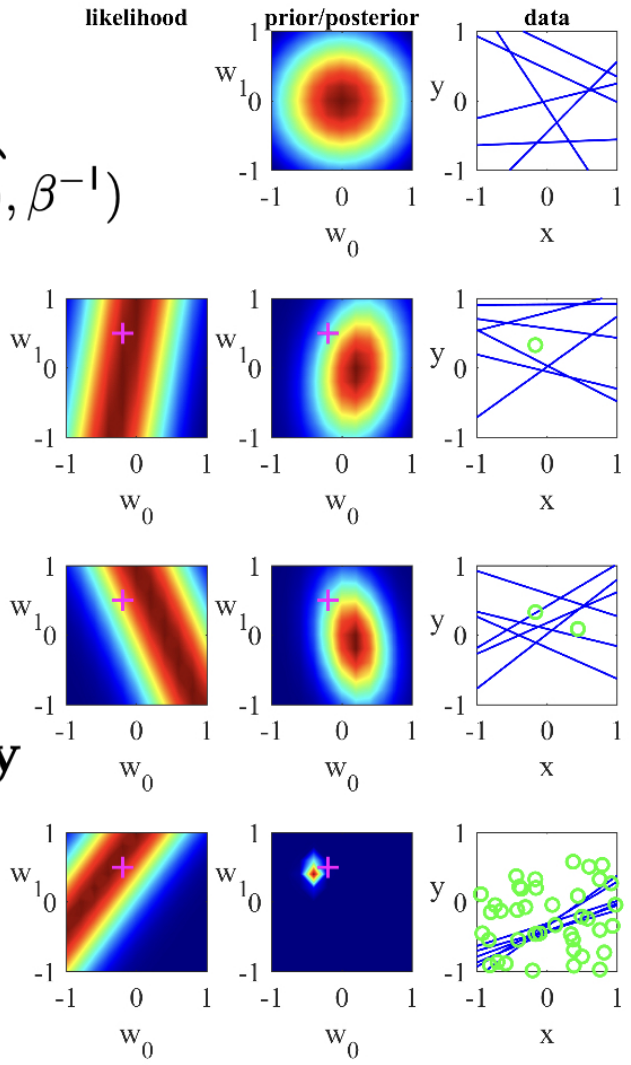
$$\mathbf{S}_0 = \alpha^{-1}\vec{\mathbf{I}}, \; \mathbf{S}_0^{-1} = \alpha\vec{\mathbf{I}}$$

$$\mathbf{m}_m = \mathbf{S}_m(\mathbf{S}_0^{-1}\mathbf{m}_0 + \beta\vec{\Phi}^\top\mathbf{y}) = \beta\mathbf{S}_m\vec{\Phi}^\top\mathbf{y} \to (\vec{\Phi}^\top\vec{\Phi})^{-1}\vec{\Phi}^\top\mathbf{y}$$

$f_{\text{phys}}(x) = -0.2 + 0.5x$

$x^{(i)} \sim \text{Unif}[-1, 1]$, $y^{(i)} \sim f_{\text{phys}}(x^{(i)}) + \mathcal{N}(0, 0.25)$

$f_{w_0,w_1}(x) = w_0 + w_1 x, \alpha = 2.0, \beta = 4.0$



*Variance reduction from data generation

$\vec{\phi}_{m+1} = \vec{\phi}(\mathbf{x}^{(m+1)})$

$$p(y^{(m+1)}|\mathbf{y},\alpha,\beta) = \int p(y^{(m+1)}|\mathbf{x}^{(m+1)},\mathbf{w},\beta)p(\mathbf{w}|\mathbf{y},\alpha,\beta)d\mathbf{w} = \mathcal{N}(y^{(m+1)}|\mathbf{m}_m^\top\vec{\phi}(\mathbf{x}^{(m+1)}),\sigma_m^2(\mathbf{x}^{(m+1)}))$$

$$\sigma_m^2(\mathbf{x}) = \overbrace{1/\beta}^{\text{noise of data sample}} + \underbrace{\vec{\phi}^\top(\mathbf{x})\mathbf{S}_m\vec{\phi}(\mathbf{x})}_{\text{uncertainty of learning parameter}}$$

$$\mathbf{S}_{m+1} = \mathbf{S}_m - \frac{\beta\mathbf{S}_m\vec{\phi}_{m+1}\vec{\phi}_{m+1}^\top\mathbf{S}_m}{1+\beta\vec{\phi}_{m+1}^\top\mathbf{S}_m\vec{\phi}_{m+1}}$$

$$\sigma_m^2(\mathbf{x}) - \sigma_{m+1}^2(\mathbf{x}) = \frac{\|\vec{\phi}^\top(\mathbf{x})\mathbf{S}_m\vec{\phi}(\mathbf{x})\|^2}{\beta^{-1}+\vec{\phi}_{m+1}^\top\mathbf{S}_m\vec{\phi}_{m+1}}$$

proof of $\sigma_{m+1}^2(\mathbf{x}) \le \sigma_m^2(\mathbf{x})$:

likelihood $\sim \mathcal{N}(y^{(m+1)}|f_w(x^{(m+1)}),\beta^{-1})$ of data "$m+1$"

exponential factor in posterior $\sim (\mathbf{w}-\mathbf{m}_m)^\top\mathbf{S}_m^{-1}(\mathbf{w}-\mathbf{m}_m) + \beta(y^{(m+1)} - \mathbf{w}^\top\vec{\phi}(\mathbf{x}^{(m+1)}))^2$

$\mathbf{S}_{m+1}^{-1} = \mathbf{S}_m^{-1} + \beta\vec{\phi}(\mathbf{x}^{(m+1)})\vec{\phi}^\top(\mathbf{x}^{(m+1)})$, $\mathbf{m}_{m+1} = \mathbf{S}_{m+1}(\mathbf{S}_m^{-1}\mathbf{m}_m + \beta\vec{\phi}(\mathbf{x}^{(m+1)})y^{(m+1)})$

Ex.: prove it!

Model selection: conceptual introduction

model posterior: $p(\mathcal{M}_\ell|\{\text{data}\}) \sim p(\mathcal{M}_\ell)p(\{\text{data}\}|\mathcal{M}_\ell)$

Bayes factor: $\dfrac{p(\{\text{data}\}|\mathcal{M}_\ell)}{p(\{\text{data}\}|\mathcal{M}_{\ell'})}$

predictive distribution: $p(y|\mathbf{x}, \{\text{data}\}) = \sum_{\ell=1}^{L} p(y|\mathbf{x}, \mathcal{M}_\ell, \{\text{data}\})p(\mathcal{M}_\ell|\{\text{data}\})$

marginal likelihood of model: $p(\{\text{data}\}|\mathcal{M}_\ell) = \int d\mathbf{w}\, \overbrace{p(\{\text{data}\}|\mathbf{w}, \mathcal{M}_\ell)}^{\text{likelihood for } \mathbf{w}} \overbrace{p(\mathbf{w}|\mathcal{M}_\ell)}^{\text{prior for } \mathbf{w}}$

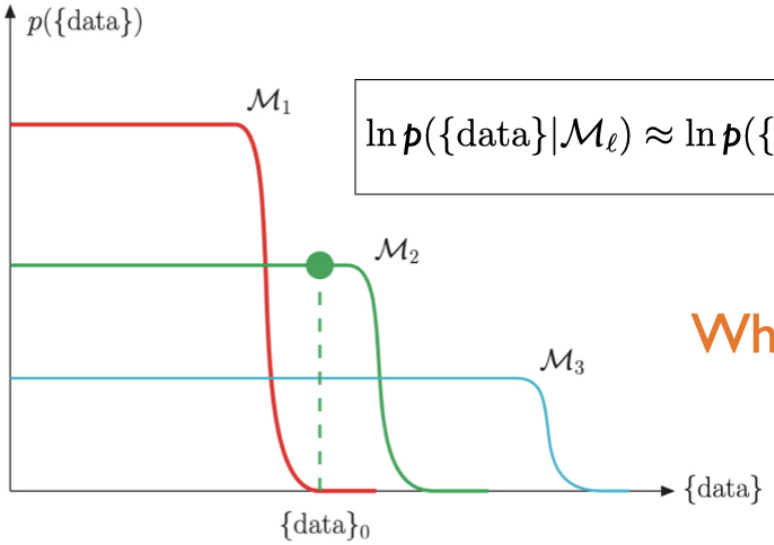
$$p(\mathbf{w}|\{\text{data}\}, \mathcal{M}_\ell) = \frac{p(\{\text{data}\}|\mathbf{w}, \mathcal{M}_\ell)p(\mathbf{w}|\mathcal{M}_\ell)}{p(\{\text{data}\}|\mathcal{M}_\ell)} = \frac{p(\{\text{data}\}|\mathbf{w}, \mathcal{M}_\ell)p(\mathbf{w}|\mathcal{M}_\ell)}{\int d\mathbf{w} p(\{\text{data}\}|\mathbf{w}, \mathcal{M}_\ell)p(\mathbf{w}|\mathcal{M}_\ell)}$$

Balance between model and data

$$p(\{\text{data}\}|\mathcal{M}_\ell) = \int dw p(\{\text{data}\}|w, \mathcal{M}_\ell) p(w|\mathcal{M}_\ell) \approx p(\{\text{data}\}|w_{\text{MAP}}, \mathcal{M}_\ell) \times \frac{\delta w_{\text{posterior}}}{\delta w_{\text{prior}}}$$

$$\rightarrow \ln p(\{\text{data}\}|\mathcal{M}_\ell) \approx \overbrace{\ln p(\{\text{data}\}|w_{\text{MAP}}, \mathcal{M}_\ell)}^{\text{information from data}} + \overbrace{\ln\left(\frac{\delta w_{\text{posterior}}}{\delta w_{\text{prior}}}\right)}^{\text{model complexity}}$$

$\delta w_{\text{posterior}} < \delta w_{\text{prior}}$: model can not be very complicated



$$\ln p(\{\text{data}\}|\mathcal{M}_\ell) \approx \ln p(\{\text{data}\}|\mathbf{w}_{\text{MAP}}, \mathcal{M}_\ell) + n \ln\left(\frac{\delta w_{\text{posterior}}}{\delta w_{\text{prior}}}\right)$$

Which is better?