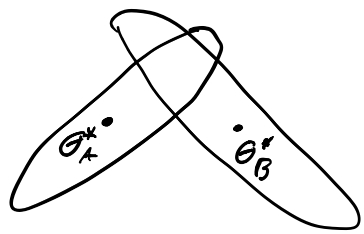
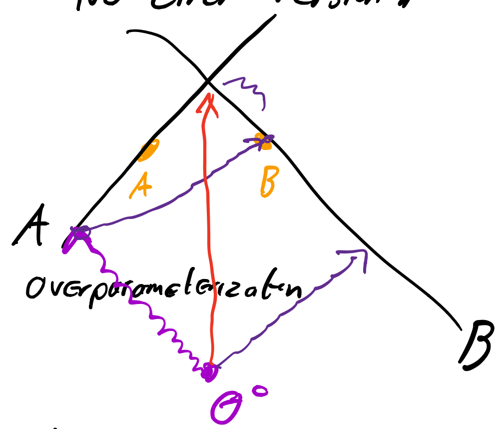


No error version:



$(x_i, y_i) \sim$ dist of data iid

$D = \{(x_i, y_i)\}_{i=1 \ldots n}$

model: $x \mapsto$ distribution on y params θ

$\max_\theta P(D|\theta)$ vs. $\max_\theta p(\theta|D)$

MLE

MAP

max a posteriori estimation

frequentist

prior $p(\theta)$

collect D

update prior $p(\theta|D)$

Bayesian

Bayes:

$$\underbrace{\log p(\theta|D)} = \underbrace{\log P(D|\theta)} + \underbrace{\log P(\theta) - \log P(D)}$$

If $p(\theta) \equiv$ const $\Rightarrow$ equivalent MLE $\Leftrightarrow$ MAP

uninformative prior

Does there exist a prior $p(\theta)$ uninformative over $\mathbb{R}^d$?

<u>Improper prior</u>

MLE training of a NN — MAP Estimation
w/ noninformative improper prior

Weight Decay

$$\min_{\theta} \underbrace{-\log P(D|\theta)}_{\ell_{CE}(D|\theta)} + \|\theta\|_2^2$$

Bayesian perspective: prior $\log P(\theta) = -\|\theta\|_2^2$

$$P(\theta) \sim N(0, I)$$

Dist $\quad P(x|\theta) \quad P_\theta(x)$

How sensitive is it to changes in θ?

$\nabla_\theta \log P(x|\theta) = 0$ at a soln to training

Instead, look at $D^2_\theta \log P(x|\theta)$

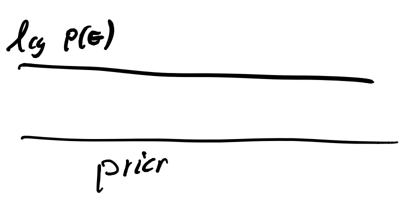
$$\mathbb{E}_{x\sim P_\theta} D^2_\theta \log P(x|\theta) = \underbrace{\mathbb{E}_{x} \nabla \log P(x|\theta)\, \nabla \log P(x|\theta)^t}_{\text{Fisher Information of } P_\theta}$$

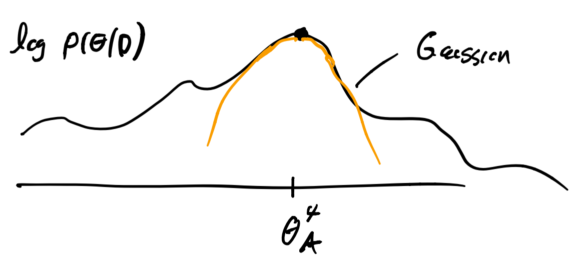
large diagonal entries have large information content

Fisher info is covariance^{-1} of $\nabla \log P$

$$\log P(\theta|D) = \log P(D_B|\theta) + \underbrace{\log P(\theta|D_A)}_{\text{approx by quadratic}} \cdots$$

Laplace approximation of a posterior





$$\log P(\theta | D_A) \approx \log P(\theta_A^* | D_A) + \frac{1}{2}(\theta - \theta_A^*)^t H (\theta - \theta_A^*)$$

$$H = \mathbb{E}\left[\nabla_\theta \log P \ \ \nabla_\theta \log P^t\right]$$

$$\mathbb{E}_x\left[\left(\nabla_\theta \log P(y_i | x_i, \theta)\right)^2\right] = \mathrm{diag}(F)$$

$$F = \mathbb{E}\left[\nabla_\theta \log P(\underset{(x_i, y_i)}{x} | \theta) \ \nabla_\theta \log P(x | \theta)^t\right]$$

$$(x_i, y_i) \qquad P((x_i, y_i) | \theta) = P(y_i | x_i, \theta) P(x_i)$$