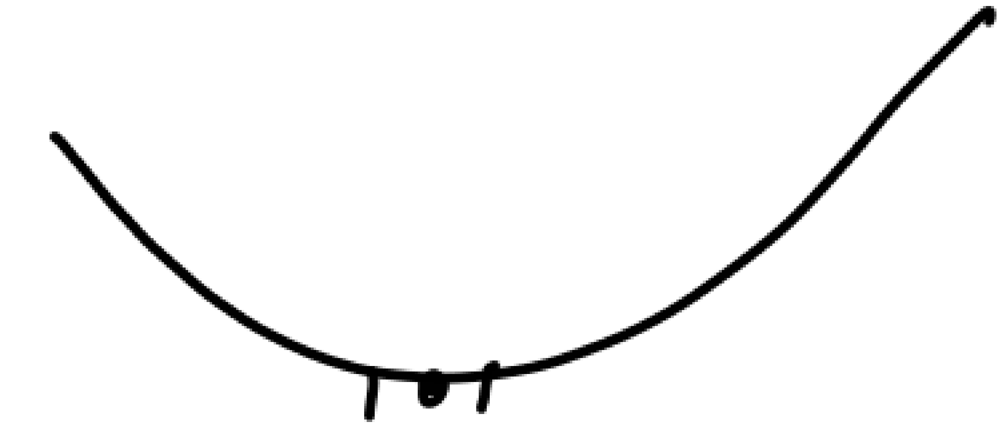
Gradient Descent



- Initialization $w \leftarrow w_0$
- Repeat until convergence $\rightarrow \|\nabla_w E\| < \epsilon$

$$w_{t+1} \leftarrow w_t - \underset{\text{learning rate}}{\eta} \nabla_w E \longrightarrow E(D, w) = \sum_{i=1}^{n} \underbrace{(w^T x_i - y_i)}^2$$

$$\nabla_w E = \sum_{i=1}^{n} \nabla_w E_i$$

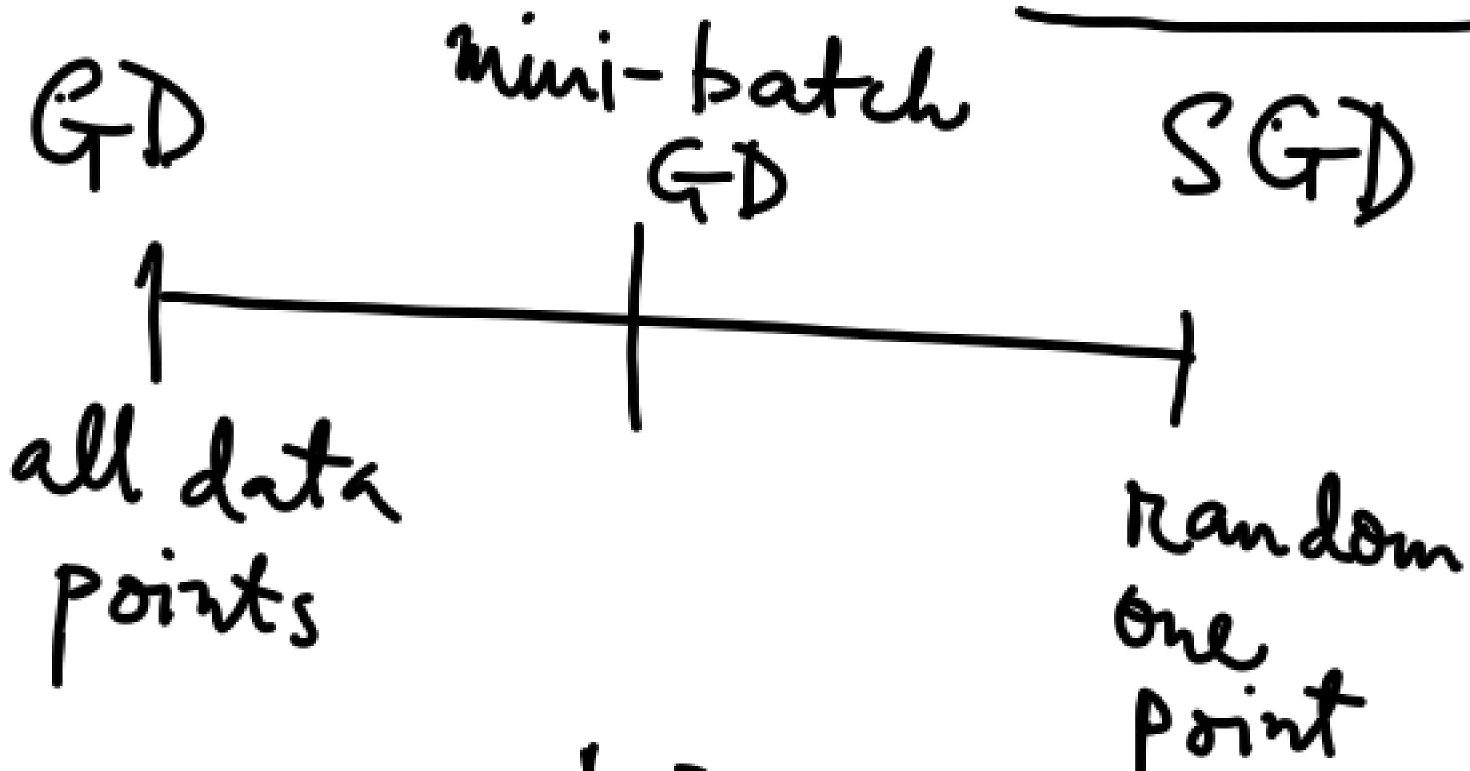
GD is excellent in accuracy
expensive in computation.

Stochastic GD

update step : $W_{t+1} \leftarrow W_t - \eta \nabla_W E(W, x_i, y_i)$

i is randomly chosen

Fast algorithm



$[n] := \{1, 2, \cdots, n\}$

$E_i = (W^T x_i - y_i)^2$

MB-GD

$E(W, (x_i, y_i)_{i \in B})$

update step: $W_{t+1} \leftarrow W_t - \eta \sum_{i \in B} \nabla_W E_i$

$B \subset \{1, \cdots, n\}$

MLE: Maximum likelihood estimate

$$D = \left\{ (x_i, y_i)_{i \in [n]} \right\} \qquad W = \theta$$

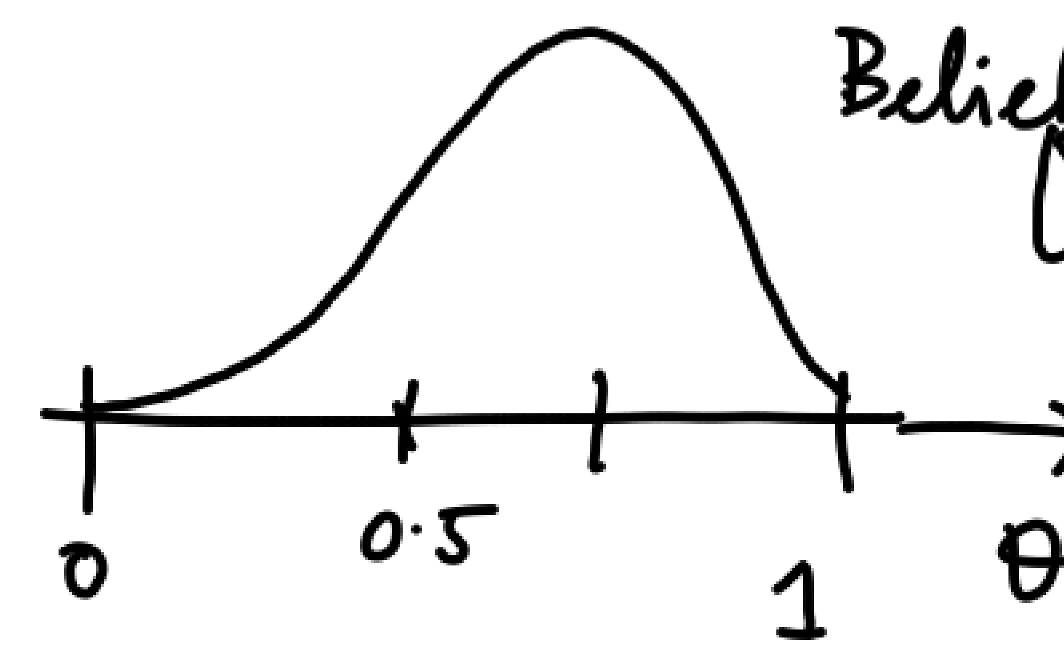
$$\underset{\theta}{\operatorname{argmax}} \; \underbrace{P(D \mid \theta)}_{\text{likelihood function}} = \theta_{MLE} \qquad y_i = w^T x_i + \epsilon_i$$

Coin-toss example: A coin is tossed n times, y_j is the j^{th} outcome. y_j is a Bernoulli RV $\begin{cases} = 1 & \text{w.p. } \theta \\ 0 & \text{w.p. } (1-\theta) \end{cases}$

$$P(y_j|\theta) = \theta^{y_j}(1-\theta)^{1-y_j}$$

Prior

Belief



$$\underbrace{P(y|\theta)}_{\text{Likelihood}} = \prod_{i=1}^{n} P(y_i|\theta)$$

$$LL(\theta) = \sum_{i=1}^{n} \log P(y_i|\theta) \Rightarrow \theta_{MLE} = \frac{1}{n}\sum_{j=1}^{n} y_j$$

$$P(D|\theta) \qquad \underline{P(\theta)}$$

Maximum Aposteriori Estimate (MAP)

$$P(\underset{\substack{\uparrow \\ \theta \\ \text{cause}}}{\text{parameters}} \mid \underset{\substack{\uparrow \\ D \\ \text{effect}}}{\text{observation}})$$

$$P(\theta \mid D) = \frac{\overbrace{P(D \mid \theta)}^{\text{likelihood}} \overbrace{P(\theta)}^{\text{prior}}}{P(D)}$$

Bayesian inference

$P(\theta \mid D)$ posterior belief

$$\theta_{MAP} \in \underset{\theta}{\operatorname{argmax}}\, P(\theta \mid D) = \underset{\theta}{\operatorname{argmax}} \left[\underline{P(D \mid \theta)}\, P(\theta) \right]$$

$$\log P(\theta|D) = \log P(D|\theta) + \log P(\theta)$$

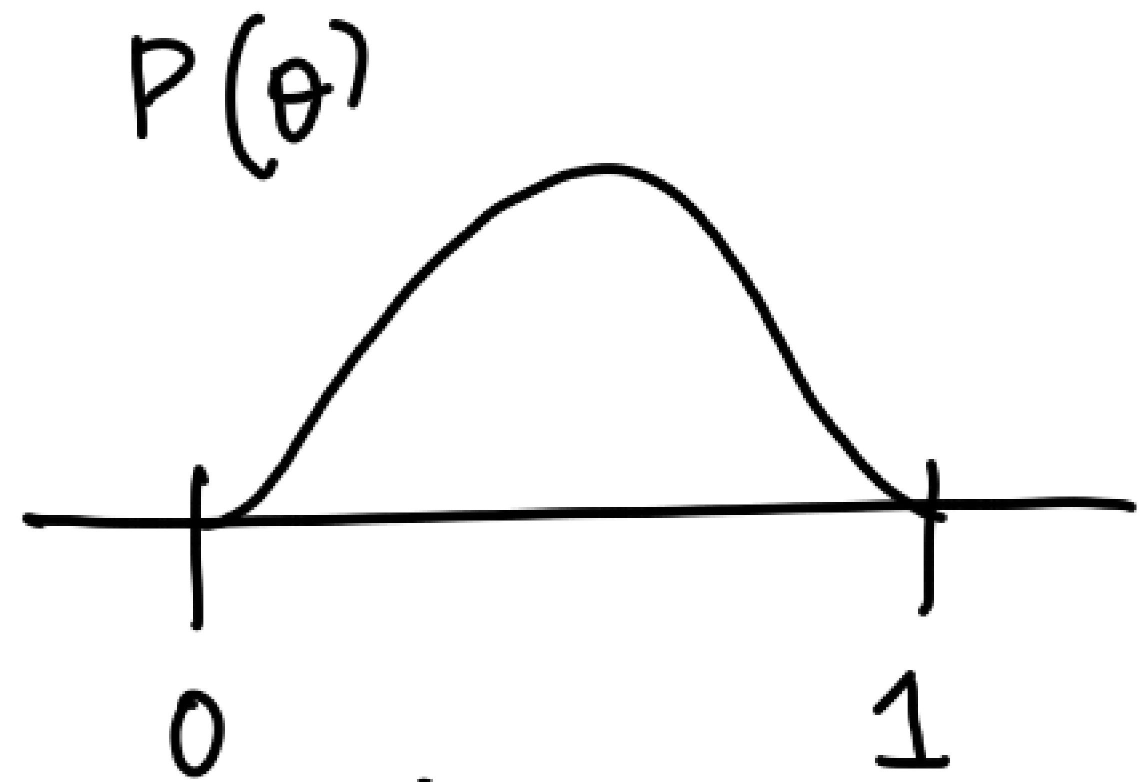
$$\theta_{MAP} \in \operatorname*{argmax}_{\theta} \left[\log P(D|\theta) + \log \underline{P(\theta)} \right]$$

$\theta_{MAP} = \theta_{MLE}$ if $P(\theta)$ is constant.

Ex. Likelihood of observing k heads in n tosses

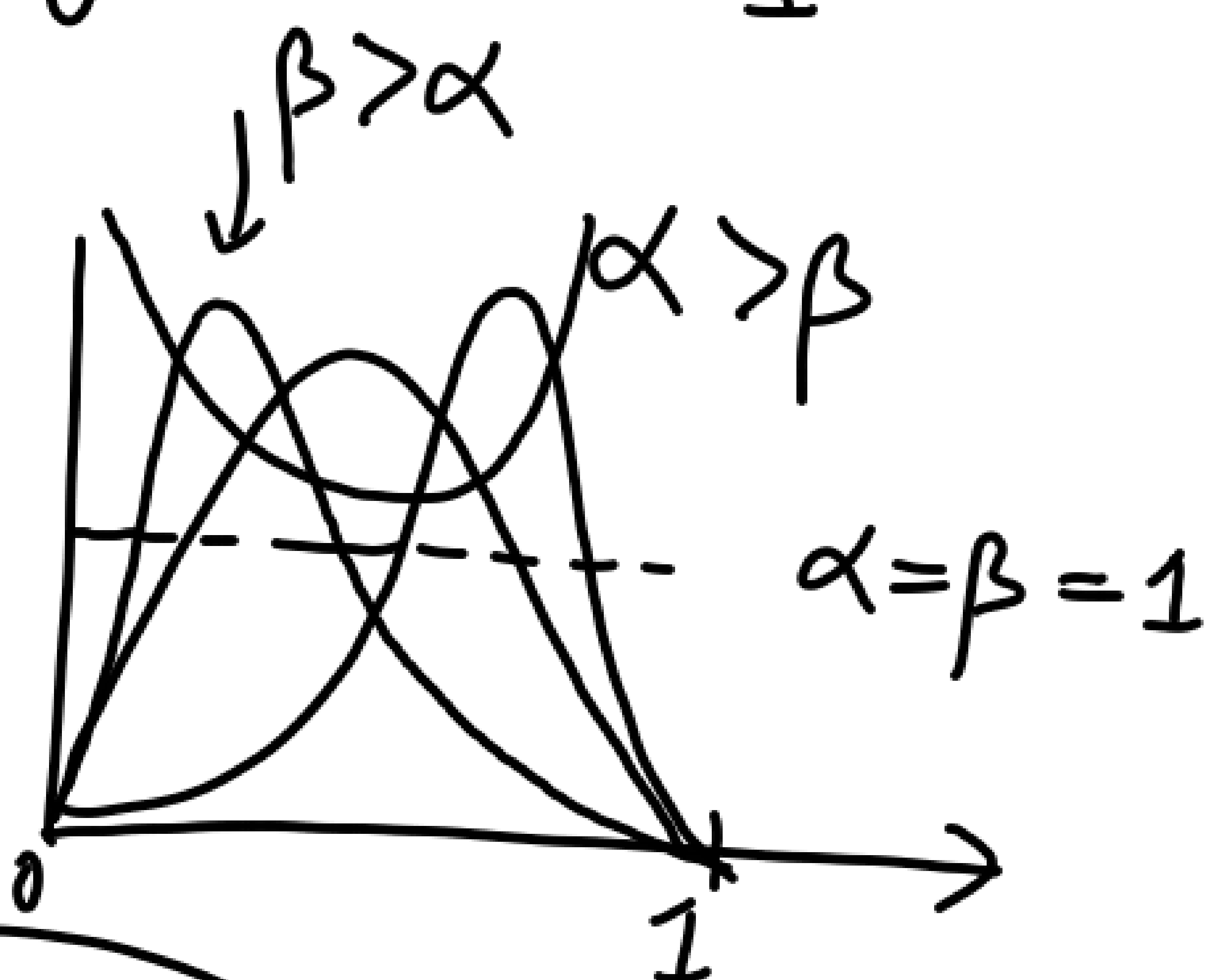
$$P(D|\theta) = \binom{n}{k} \theta^{k} (1-\theta)^{n-k} \qquad \mathrm{Bin}(n,k)$$

$$\theta_{MLE} = \frac{k}{n}$$



$$\text{Beta}(\alpha, \beta)$$

$$P(\theta) = \frac{1}{C}\, \theta^{\alpha-1} (1-\theta)^{\beta-1}$$



- Beta includes a large family of distributions in $[0,1]$
- Beta is conjugate prior of binom dist.

$P(D|\theta) \sim d_1 \quad P(\theta) \sim d_2$

$P(\theta)$ is a CP of $P(D|\theta)$ if $P(\theta|D) \sim d_2$

$$P(\theta|D) \propto P(D|\theta)\, P(\theta)$$

$$P(\theta|D) \propto \underbrace{\theta^{k}(1-\theta)^{n-k}}_{P(D|\theta)} \; \underbrace{\theta^{\alpha-1}(1-\theta)^{\beta-1}}_{P(\theta)}$$

$$\propto \theta^{\underbrace{k+\alpha-1}} (1-\theta)^{\underbrace{n-k+\beta-1}} \sim \mathrm{Beta}(k+\alpha,\ n-k+\beta)$$

$$\theta_{MAP} = \arg\max_{\theta} P(\theta|D) = \arg\max \underbrace{\log P(\theta|D)}$$

$$= \arg\max_{\theta} \left[\text{Const.} + (k+\alpha-1)\log\theta + (n-k+\beta-1)\log(1-\theta) \right]$$

$$\theta_{MAP} = \frac{k+\alpha-1}{n+\alpha+\beta-2}$$

Conjugate prior examples

1. Bernoulli / Binomial $\longleftrightarrow$ Beta
2. Geometric $\longleftrightarrow$ Beta
3. Categorical $\longleftrightarrow$ Dirichlet
4. . . .
5. Normal $\longleftrightarrow$ normal }

$$\underline{P(\theta|D)} \propto \underline{P(D|\theta)}\ \underline{P(\theta)}$$

Conjugate prior for (univariate) Gaussian with known variance

known

prior $\theta \to \mu$

likelihood $P(D|\theta) \sim N(\mu, \sigma^2)$, $P(\theta) \sim N(\mu_0, \sigma_0^2)$

$D = \{x_1, \ldots, x_n\}$

$$= \frac{1}{\sqrt{2\pi\sigma_0^2}} \exp\left\{-\frac{1}{2\sigma_0^2}(\mu-\mu_0)^2\right\}$$

$$= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2\right\}$$

$$P(\theta|D) \propto P(D|\theta)P(\theta) \propto \exp\left\{-\frac{1}{2\sigma^2}\sum(x_i-\mu)^2 - \frac{1}{2\sigma_0^2}(\mu-\mu_0)^2\right\}$$

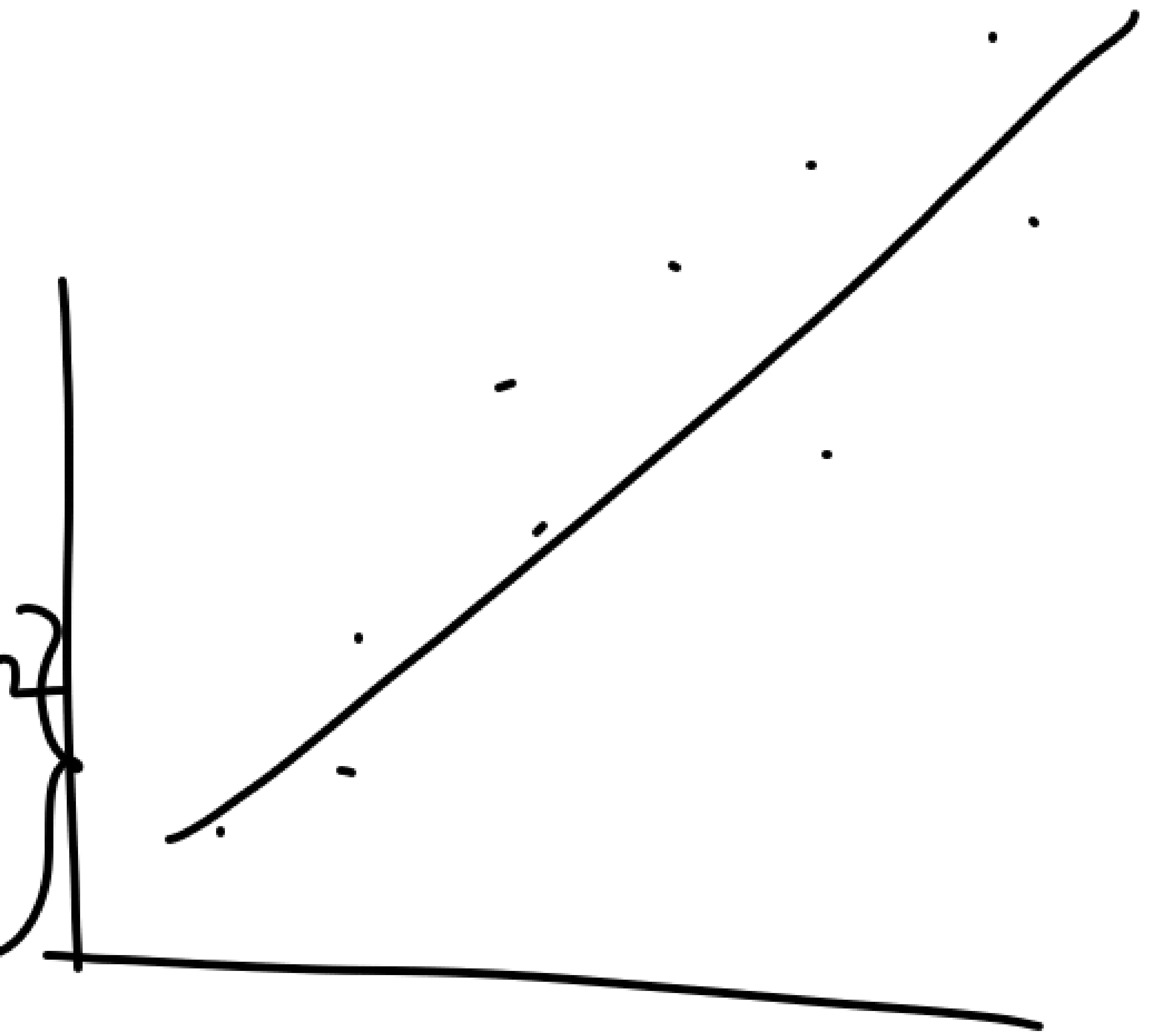
$$\propto \exp\left\{-\frac{1}{2\tilde{\sigma}^2}\sum(x_i-\tilde{\mu})^2\right\}$$

Find $\tilde{\sigma}^2$ and $\tilde{\mu}$

MAP estimate for linear regression

$$N(W^T x_i, \sigma^2) \leftarrow \underline{y_i} = W^T x_i + \underline{\epsilon_i} \sim N(0, \sigma^2)$$

$$P(D|\theta) \propto \exp\left\{-\frac{1}{2\sigma^2}\sum (y_i - W^T x_i)^2\right\}$$



$$\theta_{MLE} = \underset{W}{\arg\max} \sum_{i=1}^{n} (y_i - W^T x_i)^2$$

$$\sigma_{ij} = cov(x_i, x_j)$$

$$E[(x_i - Ex_i)(x_j - Ex_j)]$$

$d \times d$

$$P(w) \sim N\left(0, \frac{1}{\lambda} I\right) \qquad P(x) \sim N(\mu, \Sigma)$$

$$P(x_1, x_2, \dots, x_n) \qquad x \in \mathbb{R}^d$$

$\lambda > 0$

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

multivariate normal distribution.

$$P(w) = \frac{1}{(2\pi)^{d/2} \left(\frac{1}{\lambda}\right)^{d/2}} \exp\left\{ -\frac{\lambda}{2} w^T w \right\}$$

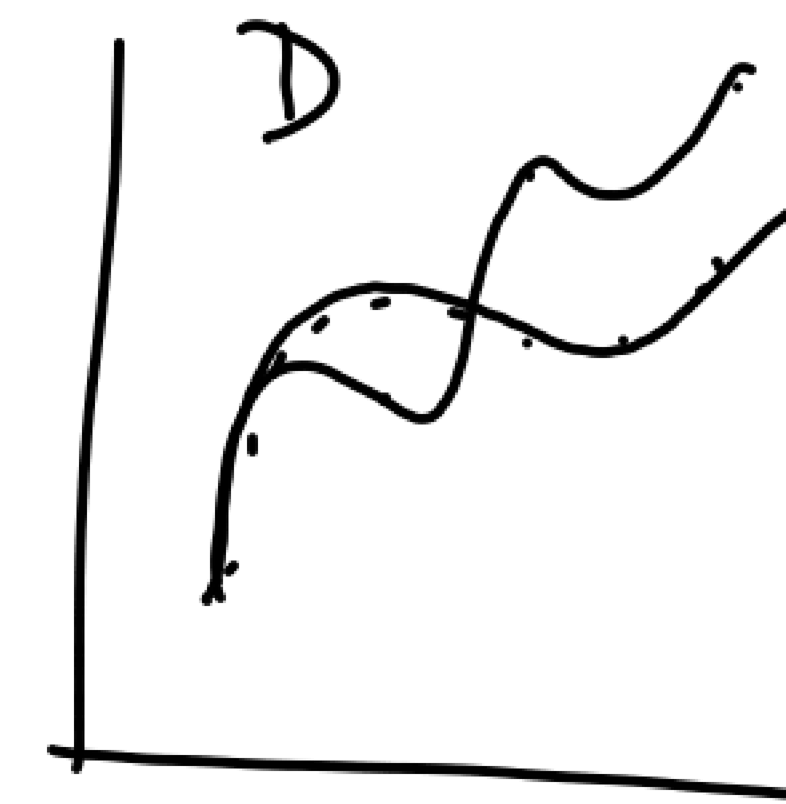
$$\sim N\left(0, \frac{1}{\lambda} I\right)$$

$$\propto \exp\left\{ -\frac{\lambda}{2} \|w\|^2 \right\}$$

$$P(w|D) \propto P(D|w)\,P(w)$$

λ = hyperparameter

$$\arg\max_{w} \left[\log P(D|w) + \log P(w) \right]$$



$$\arg\min_{w} \left\{ \underbrace{\frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - w^T x_i)^2}_{\hat{y}} + \frac{\lambda}{2} \|w\|^2 \right\}$$

$$\arg\min_{w} \left\{ \frac{1}{2\sigma^2} \|Xw - y\|^2 + \underbrace{\frac{\lambda}{2} \|w\|^2}_{\text{Regularizer}} \right\}$$

$w_0 + w_1 x + w_2 x^2 + \dots +$