MLE
(Maximum likelihood estimate).

n data points,

$D = \{x_1, x_2 \ldots x_n\}$

likelihood of these data points, given a model $\theta = \{\quad\}$

$$\log \hat{P}(D|\theta) = \hat{P}(\{x_1, x_2, \ldots x_n\}|\theta) = \prod_{k=1}^{n} \hat{P}(x_k|\theta)$$

the MLE of θ is: $\quad \theta_{MLE} = \arg\max_{\theta} \hat{P}(D|\theta)$

Coin flip
$H | T$

$$\theta = P(H) = q = \frac{\# \text{ of } H}{\text{total count}} = \frac{n_1}{n_1 + n_2}$$

$D = \{x_1, x_2 \ldots x_m\}$
$m_1 : H$
$m_2 : T$

$$P(D|\theta) = \prod_{i=1}^{n} P(x_i|\theta) = q^{n_1}(1-q)^{n_2}$$

To obtain $\hat{q} = \arg\max_{q} q^{n_1}(1-q)^{n_2} = F$

$$\frac{\partial F}{\partial q} = n_1 q^{n_1-1}(1-q)^{n_2} - q^{n_1}(1-q)^{n_2-1} \times n_2 = 0$$

$$\Rightarrow q^{n_1-1}(1-q)^{n_2-1}\left(n_1(1-q) - q n_2\right) = 0$$

$$\Rightarrow q = \frac{n_1}{n_1 + n_2}$$

$$\log \hat{P}(D|\theta) = \sum_{i=1}^{N} \log(x_i|\theta) \qquad D = \{x_1 \ldots x_m\}$$
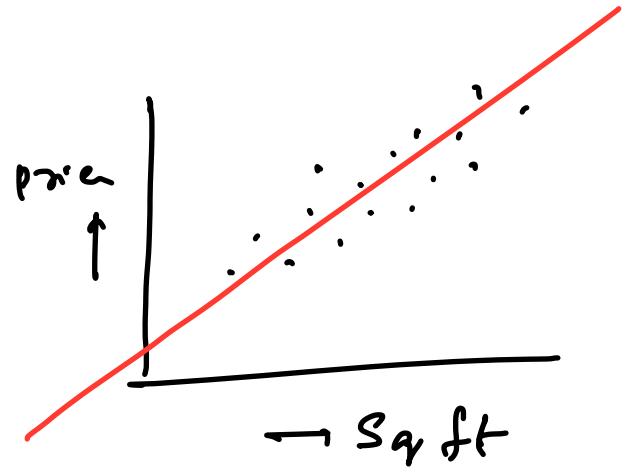
MAP : Maximum a posteriori Estimate

$$\log\left(\theta_{MAP}\right) = \arg\max_{\theta} P(\theta|D) = \arg\max_{\theta} \frac{P(D|\theta) \, P(\theta)}{}$$

Posterior prob $\nearrow$ likelihood $\rightarrow$ prior

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

$$= \log P(\theta) + \sum_{i=1}^{n} P(x_i|\theta)$$

Linear Regression

Living area (Sqft)	price ($)
2104	400
1600	330
3000	232
→ 2500	?

$|f(x) - y|$

price ↑ → Sqft

minimize the distance from all the points to the line L

$h_\theta(x) = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \cdots$

$= \theta_0 + \sum_{i=1}^{n} \theta_i x_i$

x_0

$= \sum_{i=0}^{n} \theta_i x_i$

$x_0 = 1$

$f(x_1, x_2, x_3) = x_1 + 3x_2 + 4x_3$

$(1, 2, 3)$

$f = \theta_1 x_1 + \theta_2 x_2 + \theta_3 x_3 =$

$d_i = (x_i, y_i)$

d^1
d^2
d^3
$\vdots$
d^n

Loss: $\dfrac{1}{2} \sum_{i=1}^{n} (f(x_i) - y_i)^2$

$\theta = \{ \theta_0, \theta_1, \cdots, \theta_n \}$

$J(\theta) = \sum_{i=1}^{m} (h_\theta(x^i) - y^i)^2$

$m = \#$ of training instances.

Superscript → instance

$j = $ Subscript → = parameter / feature

$\boxed{\dfrac{\partial J}{\partial \theta_j}} = \dfrac{\partial}{\partial \theta_j} \dfrac{1}{2} (h_\theta(x^i) - y^i)^2$

$= 2 \times \dfrac{1}{2} (h_\theta(x^i) - y^i) \dfrac{\partial}{\partial \theta_j} (h_\theta(x) - y)$

$= (h_\theta(x^i) - y^i) \times \dfrac{\partial}{\partial \theta_j} (\sum_{k=1}^{n} \theta_k x_k - y)$

$= (h_\theta(x^i) - y^i) x_j$

old

$\tilde{\theta}_j \leftarrow \theta_j - \textcircled{l} \dfrac{\partial J}{\partial \theta_j}$
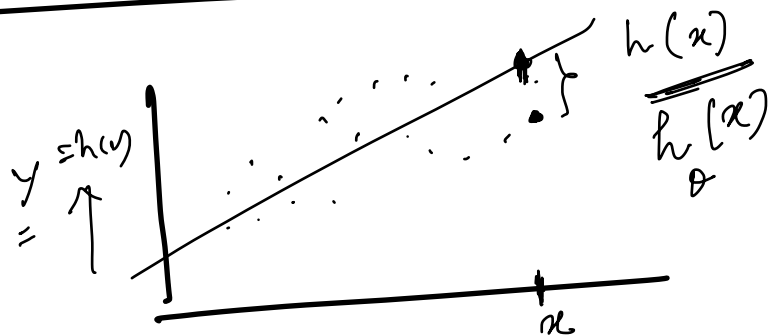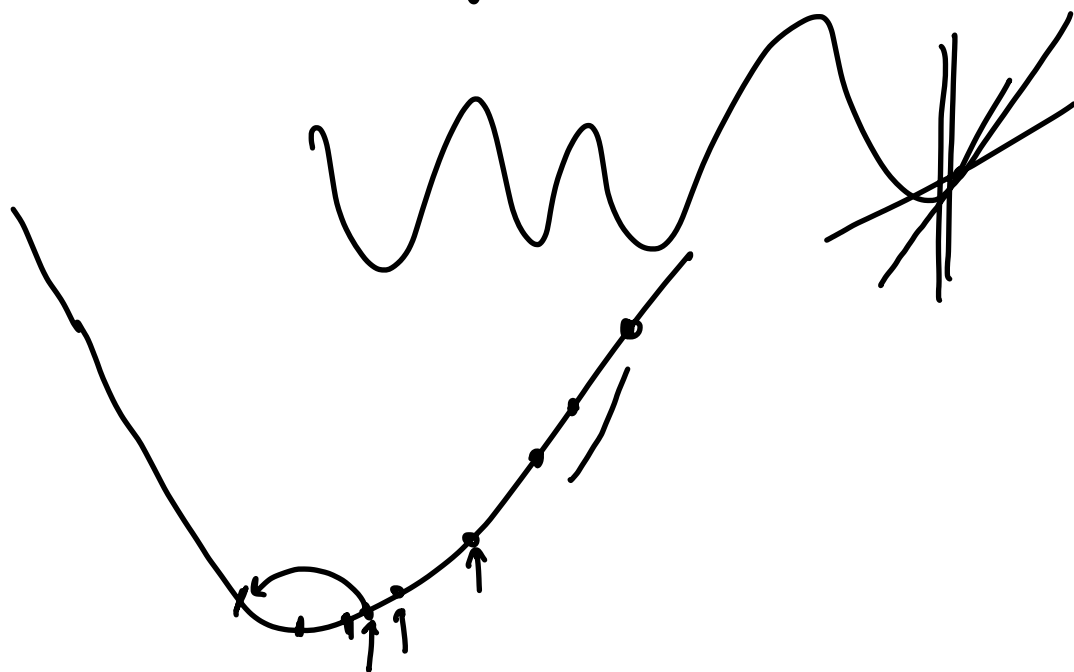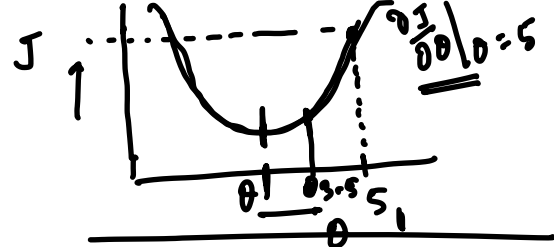
↑ new   ↑ learning rate

$\dfrac{\partial J}{\partial \theta_p} = (h_\theta(x^i) - y^i) x_p$

Batch GD | Stochastic GD | mini Batch GD

$$\hat{\theta}_j \leftarrow \theta_j - m \frac{\partial J}{\partial \theta_j}$$

$$\theta = 5$$

$$\frac{\partial J}{\partial \theta}\Big|_{\theta=5}$$

$$y = h(x)$$

$h(x)$

$\frac{h(x)}{\theta}$

$$y = w_0 + x w_1 x_1 + w_2 x_2 \cdots$$

θ

x

m = # of data points
n = # of features

$$\frac{1}{2} \sum_{i=1}^{m} \left(h_\theta(x^i) - y^i \right)^2$$

$$x = \{ \quad \text{----} \quad \}_n$$

The Normal Equation

$$f : \mathbb{R}^{m \times m} \rightarrow \mathbb{R} \qquad f(A) = \frac{3}{2} A_{11} + 5 A_{12}^2 + A_{21} A_{22}$$

$$A = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix} \qquad \nabla_A f(A) = \begin{bmatrix} \frac{\partial f}{\partial A_{11}} & \frac{\partial f}{\partial A_{12}} \\ \frac{\partial f}{\partial A_{21}} & \frac{\partial f}{\partial A_{22}} \end{bmatrix}$$

$$= \begin{bmatrix} 3/2 & 10 A_{12} \\ A_{22} & A_{21} \end{bmatrix}$$

$$tr(A) = tr\ A = \sum_{i=1}^{n} A_{ii}$$

① $tr(ABC) = tr(CAB) = tr(BCA)$

② $tr(A) = tr(A^T)$

③ $tr(A+B) = tr(A) + tr(B)$

④ $tr(aA) = a\ tr\ A$

⑤ $\nabla_A tr(AB) = B^T$ ⑥ $\nabla_{A^T} f(A) = (\nabla_A f(A))^T$

⑦ $\nabla_A tr\ ABA^T C = CAB + C^T A B^T$

$$\vec{X} = \begin{bmatrix} -x^{(1)T}- \\ -x^{(2)T}- \\ -x^{(3)T}- \\ \vdots \\ -x^{(m)T}- \end{bmatrix}_{m \times n} \qquad \vec{Y} = \begin{bmatrix} Y^1 \\ Y^2 \\ \vdots \\ y^m \end{bmatrix}_{m \times 1} \qquad \vec{\theta} = \begin{bmatrix} \ \\ \ \\ \ \end{bmatrix}_{n \times 1}$$

$$(\vec{X}\vec{\theta} - \vec{Y}) = \begin{bmatrix} h_\theta(x^1) - y^1 \\ h_\theta(x^2) - y^2 \\ \vdots \\ h_\theta(x^m) - y^m \end{bmatrix} =$$

$$J(\theta) = \frac{1}{2}\sum(h_\theta(x^i) - y^i)^2 = \frac{1}{2}(X\theta - Y)^T(X\theta - Y)$$

$$\nabla_\theta J(\theta) = \frac{1}{2}\nabla_\theta(X\theta - Y)^T(X\theta - Y)$$

$$= \frac{1}{2}\nabla_\theta(\theta^T x^T - Y^T)(X\theta - Y)$$

$$= \frac{1}{2}\nabla_\theta(\theta^T x^T X\theta - \theta^T x^T Y - Y^T X\theta + Y^T Y)$$

$$= \frac{1}{2}\left[\nabla_\theta tr(\theta^T x^T X\theta - \theta^T x^T Y - Y^T X\theta)\right]$$

$$= \frac{1}{2}\left[\nabla_\theta tr(\theta^T x^T X\theta) - tr(\theta^T x^T Y) - tr(Y^T X\theta)\right]$$

$$= \frac{1}{2}\left[\nabla_\theta tr(\theta^T x^T X\theta) - 2\ \overset{A^T}{\underset{\theta}{\nabla}} tr(Y^T X\theta)\right]^A$$

$A = Y^T X\theta$
$A^T = (Y^T X\theta)^T$
$= \theta^T x^T Y$

$$= \frac{1}{2}\left(\nabla_\theta tr(\theta^T x^T X\theta) - 2 x^T Y\right)$$

$\nabla_\theta tr(Y^T X\theta)$
$= \nabla_\theta tr(\theta Y^T X)$
$\qquad\qquad \underset{B}{}$
$= (Y^T X)^T$

$\nabla_A tr(AB)$
$= B^T$

$$\nabla_\theta tr(\underset{A}{\theta^T} \underset{B}{x^T X} \underset{A^T}{\theta})\ C = I$$

$\underset{\theta}{\theta} = A^T$

$$\nabla_\theta tr(\underset{A}{\theta}\underset{B}{\theta^T} \underset{A^T}{x^T x})\underset{C}{}$$

$\underset{I}{\vdots}$

$$\nabla_A tr\ \underset{A}{ABA^T C}$$
$$= CAB + C^T A B^T$$

$= x^T X\theta + x^T X\theta$
$= 2 x^T X\theta$

$$\downarrow \tfrac{1}{2}(2\, x^T x\, \theta - 2\, x^T Y) = 0$$

$$\boxed{x^T x\, \theta = x^T Y} \longrightarrow \text{Normal Eq}$$

$$\theta = (x^T x)^{-1} x^T Y \qquad \text{if } x^T x \text{ is invertible}$$

$$y = w_0 + w_1 x_1 + \cdots + w_n x_n$$

$$= w_0 + w_1 x_1^2 + w_2 x_2^2 + \cdots + w_m x_m^2 \leftarrow$$

linear in the parameters

$$Y = w_0 + w_1 \varphi(x_1) + w_2 \varphi(x_2) + \cdots + w_m \varphi(x_n)$$

$\varphi = $ non-linear basis funct.

$$\varphi_j(x) = x^j - \text{polynomial}$$

$$= \frac{(x-\mu)^2}{2a^2} - \text{Gaussian}$$

$$= \frac{1}{1 + \exp(-s_1 x_i)} - \text{Sigmoid}$$

$$J = \sum_{i=1}^{m} \left(y^i - \sum_{j=1}^{n} \theta_j \, \varphi(x^i) \right)^2 = \sum \left(y^i - \theta^T \underbrace{\varphi(x^i)}_{x_i} \right)^2$$

$$\frac{\partial J}{\partial \theta} = \sum \underline{(y^i - \theta^T x^i)^2} = 2 \sum (y^i - \theta^T x^i)\underline{x^{(i)}} = 0$$

$$\underline{\sum y^{(i)} x^{(i)T}} = \theta^T \sum x^{(i)T} x^{(i)}$$

$$\Rightarrow \quad x^T Y = (x^T x)\,\theta \qquad \quad \text{——} \quad \text{Normal}$$

$$\Rightarrow \boxed{\theta = (x^T x)^{-1}(x^T Y)}$$

$$x x^T \qquad (x^T x) = \text{non non-invertible}$$

$$x\Big|_{m \times n}$$

Regularizer

Ridge

L2

Ridge Regularization

$$\hat{\theta} = \arg\min \sum_{i=1}^{m} (y_i - x_i \theta)^2 + \lambda \|\theta\|_2^2$$

$$\hat{\theta} = (X^TX + \lambda I)^{-1} X^T Y$$

L1

Lasso Regularization

$$\hat{\theta} = \sum_{i=1}^{m} (y_i - x_i \theta)^2 + \lambda \|\theta\|_1$$

Prove that: $\nabla_A \, \text{tr}(ABA^TC) = C^TAB^T + CAB$

→ we use two theorems as follows:

$\text{tr}(ABCD) = \text{tr}(DABC) = \text{tr}(CDAB) = \text{tr}(BCDA)$ ——①

chain Rule of matrix differentiation:

$$d(AB) = d(A)B + A\,d(B) \quad\longleftarrow ⓘ$$

Therefore, $\nabla_A \, \text{tr}(ABA^TC) = \nabla_X \, \text{tr}(XBA^TC) + \nabla_X \, \text{tr}(ABX^TC)$

$\Rightarrow \nabla_X \, \text{tr}(XBA^TC) = (BA^TC)^T \quad (\because \nabla_A \text{tr}(AB) = B^T)$

$\qquad\qquad = C^TAB^T \quad\longrightarrow ⓘⓘ$

$\nabla_X \, \text{tr}(ABX^TC) = \nabla_X \text{tr}(X^TCAB) \quad (\text{due to } ①)$

$\qquad\qquad \Rightarrow \left((\nabla_X \, \text{tr}(X^TCAB)^T)^T \right)^T$

as $\nabla_{A^T} f(A) = (\nabla_A f(A))^T \Rightarrow \left(\nabla_{X^T} \, \text{tr}(X^TCAB)^T\right)^T$

as $\nabla_A \text{tr}(AB) = B^T \qquad \Rightarrow ((CAB)^T)^T$

$\qquad\qquad\qquad \Rightarrow CAB \quad\longrightarrow ⓘⓥ$

There from ⓘⓘ and ⓘⓥ

$$\nabla_A \, \text{tr}(ABA^TC) = C^TAB^T + CAB$$

Probabilistic Interpretation
Of Linear Reg.

$$\underline{y_\delta^i} \neq \underline{\theta^T x_\delta^i} = h_\theta(x_\delta^i)$$

$$y^i = \underline{\theta^T x^i} + \epsilon^i \underset{\text{error}}{\longleftarrow} \Rightarrow \text{unmodelled error}$$

$$E(\ell^i) = 0 \\ a\nu \qquad\qquad E(y^i) \approx E(\theta^T x^i)$$

$$P(\ell^i) = \frac{1}{\sqrt{2\pi}\, a} \exp\left(-\frac{(\ell^i)^2}{2a\nu} \right)$$

$$\underline{\underline{P(y^i | x^i ; \theta)}} = \frac{1}{\sqrt{2\pi}\, a} \exp\left(-\frac{(y^i - \theta^T x^i)^2}{2a\nu} \right) \!\!-\!\!① =$$
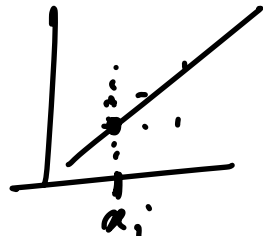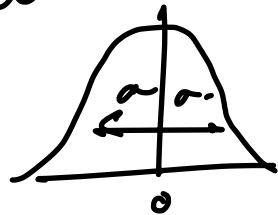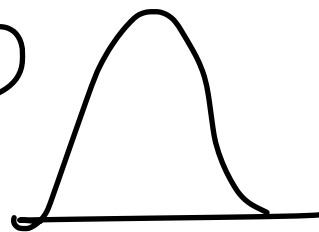
$$\underset{\text{lll}}{y^i | x^i ; \theta} \sim \mathcal{N}\left(\theta^T x^i, a\nu \right) \qquad \epsilon \sim iid$$

$$\underline{\underline{L(\theta)}} = L(\theta; \vec{X}, \vec{Y}) = P(\vec{y} | \vec{X} ; \theta) \longleftarrow$$

design matrix → ↑ vector of labels

$$= \prod_{i=1}^{m} P(y^i | x^i ; \theta) \overset{\text{from ①}}{=} \prod_{i=1}^{m} \frac{1}{\sqrt{2\pi}\, a} \exp\left(-\frac{(y^i - \theta^T x^i)^2}{2a\nu} \right)$$

maximize $L(\theta) \equiv \max \log L(\theta) = \ell(\theta)$

$$\ell(\theta) = \log L(\theta)$$

$$= \sum_{i=1}^{m} \log\left(\frac{1}{\sqrt{2\pi}\, a} \exp\left(\frac{-(y^i - \theta^T x^i)^2}{2a\nu} \right) \right)$$
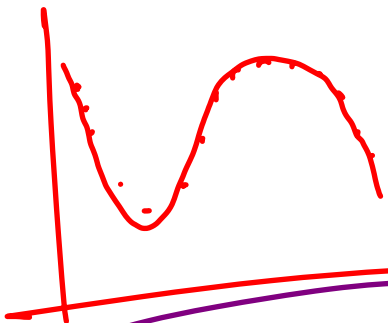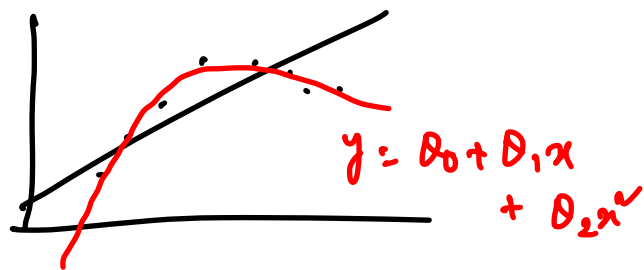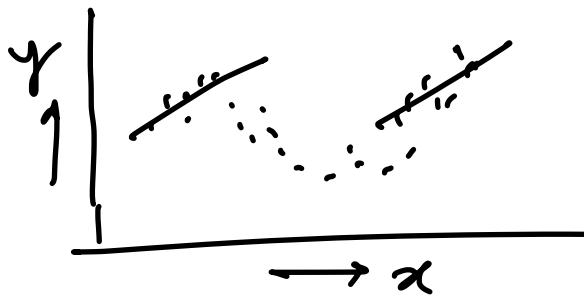
$$= \boxed{m \log \frac{1}{\sqrt{2\pi}\, a}} - \frac{1}{a\nu} \sum_{i=1}^{m} \frac{(y^i - \theta^T x^i)^2}{2}$$
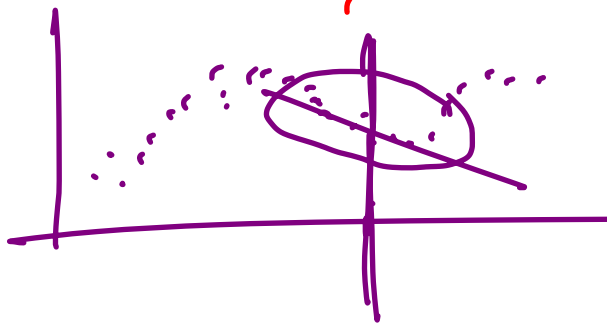
independent of θ

Locally weighted LR

$$y = \theta_0 + \theta_1 x$$
$$= \theta_0 + \theta_1 x + \theta_2 x^2$$
$$= \theta_0 + \theta_1 x + \theta_2 x^{1/2}$$

$$y = \sum_{i=1}^{5} \theta_i x^i$$

Non-parametric model

Fit θ by minimizing

$$\sum_{i=1}^{m} \underline{w^i} \left(y^i - \theta^T x^i \right)^2 \qquad w^i = \exp\left(-\frac{(x^i - x)^2}{2\tau^2} \right) \leftarrow$$

τ = bandwidth

τ =
= broad $\longrightarrow$ over smoothing / undefit
= narrow $\longrightarrow$ jagged fit / over fit