Exponential family

pdf : $p(y;\eta) = b(y) \exp\{\eta^T T(y) - a(\eta)\}$

$b(y)$ — Base measure

$y = T(y)$ — sufficient statistic

$a(\eta)$ — log-partition function

η — natural parameter

Exponential tilting.

$p(y;\eta) \propto b(y) e^{\eta^T y}$

$$p(y;\eta) = \frac{b(y) e^{\eta^T y}}{\int b(y) . e^{-\eta^T y} . dy} = \frac{b(y) . e^{\eta^T y}}{E[e^{-\eta y}] = A(\eta)}$$

mgf

$$= b(y) e^{\eta^T y - \underbrace{\log A(\eta)}_{a(\eta)}}$$

Bernoulli

$$p(y;\phi) = \phi^y (1-\phi)^{1-y}$$

$$= \exp[\log[\phi^y (1-\phi)^{1-y}]]$$

$$= \exp\{y \log \phi + (1-y) \log(1-\phi)\}$$

$$= \exp\{\log\left(\frac{\phi}{1-\phi}\right) y + \log(1-\phi)\}$$

$$p(y;\eta) = b(y) \exp\{\eta^T T(y) - a(\eta)\}$$

$$\eta = \log\left(\frac{\phi}{1-\phi}\right) \qquad b(y) = 1$$

$$\phi = \frac{1}{1+e^{-\eta}} \rightarrow \text{logistic}$$

$T(y) = y$

$a(\eta) = -\log(1-\phi) = \log(1+e^{\eta})$

⊛ Thus Bernoulli distribution belongs to exponential family.

Gaussian

$$p(y;\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2}\frac{(y-\mu)^2}{\sigma^2}\right\}$$

$\sigma^2 = 1$ (For ease)

$$p(y;\mu) = \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{1}{2}(y-\mu)^2\right\}$$

$$= \left[\frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}y^2\right)\right] \cdot \exp\left\{\mu y - \frac{1}{2}\mu^2\right\}$$

$$p(y;\eta) = b(y) \cdot \exp\left\{\eta^T T(y) - a(\eta)\right\}$$

$\eta = \mu$

$\Rightarrow \mu = \eta$

$T(y) = y$

$a(\eta) = \frac{\mu^2}{2} = \frac{\eta^2}{2}$

$$b(y) = \frac{1}{\sqrt{2\pi}} \exp\left(\frac{-y^2}{2}\right)$$

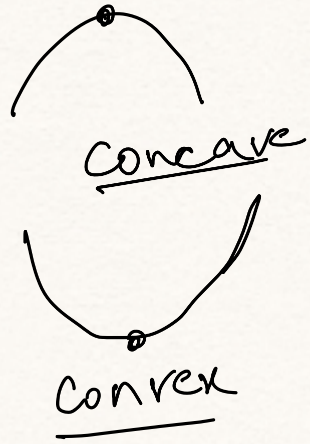
$\Rightarrow$ Thus belongs to exponential family as well!

Properties of Exponential Family

(1) $\log p(y;\eta)$ is concave in η

MLE is concave in η

$\Longleftrightarrow$ NLL is convex in η



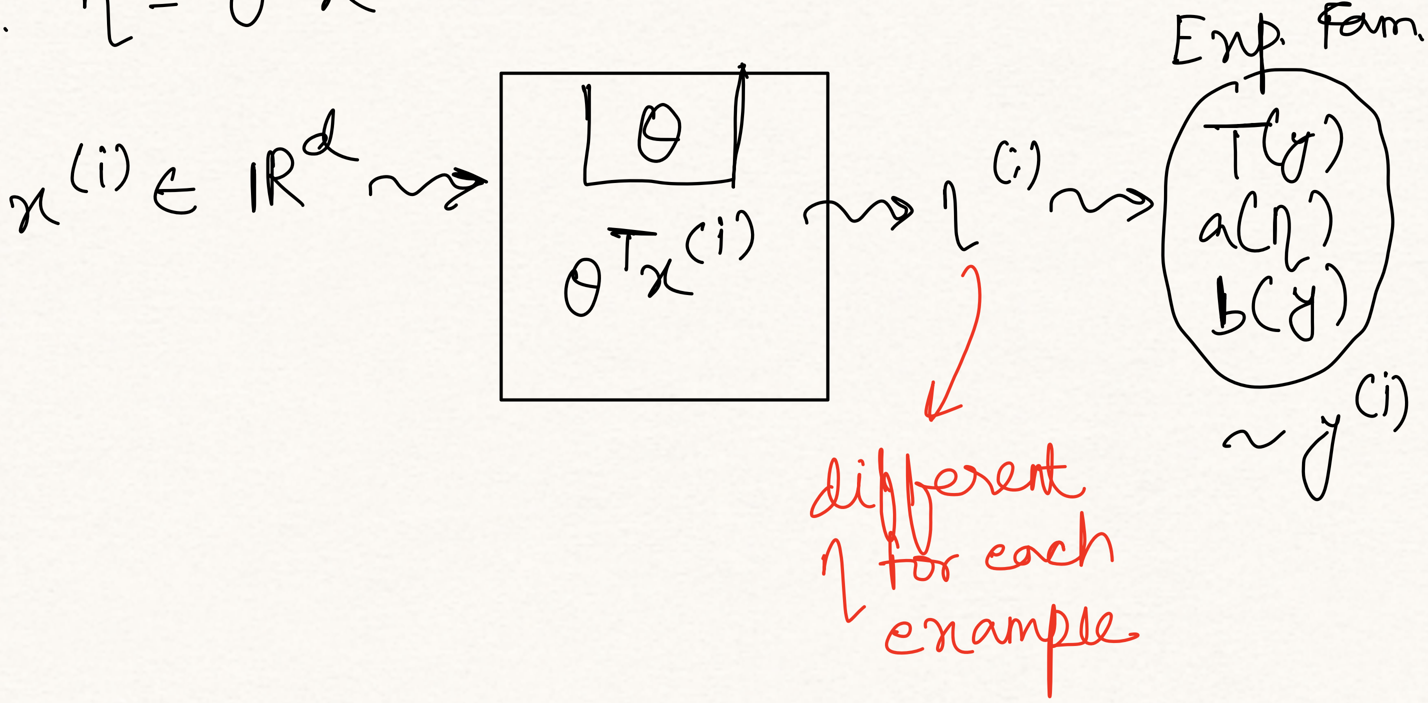
(2) $E[y;\eta] = \dfrac{\partial}{\partial\eta} a(\eta)$

(3) $\text{var}[y;\eta] = \dfrac{\partial^2}{\partial\eta^2} a(\eta)$

GLM (Generalised Linear Models)

$y \sim \text{ExpFam}(\eta)$

1. $y \mid x;\theta \sim \text{Exp Family}(\eta)$
2. $h_\theta(x) = E[y \mid x;\theta]$
3. $\eta = \theta^T x$



OLS : Ordinary Least Squares (Linear Reg.)

Exp Family is Gaussian.

$$h_\theta(x) = E[y|x;\theta]$$
$$= \mu$$
$$= \eta = \theta^T x$$

$$h_\theta(x) = \theta^T x$$

Logistic Regression

Exp. Family is Bernoulli.

$$h_\theta(x) = E[y|x;\theta]$$
$$= \phi = \frac{1}{1+e^{-\eta}} = \frac{1}{1+e^{-\theta^T x}}$$
$$= g(\theta^T x)$$

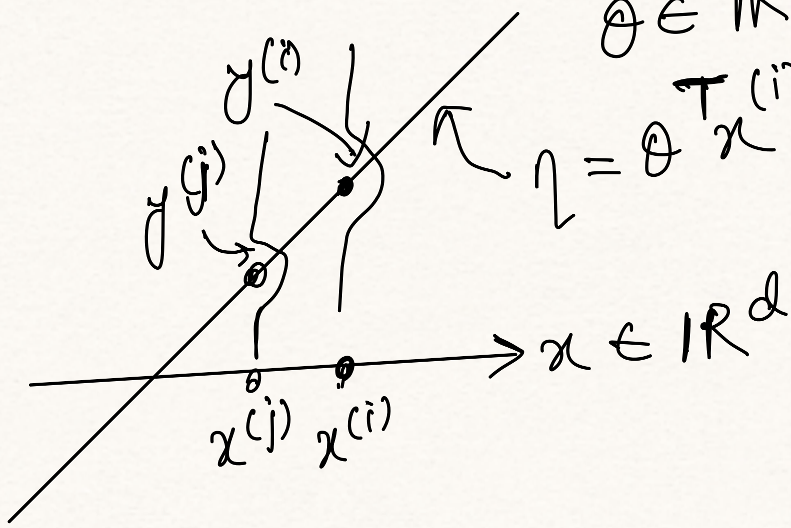
$$g(z) = \frac{1}{1+e^{-z}}$$

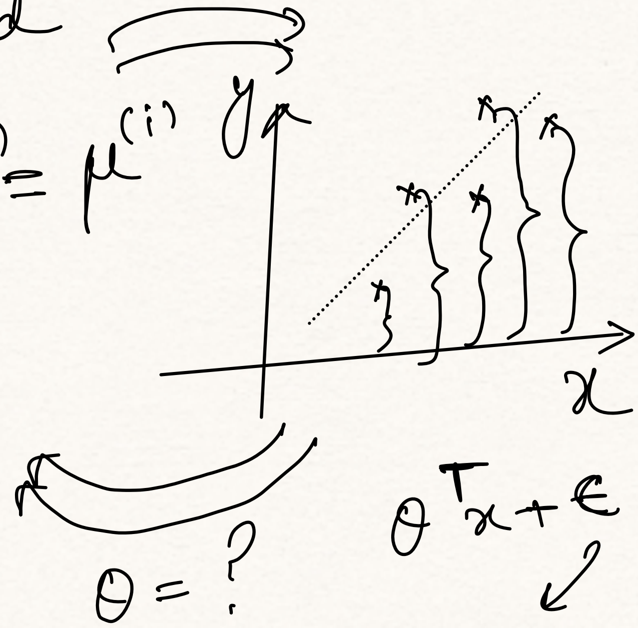
Linear Regression

Data Generation

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

$\eta = \theta^T x^{(i)} = \mu^{(i)}$



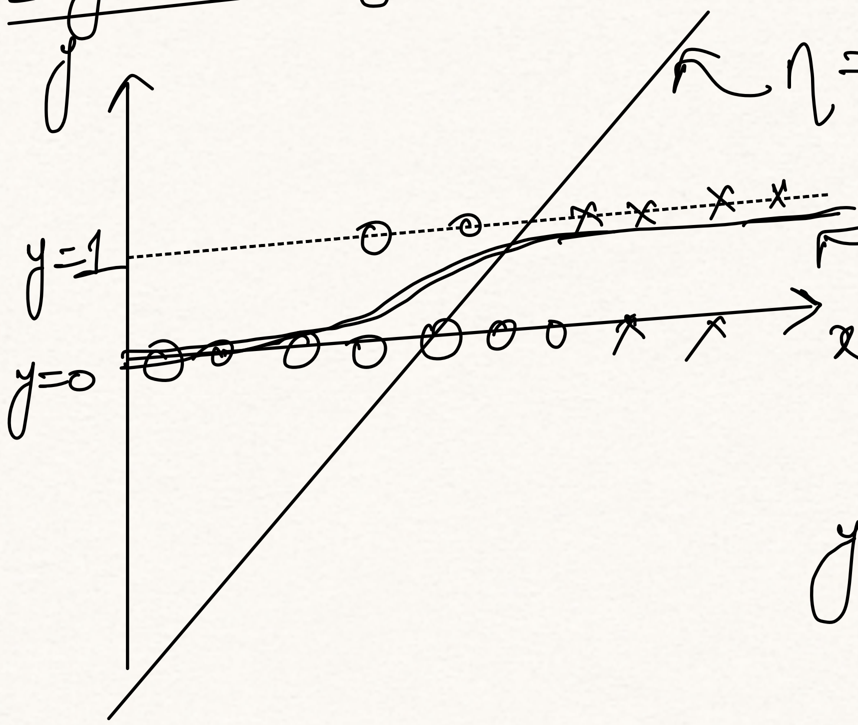


$\theta^T x + \epsilon$

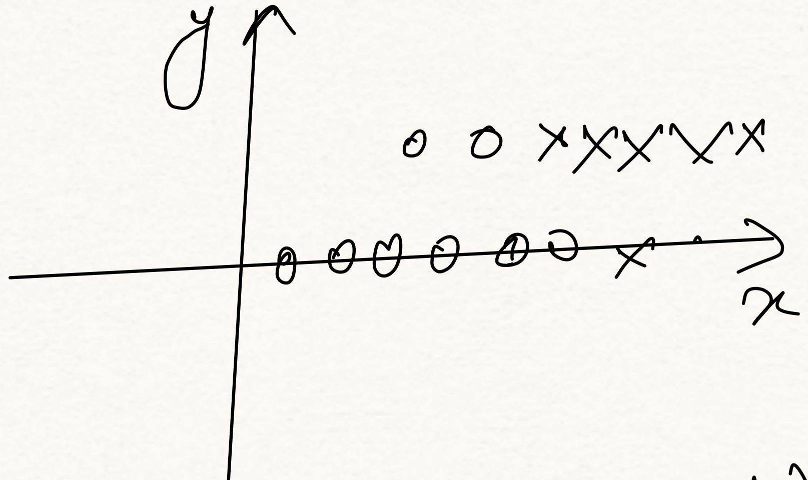
Gaussian noise $(\sim N(\theta^T x, \sigma^2))$

Logistic Regression

$\eta = \theta^T x$



$$\phi = g(\eta) = \frac{1}{1+e^{-\theta^T x}} = h_\theta(x)$$



(Mean of Y)

Model Parameters | Natural Parameter | Mean Parameter

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

$x^{(i)} \in \mathbb{R}^d$

$\theta^T x^{(i)} \rightarrow \eta^{(i)} \xrightarrow{g}$

- μ: Gaussian
- ϕ: Bernoulli
- λ: Poisson

Gaussian : $\mu = g(\eta) = \eta$

Bernoulli : $\phi = g(\eta) = \dfrac{1}{1+e^{-\eta}}$

g — canonical Response func^n.

y — response variable

g^{-1} — canonical link func^n.

$$h_\theta(x) = E[y \mid x ; \theta] = g(\theta^T x)$$

GLM ($\theta^T x$)

Data Type	Exp. Family Distrib.	Name
$\mathbb{R}$	Gaussian Laplace ⋮	Regression
$\{0,1\}$	Bernoulli	classification
$\{1,\ldots,K\}$	categorical	Multiclass* Classification
$\mathbb{N}_+$	Poisson	Count Reg. Poisson Reg.
$\mathbb{R}_+$ (time)	Exponential, Gamma	Survival analysis
Bernoulli Dist^n	Beta	---

* Also called Softmax Reg.

(i) Make a choice of dist. acc. to data type.

(ii) Express in exp. form — $a(\eta), b(y), T(y)$

$$\mu, \phi = g(\eta)$$

(iii) Hypothesis $h_\theta(x) = g(\theta^T x)$

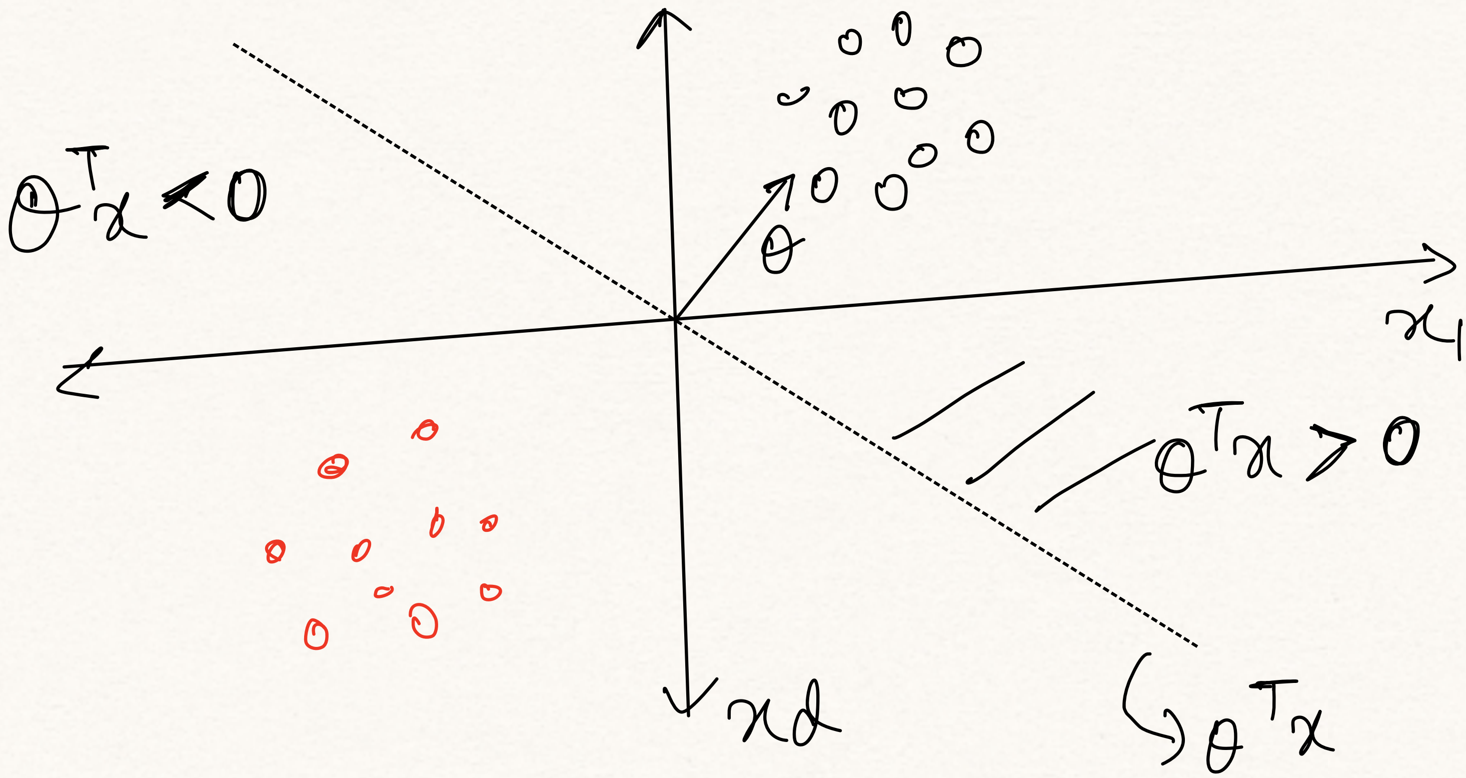
(iv) $\boxed{\theta := \theta + \alpha\left(y^{(i)} - h_\theta(x^{(i)})\right) \cdot x^{(i)}}$

This update step is true for all GLM.

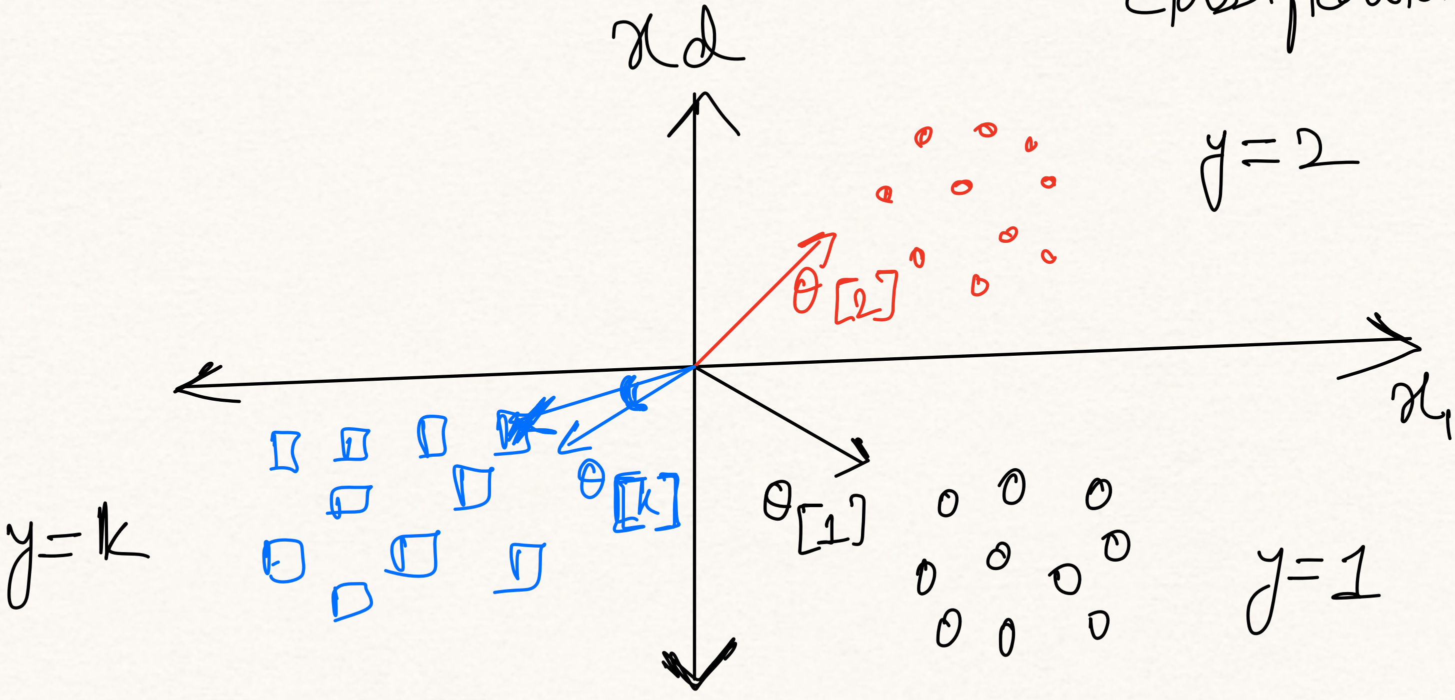
(v) Prediction : $\hat{y} = h_\theta(x^*) = g(\theta^T x^*)$

Softmax Regression

↳ Multiclass classification



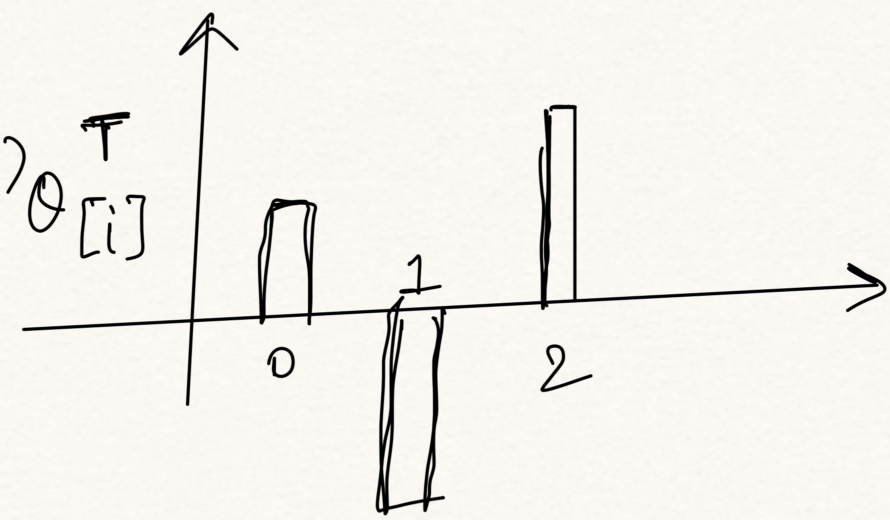
In case of 2 class classification.



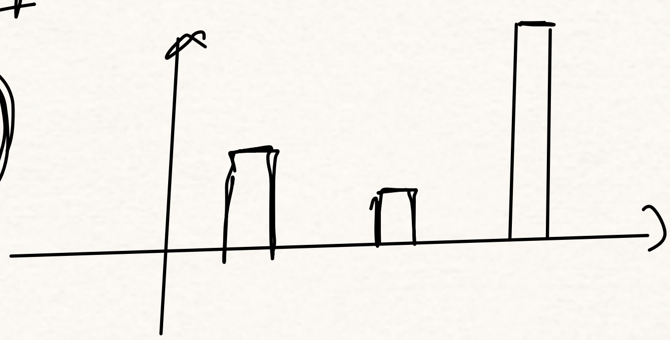
$$\left.\begin{matrix} \theta_{[1]}^T x - \mathbb{R} \\ \theta_{[2]}^T x - \mathbb{R} \\ \vdots \\ \theta_{[k]}^T x - \mathbb{R} \end{matrix}\right\} \arg\max$$

$$\theta = k\left\{ \overbrace{\begin{bmatrix} - & \theta_{[1]}^T & - \\ & \vdots & \\ - & \theta_{[k]}^T & - \end{bmatrix}}^{d} \right. \qquad x = \left.\begin{bmatrix} x \end{bmatrix}\right\} d$$

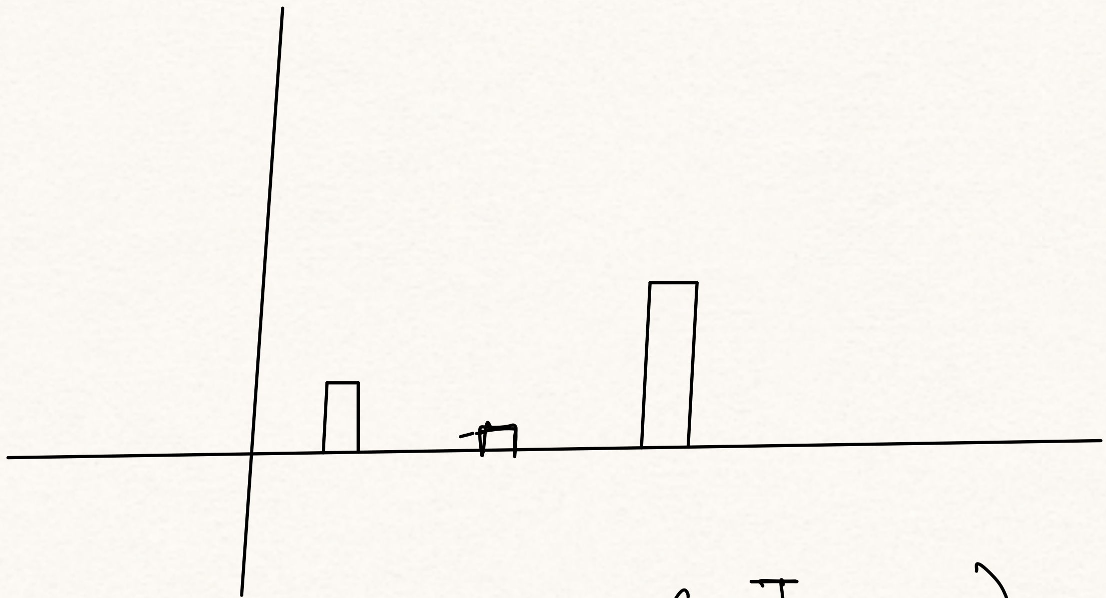
$x^{(i)}\theta_{[i]}^T$



$\exp\left(x^{(i)T}\theta_{[i]}\right)$



normalize



$$p(y=i \mid x; \theta) = \frac{\exp(x^T \theta_{[i]})}{\sum_j \exp(x^T \theta_{[j]})}$$