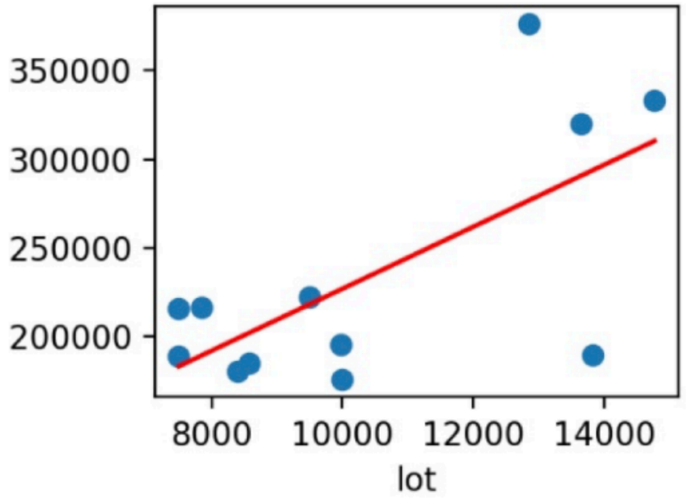
香港科技大學
THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY

COMP 5212
Machine Learning
Lecture 3

Logistic Regression, Exponential Family

Junxian He
Feb 12, 2026

Review: Another Solution — Gradient Descent



$$h_\theta(x) = \sum_{j=0}^{d} \theta_j x_j = x^T \theta$$

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

Choose

$$\theta = \underset{\theta}{\operatorname{argmin}}\, J(\theta).$$

Gradient Descent

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

Gradient Descent

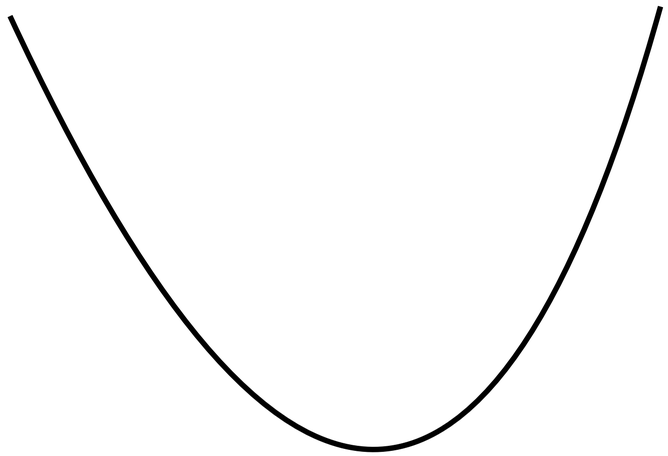
$$J(\theta) = \frac{1}{2}\sum_{i=1}^{n}\left(h_\theta(x^{(i)}) - y^{(i)}\right)^2$$

$$\theta_j := \theta_j - \alpha\frac{\partial}{\partial\theta_j}J(\theta)$$

Gradient Descent

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

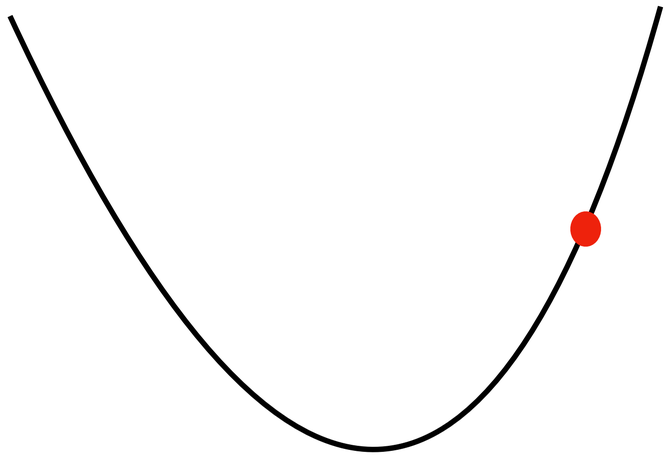
$$\theta_j := \theta_j - \alpha\frac{\partial}{\partial\theta_j}J(\theta)$$



Gradient Descent

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

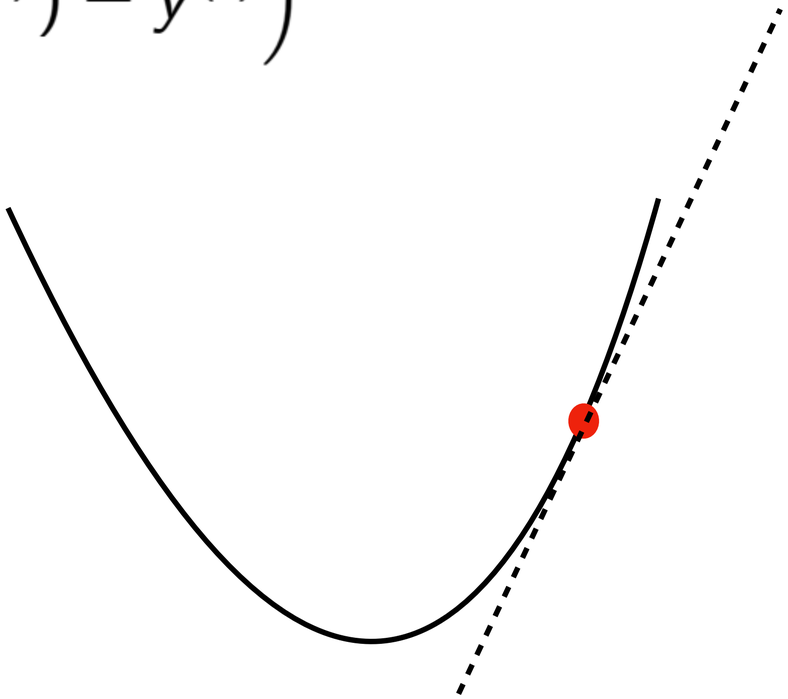
$$\theta_j := \theta_j - \alpha\frac{\partial}{\partial\theta_j}J(\theta)$$



Gradient Descent

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

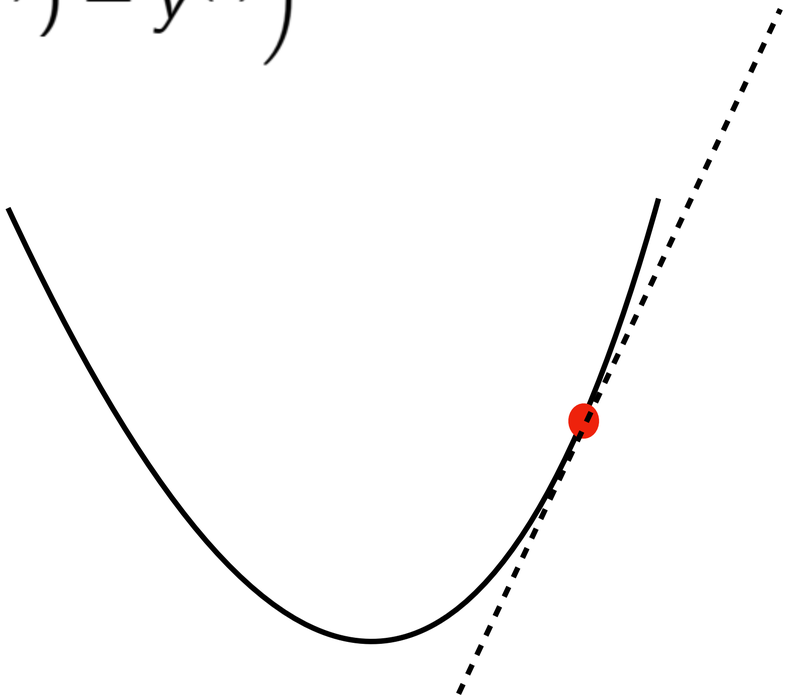
$$\theta_j := \theta_j - \alpha\frac{\partial}{\partial\theta_j}J(\theta)$$



Gradient Descent

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

$$\theta_j := \theta_j - \alpha\frac{\partial}{\partial\theta_j}J(\theta)$$



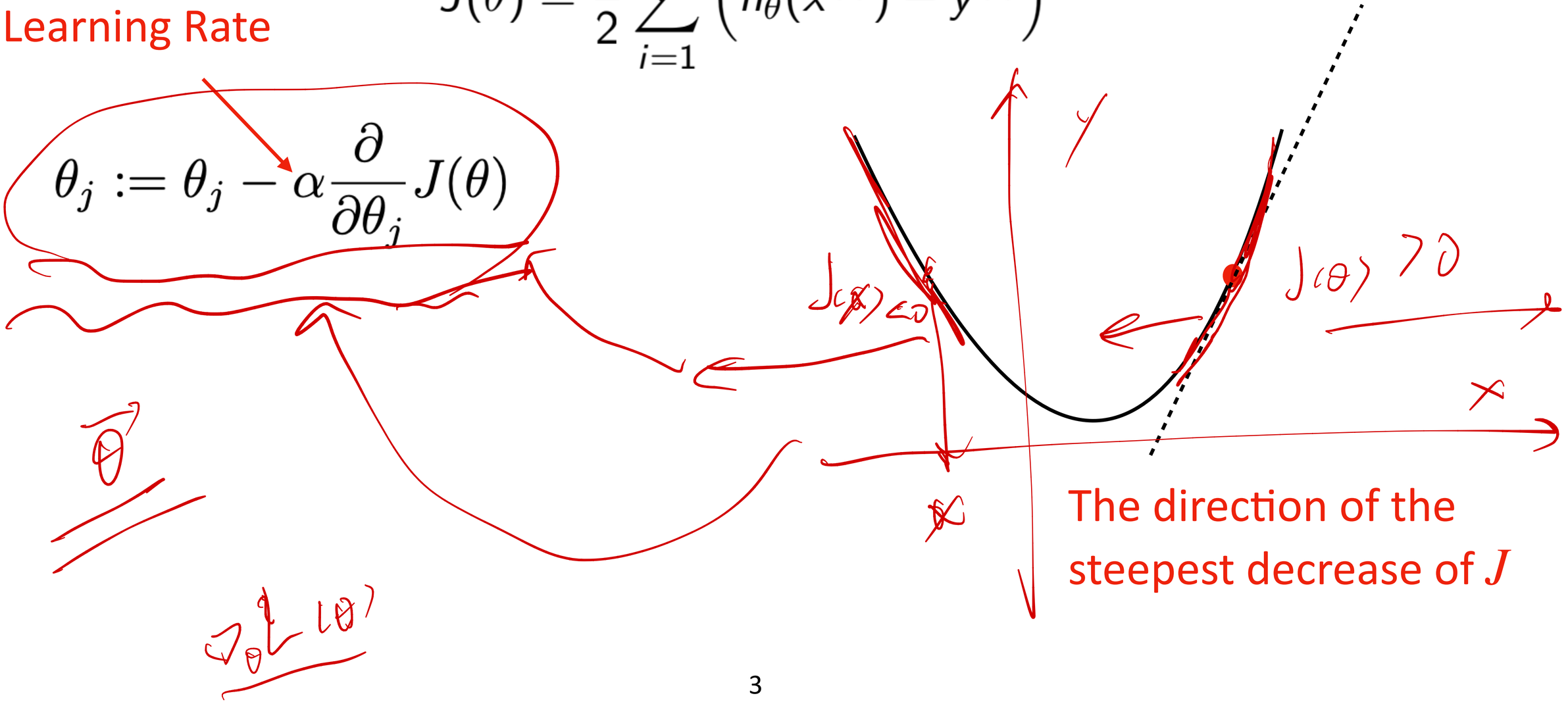
The direction of the steepest decrease of J

Gradient Descent

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

Learning Rate

$$\theta_j := \theta_j - \alpha\frac{\partial}{\partial\theta_j}J(\theta)$$



The direction of the steepest decrease of J

gradient descent:

$L(\theta)$: loss

$\vec{\theta}$: $\vec{u}$: direction vector (unit)

$\|\vec{u}\| = 1$

$\theta \leftarrow \theta - \alpha \nabla_\theta L(\theta)$ fixed

$\arg\min_{\vec{u}} L(\vec{\theta} + \vec{u} \cdot \varepsilon)$

Taylor expansion:

$L(\vec{\theta} + \vec{u} \cdot \varepsilon) \simeq L(\vec{\theta}) + \nabla_\theta L(\theta) \cdot \vec{u}$

$\nabla_\theta L(\theta) \cdot \vec{u} = \|\nabla_\theta L(\theta)\| \cdot \|\vec{u}\| \cdot \cos\theta$

constant

$\vec{u} = \dfrac{-\nabla_\theta L(\theta)}{\|\nabla_\theta L(\theta)\|}$

$\cos\theta = -1$

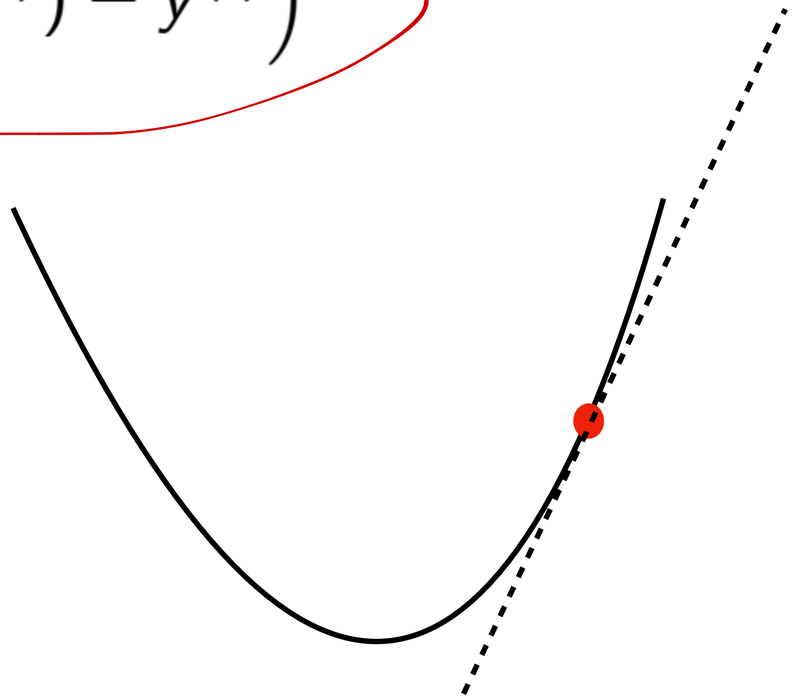
Gradient Descent

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

Learning Rate

$$\theta_j := \theta_j - \alpha\frac{\partial}{\partial\theta_j}J(\theta)$$

This update is simultaneously performed for all values of j = 0,...,d.



The direction of the steepest decrease of J

Gradient Descent

For a single training example:

$$
\begin{aligned}
\frac{\partial}{\partial \theta_j} J(\theta) &= \frac{\partial}{\partial \theta_j} \frac{1}{2} \left(h_\theta(x) - y\right)^2 \\
&= 2 \cdot \frac{1}{2} \left(h_\theta(x) - y\right) \cdot \frac{\partial}{\partial \theta_j} \left(h_\theta(x) - y\right) \\
&= \left(h_\theta(x) - y\right) \cdot \frac{\partial}{\partial \theta_j} \left(\sum_{i=0}^{d} \theta_i x_i - y \right) \\
&= \left(h_\theta(x) - y\right) x_j
\end{aligned}
$$

Gradient Descent

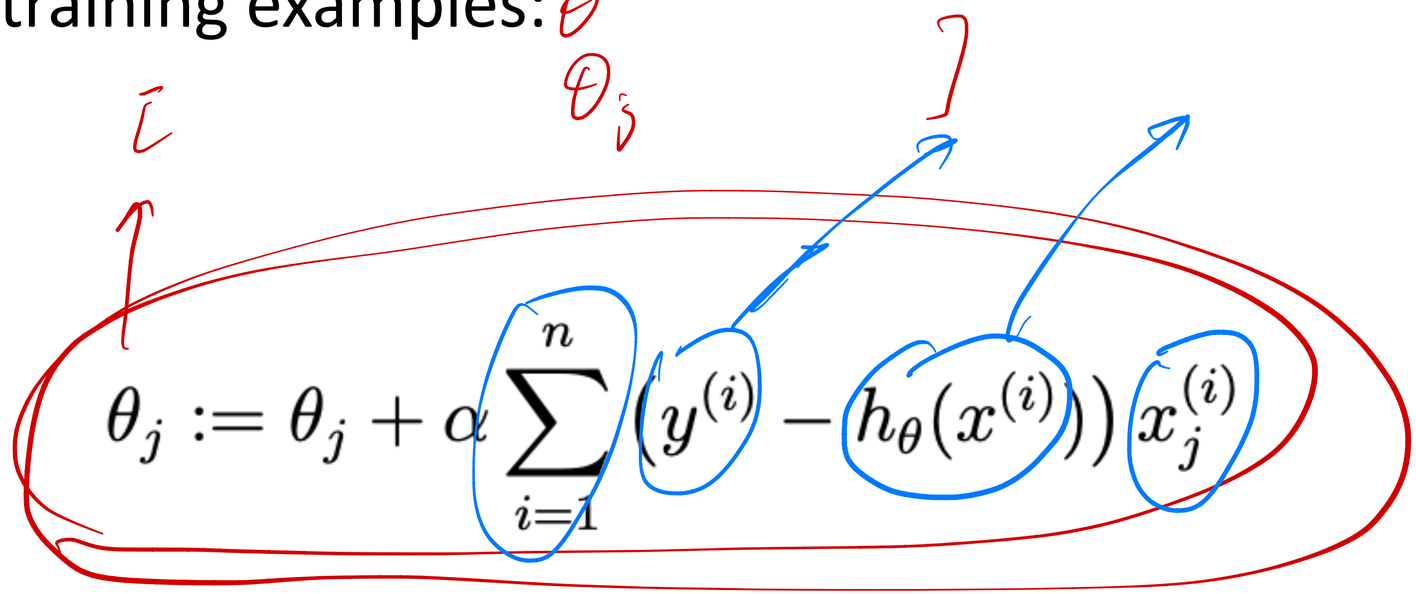
For a single training example:

$$
\begin{aligned}
\frac{\partial}{\partial \theta_j} J(\theta) &= \frac{\partial}{\partial \theta_j} \frac{1}{2} \left(h_\theta(x) - y \right)^2 \\
&= 2 \cdot \frac{1}{2} \left(h_\theta(x) - y \right) \cdot \frac{\partial}{\partial \theta_j} \left(h_\theta(x) - y \right) \\
&= \left(h_\theta(x) - y \right) \cdot \frac{\partial}{\partial \theta_j} \left(\sum_{i=0}^{d} \theta_i x_i - y \right) \\
&= \left(h_\theta(x) - y \right) x_j
\end{aligned}
$$

LMS (Least Mean Square) Update Rule

Batch Gradient Descent

For a multiple training examples:



$$\theta_j := \theta_j + \alpha \sum_{i=1}^{n} \left(y^{(i)} - h_\theta(x^{(i)}) \right) x_j^{(i)}$$

Repeat until convergence

$\theta_j^{new} x_j = \theta_j^{old} x_j + \alpha (y - h_\theta(x)) x_j^2 \neq 0$

$\theta_j := \theta_j + \alpha (y - h_\theta(x)) x_j$

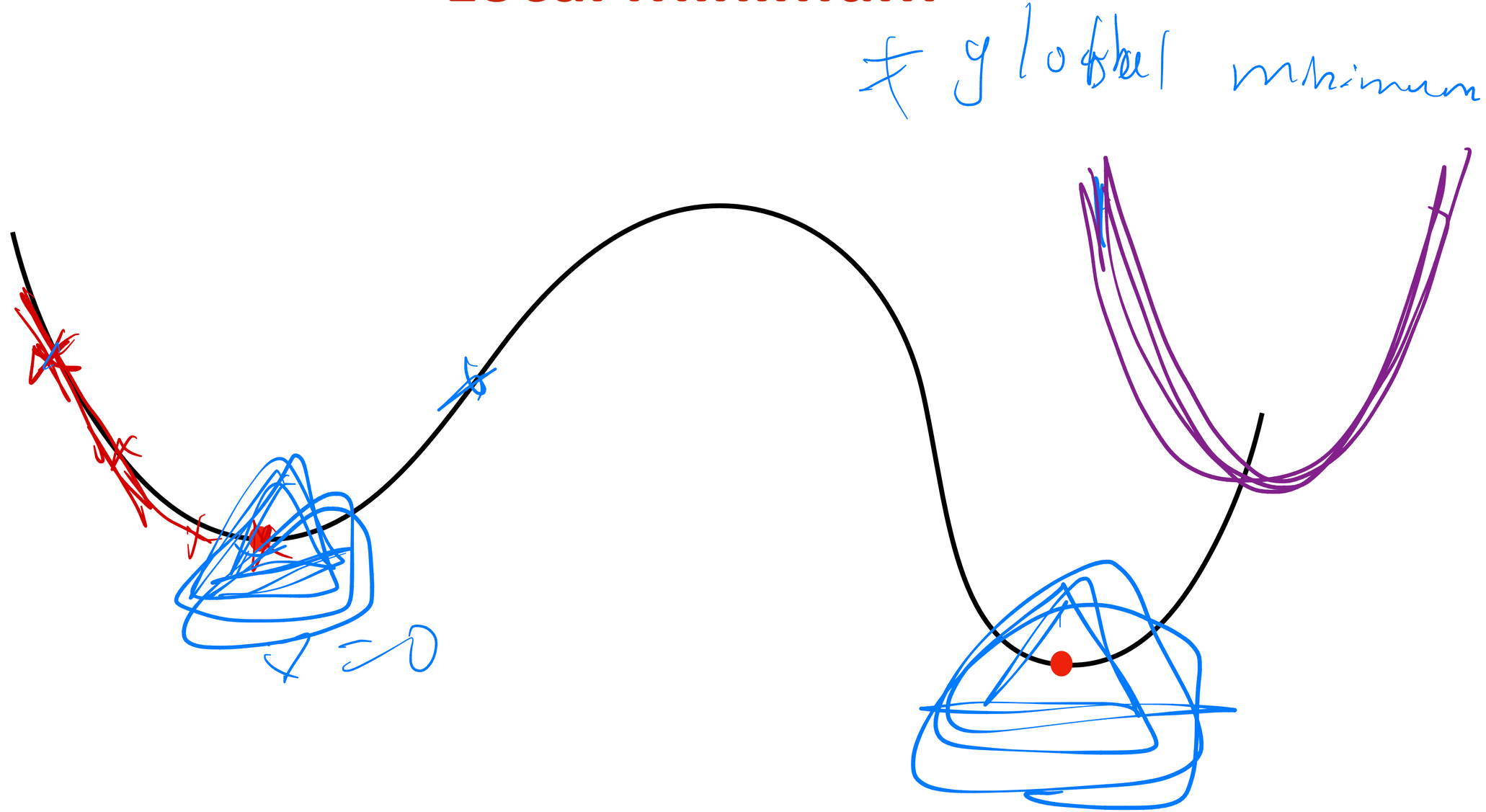
when $y - h_\theta(x) > 0$, $\theta_j x_j$ becomes lar

when $y - h_\theta(x) < 0$, $\theta_j x_j$ be comes smaller

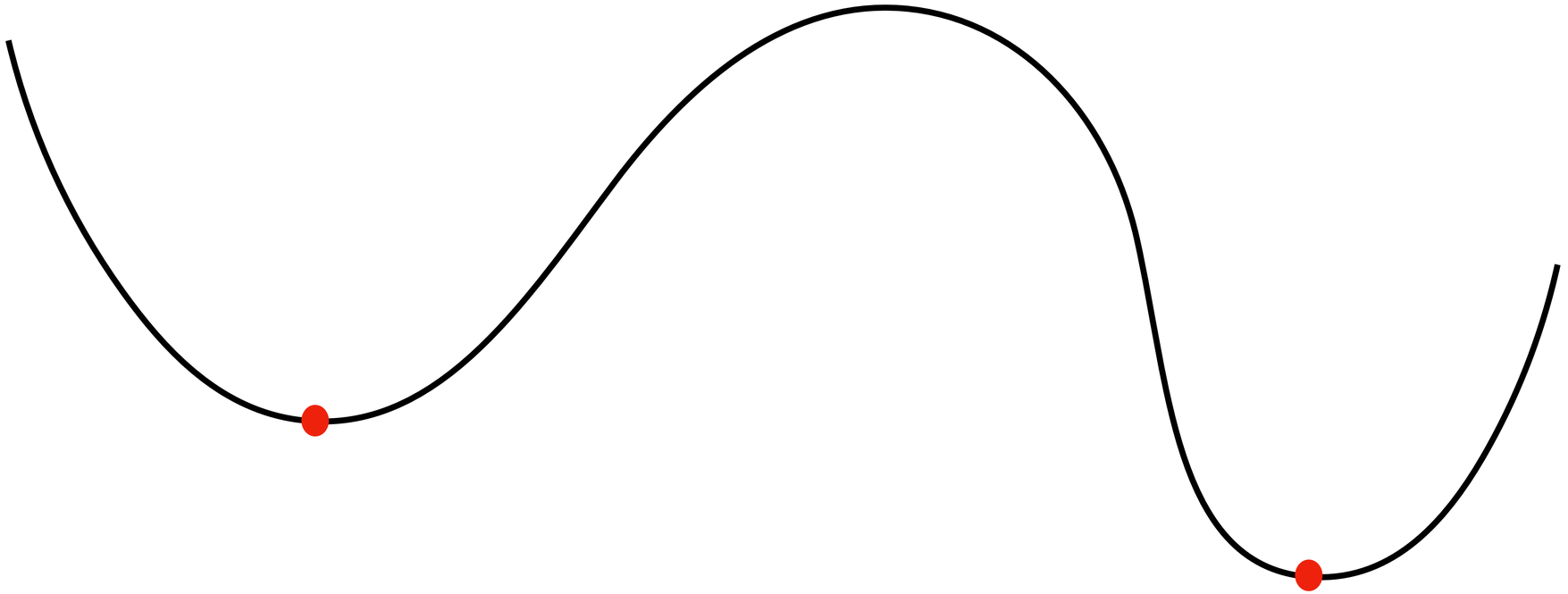
$h_\theta(x)$: $\theta_j x_j$ $\theta_j x_j$

$\theta_j^{old} x_j$

Local Minimum



Local Minimum

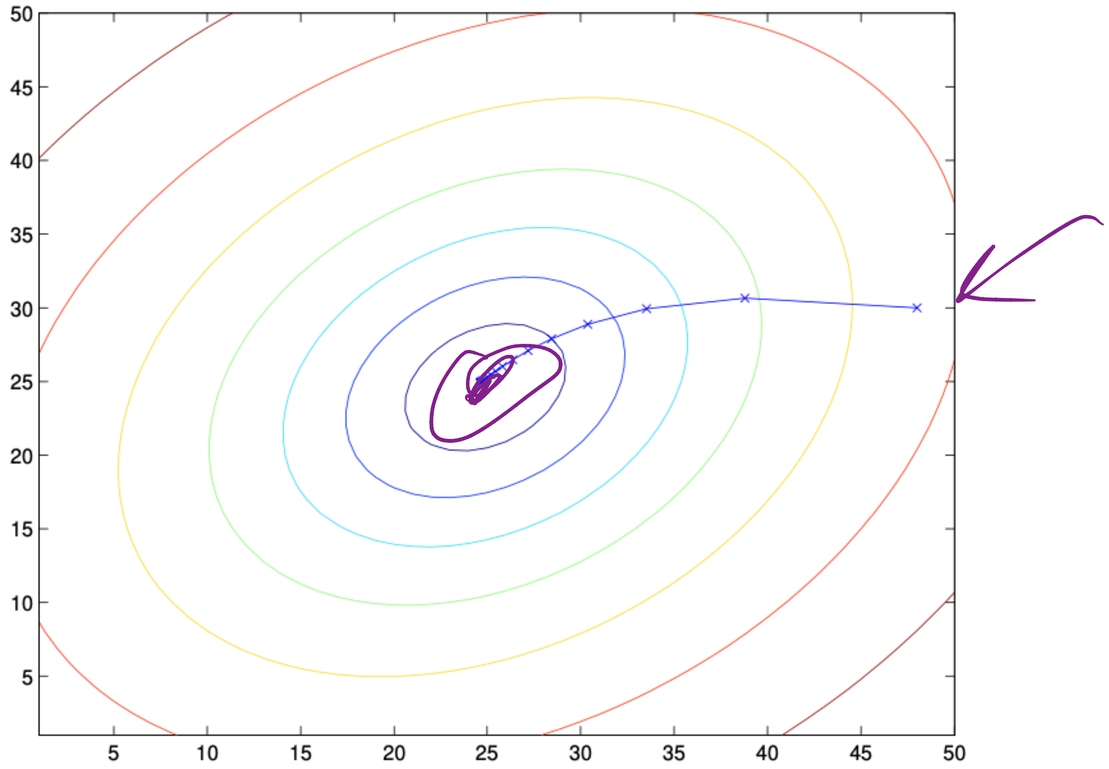


For least square optimization, are we likely to get local minima rather than the global minima through gradient descent?

No

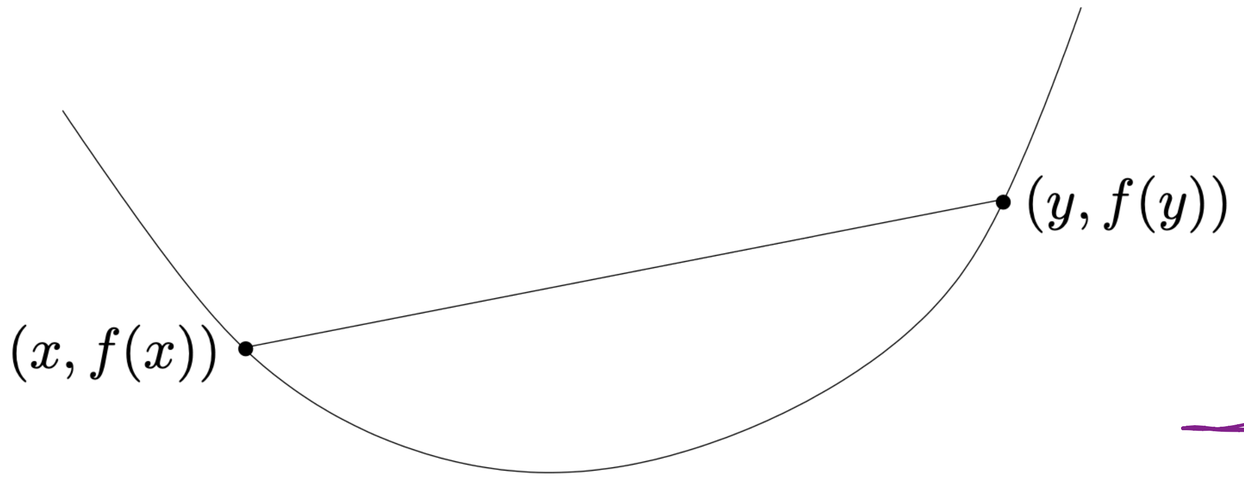
J is a convex quadratic function

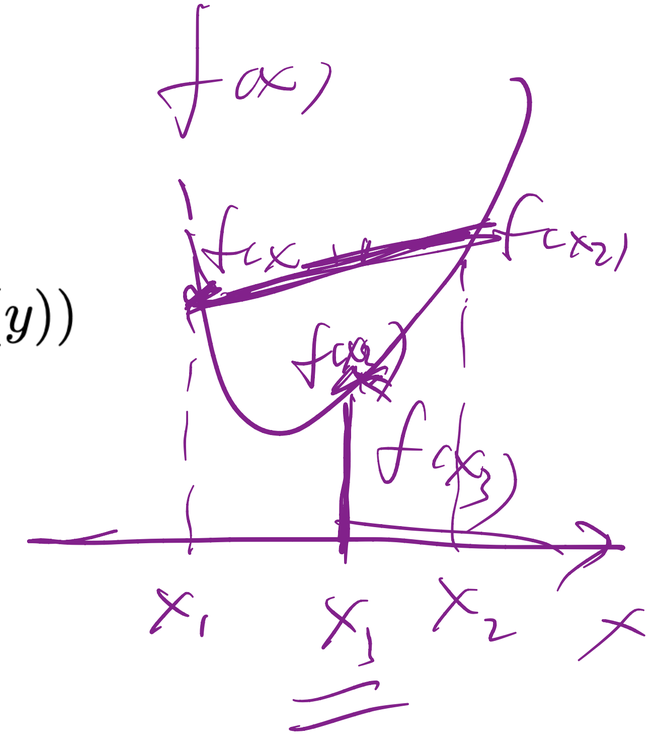
There is only one local minima for J



Convex Function

$$f(tx + (1-t)y) \leq tf(x) + (1-t)f(y) \quad \text{for } 0 \leq t \leq 1$$





Classification

CAT

Labels are discrete

Logistic Regression

Logistic Regression

Given a training set $\{(x^{(i)}, y^{(i)}) \text{ for } i = 1, \ldots, n\}$ let $y^{(i)} \in \{0, 1\}$.
Want $h_\theta(x) \in [0, 1]$ Let's pick a smooth function:

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

g map $\mathbb{R} \to [0,1]$

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

$h_\theta(x) = g = [0,1]$

$[0,1]$

$\in \mathbb{R}$

$\theta^T x$

$\text{prob}(y=1) = h_\theta(x)$

x_0

Logistic Regression

Given a training set $\{(x^{(i)}, y^{(i)}) \text{ for } i = 1, \ldots, n\}$ let $y^{(i)} \in \{0, 1\}$.
Want $h_\theta(x) \in [0, 1]$. Let's pick a smooth function:

$$h_\theta(x) = g(\theta^T x) \quad \text{Link Function}$$

Logistic Regression

Given a training set $\{(x^{(i)}, y^{(i)}) \text{ for } i = 1, \ldots, n\}$ let $y^{(i)} \in \{0, 1\}$.
Want $h_\theta(x) \in [0, 1]$. Let's pick a smooth function:

$$h_\theta(x) = g(\theta^T x) \quad \text{Link Function}$$

There are many options of g....

Logistic Regression

Given a training set $\{(x^{(i)}, y^{(i)}) \text{ for } i = 1, \ldots, n\}$ let $y^{(i)} \in \{0, 1\}$.
Want $h_\theta(x) \in [0, 1]$. Let's pick a smooth function:

$$h_\theta(x) = g(\theta^T x) \quad \text{Link Function}$$

There are many options of g....

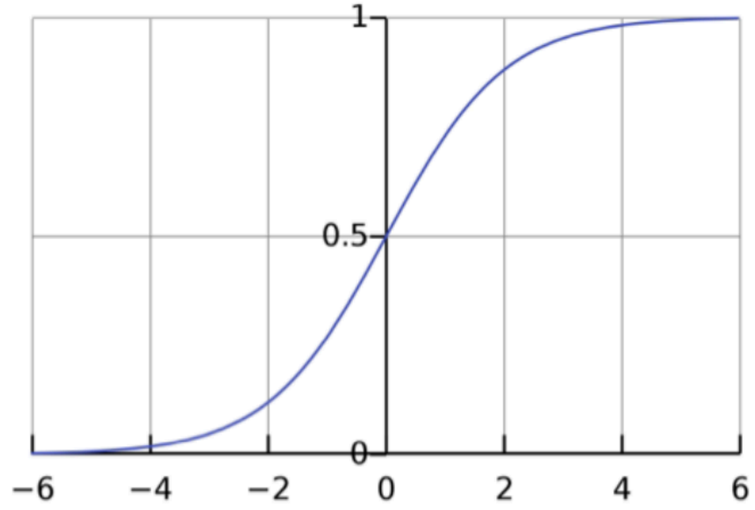
$$g(z) = \frac{1}{1 + e^{-z}}.$$

Logistic Regression

Given a training set $\{(x^{(i)}, y^{(i)}) \text{ for } i = 1, \ldots, n\}$ let $y^{(i)} \in \{0, 1\}$.
Want $h_\theta(x) \in [0, 1]$. Let's pick a smooth function:

$$h_\theta(x) = g(\theta^T x) \quad \text{Link Function}$$

There are many options of g....



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

Logistic Regression

Given a training set $\{(x^{(i)}, y^{(i)}) \text{ for } i = 1, \ldots, n\}$ let $y^{(i)} \in \{0, 1\}$.
Want $h_\theta(x) \in [0, 1]$. Let's pick a smooth function:

$$h_\theta(x) = g(\theta^T x)$$ Link Function

There are many options of g....

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

How do we interpret $h_\theta(x)$?

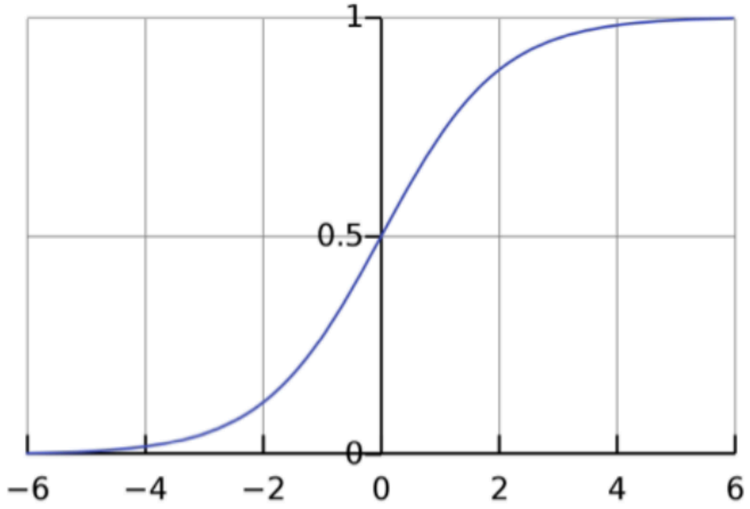
$$P(y = 1 \mid x; \theta) = h_\theta(x)$$
$$P(y = 0 \mid x; \theta) = 1 - h_\theta(x)$$

Logistic Regression

Given a training set $\{(x^{(i)}, y^{(i)}) \text{ for } i = 1, \ldots, n\}$ let $y^{(i)} \in \{0, 1\}$.
Want $h_\theta(x) \in [0, 1]$. Let's pick a smooth function:

$$h_\theta(x) = g(\theta^T x)$$ Link Function

There are many options of g....



Logistic Function

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

How do we interpret $h_\theta(x)$?

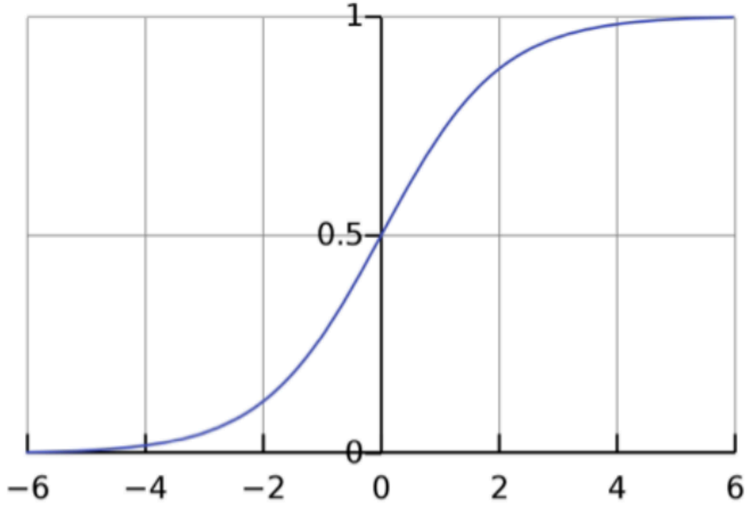
$$P(y = 1 \mid x; \theta) = h_\theta(x)$$
$$P(y = 0 \mid x; \theta) = 1 - h_\theta(x)$$

Logistic Regression

Given a training set $\{(x^{(i)}, y^{(i)}) \text{ for } i = 1, \ldots, n\}$ let $y^{(i)} \in \{0, 1\}$.
Want $h_\theta(x) \in [0, 1]$. Let's pick a smooth function:

$$h_\theta(x) = g(\theta^T x)$$ Link Function

There are many options of g....



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

Logistic Function

Sigmoid Function

How do we interpret $h_\theta(x)$?

$$P(y = 1 \mid x; \theta) = h_\theta(x)$$
$$P(y = 0 \mid x; \theta) = 1 - h_\theta(x)$$

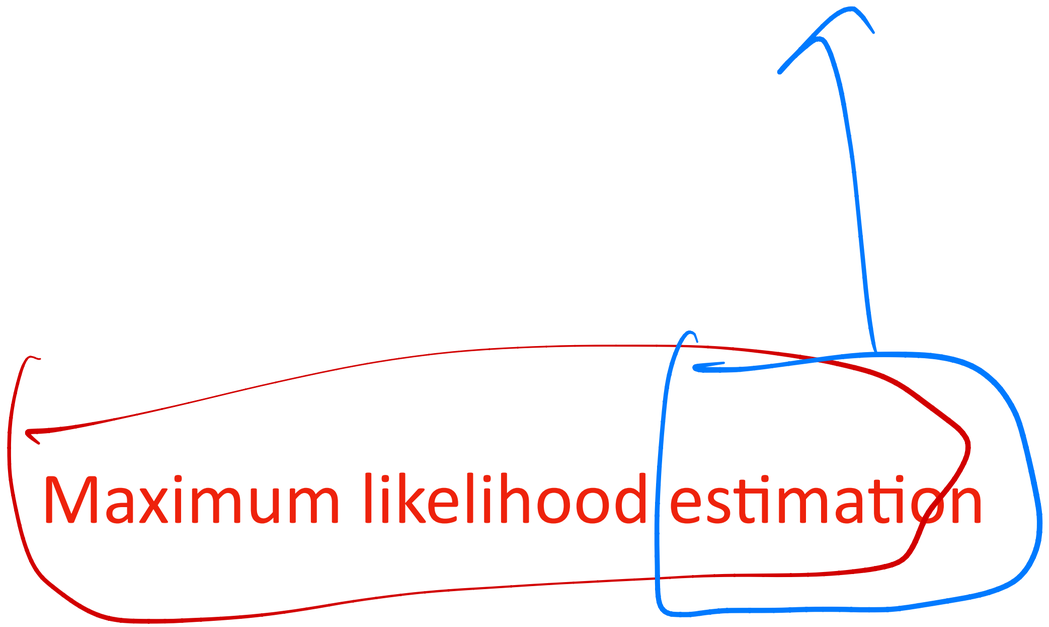
Logistic Regression

Let's write the Likelihood function. Recall:

$$P(y = 1 \mid x; \theta) = h_\theta(x)$$
$$P(y = 0 \mid x; \theta) = 1 - h_\theta(x)$$

$(x^{(i)}, y^{(i)})$

parameter



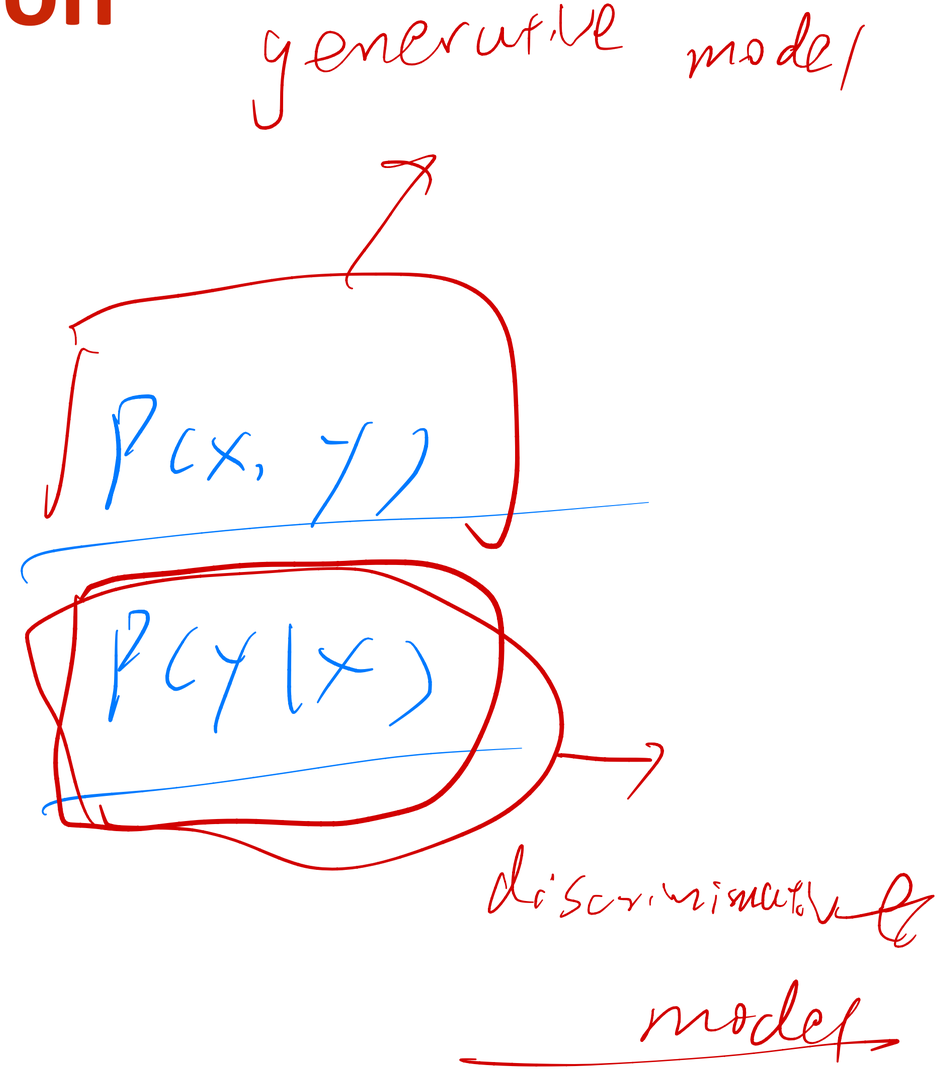
Logistic Regression

Let's write the Likelihood function. Recall:

$$P(y = 1 \mid x; \theta) = h_\theta(x)$$
$$P(y = 0 \mid x; \theta) = 1 - h_\theta(x)$$

Then,

$$L(\theta) = P(y \mid X; \theta) = \prod_{i=1}^{n} p(y^{(i)} \mid x^{(i)}; \theta)$$



Maximum likelihood estimation

Logistic Regression

Let's write the Likelihood function. Recall:

$$P(y = 1 \mid x; \theta) = h_\theta(x)$$
$$P(y = 0 \mid x; \theta) = 1 - h_\theta(x)$$

$P(y|x;\theta) = \begin{cases} h_\theta(x) & \text{if } y=1 \\ 1-h_\theta(x) & \text{if } y=0 \end{cases}$

Then,

$$L(\theta) = P(y \mid X; \theta) = \prod_{i=1}^{n} p(y^{(i)} \mid x^{(i)}; \theta)$$

We want to express "if-then" logics, how?

Maximum likelihood estimation

Logistic Regression

Let's write the Likelihood function. Recall:

$$P(y = 1 \mid x; \theta) = h_\theta(x)$$
$$P(y = 0 \mid x; \theta) = 1 - h_\theta(x)$$

Then,

$$L(\theta) = P(y \mid X; \theta) = \prod_{i=1}^{n} p(y^{(i)} \mid x^{(i)}; \theta)$$
$$= \prod_{i=1}^{n} h_\theta(x^{(i)})^{y^{(i)}} (1 - h_\theta(x^{(i)}))^{1-y^{(i)}}$$

We want to express "if-then" logics, how?

$$= \begin{cases} h_\theta(x) & \text{if } y=1 \\ 1-h_\theta(x) & \text{if } y=0 \end{cases}$$

Maximum likelihood estimation

Logistic Regression

Let's write the Likelihood function. Recall:

$$P(y = 1 \mid x; \theta) = h_\theta(x)$$
$$P(y = 0 \mid x; \theta) = 1 - h_\theta(x)$$

exp

$= \arg\min_\theta \exp L(\theta)$

$\arg\min_\theta L(\theta) = \arg\min_\theta \log L(\theta)$

Then,

$$L(\theta) = P(y \mid X; \theta) = \prod_{i=1}^{n} p(y^{(i)} \mid x^{(i)}; \theta)$$
$$= \prod_{i=1}^{n} h_\theta(x^{(i)})^{y^{(i)}} (1 - h_\theta(x^{(i)}))^{1-y^{(i)}}$$

We want to express "if-then" logics, how?

$h(x) = g(\theta^T x)$

Taking logs to compute the log likelihood $\ell(\theta)$ we have:

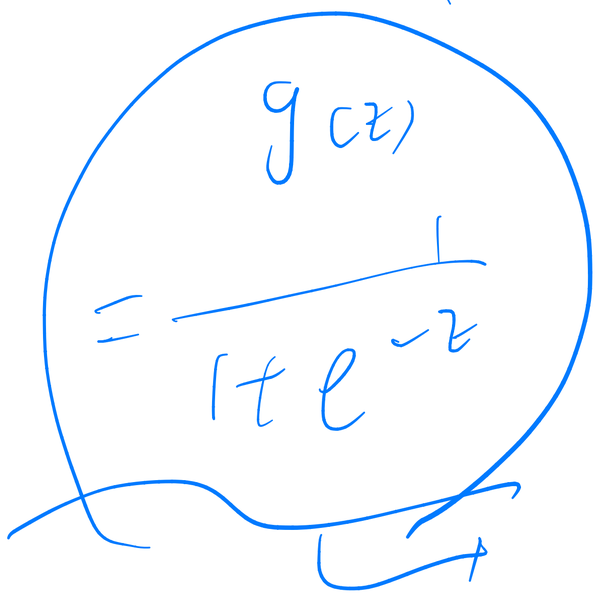
$$\ell(\theta) = \log L(\theta) = \sum_{i=1}^{n} y^{(i)} \log h_\theta(x^{(i)}) + (1 - y^{(i)}) \log(1 - h_\theta(x^{(i)}))$$

log likelihood

Maximum likelihood estimation

Derivative of Logistic Function

$$
\begin{aligned}
g'(z) &= \frac{d}{dz}\frac{1}{1+e^{-z}} \\
&= \frac{1}{(1+e^{-z})^2}\left(e^{-z}\right) \\
&= \frac{1}{(1+e^{-z})}\cdot\left(1-\frac{1}{(1+e^{-z})}\right) \\
&= g(z)(1-g(z)).
\end{aligned}
$$



= g(z)(1 − g(z))

Gradient Descent

$$
\begin{aligned}
\frac{\partial}{\partial \theta_j} \ell(\theta) &= \left(y \frac{1}{g(\theta^T x)} - (1-y) \frac{1}{1 - g(\theta^T x)} \right) \frac{\partial}{\partial \theta_j} g(\theta^T x) \\
&= \left(y \frac{1}{g(\theta^T x)} - (1-y) \frac{1}{1 - g(\theta^T x)} \right) g(\theta^T x)(1 - g(\theta^T x)) \frac{\partial}{\partial \theta_j} \theta^T x \\
&= \left(y(1 - g(\theta^T x)) - (1-y) g(\theta^T x) \right) x_j \\
&= (y - h_\theta(x))\, x_j
\end{aligned}
$$

$$
\theta_j := \theta_j + \alpha \left(y^{(i)} - h_\theta(x^{(i)}) \right) x_j^{(i)}
$$

$h_\theta(x^{(i)}) = g(\theta^T x)$

linear regression: $h_\theta(x^{(i)}) = \theta^T x$

Gradient Descent

$$
\begin{aligned}
\frac{\partial}{\partial \theta_j} \ell(\theta) &= \left(y \frac{1}{g(\theta^T x)} - (1-y) \frac{1}{1 - g(\theta^T x)} \right) \frac{\partial}{\partial \theta_j} g(\theta^T x) \\
&= \left(y \frac{1}{g(\theta^T x)} - (1-y) \frac{1}{1 - g(\theta^T x)} \right) g(\theta^T x)(1 - g(\theta^T x)) \frac{\partial}{\partial \theta_j} \theta^T x \\
&= \left(y(1 - g(\theta^T x)) - (1-y) g(\theta^T x) \right) x_j \\
&= \left(y - h_\theta(x) \right) x_j
\end{aligned}
$$

$$
\theta_j := \theta_j + \alpha \left(y^{(i)} - h_\theta(x^{(i)}) \right) x_j^{(i)}
$$

Looks identical to LMS update rule in linear regression

Gradient Descent

$$
\begin{aligned}
\frac{\partial}{\partial \theta_j}\ell(\theta) &= \left(y\frac{1}{g(\theta^T x)} - (1-y)\frac{1}{1-g(\theta^T x)}\right)\frac{\partial}{\partial \theta_j} g(\theta^T x) \\
&= \left(y\frac{1}{g(\theta^T x)} - (1-y)\frac{1}{1-g(\theta^T x)}\right) g(\theta^T x)(1-g(\theta^T x))\frac{\partial}{\partial \theta_j}\theta^T x \\
&= \left(y(1-g(\theta^T x)) - (1-y)g(\theta^T x)\right) x_j \\
&= \left(y - h_\theta(x)\right) x_j
\end{aligned}
$$

$$
\theta_j := \theta_j + \alpha \left(y^{(i)} - h_\theta(x^{(i)})\right) x_j^{(i)}
$$

Looks identical to LMS update rule in linear regression

Is this coincidence?

Multi-Label Classification

{Cat, dog, dragon, fish, pig}

Language Model

Multi-Label Classification

Given a training set $\{(x^{(1)}, y^{(1)}), \cdots, (x^{(n)}, y^{(n)})\}$, $y^{(i)} \in \{1,2,\cdots,k\}$,
we aim to model the distribution $p(y \mid x; \theta)$

$P(y=1 \mid x; \theta) = \sigma(\theta^T x)$

$P(y=0 \mid x; \theta) = 1 - \sigma(\theta^T x)$

$P(y \mid x; \theta) \in [0,1]$

$\theta^T x$ map $\theta^T x \longrightarrow [0,1]$

$P(y=1 \mid x) + P(y=2 \mid x) + \cdots + P(y=k \mid x) = 1$

$P(y=1 \mid x) = \sigma(\theta_1^T x)$

$P(y=2 \mid x) = \sigma(\theta_2^T x)$

Multi-Label Classification

Given a training set $\{(x^{(1)}, y^{(1)}), \cdots, (x^{(n)}, y^{(n)})\}$, $y^{(i)} \in \{1,2,\cdots,k\}$,
we aim to model the distribution $p(y \mid x; \theta)$

Categorical distribution, $p(y = k \mid x; \theta) = \phi_k$

$$\text{s.t.} \sum_{i=1}^{k} \phi_i = 1$$

Multi-Label Classification

Given a training set $\{(x^{(1)}, y^{(1)}), \cdots, (x^{(n)}, y^{(n)})\}$, $y^{(i)} \in \{1,2,\cdots,k\}$,
we aim to model the distribution $p(y \mid x; \theta)$

Categorical distribution, $p(y = k \mid x; \theta) = \phi_k$

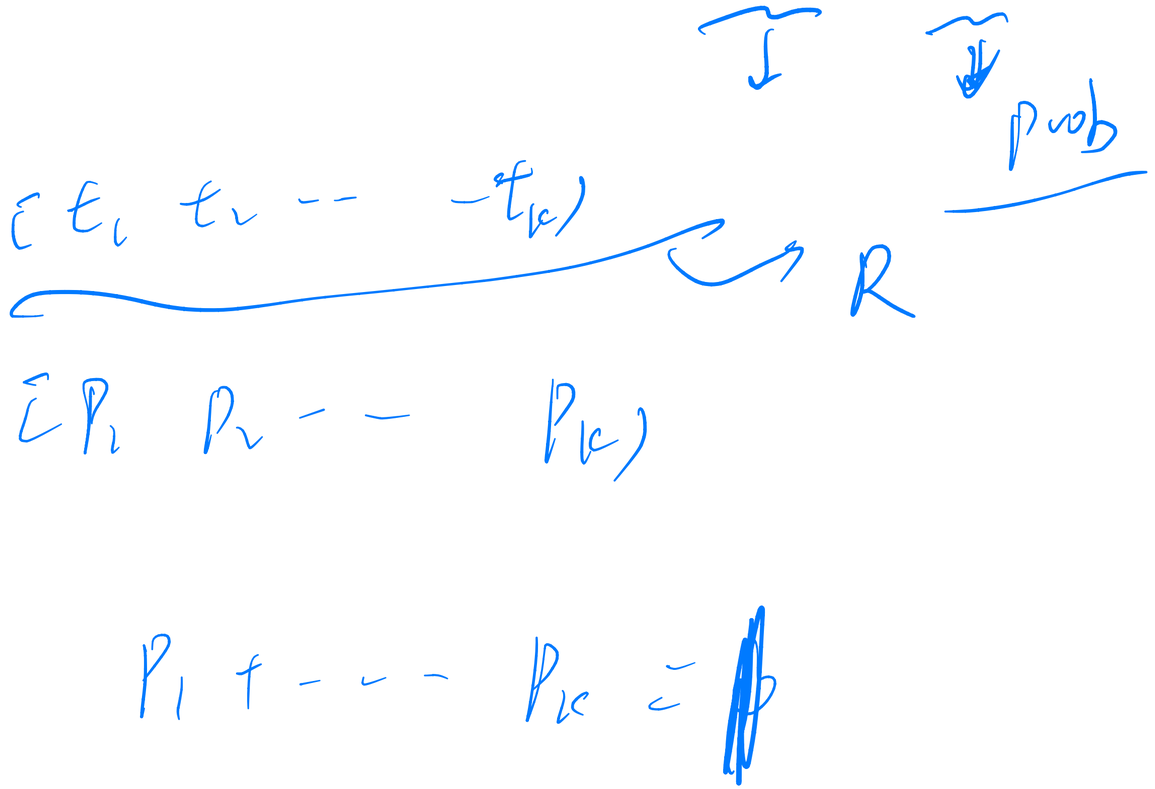
$$\text{s.t.} \sum_{i=1}^{k} \phi_i = 1$$

$\phi_i = \theta_i^T x$?

Softmax Function

Softmax Function

Softmax: $\mathbb{R}^k \rightarrow \mathbb{R}^k$



Softmax Function

Softmax: $\mathbb{R}^k \to \mathbb{R}^k$

$$\text{softmax}(t_1, \ldots, t_k) = \begin{bmatrix} \frac{\exp(t_1)}{\sum_{j=1}^{k} \exp(t_j)} \\ \vdots \\ \frac{\exp(t_k)}{\sum_{j=1}^{k} \exp(t_j)} \end{bmatrix}$$

Softmax Function

Softmax: $\mathbb{R}^k \to \mathbb{R}^k$

$$\text{softmax}(t_1, \ldots, t_k) = \begin{bmatrix} \frac{\exp(t_1)}{\sum_{j=1}^{k} \exp(t_j)} \\ \vdots \\ \frac{\exp(t_k)}{\sum_{j=1}^{k} \exp(t_j)} \end{bmatrix}$$

The denominator is a normalization constant

Softmax VS. Sigmoid

$z = t_1 - t_2$

$\frac{1}{1+e^{-z}}$

2-label

Softmax

$y=1$ $y=0$

$t_1 = \theta_1^T x$

$t_2 = \theta_2^T x$

$P(y=1|x) = \frac{e^{t_1}}{e^{t_1}+e^{t_2}} = \frac{1}{1+e^{-(t_1-t_2)}}$

$P(y=0|x) = \frac{e^{t_2}}{e^{t_1}+e^{t_2}}$

Multi-Label Classification

Multi-Label Classification

$$\text{Let } (t_1, \ldots, t_k) = (\theta_1^\top x, \cdots, \theta_k^\top x)$$

Multi-Label Classification

$$\text{Let } (t_1, \ldots, t_k) = (\theta_1^\top x, \cdots, \theta_k^\top x)$$

$$\begin{bmatrix} P(y=1 \mid x;\theta) \\ \vdots \\ P(y=k \mid x;\theta) \end{bmatrix} = \text{softmax}(t_1, \cdots, t_k) = \begin{bmatrix} \frac{\exp(\theta_1^\top x)}{\sum_{j=1}^k \exp(\theta_j^\top x)} \\ \vdots \\ \frac{\exp(\theta_k^\top x)}{\sum_{j=1}^k \exp(\theta_j^\top x)} \end{bmatrix}$$

Multi-Label Classification

$$\text{Let } (t_1, \ldots, t_k) = (\theta_1^\top x, \cdots, \theta_k^\top x)$$

$$\begin{bmatrix} P(y=1 \mid x;\theta) \\ \vdots \\ P(y=k \mid x;\theta) \end{bmatrix} = \text{softmax}(t_1, \cdots, t_k) = \begin{bmatrix} \frac{\exp(\theta_1^\top x)}{\sum_{j=1}^k \exp(\theta_j^\top x)} \\ \vdots \\ \frac{\exp(\theta_k^\top x)}{\sum_{j=1}^k \exp(\theta_j^\top x)} \end{bmatrix}$$

$$P(y=i \mid x;\theta) = \phi_i = \frac{\exp(t_i)}{\sum_{j=1}^k \exp(t_j)} = \frac{\exp(\theta_i^\top x)}{\sum_{j=1}^k \exp(\theta_j^\top x)}$$

Multi-Label Classification

Multi-Label Classification

$$-\log p(y \mid x, \theta) = -\log\left(\frac{\exp(t_y)}{\sum_{j=1}^{k}\exp(t_j)}\right) = -\log\left(\frac{\exp(\theta_y^\top x)}{\sum_{j=1}^{k}\exp(\theta_j^\top x)}\right)$$

Multi-Label Classification

$$-\log p(y \mid x, \theta) = -\log\left(\frac{\exp(t_y)}{\sum_{j=1}^{k}\exp(t_j)}\right) = -\log\left(\frac{\exp(\theta_y^\top x)}{\sum_{j=1}^{k}\exp(\theta_j^\top x)}\right)$$

$$\ell(\theta) = \sum_{i=1}^{n} -\log\left(\frac{\exp(\theta_{y^{(i)}}^\top x^{(i)})}{\sum_{j=1}^{k}\exp(\theta_j^\top x^{(i)})}\right)$$

$y^{(i)} \in \{1, 2, 3 \cdots k\}$

sample index

Multi-Label Classification

$$-\log p(y \mid x, \theta) = -\log\left(\frac{\exp(t_y)}{\sum_{j=1}^{k}\exp(t_j)}\right) = -\log\left(\frac{\exp(\theta_y^\top x)}{\sum_{j=1}^{k}\exp(\theta_j^\top x)}\right)$$

$$\ell(\theta) = \sum_{i=1}^{n} -\log\left(\frac{\exp(\theta_{y^{(i)}}^\top x^{(i)})}{\sum_{j=1}^{k}\exp(\theta_j^\top x^{(i)})}\right)$$

Negative log likelihood

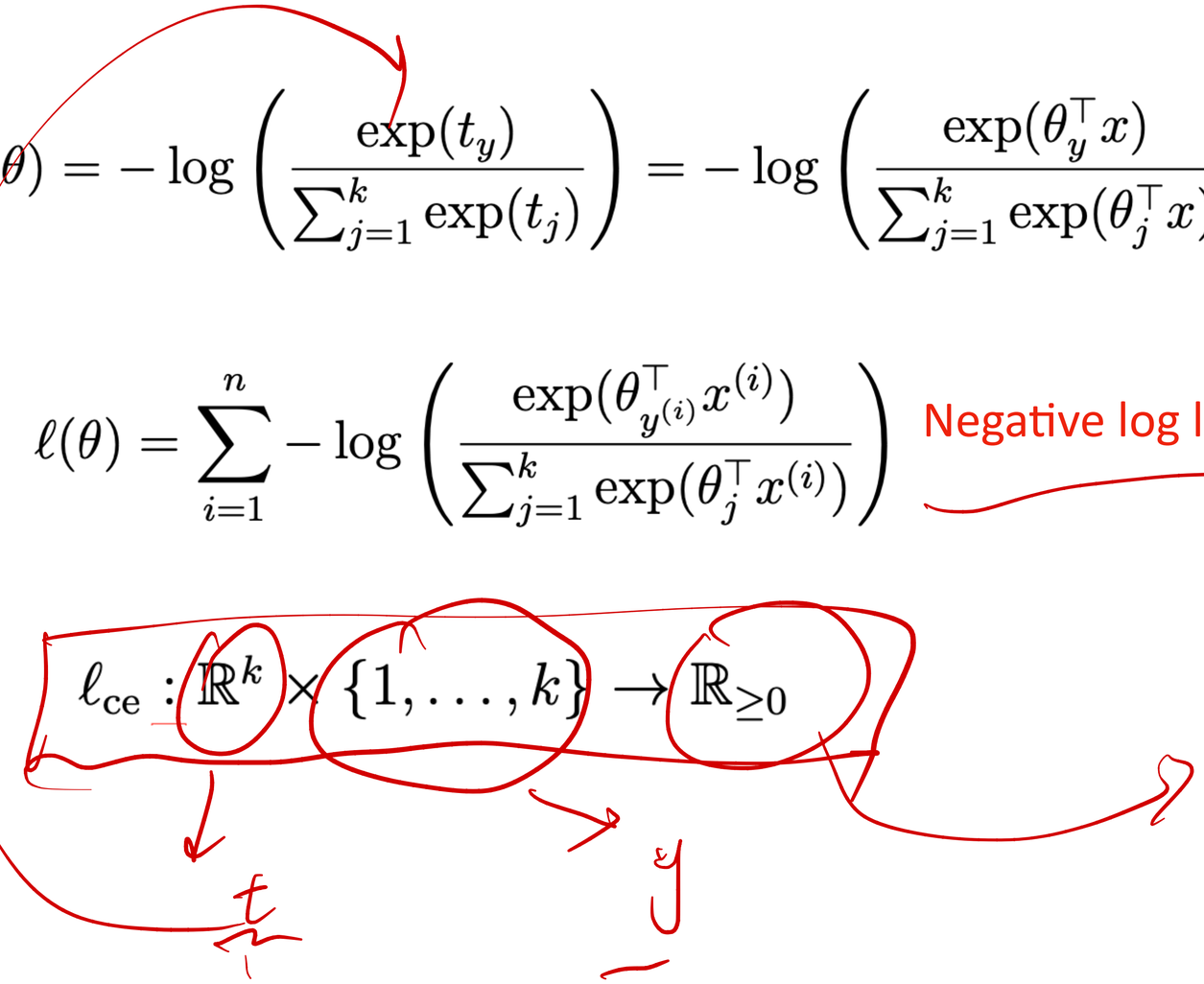
Multi-Label Classification

$$-\log p(y \mid x, \theta) = -\log\left(\frac{\exp(t_y)}{\sum_{j=1}^{k}\exp(t_j)}\right) = -\log\left(\frac{\exp(\theta_y^\top x)}{\sum_{j=1}^{k}\exp(\theta_j^\top x)}\right)$$

$$\ell(\theta) = \sum_{i=1}^{n} -\log\left(\frac{\exp(\theta_{y^{(i)}}^\top x^{(i)})}{\sum_{j=1}^{k}\exp(\theta_j^\top x^{(i)})}\right)$$

Negative log likelihood

Cross-entropy loss $\ell_{\text{ce}} : \mathbb{R}^k \times \{1, \ldots, k\} \to \mathbb{R}_{\geq 0}$



Multi-Label Classification

$$-\log p(y \mid x, \theta) = -\log\left(\frac{\exp(t_y)}{\sum_{j=1}^{k}\exp(t_j)}\right) = -\log\left(\frac{\exp(\theta_y^\top x)}{\sum_{j=1}^{k}\exp(\theta_j^\top x)}\right)$$

$$\ell(\theta) = \sum_{i=1}^{n} -\log\left(\frac{\exp(\theta_{y^{(i)}}^\top x^{(i)})}{\sum_{j=1}^{k}\exp(\theta_j^\top x^{(i)})}\right)$$

Negative log likelihood

Cross-entropy loss

$$\ell_{\text{ce}} : \mathbb{R}^k \times \{1, \ldots, k\} \to \mathbb{R}_{\geq 0}$$

$$\ell_{\text{ce}}((t_1, \ldots, t_k), y) = -\log\left(\frac{\exp(t_y)}{\sum_{j=1}^{k}\exp(t_j)}\right)$$

Multi-Label Classification

$$-\log p(y \mid x, \theta) = -\log\left(\frac{\exp(t_y)}{\sum_{j=1}^{k}\exp(t_j)}\right) = -\log\left(\frac{\exp(\theta_y^\top x)}{\sum_{j=1}^{k}\exp(\theta_j^\top x)}\right)$$

$$\ell(\theta) = \sum_{i=1}^{n} -\log\left(\frac{\exp(\theta_{y^{(i)}}^\top x^{(i)})}{\sum_{j=1}^{k}\exp(\theta_j^\top x^{(i)})}\right)$$

Negative log likelihood

Cross-entropy loss $\ell_{\text{ce}} : \mathbb{R}^k \times \{1, \ldots, k\} \to \mathbb{R}_{\geq 0}$

$$\ell_{\text{ce}}((t_1, \ldots, t_k), y) = -\log\left(\frac{\exp(t_y)}{\sum_{j=1}^{k}\exp(t_j)}\right)$$

$$\ell(\theta) = \sum_{i=1}^{n} \ell_{\text{ce}}((\theta_1^\top x^{(i)}, \ldots, \theta_k^\top x^{(i)}), y^{(i)})$$

$t_j = \theta_j^\top x$

$t_j = NN(x)$

$t_j = GPT(x)$

The Derivative

The Derivative

$$\frac{\partial \ell_{\text{ce}}(t, y)}{\partial t_i} = \phi_i - 1\{y = i\}$$

$$1\{y=i\} = \begin{cases} 1 & \text{if } y=i \\ 0 & \text{if } y \neq i \end{cases}$$

ϕ_i : Prob($y=i$)

t_i

The Derivative

$$\frac{\partial \ell_{\text{ce}}(t, y)}{\partial t_i} = \phi_i - 1\{y = i\}$$

$$\phi_i = \frac{\exp(t_i)}{\sum_{j=1}^{k} \exp(t_j)}$$

The Derivative

$$\frac{\partial \ell_{\text{ce}}(t, y)}{\partial t_i} = \phi_i - 1\{y = i\} \qquad \phi_i = \frac{\exp(t_i)}{\sum_{j=1}^{k} \exp(t_j)}$$

Chain rule

$$\frac{\partial \ell_{\text{ce}}((\theta_1^\top x, \ldots, \theta_k^\top x), y)}{\partial \theta_i} = \frac{\partial \ell(t, y)}{\partial t_i} \cdot \frac{\partial t_i}{\partial \theta_i} = (\phi_i - 1\{y = i\}) \cdot x$$

$t_i = \theta_i^\top x$

$= x$

θ_i

The Derivative

$$\frac{\partial \ell_{\text{ce}}(t, y)}{\partial t_i} = \phi_i - 1\{y = i\} \qquad \phi_i = \frac{\exp(t_i)}{\sum_{j=1}^{k} \exp(t_j)}$$

Chain rule

$$\frac{\partial \ell_{\text{ce}}((\theta_1^\top x, \ldots, \theta_k^\top x), y)}{\partial \theta_i} = \frac{\partial \ell(t, y)}{\partial t_i} \cdot \frac{\partial t_i}{\partial \theta_i} = (\phi_i - 1\{y = i\}) \cdot x$$

$$\frac{\partial \ell(\theta)}{\partial \theta_i} = \sum_{j=1}^{n} (\phi_i^{(j)} - 1\{y^{(j)} = i\}) \cdot x^{(j)}$$

The Derivative

$$\frac{\partial \ell_{\text{ce}}(t, y)}{\partial t_i} = \phi_i - 1\{y = i\} \qquad \phi_i = \frac{\exp(t_i)}{\sum_{j=1}^{k} \exp(t_j)}$$

Chain rule

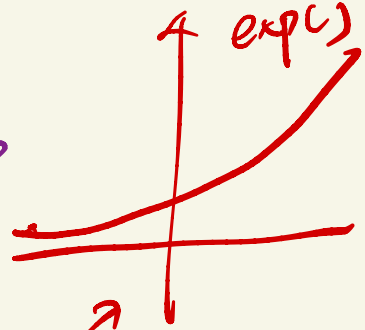
$$\frac{\partial \ell_{\text{ce}}((\theta_1^\top x, \ldots, \theta_k^\top x), y)}{\partial \theta_i} = \frac{\partial \ell(t, y)}{\partial t_i} \cdot \frac{\partial t_i}{\partial \theta_i} = (\phi_i - 1\{y = i\}) \cdot x$$

$$\frac{\partial \ell(\theta)}{\partial \theta_i} = \sum_{j=1}^{n} (\phi_i^{(j)} - 1\{y^{(j)} = i\}) \cdot x^{(j)}$$

Intuitive explanation of the rule?

$$\vec{\theta}_i \longleftarrow \vec{\theta}_i + [\,1(y=i) - \phi_i\,]\,\vec{x}$$

$\vec{\theta}_i^{\,old}$ $\vec{\theta}_i^{\,new}$



$t_i = \vec{\theta}_i^{\,T}\vec{x}$

$\text{Prob} \propto \exp(t_i)$

$\text{Prob}(y=i) = \dfrac{\exp(t_i)}{\sum \exp \cdots}$

$\vec{\theta}_i^{\,new}\vec{x} = \vec{\theta}_i^{\,old}\vec{x} + [\,1(y=i) - \phi_i]\,\vec{x}^{\,T}\vec{x} \geq 0$

if $y=i$, $1(y=i) - \phi_i > 0$, $\vec{\theta}_i^{\,new}\vec{x} > \vec{\theta}_i^{\,old}\vec{x}$

if $y \neq i$, $\vec{\theta}_i^{\,new}\vec{x} < \vec{\theta}_i^{\,old}\vec{x}$

Another Optimization Method — Newton's Method

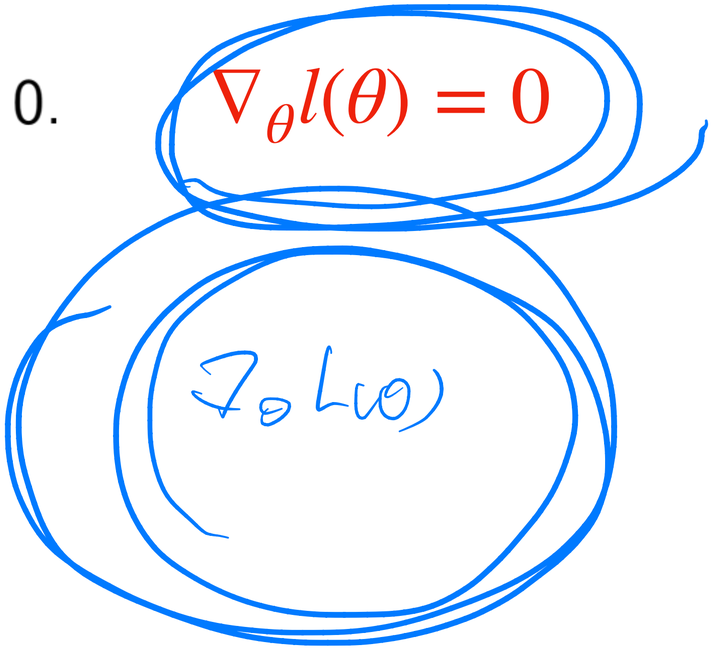
Another Optimization Method — Newton's Method

Given $f : \mathbb{R}^d \to \mathbb{R}$ find x s.t. $f(x) = 0$.

$f(x) = 0$

Another Optimization Method — Newton's Method

Given $f : \mathbb{R}^d \to \mathbb{R}$ find x s.t. $f(x) = 0$.



Another Optimization Method — Newton's Method

Given $f : \mathbb{R}^d \to \mathbb{R}$ find x s.t. $f(x) = 0$.

- This is the update rule in 1d

$$x^{(t+1)} = x^{(t)} - \frac{f(x^{(t)})}{f'(x^{(t)})}$$

$\nabla_\theta l(\theta) = 0$

$f(x) = \nabla_x L(x)$

Another Optimization Method — Newton's Method

Given $f : \mathbb{R}^d \to \mathbb{R}$ find x s.t. $f(x) = 0$. $\quad \nabla_\theta l(\theta) = 0$

- This is the update rule in 1d

$$x^{(t+1)} = x^{(t)} - \frac{f(x^{(t)})}{f'(x^{(t)})}$$

↓ Solution to a linear equation

$$f'(x^{(t)})x^{(t+1)} + f(x^{(t)}) - x^{(t)}f'(x^{(t)}) = 0$$

Another Optimization Method — Newton's Method

Given $f : \mathbb{R}^d \to \mathbb{R}$ find x s.t. $f(x) = 0$. $\quad \nabla_\theta l(\theta) = 0$

- This is the update rule in 1d

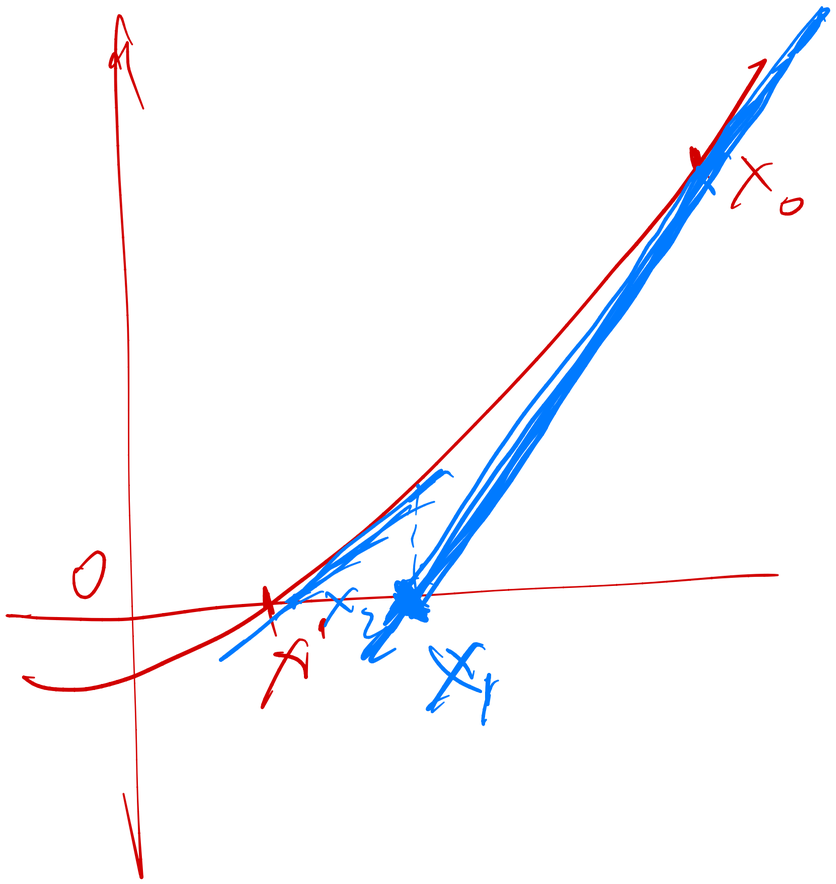
$$x^{(t+1)} = x^{(t)} - \frac{f(x^{(t)})}{f'(x^{(t)})}$$

↓ Solution to a linear equation

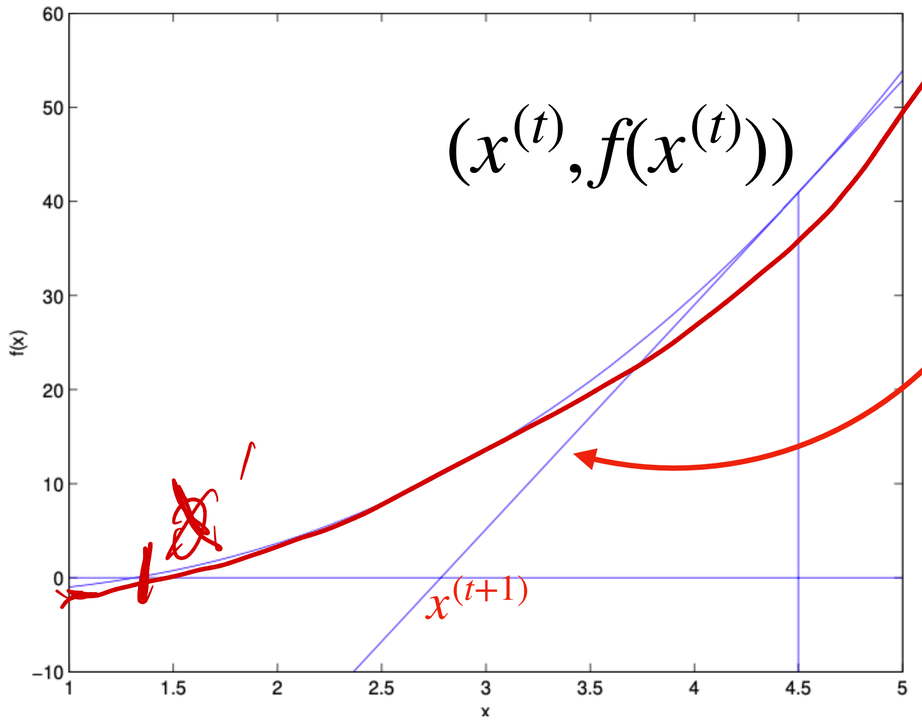
$$f'(x^{(t)})x^{(t+1)} + f(x^{(t)}) - x^{(t)}f'(x^{(t)}) = 0$$

View it as a equation of $x^{(t+1)}$, and $x^{(t)}$ is a constant

Another Optimization Method — Newton's Method

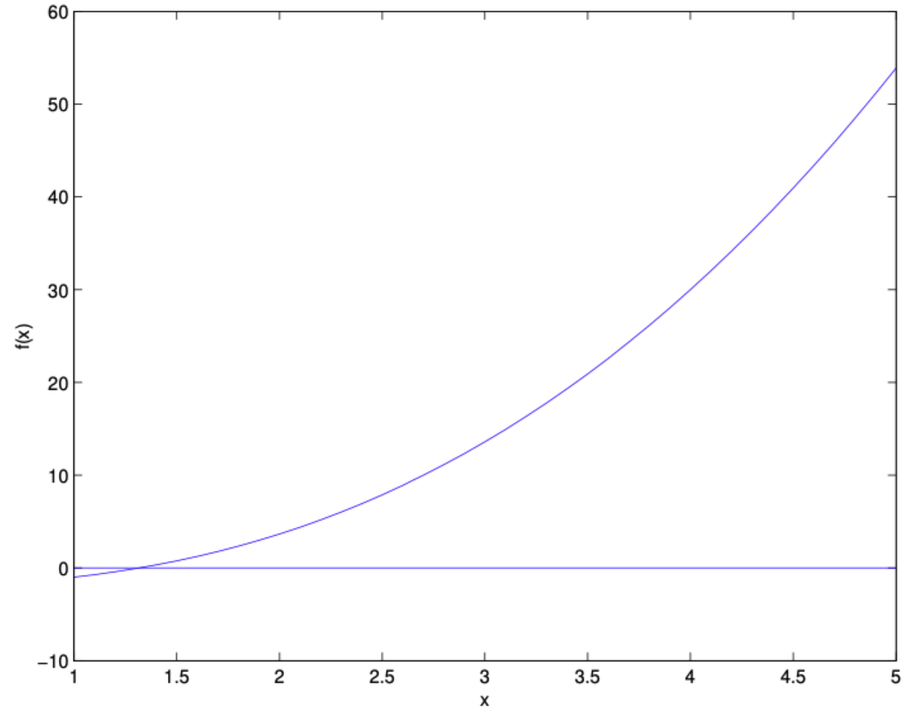


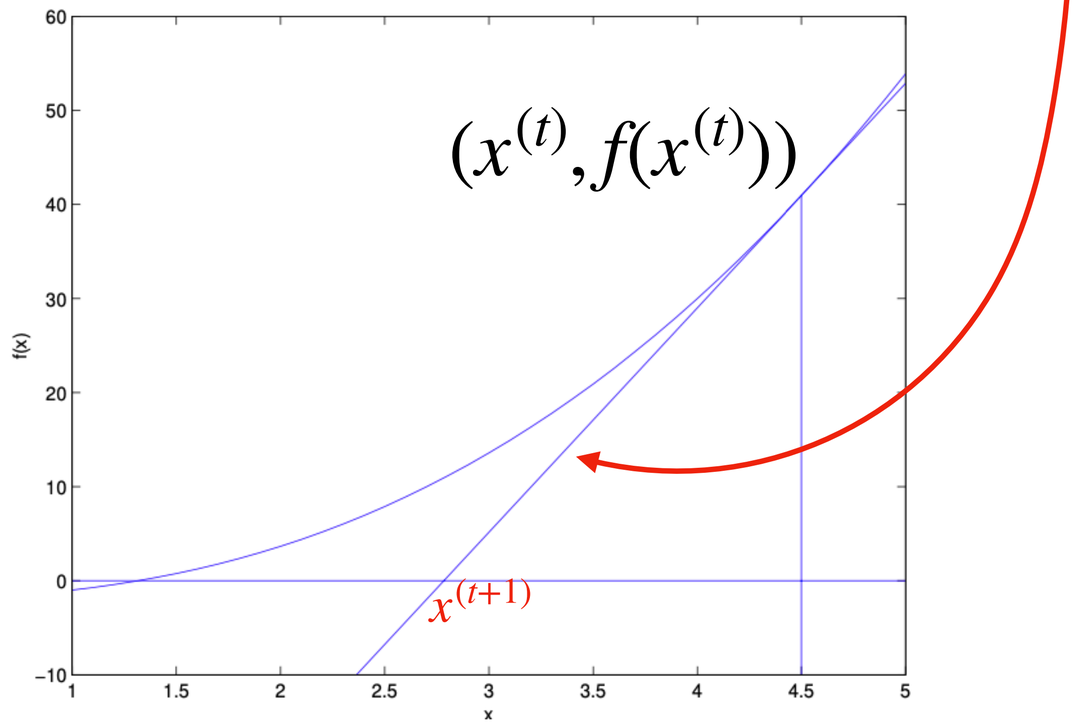
$$f'(x^{(t)})x + f(x^{(t)}) - x^{(t)}f'(x^{(t)}) = y$$



Another Optimization Method — Newton's Method

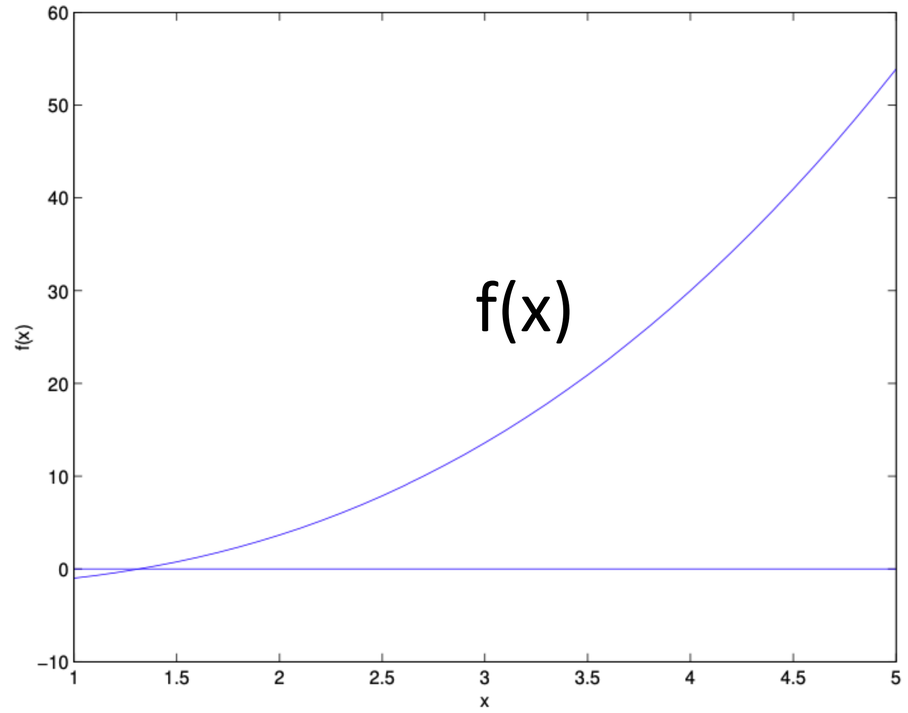
$$f'(x^{(t)})x + f(x^{(t)}) - x^{(t)}f'(x^{(t)}) = y$$

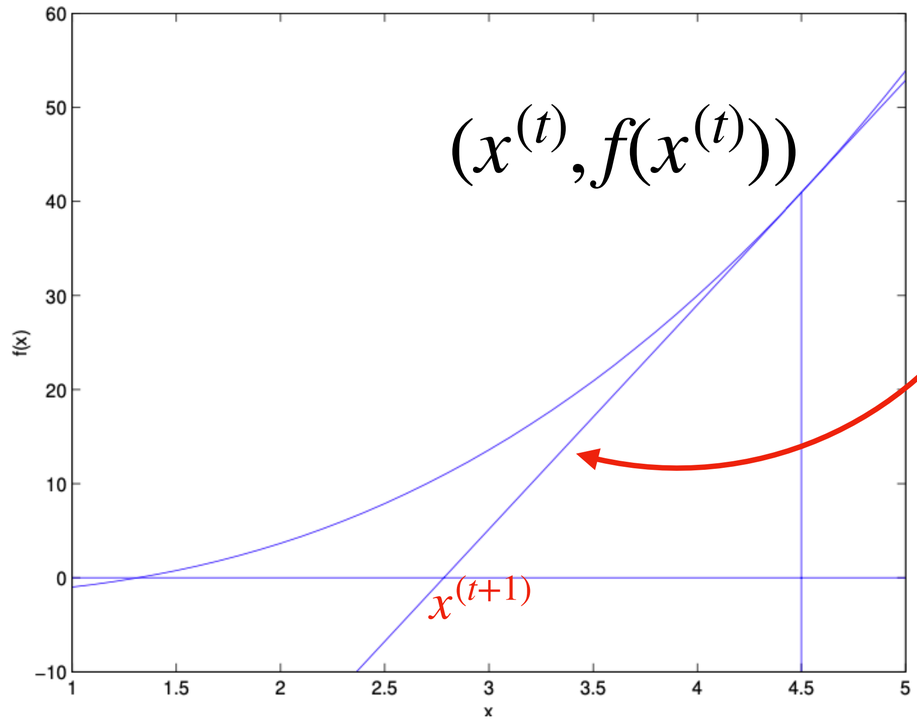




Another Optimization Method — Newton's Method

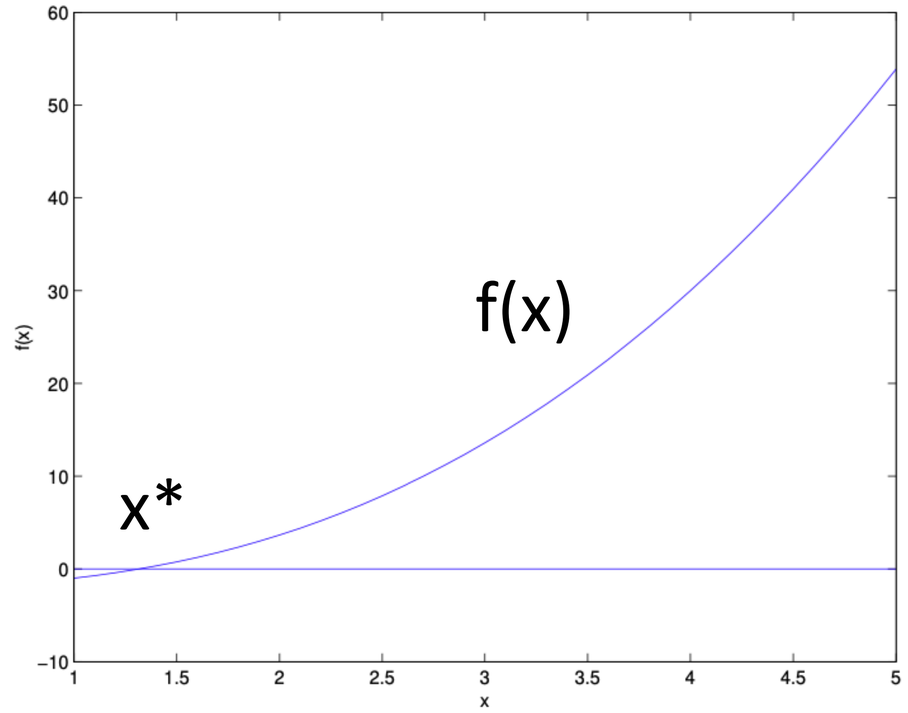
$$f'(x^{(t)})x + f(x^{(t)}) - x^{(t)}f'(x^{(t)}) = y$$

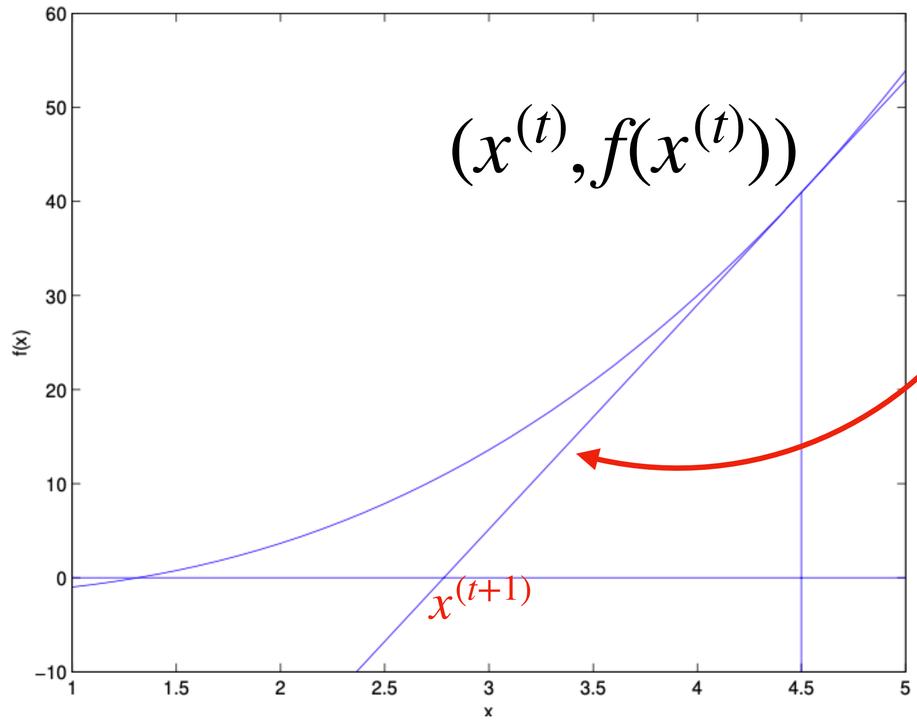




Another Optimization Method — Newton's Method

$$f'(x^{(t)})x + f(x^{(t)}) - x^{(t)}f'(x^{(t)}) = y$$





Another Optimization Method — Newton's Method

Given $f : \mathbb{R}^d \to \mathbb{R}$ find x s.t. $f(x) = 0$. $\nabla_\theta l(\theta) = 0$

- This is the update rule in 1d

$$x^{(t+1)} = x^{(t)} - \frac{f(x^{(t)})}{f'(x^{(t)})} \qquad \theta := \theta - \frac{\ell'(\theta)}{\ell''(\theta)}.$$

Another Optimization Method — Newton's Method

Given $f : \mathbb{R}^d \to \mathbb{R}$ find x s.t. $f(x) = 0$. $\nabla_\theta l(\theta) = 0$

- This is the update rule in 1d

$$x^{(t+1)} = x^{(t)} - \frac{f(x^{(t)})}{f'(x^{(t)})} \qquad \theta := \theta - \frac{\ell'(\theta)}{\ell''(\theta)}.$$

- It may converge *very* fast (quadratic local convergence!) Requires fewer iterations

Another Optimization Method — Newton's Method

Given $f : \mathbb{R}^d \to \mathbb{R}$ find x s.t. $f(x) = 0$. $\nabla_\theta l(\theta) = 0$

- This is the update rule in 1d

$$x^{(t+1)} = x^{(t)} - \frac{f(x^{(t)})}{f'(x^{(t)})} \qquad \theta := \theta - \frac{\ell'(\theta)}{\ell''(\theta)}.$$

- It may converge *very* fast (quadratic local convergence!) Requires fewer iterations
- For the likelihood, i.e., $f(\theta) = \nabla_\theta \ell(\theta)$ we need to generalize to a vector-valued function which has:

$$\theta^{(t+1)} = \theta^{(t)} - \left(H(\theta^{(t)})\right)^{-1} \nabla_\theta \ell(\theta^{(t)}).$$

in which $H_{i,j}(\theta) = \frac{\partial}{\partial\theta_i \partial\theta_j} \ell(\theta)$.

expensive

second-order gradient

Hessian

Exponential Family

Exponential Family

- Exponential family unifies inference and learning for many important models

Exponential Family

Exponential Family

Rough Idea *"If P has a a special form, then inference and learning come for free"*

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

Here y, $a(\eta)$, and $b(y)$ are scalars. $T(y)$ same dimension as η.

Exponential Family

Rough Idea *"If P has a a special form, then inference and learning come for free"*

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

η: natural parameter or canonical parameter

Here y, $a(\eta)$, and $b(y)$ are scalars. $T(y)$ same dimension as η.

Exponential Family

Rough Idea *"If P has a a special form, then inference and learning come for free"*

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

η: natural parameter or canonical parameter

Here y, $a(\eta)$, and $b(y)$ are scalars. $T(y)$ same dimension as η.

$T(y)$ is called the **sufficient statistic**.

$b(y)$ is called the **base measure** – does *not* depend on η.

$a(\eta)$ is called the **log partition function** – does *not* depend on y.

Exponential Family

Rough Idea *"If P has a a special form, then inference and learning come for free"*

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

η: natural parameter or canonical parameter

Here y, $a(\eta)$, and $b(y)$ are scalars. $T(y)$ same dimension as η.

$T(y)$ is called the **sufficient statistic**. holds all information the data provides with regard to the unknown parameter values

$b(y)$ is called the **base measure** – does *not* depend on η.

$a(\eta)$ is called the **log partition function** – does *not* depend on y.

Exponential Family

Rough Idea *"If P has a a special form, then inference and learning come for free"*

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

η: natural parameter or canonical parameter

Here y, $a(\eta)$, and $b(y)$ are scalars. $T(y)$ same dimension as η.

$T(y)$ is called the **sufficient statistic**. holds all information the data provides with regard to the unknown parameter values

$b(y)$ is called the **base measure** – does *not* depend on η.

$a(\eta)$ is called the **log partition function** – does *not* depend on y.

$$1 = \sum_y P(y;\eta) = e^{-a(\eta)} \sum_y b(y)\exp\left\{\eta^T T(y)\right\}$$

$$\implies a(\eta) = \log \sum_y b(y)\exp\left\{\eta^T T(y)\right\}$$

Example: Bernoulli

Bernoulli random variable is an event (say flipping a coin) then:

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

Example: Bernoulli

Bernoulli random variable is an event (say flipping a coin) then:

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

How do we put it in the required form?

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

Example: Bernoulli

Bernoulli random variable is an event (say flipping a coin) then:

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

How do we put it in the required form?

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

$$\begin{aligned} p(y;\phi) &= \phi^y(1-\phi)^{1-y} \\ &= \exp(y\log\phi + (1-y)\log(1-\phi)) \\ &= \exp\left(\left(\log\left(\frac{\phi}{1-\phi}\right)\right)y + \log(1-\phi)\right) \end{aligned}$$

Example: Bernoulli

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

$$\begin{aligned} p(y;\phi) &= \phi^y(1-\phi)^{1-y} \\ &= \exp(y\log\phi + (1-y)\log(1-\phi)) \\ &= \exp\left(\left(\log\left(\frac{\phi}{1-\phi}\right)\right)y + \log(1-\phi)\right) \end{aligned}$$

Example: Bernoulli

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

$$\begin{aligned} p(y;\phi) &= \phi^y(1-\phi)^{1-y} \\ &= \exp(y\log\phi + (1-y)\log(1-\phi)) \\ &= \exp\left(\left(\log\left(\frac{\phi}{1-\phi}\right)\right)y + \log(1-\phi)\right) \end{aligned}$$

So then:

$$\eta = \log\frac{\phi}{1-\phi},\ T(y) = y,\ a(\eta) = -\log(1-\phi).$$

$$b(y) = 1$$

Example: Bernoulli

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

$$\begin{aligned} p(y;\phi) &= \phi^y(1-\phi)^{1-y} \\ &= \exp(y\log\phi + (1-y)\log(1-\phi)) \\ &= \exp\left(\left(\log\left(\frac{\phi}{1-\phi}\right)\right)y + \log(1-\phi)\right) \end{aligned}$$

So then:

$$\eta = \log\frac{\phi}{1-\phi},\ T(y) = y,\ a(\eta) = -\log(1-\phi).$$

$$b(y) = 1$$

We need to show $a(\eta)$ is a function of $\log\frac{\phi}{1-\phi}$

Example: Bernoulli

Example: Bernoulli

We first observe that:

$$\eta = \log \frac{\phi}{1-\phi} \implies e^{\eta}(1-\phi) = \phi$$

$$e^{\eta} = (e^{\eta}+1)\phi \implies \phi = \frac{1}{1+e^{-\eta}}$$

Example: Bernoulli

We first observe that:

$$\eta = \log \frac{\phi}{1-\phi} \implies e^{\eta}(1-\phi) = \phi$$

$$e^{\eta} = (e^{\eta}+1)\phi \implies \phi = \frac{1}{1+e^{-\eta}}$$

Now, we plug into $\log(1-\phi)$ and we verify:

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

Example: Bernoulli

We first observe that:

$$\eta = \log \frac{\phi}{1-\phi} \implies e^{\eta}(1-\phi) = \phi$$

$$e^{\eta} = (e^{\eta}+1)\phi \implies \phi = \frac{1}{1+e^{-\eta}}$$

Now, we plug into $\log(1-\phi)$ and we verify:

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

We have verified Bernoulli distribution is in the exponential family

Example: Gaussian with Fixed Variance $\sigma^2 = 1$

Example: Gaussian with Fixed Variance $\sigma^2 = 1$

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

Can we put it in the exponential family form?

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

Example: Gaussian with Fixed Variance $\sigma^2 = 1$

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

Can we put it in the exponential family form?

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

Multiply out the square and group terms:

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

Example: Gaussian with Fixed Variance $\sigma^2 = 1$

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

Can we put it in the exponential family form?

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

Multiply out the square and group terms:

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

$$\eta = \mu, T(y) = y, a(\eta) = \frac{1}{2}\eta^2.$$

Example: Gaussian with Fixed Variance $\sigma^2 = 1$

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

Can we put it in the exponential family form?

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

Multiply out the square and group terms:

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

$$\eta = \mu,\ T(y) = y,\ a(\eta) = \frac{1}{2}\eta^2.$$

In all the exponential family distribution we work with in the course, T(y) = y

An Observation

An Observation

Notice that for a Gaussian with mean μ we had

$$\eta = \mu, T(y) = y, a(\eta) = \frac{1}{2}\eta^2.$$

An Observation

Notice that for a Gaussian with mean μ we had

$$\eta = \mu, T(y) = y, a(\eta) = \frac{1}{2}\eta^2.$$

$$\partial_\eta a(\eta) = \eta = \mu = \mathbb{E}[y] \text{ and } \partial_\eta^2 a(\eta) = 1 = \sigma^2 = \text{var}(y)$$

An Observation

Notice that for a Gaussian with mean μ we had

$$\eta = \mu, T(y) = y, a(\eta) = \frac{1}{2}\eta^2.$$

$$\partial_\eta a(\eta) = \eta = \mu = \mathbb{E}[y] \text{ and } \partial_\eta^2 a(\eta) = 1 = \sigma^2 = \text{var}(y)$$

Is this true for general?

Log Partition Function

Yes! Recall that

$$a(\eta) = \log \sum_{y} b(y) \exp \left\{ \eta^T T(y) \right\}$$

Log Partition Function

Yes! Recall that

$$a(\eta) = \log \sum_y b(y) \exp\left\{\eta^T T(y)\right\}$$

Then, taking derivatives

$$\nabla_\eta a(\eta) = \frac{\sum_y T(y) b(y) \exp\left\{\eta^T T(y)\right\}}{\sum_y b(y) \exp\left\{\eta^T T(y)\right\}} = \mathbb{E}[T(y); \eta]$$

Many Other Exponential Models

- There are many canonical exponential family models:
  - Binary $\mapsto$ Bernoulli
  - Multiple Classses $\mapsto$ Multinomial
  - Real $\mapsto$ Gaussian
  - Counts $\mapsto$ Poisson
  - $\mathbb{R}_+ \mapsto$ Gamma, Exponential
  - Distributions $\mapsto$ Dirichlet

Thank You!
Q & A