



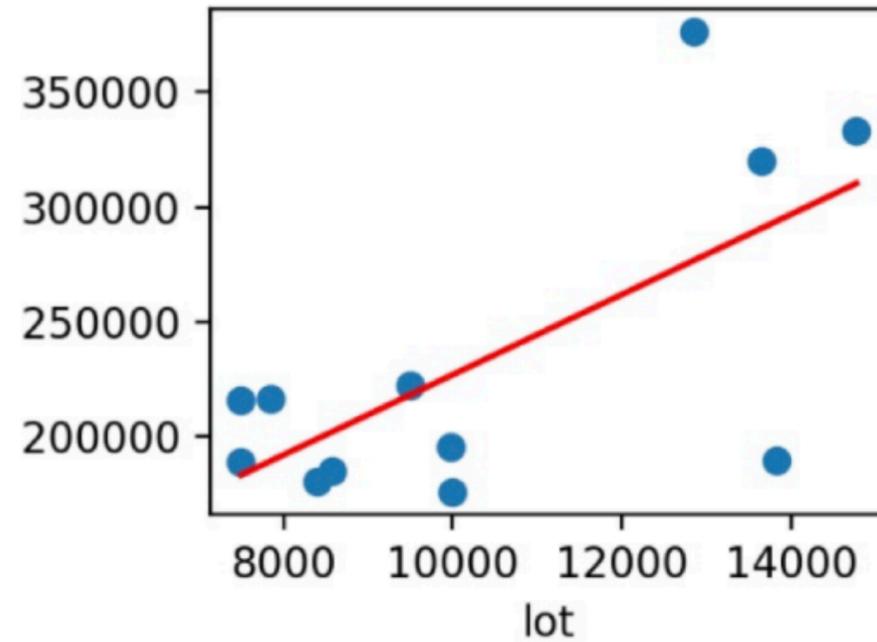
香港科技大學  
THE HONG KONG  
UNIVERSITY OF SCIENCE  
AND TECHNOLOGY

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

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

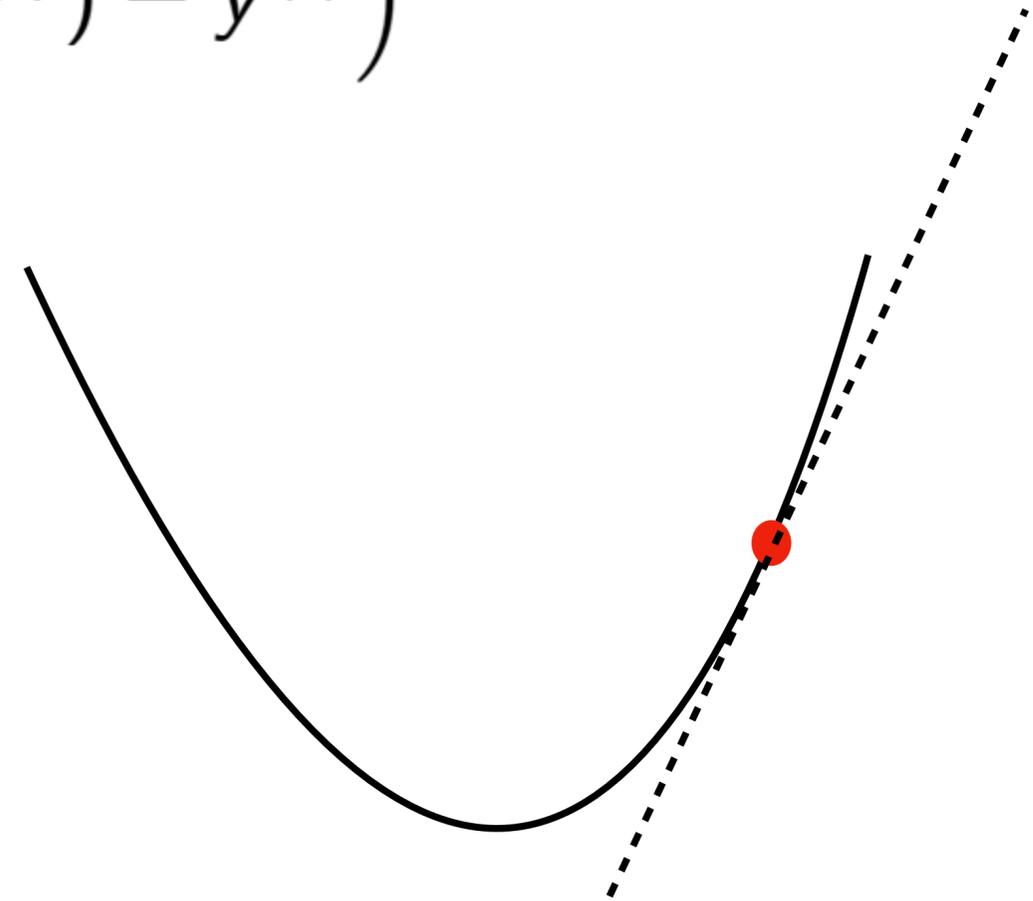
# Gradient Descent

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

Learning Rate

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

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



The direction of the steepest decrease of  $J$

# Gradient Descent

For a single training example:

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

LMS (Least Mean Square) Update Rule

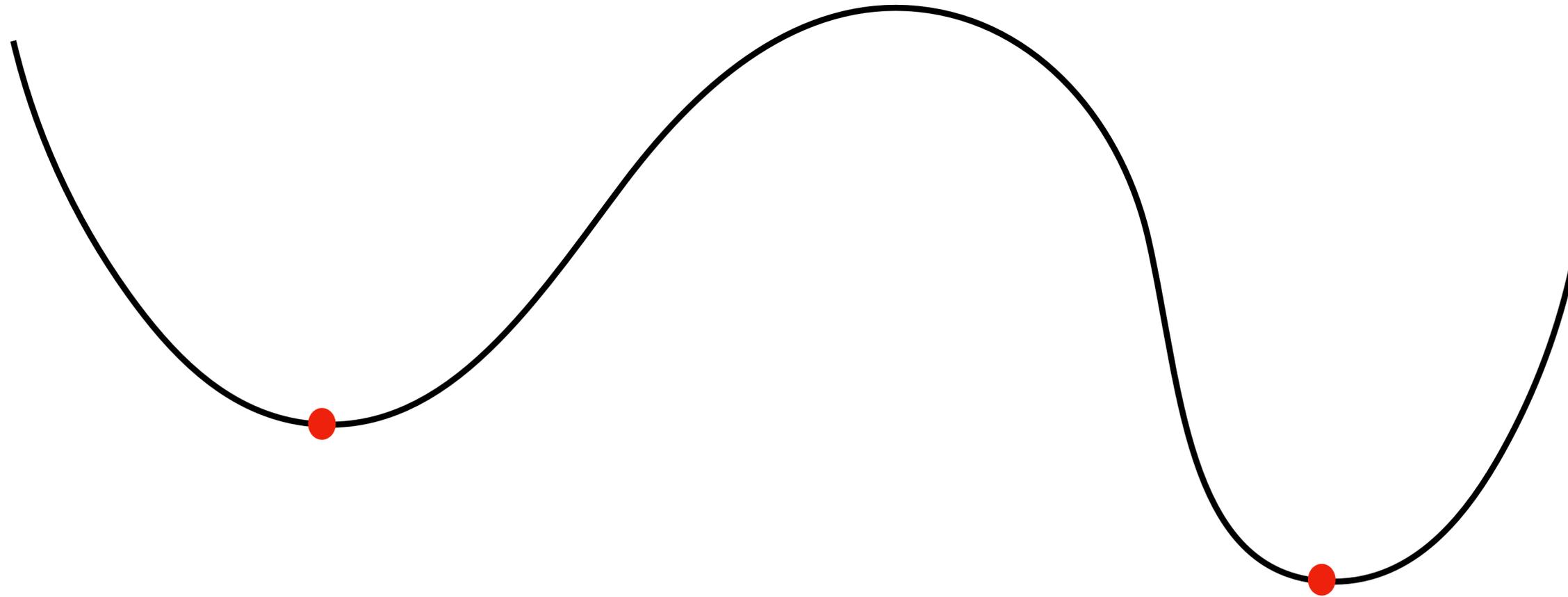
# Batch Gradient Descent

For a multiple training examples:

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

Repeat until convergence

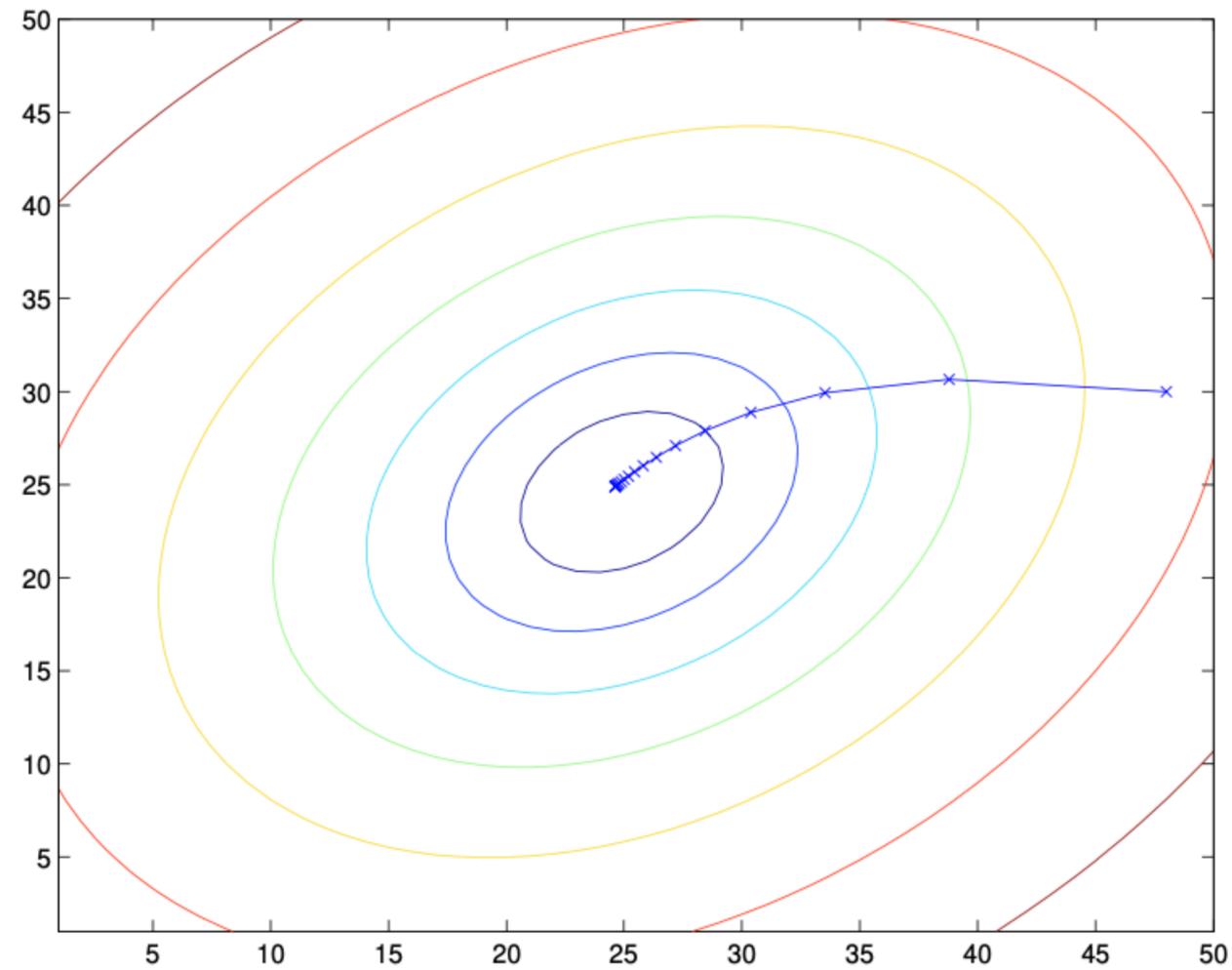
# Local Minimum



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

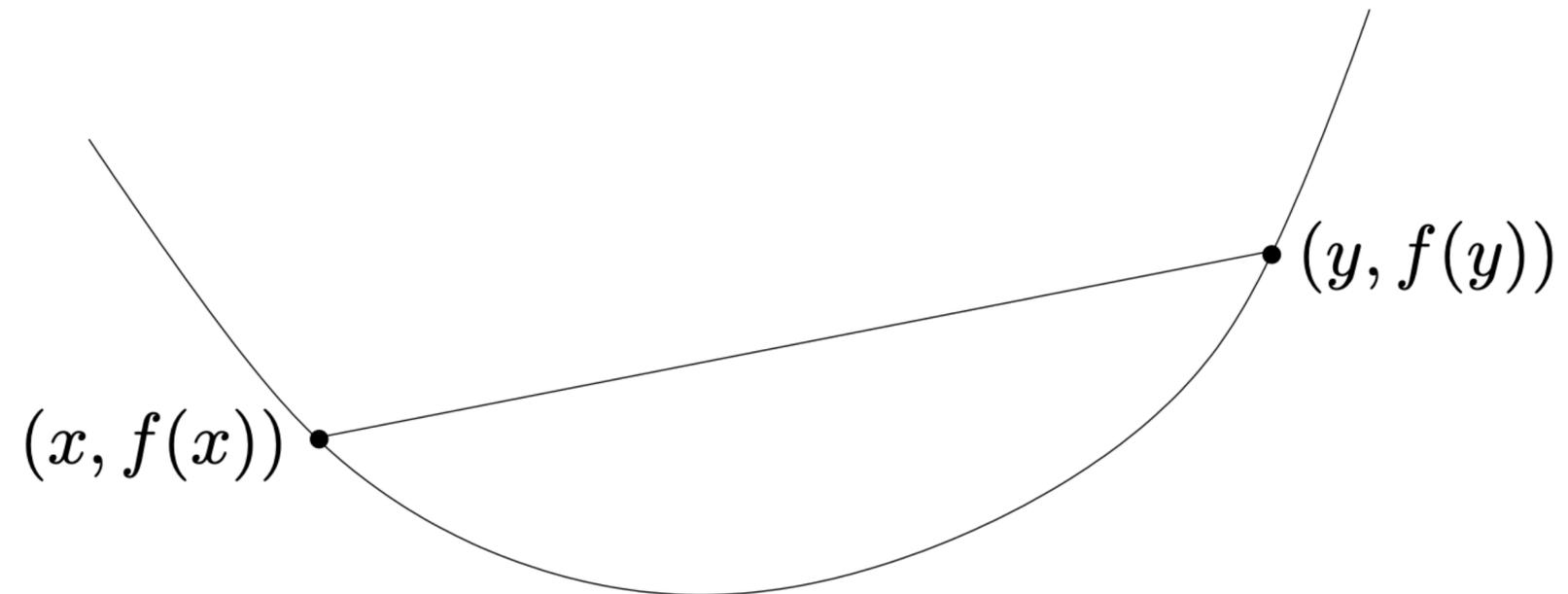
# $J$ is a convex quadratic function

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

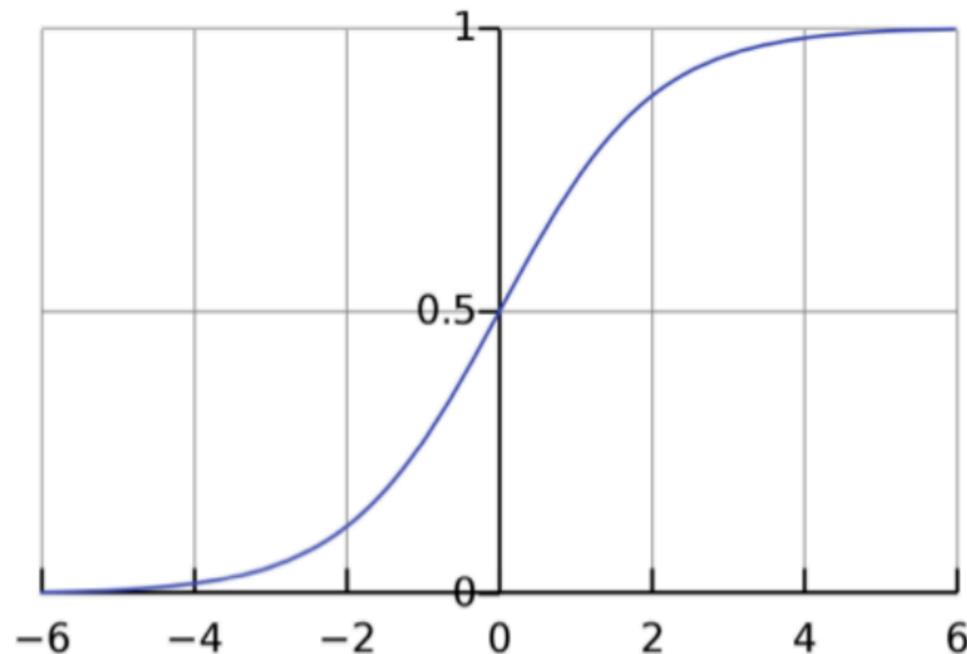
Labels are discrete

# Logistic Regression

Given a training set  $\{(x^{(i)}, y^{(i)}) \text{ for } i = 1, \dots, n\}$  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 Function

$$g(z) = \frac{1}{1 + e^{-z}} \cdot \text{Sigmoid Function}$$

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:

$$P(y = 1 \mid x; \theta) = h_{\theta}(x)$$

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

Then,

$$\begin{aligned} L(\theta) = P(y \mid X; \theta) &= \prod_{i=1}^n p(y^{(i)} \mid x^{(i)}; \theta) \quad \text{We want to express "if-then" logics, how?} \\ &= \prod_{i=1}^n h_{\theta}(x^{(i)})^{y^{(i)}} (1 - h_{\theta}(x^{(i)}))^{1-y^{(i)}} \end{aligned}$$

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

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

# Derivative of Logistic Function

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

# Gradient Descent

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

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

Looks identical to LMS update rule in linear regression

Is this coincidence?

# Multi-Label Classification



{Cat, dog, dragon, fish, pig}

# Multi-Label Classification

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

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

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

$$\phi_i = \theta_i^T x ?$$

# Softmax Function

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

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

The denominator is a normalization constant

# Multi-Label Classification

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

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

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

# Multi-Label Classification

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

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

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

$$\ell_{\text{ce}}((t_1, \dots, t_k), y) = -\log \left( \frac{\exp(t_y)}{\sum_{j=1}^k \exp(t_j)} \right) \quad \ell(\theta) = \sum_{i=1}^n \ell_{\text{ce}}((\theta_1^\top x^{(i)}, \dots, \theta_k^\top x^{(i)}), y^{(i)})$$

# The Derivative

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

## Chain rule

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

$$\frac{\partial \ell(\theta)}{\partial \theta_i} = \sum_{j=1}^n (\phi_i^{(j)} - 1\{y^{(j)} = i\}) \cdot x^{(j)} \quad \text{Intuitive explanation of the rule?}$$

# Another Optimization Method — Newton's Method

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

- ▶ This is the update rule in 1d

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



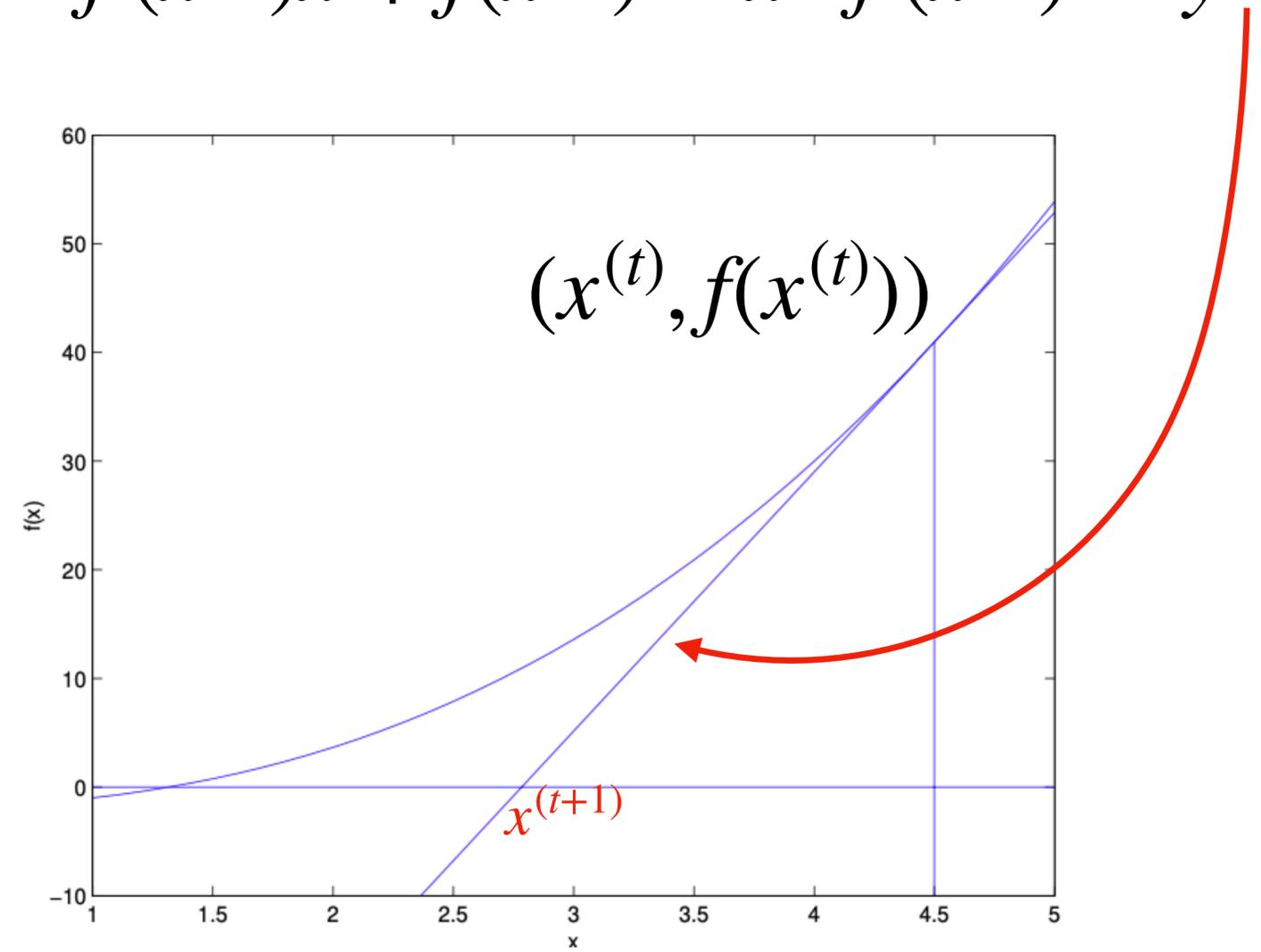
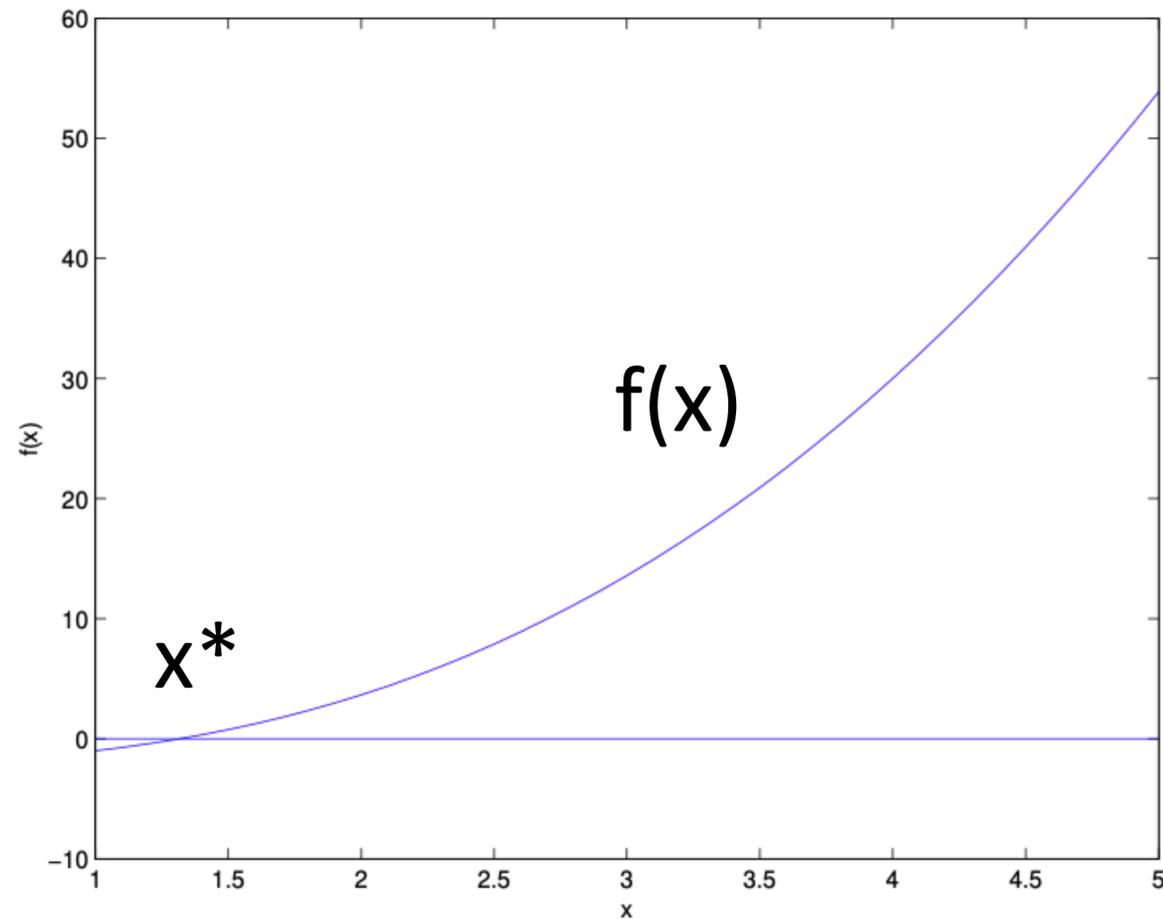
Solution to a linear equation

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

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

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

- ▶ This is the update rule in 1d

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

- ▶ 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  $H_{i,j}(\theta) = \frac{\partial^2}{\partial \theta_i \partial \theta_j} \ell(\theta)$ .

# Exponential Family

# Exponential Family

- Exponential family unifies inference and learning for many important models

# Exponential Family

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

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

$\eta$ : natural parameter or canonical parameter

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

$T(y)$  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  $\eta$ .

$a(\eta)$  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:

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

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

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

So then:

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

$$b(y) = 1$$

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

# Example: Bernoulli

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 $\sigma^2 = 1$

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

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

Notice that for a Gaussian with mean  $\mu$  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

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

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 Classes  $\mapsto$  Multinomial
  - ▶ Real  $\mapsto$  Gaussian
  - ▶ Counts  $\mapsto$  Poisson
  - ▶  $\mathbb{R}_+$   $\mapsto$  Gamma, Exponential
  - ▶ Distributions  $\mapsto$  Dirichlet

**Thank You!**  
**Q & A**