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
Sep 12, 2024

Classification

CAT

Labels are discrete

Logistic Regression

Logistic Regression

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

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

Logistic Regression

Given a training set $\{(x^{(i)}, y^{(i)})$ 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

Logistic Regression

Given a training set $\{(x^{(i)}, y^{(i)})$ 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 Regression

Given a training set $\{(x^{(i)}, y^{(i)})$ 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....

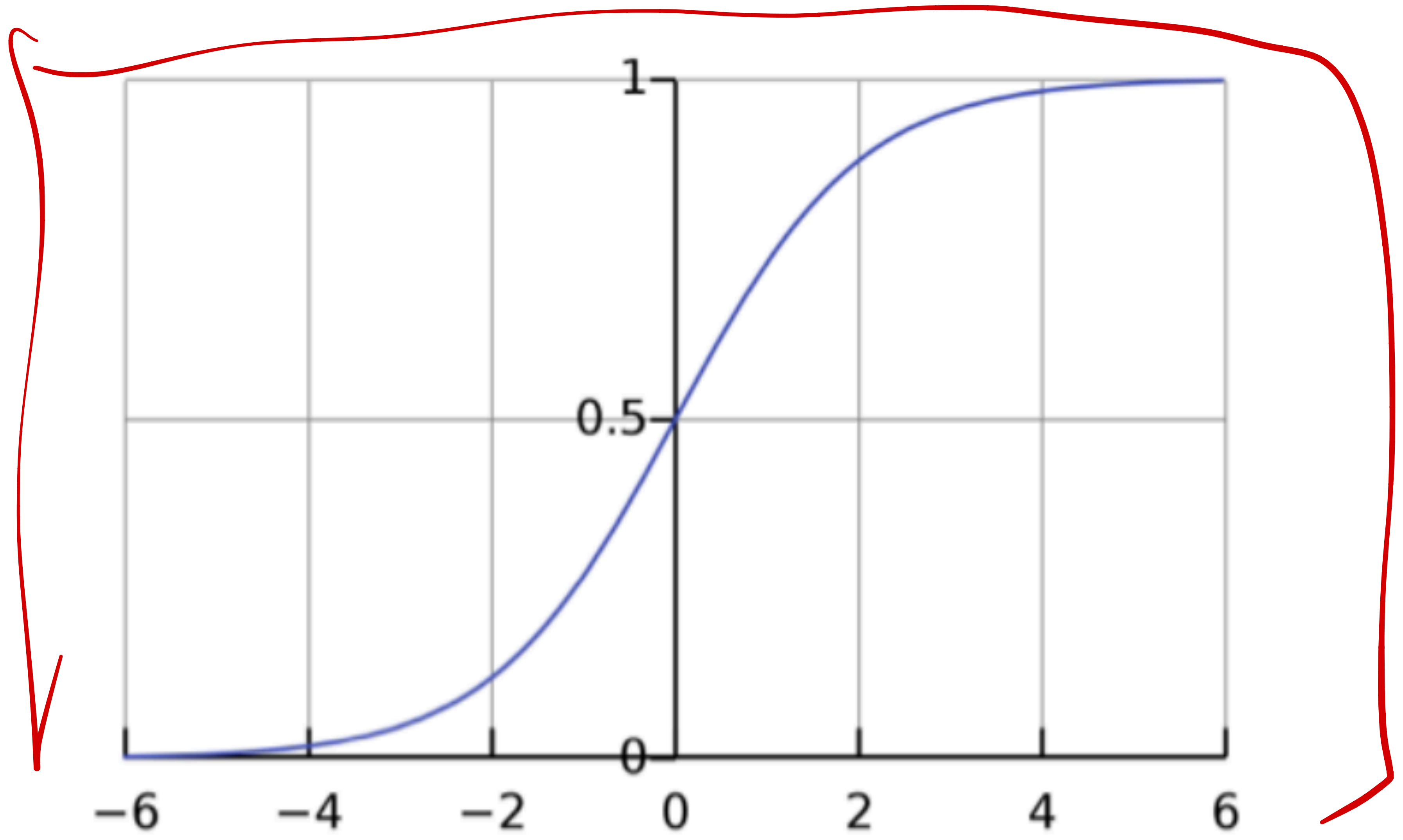
$$g(z) = \frac{1}{1 + e^{-z}}.$$

Logistic 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....



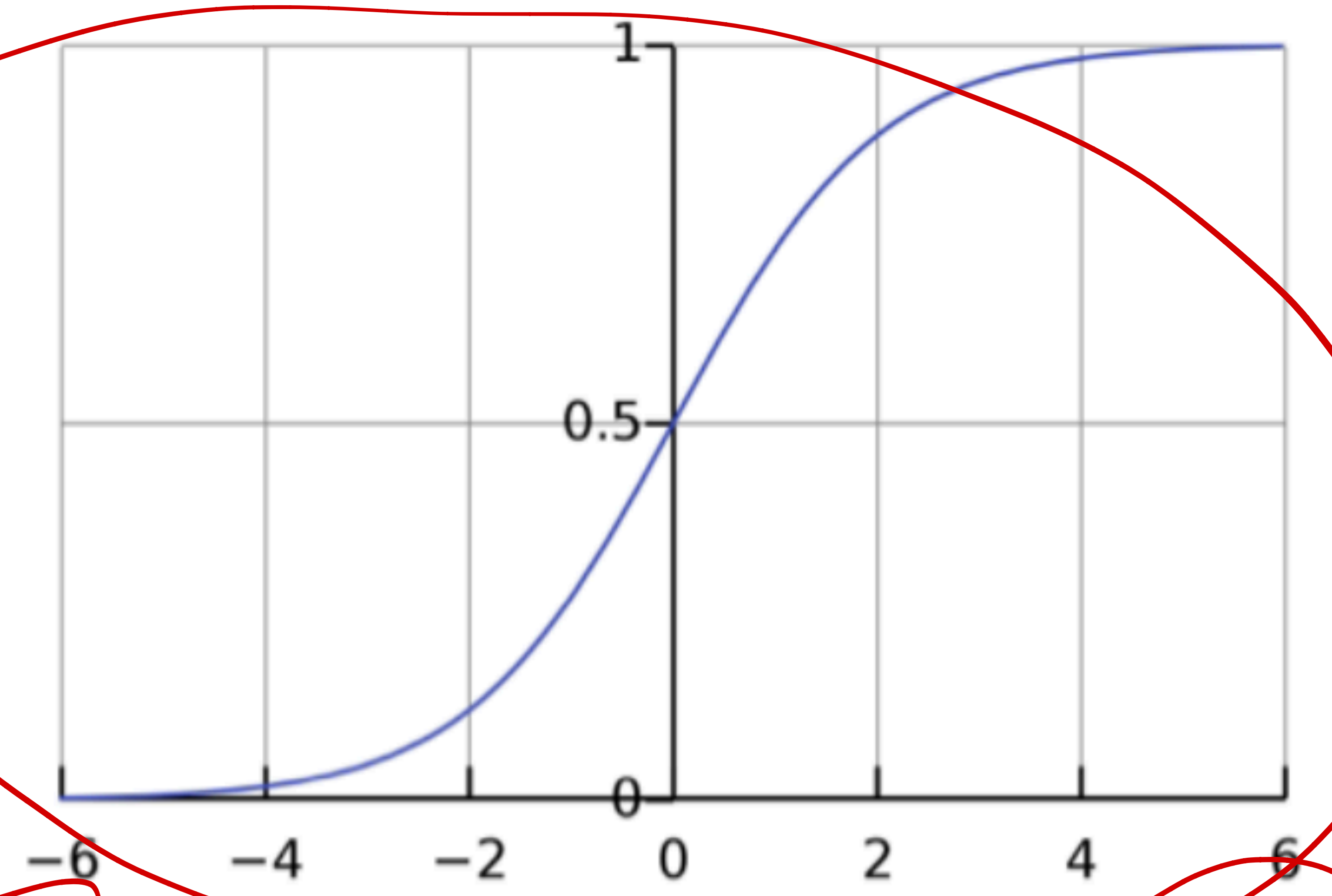
$$g(z) = \frac{1}{1 + e^{-z}}.$$

Logistic Regression

Given a training set $\{(x^{(i)}, y^{(i)})$ 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 $h_\theta(x)$?

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

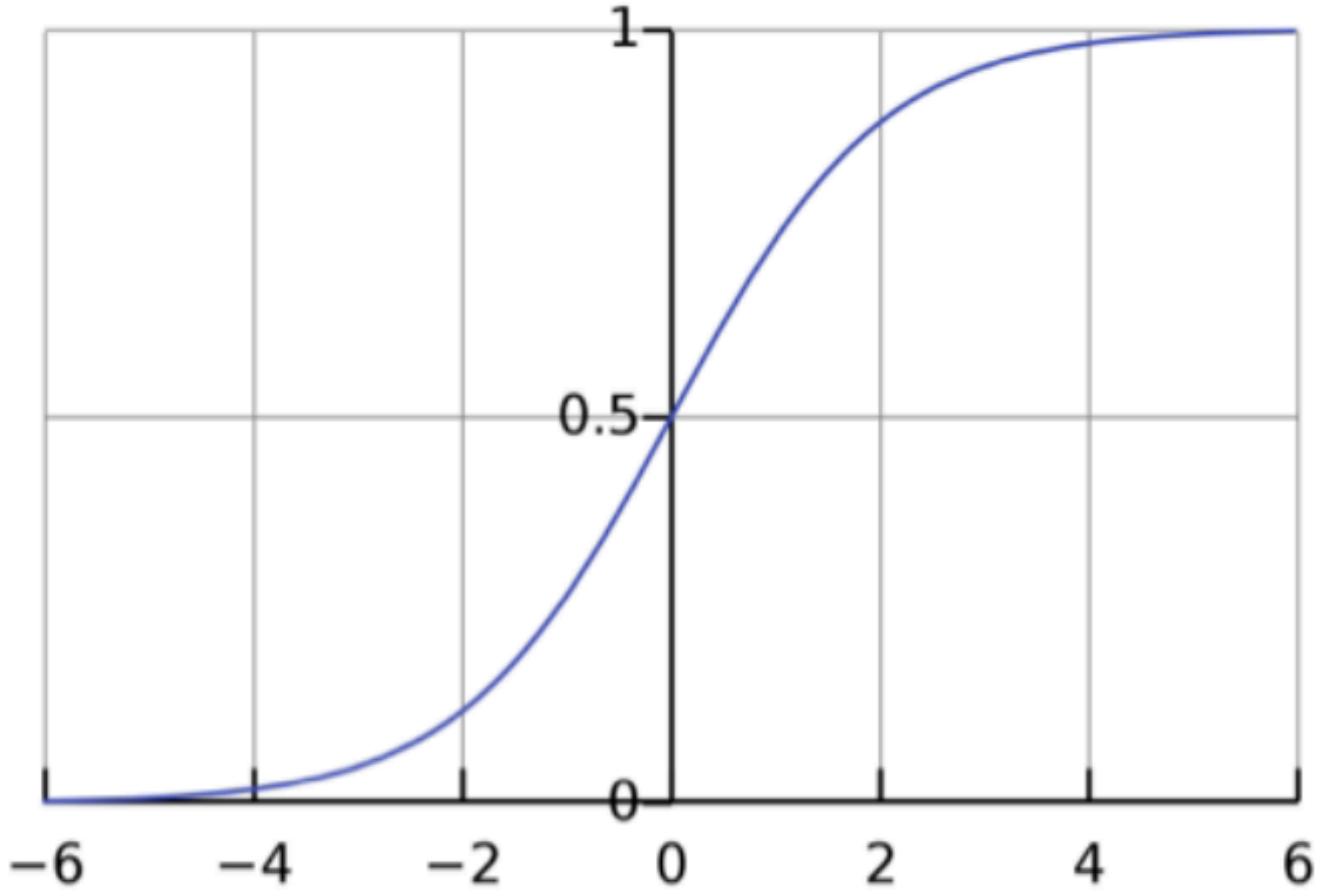
Logistic Regression

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 $h_\theta(x)$?

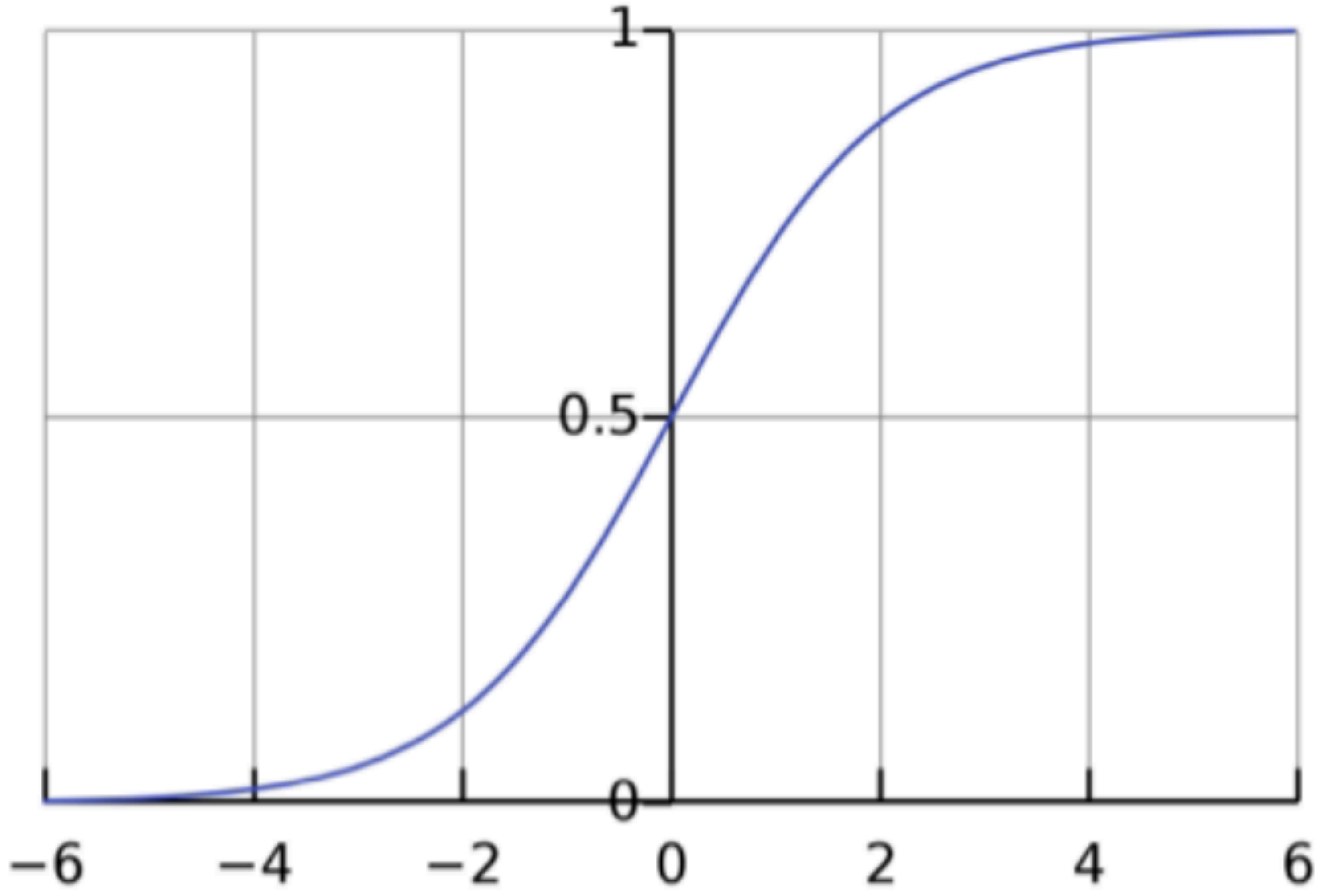
$$P(y = 1 \mid x; \theta) = h_\theta(x)$$
$$P(y = 0 \mid x; \theta) = 1 - h_\theta(x)$$

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:

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

Least Mean Square

Maximum likelihood estimation

Logistic Regression

Let's write the Likelihood function. Recall:

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

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

Then,

IID

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

Maximum likelihood estimation

Logistic Regression

Let's write the Likelihood function. Recall:

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

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

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,

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

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

y=1 y=0

$$= \prod_{i=1}^{n} h_\theta(x^{(i)})^{y^{(i)}} (1 - h_\theta(x^{(i)}))^{1-y^{(i)}}$$

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,

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

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

$$= \prod_{i=1}^{n} h_\theta(x^{(i)})^{y^{(i)}} (1 - h_\theta(x^{(i)}))^{1-y^{(i)}}$$

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

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

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

Maximum likelihood estimation

Derivative of Logistic Function

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

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

$g(z)(1-g(z))$ $\frac{1}{g(z)}$

$$l(\theta) = y \log h_\theta(x) + (1-y)\log[1-h_\theta(x)]$$

$h_\theta(x)$ logistic function

$= g(\theta^T x)$

$$\frac{\partial}{\partial \theta_j} l(\theta) = y \cdot \frac{1}{g(\theta^T x)} \frac{\partial}{\partial \theta_j} g(\theta^T x)$$

$$+ (1-y)\frac{1}{1-g(\theta^T x)}\left(-\frac{\partial g(\theta^T x)}{\partial \theta_j}\right)$$

$g'(z) = g(z)(1-g(z))$

$$= \left(y \cdot \frac{1}{g(\theta^T x)} - (1-y)\frac{1}{1-g(\theta^T x)}\right] \frac{\partial}{\partial \theta_j} g(\theta^T x)$$

$$\frac{\partial \theta^T x}{\partial \theta_j} = x_j$$

$$= \left(y \cdot \frac{1}{g(\theta^T x)} - (1-y)\frac{1}{1-g(\theta^T x)}\right] g(\theta^T x)(1-g(\theta^T x)]\, x_j$$

$$= (y(1-g(\theta^T x)] - (1-y)g(\theta^T x)]\, x_j = [y - g(\theta^T x)]\, x_j$$

Gradient Descent

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

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

Gradient Descent

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

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

Looks identical to LMS update rule in linear regression

Gradient Descent

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

GLMs

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

Looks identical to LMS update rule in linear regression

Is this coincidence?

Multi-Label Classification

{Cat, dog, dragon, fish, pig}

Language Models

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 $p(y \mid x; \theta)$

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

ϕ_k

$\phi_k \in [0,1]$

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

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

$\phi_i = \theta_i^T x$? $\notin [0, 1]$

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

$k^* = \arg\max \{t_1, t_2, \ldots, t_k\}$

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

logit

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

$\sum \phi_i = 1$

exp() is monotonic

$t_i > t_j, \ \phi_i > \phi_j$

The denominator is a normalization constant

$\exp(t_i) > 0 \quad 0 < \phi_i \leq 1$

Multi-Label Classification

Multi-Label Classification

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

$NN(x) = (t_1 \cdots t_k)$

$$\begin{bmatrix} \vec{\theta}_1^\top \\ \vec{\theta}_2^\top \\ \vdots \\ \vec{\theta}_k^\top \end{bmatrix} \qquad \vec{\theta}$$

Multi-Label Classification

CNN LSTM transformer

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

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

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

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

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

n # data samples

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

$t_1 \ldots t_k, y$

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

The Derivative

The Derivative

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

g(z)

$$\mathbb{1}(e) = \begin{cases} 1 & e \text{ is true} \\ 0 & e \text{ is false} \end{cases}$$

$$l_{ce}([t_1 \cdots t_k], y) = -\log \left(\frac{\exp(t_y)}{\sum_{j=1}^{k} \exp(t_j)} \right)$$

$$\phi_i = \frac{\exp(t_i)}{\sum \cdots}$$

$$\frac{\partial l_{ce}}{\partial t_i} = \frac{1}{-\phi_y} \cdot \frac{\partial}{\partial t_i} \frac{\exp(t_y)}{\sum_{j=1}^{k} \exp(t_j)}$$

$$= \frac{1}{-\phi_y} \left[\frac{\partial}{\partial t_i} \exp(t_y) \times \frac{1}{\sum_{j=1}^{k} \exp(t_j)} + \exp(t_y) \cdot \frac{\partial}{\partial t_i} \left[\frac{1}{\sum_{j=1}^{k} \exp(t_j)} \right] \right]$$

if - then $y = i$

$$= -\frac{1}{\phi_y} \left[\exp(t_y) \cdot \frac{-\exp(t_i)}{\left(\sum_{j=1}^{k} \exp(t_j)\right)^2} + \begin{cases} \frac{\exp(t_i)}{\sum_{j=1}^{k} \exp(t_j)} & i = y \\ 0 & \text{if } y \end{cases} \right]$$

$$-\phi_y \cdot \phi_i \quad + \begin{cases} \phi_i & i = y \\ 0 & i \neq y \end{cases}$$

$$= \begin{cases} \phi_i - 1 & i = y \\ \phi_i & i \neq y \end{cases}$$

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

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

Chain rule

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

$t_i = \theta_i^T x$

back propagation

θ_i is only related to t_i

$t_i = \frac{\theta_i^T x}{\sum \theta_i^T x}$

$$\frac{\partial l_{ce}(t_1 \cdots t_k)}{\partial \theta_i} = \frac{\partial l_{ce}}{\partial t_i} \cdot \frac{\partial t_i}{\theta_i} \qquad t_i = f(\theta_i)$$

For every j, $t_j = g_j(\theta_i)$

$$\frac{\partial l_{ce}}{\partial \theta_i} = \sum_j \frac{\partial l_{ce}}{\partial t_j} \cdot \frac{\partial t_j}{\partial \theta_i}$$

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

GLM

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

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

$$\theta_i \leftarrow \theta_i + \alpha \sum_{j=1}^{n} (1\{y^{(j)} = i\} - \phi_i^{(j)}) \cdot x^{(j)}$$

$$\theta_i \leftarrow \theta_i + \alpha \sum_{j=1}^{n} (y^{(j)} - h_\theta(x^{(j)})) x_i$$

The Derivative

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

Chain rule

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

$$\theta^{new} = \theta^{old} - \alpha \cdot \text{coefficient} \cdot \vec{x}$$

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

Intuitive explanation of the rule?

$\vec{\theta}_i$ to label i

$y^{(j)} = i \quad (\phi_i^{(j)} - 1(y^{(j)} = i)) < 0$

$y^{(j)} \neq i \quad$ coeffi... ≥ 0

$$\theta_i^{new} = \theta_i^{old} - [\phi_i^{old} - 1(y=i)]x$$

$$t_i = \theta_i^T x$$

$$\begin{cases} t_i^{new} = \theta_i^{old^T} x - [\phi_i^{old} - 1(y=i)] x^T x \\ t_i^{old} = \theta_i^{old^T} x \end{cases}$$

$x^T x > 0$

$[\phi_i^{old} - 1(y=i)]$:

$$\begin{cases} y=i & <0 \\ y \neq i & >0 \end{cases}$$

$$t_i^{new} > t_i^{old}$$

$$\begin{cases} t_i^{new} > t_i^{old} & y=i \\ t_i^{new} < t_i^{old} & y \neq i \end{cases}$$

Another Optimization Method — Newton's Method

Another Optimization Method — Newton's Method

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

Another Optimization Method — Newton's Method

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

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

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

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

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

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

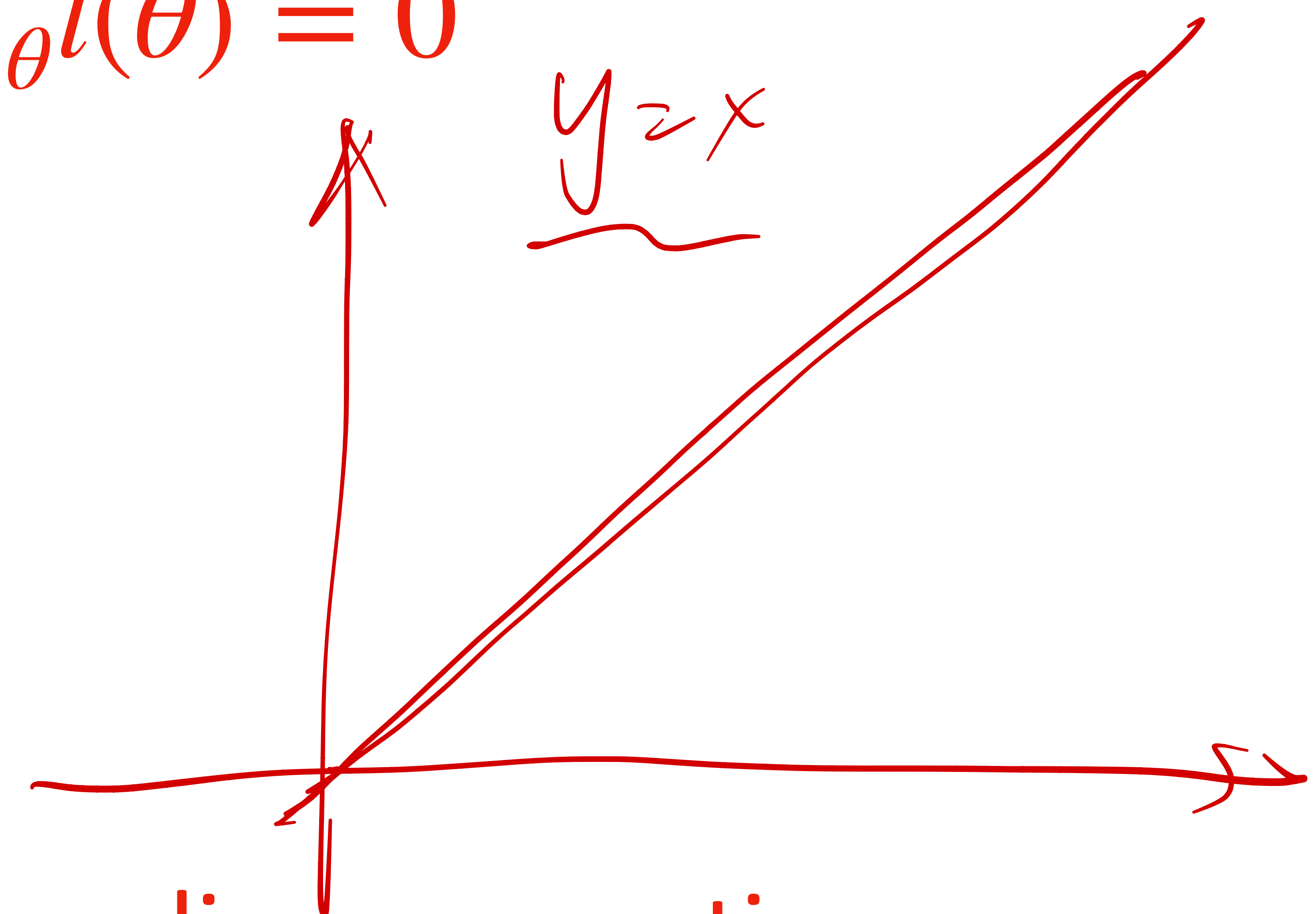
$\downarrow$ Solution to a linear equation

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

Another Optimization Method — Newton's Method

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

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



- This is the update rule in 1d

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

Solution to a linear equation

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

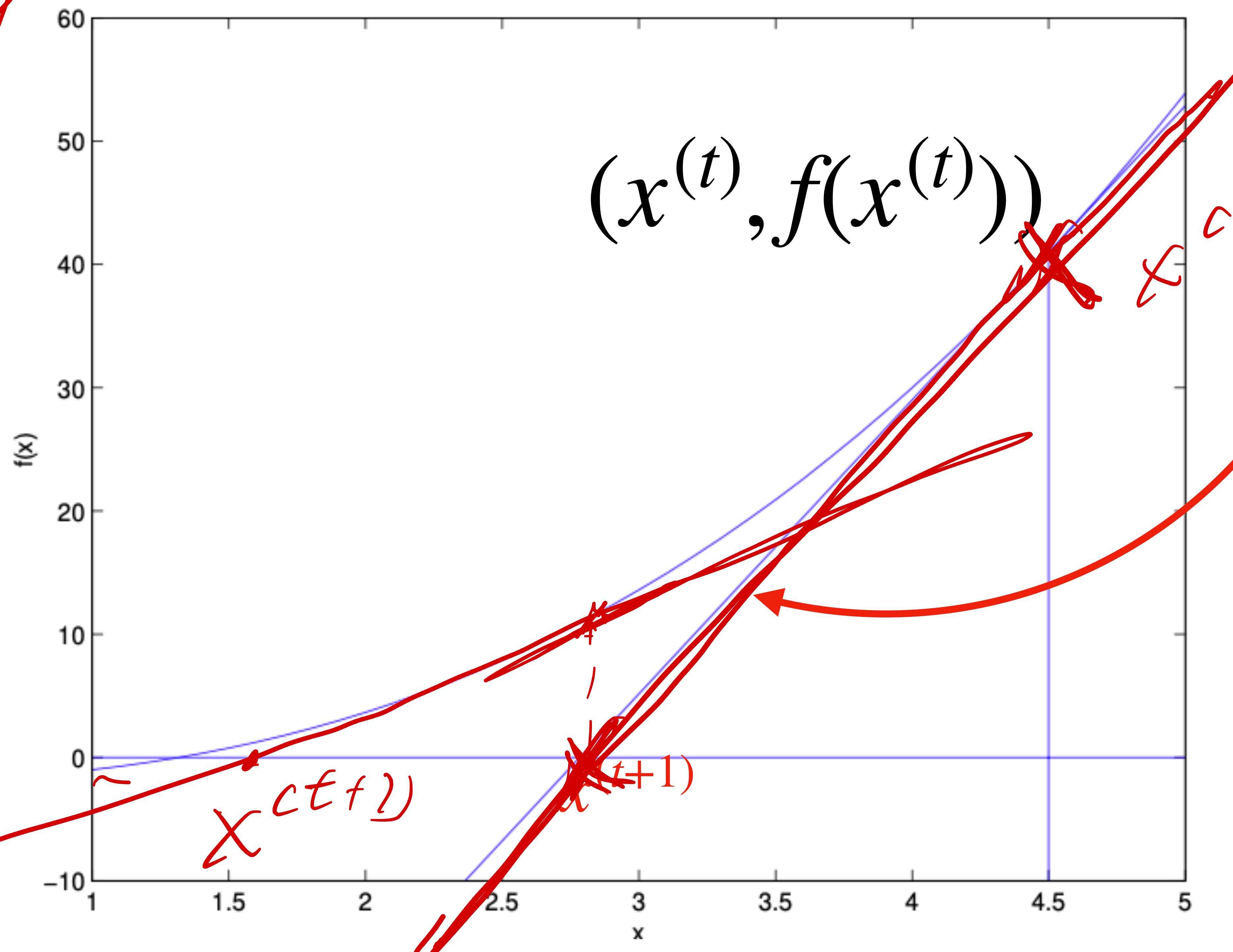
$y = 0$

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

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

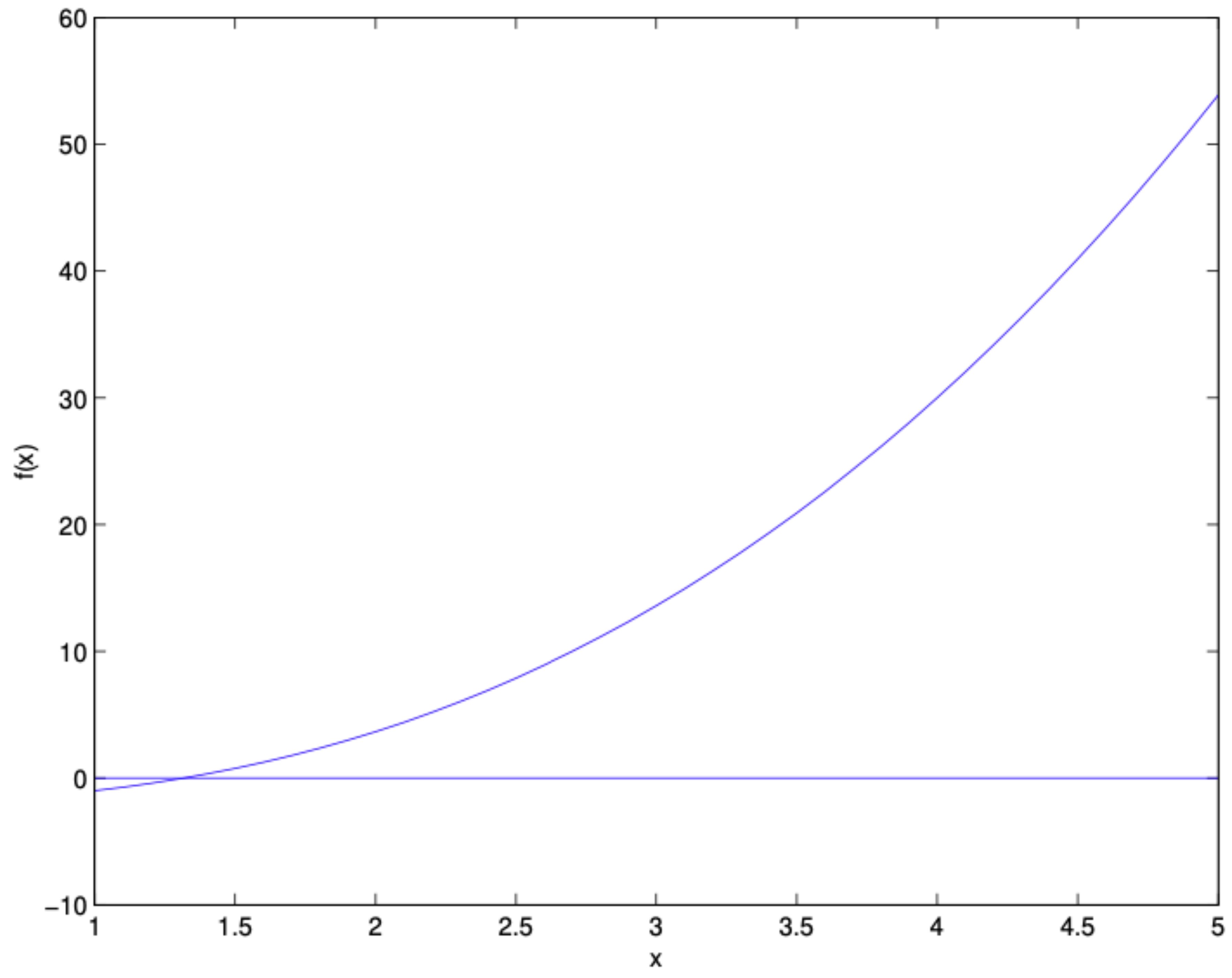
Another Optimization Method — Newton's Method

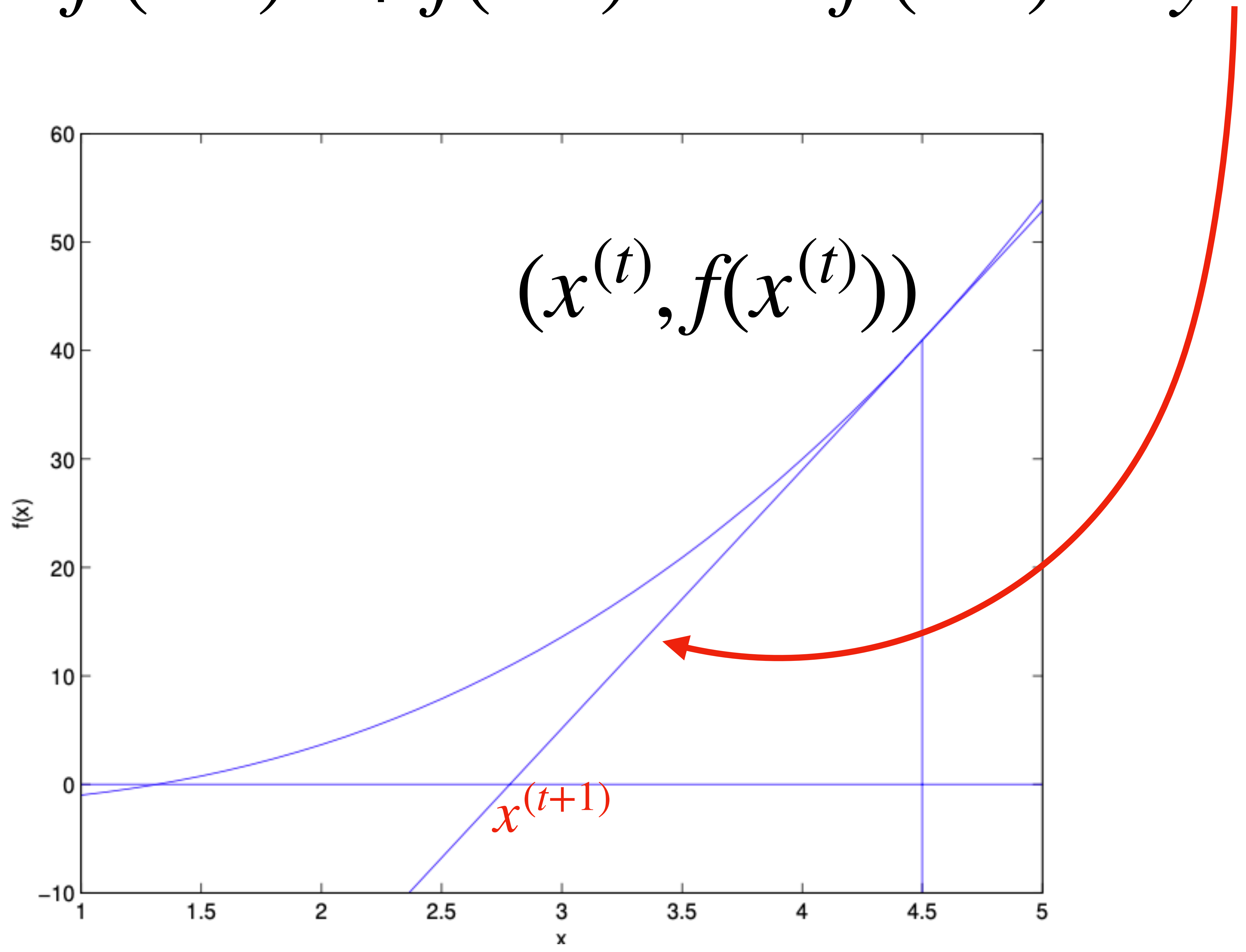
$$f'(x^{(t)})x + f(x^{(t)}) - x^{(t)}f'(x^{(t)}) = y$$



Another Optimization Method — Newton's Method

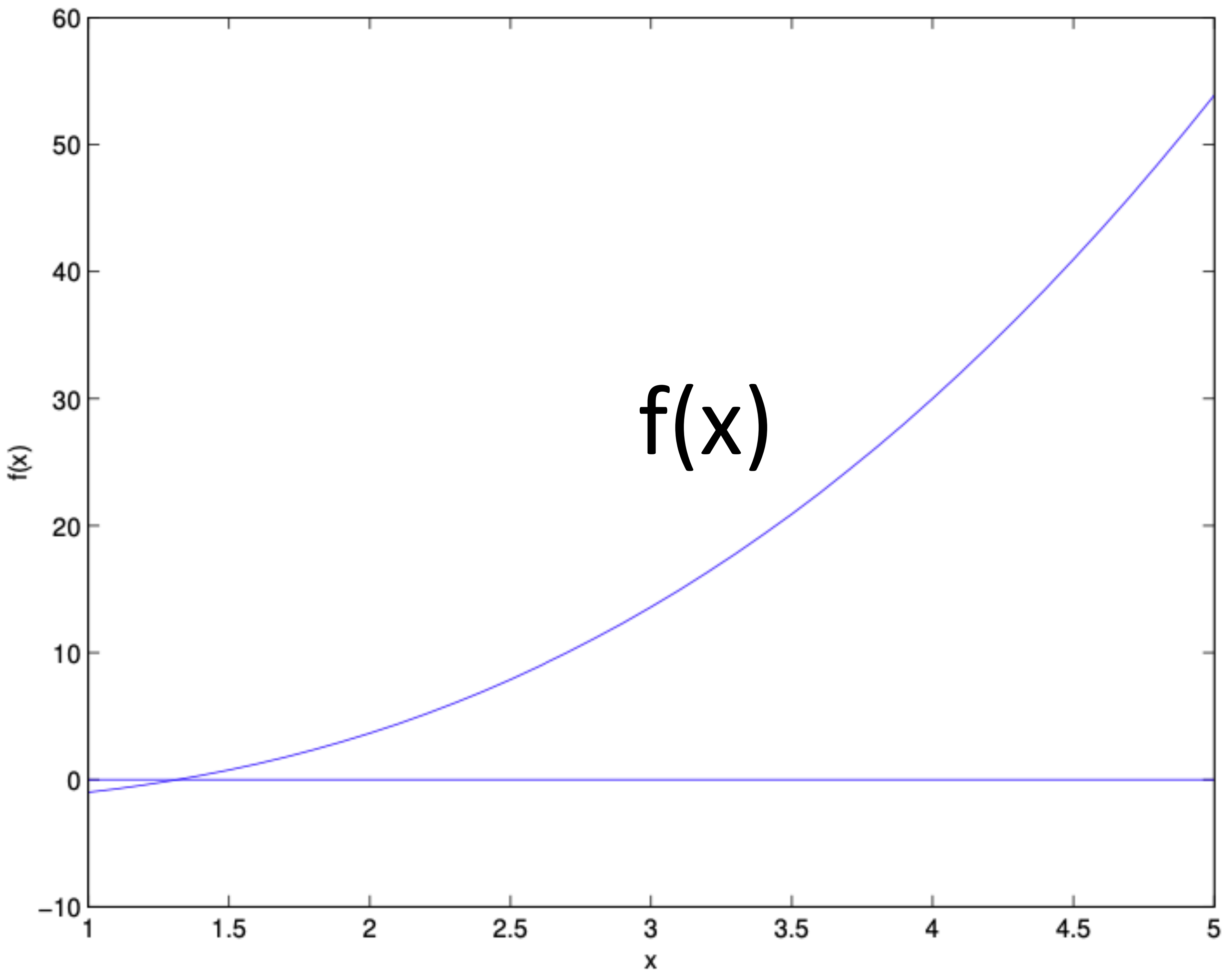
$$f'(x^{(t)})x + f(x^{(t)}) - x^{(t)}f'(x^{(t)}) = y$$

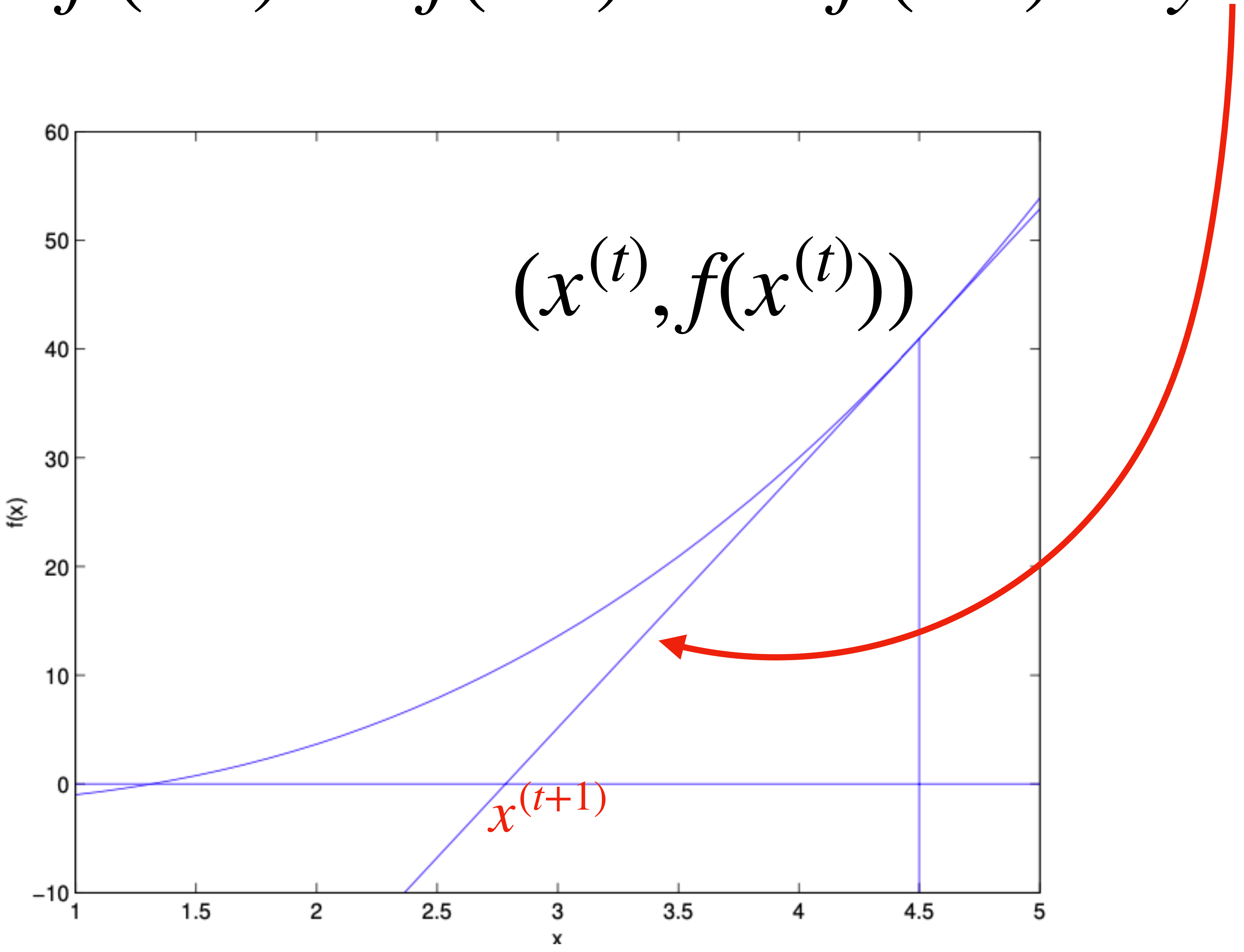




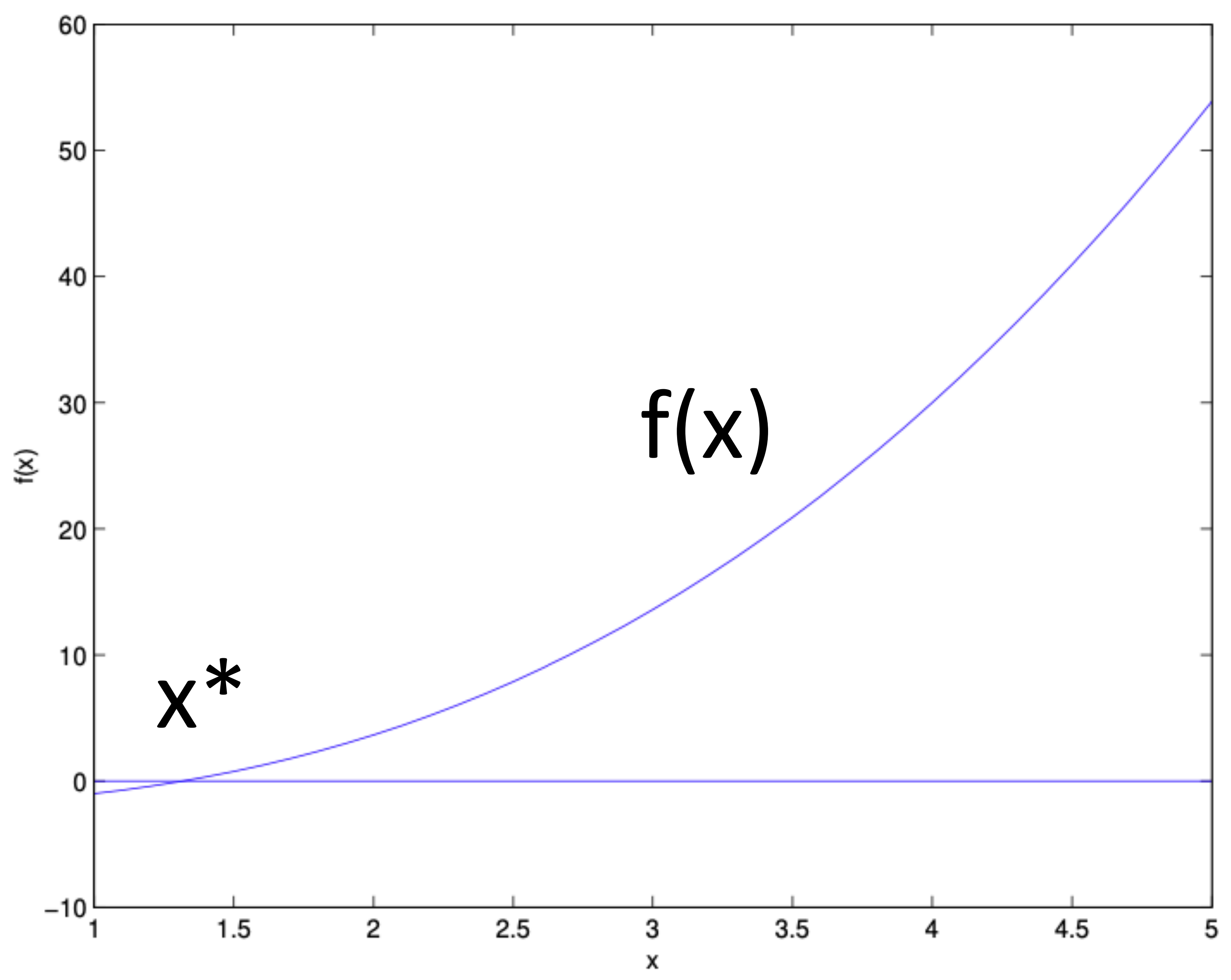
Another Optimization Method — Newton's Method

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

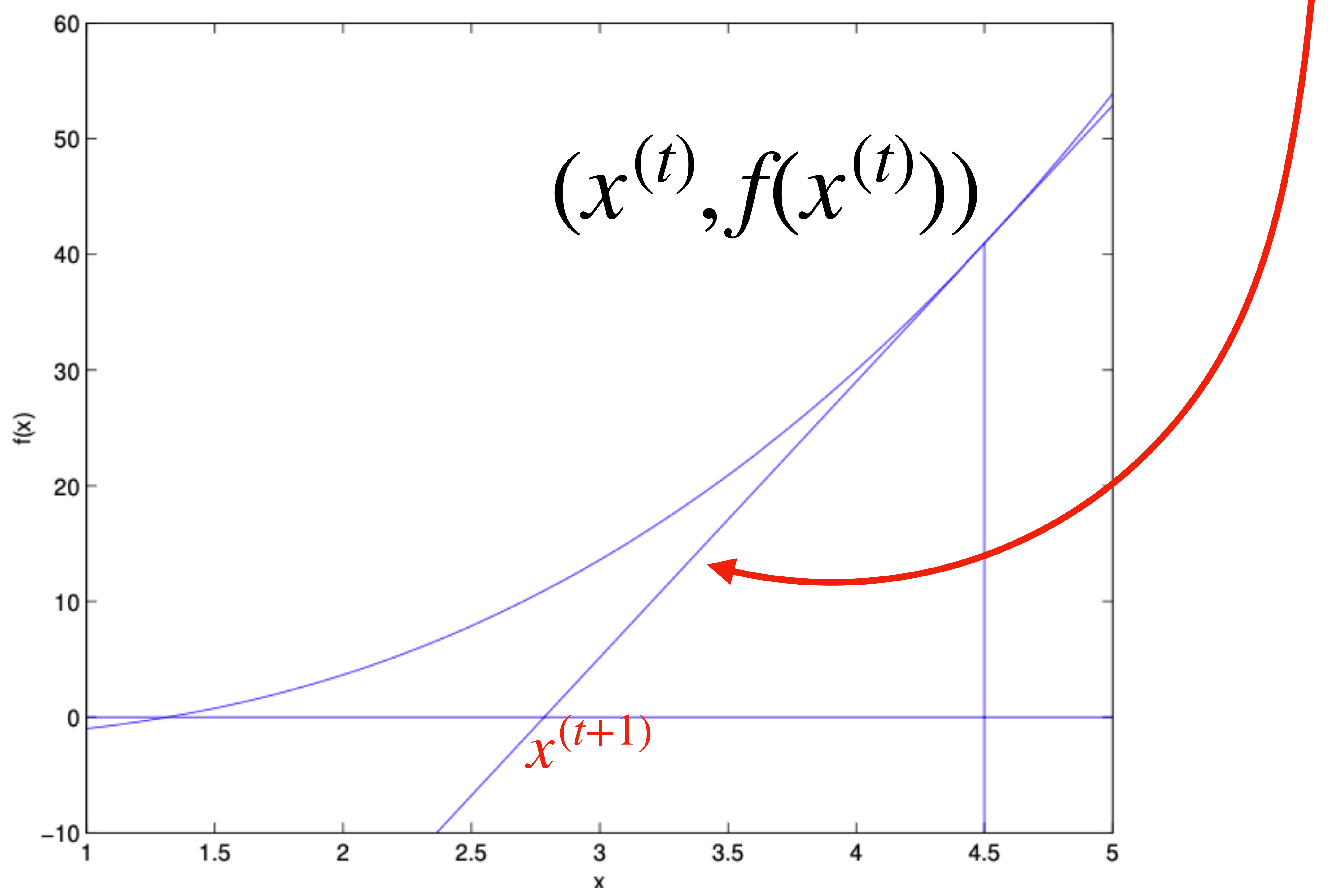


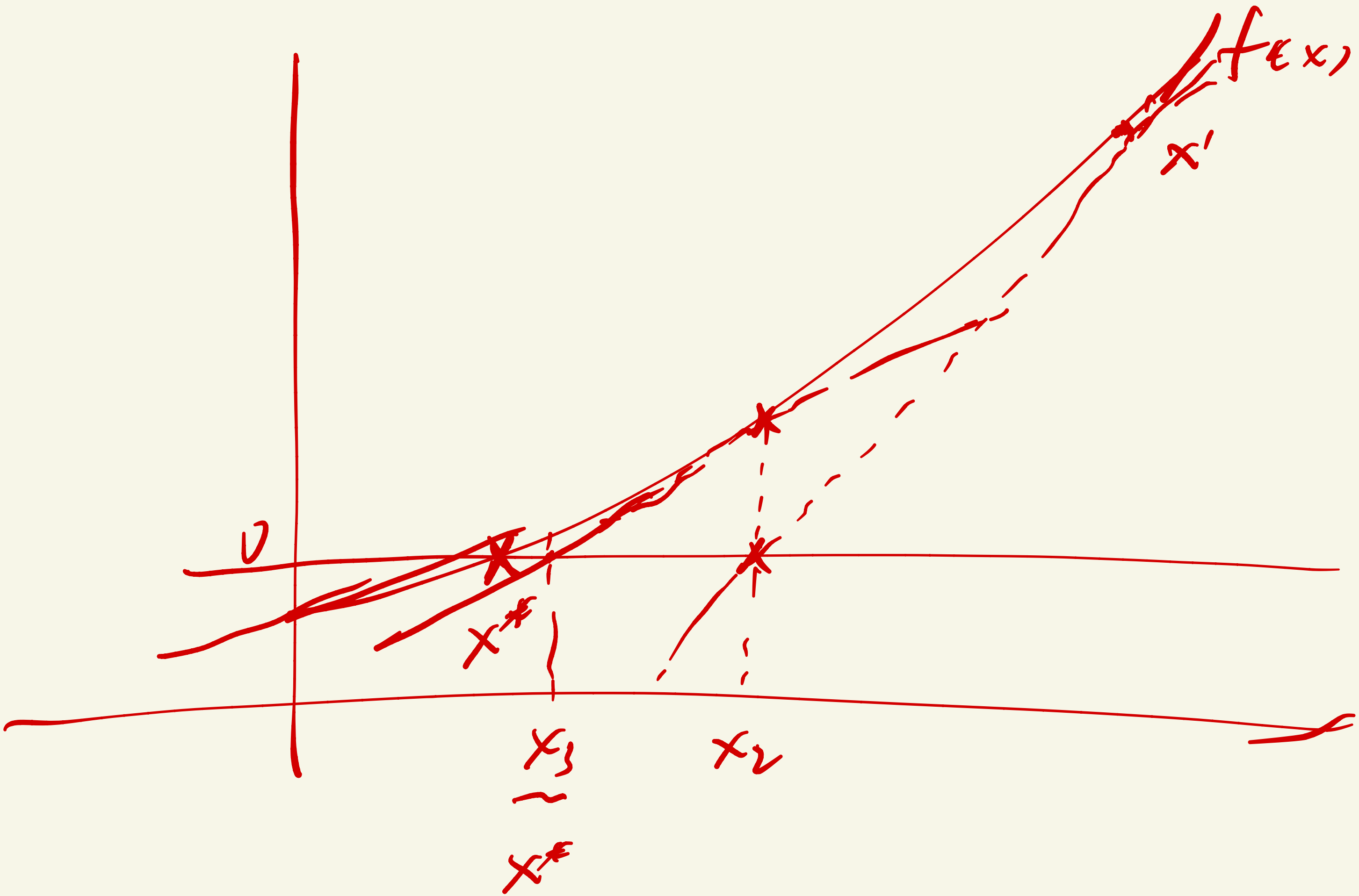


Another Optimization Method — Newton's Method



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




f(x)
x'
0
x*
x₃
x₂
x*

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

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

$$\theta := \theta - \frac{\ell'(\theta)}{\ell''(\theta)}.$$

second-order derivative

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. $f(x) = 0$. $\nabla_\theta l(\theta) = 0$

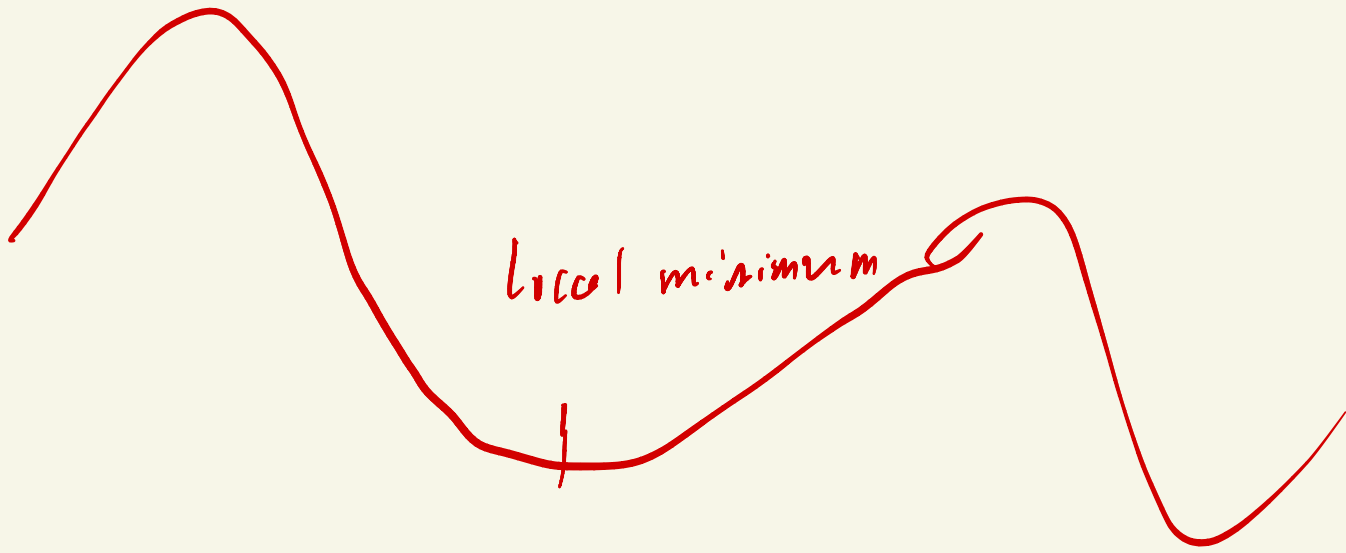
- This is the update rule in 1d

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

- 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 $H_{i,j}(\theta) = \frac{\partial}{\partial \theta_i \partial \theta_j} \ell(\theta)$. second-order


local minimum

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"*

y not related

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

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.

normalization

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:

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

Logistic binary

$$P(y|x) = h_\theta(x)^y (1-h_\theta(x))^{1-y}$$

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

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

$\eta = \theta^T x$ GLM

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

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

So then:

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

$$b(y) = 1$$

$\phi = \frac{1}{1+e^{-\theta^T x}}$

Generalized linear models

$\eta = \log\frac{\phi}{1-\phi} = \eta = \theta^T x$

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:

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

$$b(y) = 1$$

η

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:

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

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

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

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

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

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

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

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