# CS189/CS289A
# Introduction to Machine Learning
# Lecture 11: Logistic Regression

Peter Bartlett

February 24, 2015

# Outline

**First: A Bayesian view of linear regression**

Logistic Regression:

- Gaussian generative to logistic discriminative models.
- Parameter estimates for logistic models.
  - Maximum likelihood: coupled non-linear equations.
  - Gradient ascent.
  - Stochastic gradient.
  - (Detour: stochastic gradient for linear regression.)
  - Newton's method: iteratively reweighted least squares.

# Bayesian linear regression

## Linear model

# Bayesian linear regression

## Linear model

$$P(Y|X = x) = \mathcal{N}(x'\beta, \sigma^2).$$

# Bayesian linear regression

## Linear model

$$P(Y|X = x) = \mathcal{N}(x'\beta, \sigma^2).$$

Equivalently: $Y = x'\beta + \epsilon$, where $\epsilon \sim \mathcal{N}(0, \sigma^2)$.

# Bayesian linear regression

## Linear model

$$P(Y|X = x) = \mathcal{N}(x'\beta, \sigma^2).$$

Equivalently: $Y = x'\beta + \epsilon$, where $\epsilon \sim \mathcal{N}(0, \sigma^2)$.

## Bayesian Analysis

# Bayesian linear regression

## Linear model

$$P(Y|X = x) = \mathcal{N}(x'\beta, \sigma^2).$$

Equivalently: $Y = x'\beta + \epsilon$, where $\epsilon \sim \mathcal{N}(0, \sigma^2)$.

## Bayesian Analysis

- Model $\beta$ as a random variable.

# Bayesian linear regression

## Linear model

$$P(Y|X = x) = \mathcal{N}(x'\beta, \sigma^2).$$

Equivalently: $Y = x'\beta + \epsilon$, where $\epsilon \sim \mathcal{N}(0, \sigma^2)$.

## Bayesian Analysis

- Model $\beta$ as a random variable.
- e.g., $\beta \sim \mathcal{N}(0, \tau^2 I)$.

# Bayesian linear regression

## Linear model

$$P(Y|X = x) = \mathcal{N}(x'\beta, \sigma^2).$$

Equivalently: $Y = x'\beta + \epsilon$, where $\epsilon \sim \mathcal{N}(0, \sigma^2)$.

## Bayesian Analysis

- Model $\beta$ as a random variable.
- e.g., $\beta \sim \mathcal{N}(0, \tau^2 I)$. $\leftarrow$ prior distribution of $\beta$.

# Bayesian linear regression

## Linear model

$$P(Y|X = x) = \mathcal{N}(x'\beta, \sigma^2).$$

Equivalently: $Y = x'\beta + \epsilon$, where $\epsilon \sim \mathcal{N}(0, \sigma^2)$.

## Bayesian Analysis

- Model $\beta$ as a random variable.
- e.g., $\beta \sim \mathcal{N}(0, \tau^2 I)$. $\leftarrow$ prior distribution of $\beta$.
- Then compute posterior distribution $P(\beta|X, Y)$:

# Bayesian linear regression

## Linear model

$$P(Y|X = x) = \mathcal{N}(x'\beta, \sigma^2).$$

Equivalently: $Y = x'\beta + \epsilon$, where $\epsilon \sim \mathcal{N}(0, \sigma^2)$.

## Bayesian Analysis

- Model $\beta$ as a random variable.
- e.g., $\beta \sim \mathcal{N}(0, \tau^2 I)$. $\leftarrow$ prior distribution of $\beta$.
- Then compute posterior distribution $P(\beta|X, Y)$:

$$P(\beta|X, Y) = \frac{P(Y|\beta, X)P(\beta)}{P(Y|X)}$$

# Bayesian linear regression

## Linear model

$$P(Y|X=x) = \mathcal{N}(x'\beta, \sigma^2).$$

Equivalently: $Y = x'\beta + \epsilon$, where $\epsilon \sim \mathcal{N}(0, \sigma^2)$.

## Bayesian Analysis

- Model $\beta$ as a random variable.
- e.g., $\beta \sim \mathcal{N}(0, \tau^2 I)$. $\leftarrow$ prior distribution of $\beta$.
- Then compute posterior distribution $P(\beta|X,Y)$:

$$P(\beta|X,Y) = \frac{P(Y|\beta,X)P(\beta)}{P(Y|X)} \propto \underbrace{P(Y|\beta,X)}_{\text{likelihood}}\underbrace{P(\beta)}_{\text{prior}}.$$

# Bayesian linear regression

## Linear model

$$P(Y|X = x) = \mathcal{N}(x'\beta, \sigma^2).$$

Equivalently: $Y = x'\beta + \epsilon$, where $\epsilon \sim \mathcal{N}(0, \sigma^2)$.

## Bayesian Analysis

- Model $\beta$ as a random variable.
- e.g., $\beta \sim \mathcal{N}(0, \tau^2 I)$. $\leftarrow$ prior distribution of $\beta$.
- Then compute posterior distribution $P(\beta|X, Y)$:

$$P(\beta|X, Y) = \frac{P(Y|\beta, X)P(\beta)}{P(Y|X)} \propto \underbrace{P(Y|\beta, X)}_{\text{likelihood}}\,\underbrace{P(\beta)}_{\text{prior}}.$$

$$P(\beta|X_1, Y_1, X_2, Y_2) \propto \underbrace{P(Y_2|\beta, X_2)P(Y_1|\beta, X_1)}\,P(\beta)$$

# Bayesian linear regression

## Linear model

$$P(Y|X = x) = \mathcal{N}(x'\beta, \sigma^2).$$

Equivalently: $Y = x'\beta + \epsilon$, where $\epsilon \sim \mathcal{N}(0, \sigma^2)$.

## Bayesian Analysis

- Model $\beta$ as a random variable.
- e.g., $\beta \sim \mathcal{N}(0, \tau^2 I)$. $\leftarrow$ prior distribution of $\beta$.
- Then compute posterior distribution $P(\beta|X, Y)$:

$$P(\beta|X, Y) = \frac{P(Y|\beta, X)P(\beta)}{P(Y|X)} \propto \underbrace{P(Y|\beta, X)}_{\text{likelihood}} \underbrace{P(\beta)}_{\text{prior}}.$$

$$\begin{aligned} P(\beta|X_1, Y_1, X_2, Y_2) &\propto \underbrace{P(Y_2|\beta, X_2)P(Y_1|\beta, X_1)} P(\beta) \\ &= P(Y_2|\beta, X_2) \underbrace{P(Y_1|\beta, X_1)P(\beta)}. \end{aligned}$$

## Bayesian Analysis

- Prior $\beta \sim \mathcal{N}(0, \tau^2 I)$.
- Posterior distribution $P(\beta|X, Y)$:

$$P(\beta|X, Y) \propto P(Y|\beta, X)P(\beta)$$

# Bayesian linear regression

## Bayesian Analysis

- Prior $\beta \sim \mathcal{N}(0, \tau^2 I)$.
- Posterior distribution $P(\beta|X, Y)$:

$$
\begin{aligned}
P(\beta|X, Y) &\propto P(Y|\beta, X)P(\beta) \\
&\propto \exp\left(-\frac{(Y - X'\beta)^2}{2\sigma^2} - \frac{\|\beta\|^2}{2\tau^2}\right)
\end{aligned}
$$

# Bayesian linear regression

## Bayesian Analysis

- Prior $\beta \sim \mathcal{N}(0, \tau^2 I)$.
- Posterior distribution $P(\beta|X, Y)$:

$$\begin{aligned} P(\beta|X, Y) &\propto P(Y|\beta, X)P(\beta) \\ &\propto \exp\left(-\frac{(Y - X'\beta)^2}{2\sigma^2} - \frac{\|\beta\|^2}{2\tau^2}\right) \end{aligned}$$

$$P(\beta|X_1, Y_1, \ldots, X_n, Y_n) \propto P(Y_1, \ldots, Y_n|\beta, X_1, \ldots, X_n)P(\beta)$$

# Bayesian linear regression

## Bayesian Analysis

- Prior $\beta \sim \mathcal{N}(0, \tau^2 I)$.
- Posterior distribution $P(\beta|X, Y)$:

$$
\begin{aligned}
P(\beta|X, Y) &\propto P(Y|\beta, X)P(\beta) \\
&\propto \exp\left(-\frac{(Y - X'\beta)^2}{2\sigma^2} - \frac{\|\beta\|^2}{2\tau^2}\right) \\
P(\beta|X_1, Y_1, \ldots, X_n, Y_n) &\propto P(Y_1, \ldots, Y_n|\beta, X_1, \ldots, X_n)P(\beta) \\
&\propto \exp\left(-\frac{1}{2}\left(\sum_{i=1}^{n}\frac{(Y_i - X_i'\beta)^2}{\sigma^2} + \frac{1}{\tau^2}\|\beta\|^2\right)\right).
\end{aligned}
$$

# Bayesian linear regression

## Bayesian Analysis

# Bayesian linear regression

## Bayesian Analysis

- Maximum a posteriori probability estimate: estimate $\beta$ as the mode of the posterior distribution.

# Bayesian linear regression

## Bayesian Analysis

- Maximum a posteriori probability estimate: estimate $\beta$ as the mode of the posterior distribution.
- Ridge regression (least squares with a squared Euclidean norm penalty) is equivalent to a MAP estimate with a Gaussian prior.

# Bayesian linear regression

## Bayesian Analysis

- Maximum a posteriori probability estimate: estimate $\beta$ as the mode of the posterior distribution.
- Ridge regression (least squares with a squared Euclidean norm penalty) is equivalent to a MAP estimate with a Gaussian prior.
- Consider a Laplace prior:

$$P(\beta) \propto \exp\left(-\lambda \|\beta\|_1\right).$$

# Bayesian linear regression

## Bayesian Analysis

- Maximum a posteriori probability estimate: estimate $\beta$ as the mode of the posterior distribution.
- Ridge regression (least squares with a squared Euclidean norm penalty) is equivalent to a MAP estimate with a Gaussian prior.
- Consider a Laplace prior:

$$P(\beta) \propto \exp\left(-\lambda\|\beta\|_1\right).$$

- Lasso (least squares with a squared one-norm penalty) is equivalent to a MAP estimate with a Laplace prior.

# Outline

**Logistic Regression:**

# Outline

**Logistic Regression:**

- **Gaussian generative to logistic discriminative models.**

# Outline

**Logistic Regression:**

- **Gaussian generative to logistic discriminative models.**
- Parameter estimates for logistic models.

# Outline

**Logistic Regression:**

- **Gaussian generative to logistic discriminative models.**
- Parameter estimates for logistic models.
  - Maximum likelihood: coupled non-linear equations.

# Outline

**Logistic Regression:**

- **Gaussian generative to logistic discriminative models.**
- Parameter estimates for logistic models.
  - Maximum likelihood: coupled non-linear equations.
  - Gradient ascent.

# Outline

**Logistic Regression:**

- **Gaussian generative to logistic discriminative models.**
- Parameter estimates for logistic models.
  - Maximum likelihood: coupled non-linear equations.
  - Gradient ascent.
  - Stochastic gradient.

**Logistic Regression:**

- **Gaussian generative to logistic discriminative models.**
- Parameter estimates for logistic models.
  - Maximum likelihood: coupled non-linear equations.
  - Gradient ascent.
  - Stochastic gradient.
  - (Detour: stochastic gradient for linear regression.)

# Outline

**Logistic Regression:**

- **Gaussian generative to logistic discriminative models.**
- Parameter estimates for logistic models.
  - Maximum likelihood: coupled non-linear equations.
  - Gradient ascent.
  - Stochastic gradient.
  - (Detour: stochastic gradient for linear regression.)
  - Newton's method: iteratively reweighted least squares.

# Logistic regression

## Class conditionals to posterior

For Gaussian class conditional densities $P(X|Y=1)$, $P(X|Y=0)$ (with the same variance), the posterior probability is logistic

# Logistic regression

## Class conditionals to posterior

For Gaussian class conditional densities $P(X|Y=1)$, $P(X|Y=0)$ (with the same variance), the posterior probability is logistic:

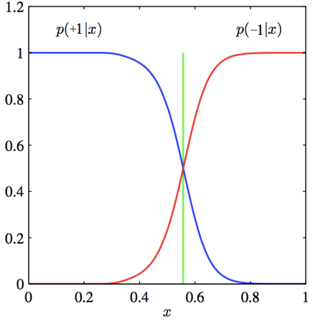
$$P(Y=1|x) = \frac{1}{1+\exp\left(-x\cdot\beta-\beta_0\right)}.$$

# Logistic regression

## Class conditionals to posterior

For Gaussian class conditional densities $P(X|Y=1)$, $P(X|Y=0)$ (with the same variance), the posterior probability is logistic:

$$P(Y=1|x) = \frac{1}{1+\exp\left(-x\cdot\beta-\beta_0\right)}.$$



# Gaussian generative to logistic discriminative models

- Suppose the class conditional distributions are Gaussian:

$$p(x|y=1) = \mathcal{N}(\mu_1, \Sigma), \qquad p(x|y=0) = \mathcal{N}(\mu_0, \Sigma)$$

# Gaussian generative to logistic discriminative models

- Suppose the class conditional distributions are Gaussian:

$$p(x|y=1) = \mathcal{N}(\mu_1, \Sigma), \qquad p(x|y=0) = \mathcal{N}(\mu_0, \Sigma)$$

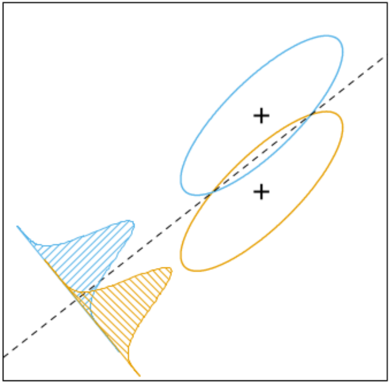
$$p(x|y=1) = \frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(x-\mu_1)'\Sigma^{-1}(x-\mu_1)\right).$$

# Gaussian generative to logistic discriminative models

- Suppose the class conditional distributions are Gaussian:

$$p(x|y=1) = \mathcal{N}(\mu_1, \Sigma), \qquad p(x|y=0) = \mathcal{N}(\mu_0, \Sigma)$$

$$p(x|y=1) = \frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(x-\mu_1)'\Sigma^{-1}(x-\mu_1)\right).$$



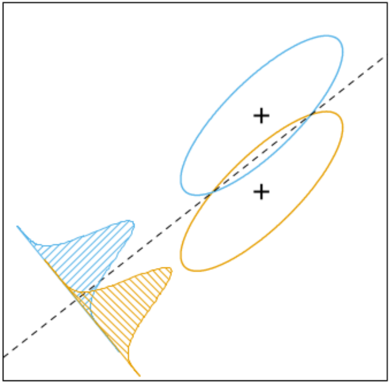
# Gaussian generative to logistic discriminative models

- Suppose the class conditional distributions are Gaussian:

$$p(x|y=1) = \mathcal{N}(\mu_1, \Sigma), \qquad p(x|y=0) = \mathcal{N}(\mu_0, \Sigma)$$

$$p(x|y=1) = \frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(x-\mu_1)'\Sigma^{-1}(x-\mu_1)\right).$$

- Class 1 has mean $\mu_1$.
  Class 0 has mean $\mu_0$.



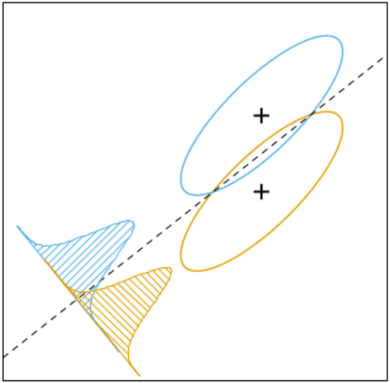
# Gaussian generative to logistic discriminative models

- Suppose the class conditional distributions are Gaussian:

$$p(x|y=1) = \mathcal{N}(\mu_1, \Sigma), \qquad p(x|y=0) = \mathcal{N}(\mu_0, \Sigma)$$

$$p(x|y=1) = \frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(x-\mu_1)'\Sigma^{-1}(x-\mu_1)\right).$$

- Class $1$ has mean $\mu_1$.
  Class $0$ has mean $\mu_0$.
- Both have covariance matrix $\Sigma$.



# Gaussian generative to logistic discriminative models

$$p(x|Y=1) = \frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(x-\mu_1)'\Sigma^{-1}(x-\mu_1)\right)$$

# Gaussian generative to logistic discriminative models

$$p(x|Y=1) = \frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(x-\mu_1)'\Sigma^{-1}(x-\mu_1)\right)$$

$$\log \frac{P(Y=1|x)}{P(Y=0|x)} = \log \frac{P(x|Y=1)P(Y=1)}{P(x|Y=0)P(Y=0)}$$

# Gaussian generative to logistic discriminative models

$$p(x|Y=1) = \frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(x-\mu_1)'\Sigma^{-1}(x-\mu_1)\right)$$

$$\begin{aligned}
\log\frac{P(Y=1|x)}{P(Y=0|x)} &= \log\frac{P(x|Y=1)P(Y=1)}{P(x|Y=0)P(Y=0)} \\
&= \frac{1}{2}\Big[(x-\mu_0)'\Sigma^{-1}(x-\mu_0) \\
&\qquad - (x-\mu_1)'\Sigma^{-1}(x-\mu_1)\Big] + \log\frac{P(Y=1)}{P(Y=0)}
\end{aligned}$$

# Gaussian generative to logistic discriminative models

$$p(x|Y=1) = \frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(x-\mu_1)'\Sigma^{-1}(x-\mu_1)\right)$$

$$\begin{aligned}
\log\frac{P(Y=1|x)}{P(Y=0|x)} &= \log\frac{P(x|Y=1)P(Y=1)}{P(x|Y=0)P(Y=0)} \\
&= \frac{1}{2}\big[(x-\mu_0)'\Sigma^{-1}(x-\mu_0) \\
&\qquad -(x-\mu_1)'\Sigma^{-1}(x-\mu_1)\big] + \log\frac{P(Y=1)}{P(Y=0)} \\
&= \frac{1}{2}\left(\mu_0'\Sigma^{-1}\mu_0 - \mu_1'\Sigma^{-1}\mu_1\right) + \log\frac{P(Y=1)}{P(Y=0)} \\
&\qquad + (\mu_1-\mu_0)'\Sigma^{-1}x
\end{aligned}$$

# Gaussian generative to logistic discriminative models

$$p(x|Y=1) = \frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(x-\mu_1)'\Sigma^{-1}(x-\mu_1)\right)$$

$$\begin{aligned}
\log \frac{P(Y=1|x)}{P(Y=0|x)} &= \log \frac{P(x|Y=1)P(Y=1)}{P(x|Y=0)P(Y=0)} \\
&= \frac{1}{2}\left[(x-\mu_0)'\Sigma^{-1}(x-\mu_0)\right. \\
&\qquad \left. -(x-\mu_1)'\Sigma^{-1}(x-\mu_1)\right] + \log \frac{P(Y=1)}{P(Y=0)} \\
&= \frac{1}{2}\left(\mu_0'\Sigma^{-1}\mu_0 - \mu_1'\Sigma^{-1}\mu_1\right) + \log \frac{P(Y=1)}{P(Y=0)} \\
&\qquad + (\mu_1-\mu_0)'\Sigma^{-1}x \\
&= \beta_0 + \beta' x.
\end{aligned}$$

# Gaussian generative to logistic discriminative models

$$\log \frac{P(Y=1|x)}{P(Y=0|x)} = \beta_0 + \beta' x.$$

# Gaussian generative to logistic discriminative models

$$\log \frac{P(Y=1|x)}{P(Y=0|x)} = \beta_0 + \beta' x.$$

For $p = P(Y=1|x)$, we have

$$\log \frac{p}{1-p} = \beta_0 + \beta' x,$$

$$\log \frac{P(Y = 1|x)}{P(Y = 0|x)} = \beta_0 + \beta' x.$$

For $p = P(Y = 1|x)$, we have

$$\begin{aligned} \log \frac{p}{1-p} &= \beta_0 + \beta' x, \\ \frac{p}{1-p} &= \exp\left(\beta_0 + \beta' x\right), \end{aligned}$$

# Gaussian generative to logistic discriminative models

$$\log \frac{P(Y=1|x)}{P(Y=0|x)} = \beta_0 + \beta' x.$$

For $p = P(Y=1|x)$, we have

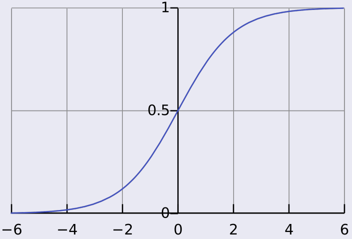
$$\begin{aligned}
\log \frac{p}{1-p} &= \beta_0 + \beta' x, \\
\frac{p}{1-p} &= \exp\left(\beta_0 + \beta' x\right), \\
P(Y=1|X) = p &= \frac{1}{1+\exp\left(-\left(\beta_0 + \beta' x\right)\right)}.
\end{aligned}$$

# Gaussian generative to logistic discriminative models

Logistic model: $$P(Y=1|x) = \frac{1}{1+\exp\left(-\beta \cdot x - \beta_0\right)}.$$

# Gaussian generative to logistic discriminative models

Logistic model: $$P(Y=1|x) = \frac{1}{1+\exp(-\beta \cdot x - \beta_0)}.$$

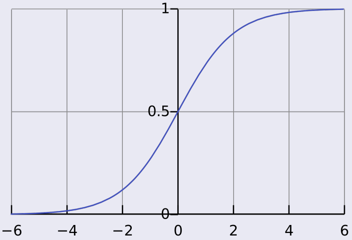



The logistic function $\frac{1}{1+e^{-\alpha}}$.

# Gaussian generative to logistic discriminative models

Logistic model: $$P(Y=1|x) = \frac{1}{1+\exp\left(-\beta \cdot x - \beta_0\right)}.$$

$$\beta = \Sigma^{-1}(\mu_1 - \mu_0), \qquad \beta_0 = \frac{\mu_0'\Sigma^{-1}\mu_0 - \mu_1'\Sigma^{-1}\mu_1}{2} + \log\frac{P(Y=1)}{P(Y=0)}.$$



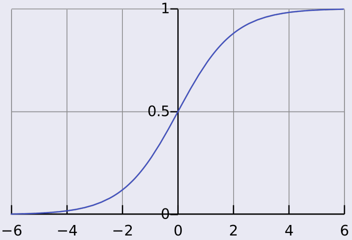

The logistic function $\dfrac{1}{1+e^{-\alpha}}$.

# Gaussian generative to logistic discriminative models

Logistic model: $$P(Y=1|x) = \frac{1}{1+\exp\left(-\beta\cdot x - \beta_0\right)}.$$

$$\beta = \Sigma^{-1}(\mu_1 - \mu_0), \qquad \beta_0 = \frac{\mu_0'\Sigma^{-1}\mu_0 - \mu_1'\Sigma^{-1}\mu_1}{2} + \log\frac{P(Y=1)}{P(Y=0)}.$$

- $P(Y=1|x)$ increases as



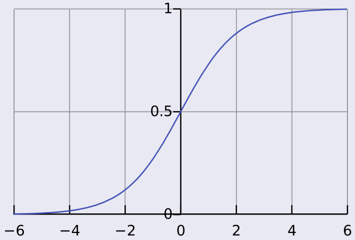

The logistic function $\frac{1}{1+e^{-\alpha}}$.

## Gaussian generative to logistic discriminative models

Logistic model: $$P(Y=1|x) = \frac{1}{1+\exp\left(-\beta\cdot x - \beta_0\right)}.$$

$$\beta = \Sigma^{-1}(\mu_1 - \mu_0), \qquad \beta_0 = \frac{\mu_0'\Sigma^{-1}\mu_0 - \mu_1'\Sigma^{-1}\mu_1}{2} + \log\frac{P(Y=1)}{P(Y=0)}.$$

- $P(Y=1|x)$ increases as
  - $x$ moves from $\mu_0$ to $\mu_1$,



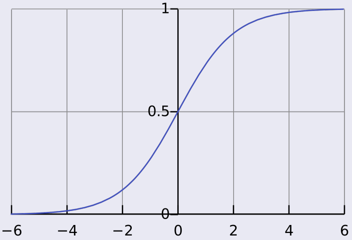

The logistic function $\dfrac{1}{1+e^{-\alpha}}$.

# Gaussian generative to logistic discriminative models

Logistic model: $P(Y = 1|x) = \dfrac{1}{1 + \exp(-\beta \cdot x - \beta_0)}$.

$$\beta = \Sigma^{-1}(\mu_1 - \mu_0), \qquad \beta_0 = \frac{\mu_0' \Sigma^{-1} \mu_0 - \mu_1' \Sigma^{-1} \mu_1}{2} + \log \frac{P(Y = 1)}{P(Y = 0)}.$$

- $P(Y = 1|x)$ increases as
  - $x$ moves from $\mu_0$ to $\mu_1$,
  - $P(Y = 1)$ increases.




The logistic function $\dfrac{1}{1 + e^{-\alpha}}$.

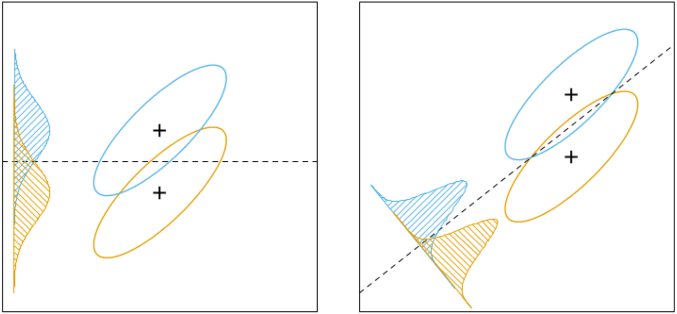


**FIGURE 4.9.** *Although the line joining the centroids defines the direction of greatest centroid spread, the projected data overlap because of the covariance (left panel). The discriminant direction minimizes this overlap for Gaussian data (right panel).*

- This is one motivation for logistic regression.

- This is one motivation for logistic regression.
  But logistic regression does not *require* an assumption of Gaussian class conditionals!

# Gaussian generative to logistic discriminative models

- This is one motivation for logistic regression.
  But logistic regression does not *require* an assumption of Gaussian class conditionals!
- Logistic regression: Model log odds ($\log p/(1-p)$) as an affine function of $x$.

# Outline

**Logistic Regression:**

- Gaussian generative to logistic discriminative models.
- **Parameter estimates for logistic models.**
  - Maximum likelihood: coupled non-linear equations.
  - Gradient ascent.
  - Stochastic gradient.
  - (Detour: stochastic gradient for linear regression.)
  - Newton's method: iteratively reweighted least squares.

# Logistic regression

## Logistic Model

$$P(Y = 1|x) = \frac{1}{1 + \exp(-\beta' x)}. \qquad \text{(incorporate } \beta_0 \text{ into } \beta\text{)}$$

# Logistic regression

## Logistic Model

$$P(Y = 1|x) = \frac{1}{1 + \exp(-\beta' x)}. \qquad \text{(incorporate } \beta_0 \text{ into } \beta\text{)}$$

## Logistic Regression

Given data $(X_1, Y_1), \ldots, (X_n, Y_n) \in \mathbb{R}^p \times \{0, 1\}$,
estimate $\beta$

# Logistic regression

## Logistic Model

$$P(Y = 1|x) = \frac{1}{1 + \exp(-\beta' x)}. \qquad \text{(incorporate } \beta_0 \text{ into } \beta)$$

## Logistic Regression

Given data $(X_1, Y_1), \ldots, (X_n, Y_n) \in \mathbb{R}^p \times \{0, 1\}$,
estimate $\beta$: maximum likelihood.

### Maximum likelihood estimation

# Logistic regression

## Maximum likelihood estimation

Log likelihood: $\ell(\beta) = \log P(y_1, \ldots, y_n | x_1, \ldots, x_n; \beta)$

# Logistic regression

## Maximum likelihood estimation

Log likelihood:

$$
\begin{aligned}
\ell(\beta) &= \log P(y_1, \dots, y_n | x_1, \dots, x_n; \beta) \\
&= \sum_{i=1}^{n} y_i \log \mu_i(\beta) + (1 - y_i) \log(1 - \mu_i(\beta)),
\end{aligned}
$$

# Logistic regression

## Maximum likelihood estimation

Log likelihood:
$$\begin{aligned}\ell(\beta) &= \log P(y_1, \ldots, y_n | x_1, \ldots, x_n; \beta) \\ &= \sum_{i=1}^{n} y_i \log \mu_i(\beta) + (1 - y_i) \log(1 - \mu_i(\beta)),\end{aligned}$$

where
$$\mu_i(\beta) = P(Y = 1 | X = x_i, \beta) = \frac{1}{1 + \exp(-\beta' x_i)}.$$

# Logistic regression

## Maximum likelihood estimation

Log likelihood:

$$\begin{aligned}\ell(\beta) &= \log P(y_1, \ldots, y_n | x_1, \ldots, x_n; \beta) \\ &= \sum_{i=1}^{n} y_i \log \mu_i(\beta) + (1 - y_i) \log(1 - \mu_i(\beta)),\end{aligned}$$

where $\mu_i(\beta) = P(Y = 1 | X = x_i, \beta) = \dfrac{1}{1 + \exp(-\beta' x_i)}$.

NB: $\nabla_\beta \mu_i(\beta) = \mu_i(\beta)(1 - \mu_i(\beta)) x_i$.

# Logistic regression

## Maximum likelihood estimation

Log likelihood: $$\ell(\beta) = \sum_{i=1}^{n} y_i \log \mu_i(\beta) + (1 - y_i) \log(1 - \mu_i(\beta)).$$

# Logistic regression

## Maximum likelihood estimation

Log likelihood: $$\ell(\beta) = \sum_{i=1}^{n} y_i \log \mu_i(\beta) + (1 - y_i) \log(1 - \mu_i(\beta)).$$

First derivative: $$\nabla_\beta \ell(\beta) = \sum_{i=1}^{n} \left( \left( \frac{y_i}{\mu_i(\beta)} - \frac{1 - y_i}{1 - \mu_i(\beta)} \right) \nabla_\beta \mu_i(\beta) \right)$$

## Maximum likelihood estimation

Log likelihood: $$\ell(\beta) = \sum_{i=1}^{n} y_i \log \mu_i(\beta) + (1 - y_i) \log(1 - \mu_i(\beta)).$$

First derivative: $$\nabla_\beta \ell(\beta) = \sum_{i=1}^{n} \left( \left( \frac{y_i}{\mu_i(\beta)} - \frac{1 - y_i}{1 - \mu_i(\beta)} \right) \nabla_\beta \mu_i(\beta) \right)$$
$$= \sum_{i=1}^{n} (y_i - \mu_i(\beta))\, x_i.$$

# Logistic regression

## Maximum likelihood estimation

Log likelihood: $$\ell(\beta) = \sum_{i=1}^{n} y_i \log \mu_i(\beta) + (1 - y_i) \log(1 - \mu_i(\beta)).$$

First derivative: $$\nabla_\beta \ell(\beta) = \sum_{i=1}^{n} \left( \left( \frac{y_i}{\mu_i(\beta)} - \frac{1 - y_i}{1 - \mu_i(\beta)} \right) \nabla_\beta \mu_i(\beta) \right)$$
$$= \sum_{i=1}^{n} (y_i - \mu_i(\beta))\, x_i.$$

Second derivative: $$\nabla_\beta^2 \ell(\beta) = \sum_{i=1}^{n} -\mu_i(\beta)(1 - \mu_i(\beta)) x_i x_i'.$$

# Logistic regression

## Maximum likelihood estimate

$\hat{\beta}^{ml}$ solves: $$0 = \sum_{i=1}^{n} \left(y_i - \mu_i(\beta)\right) x_i$$

# Logistic regression

## Maximum likelihood estimate

$\hat{\beta}^{ml}$ solves: $$0 = \sum_{i=1}^{n} (y_i - \mu_i(\beta))\, x_i = \sum_{i=1}^{n} \left( y_i - \frac{1}{1+\exp(-\beta' x_i)} \right) x_i$$

## Maximum likelihood estimate

$\hat{\beta}^{ml}$ solves:

$$0 = \sum_{i=1}^{n} (y_i - \mu_i(\beta))\, x_i = \sum_{i=1}^{n} \left( y_i - \frac{1}{1 + \exp(-\beta' x_i)} \right) x_i$$

$$\Leftrightarrow \sum_{i=1}^{n} y_i x_i = \sum_{i=1}^{n} \mu_i(\beta) x_i .$$

# Logistic regression

## Maximum likelihood estimate

$\hat{\beta}^{ml}$ solves:

$$0 = \sum_{i=1}^{n} (y_i - \mu_i(\beta))\, x_i = \sum_{i=1}^{n} \left( y_i - \frac{1}{1 + \exp(-\beta' x_i)} \right) x_i$$

$$\Leftrightarrow \sum_{i=1}^{n} y_i x_i = \sum_{i=1}^{n} \mu_i(\beta) x_i. \qquad \text{(moment matching)}$$

# Logistic regression

**Maximum likelihood estimate**

$\hat{\beta}^{ml}$ solves:

$$0 = \sum_{i=1}^{n} (y_i - \mu_i(\beta))\, x_i = \sum_{i=1}^{n} \left( y_i - \frac{1}{1+\exp(-\beta' x_i)} \right) x_i$$

$$\Leftrightarrow \sum_{i=1}^{n} y_i x_i = \sum_{i=1}^{n} \mu_i(\beta) x_i. \qquad \text{(moment matching)}$$

- System of $p$ coupled non-linear equations.

# Logistic regression

## Maximum likelihood estimate

$\hat{\beta}^{ml}$ solves:

$$0 = \sum_{i=1}^{n} (y_i - \mu_i(\beta))\, x_i = \sum_{i=1}^{n} \left( y_i - \frac{1}{1 + \exp(-\beta' x_i)} \right) x_i$$

$$\Leftrightarrow \sum_{i=1}^{n} y_i x_i = \sum_{i=1}^{n} \mu_i(\beta) x_i. \qquad \text{(moment matching)}$$

- System of $p$ coupled non-linear equations.
- No closed-form solution.

# Logistic regression

## Maximum likelihood estimate

$\hat{\beta}^{ml}$ solves:

$$0 = \sum_{i=1}^{n} (y_i - \mu_i(\beta))\, x_i = \sum_{i=1}^{n} \left( y_i - \frac{1}{1 + \exp(-\beta' x_i)} \right) x_i$$

$$\Leftrightarrow \sum_{i=1}^{n} y_i x_i = \sum_{i=1}^{n} \mu_i(\beta) x_i. \qquad \text{(moment matching)}$$

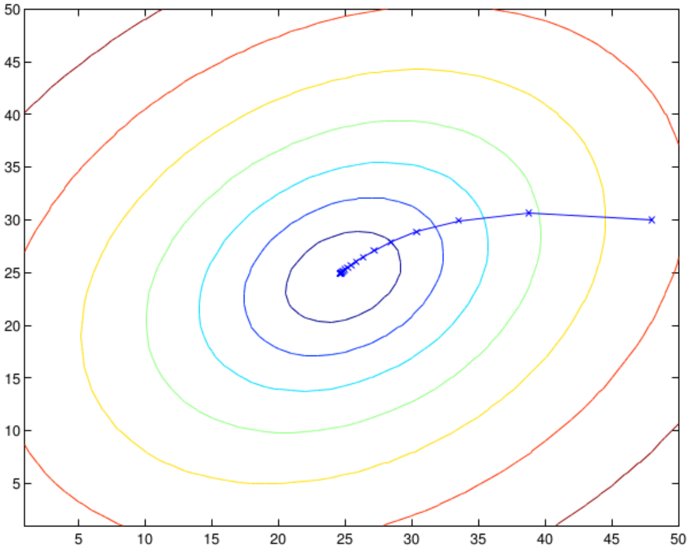
- System of $p$ coupled non-linear equations.
- No closed-form solution.
- But $\ell$ is *concave*, so we can find the solution.

# Outline

**Logistic Regression:**

- Gaussian generative to logistic discriminative models.
- Parameter estimates for logistic models.
    - Maximum likelihood: coupled non-linear equations.
    - **Gradient ascent.**
    - Stochastic gradient.
    - (Detour: stochastic gradient for linear regression.)
    - Newton's method: iteratively reweighted least squares.

# Gradient Ascent



# Logistic regression

## Gradient ascent

$$\beta^{(t+1)} = \beta^{(t)} + \eta \nabla_\beta \ell\left(\beta^{(t)}\right)$$

# Logistic regression

## Gradient ascent

$$
\begin{aligned}
\beta^{(t+1)} &= \beta^{(t)} + \eta \nabla_{\beta} \ell \left(\beta^{(t)}\right) \\
&= \beta^{(t)} + \eta \sum_{i=1}^{n} \left(y_i - \mu_i(\beta^{(t)})\right) x_i .
\end{aligned}
$$

# Logistic regression

## Gradient ascent

$$\beta^{(t+1)} = \beta^{(t)} + \eta \nabla_\beta \ell\left(\beta^{(t)}\right)$$
$$= \beta^{(t)} + \eta \sum_{i=1}^{n} \left(y_i - \mu_i(\beta^{(t)})\right) x_i.$$

- $\eta$ is a step-size parameter.

# Logistic regression

## Gradient ascent

$$\beta^{(t+1)} = \beta^{(t)} + \eta \nabla_{\beta} \ell \left( \beta^{(t)} \right)$$
$$= \beta^{(t)} + \eta \sum_{i=1}^{n} \left( y_i - \mu_i(\beta^{(t)}) \right) x_i.$$

- $\eta$ is a step-size parameter.
- Since the $\mu_i$ depend on $\beta$, the gradient changes as $\beta$ changes.

# Logistic regression

## Gradient ascent

$$\beta^{(t+1)} = \beta^{(t)} + \eta \nabla_\beta \ell\left(\beta^{(t)}\right)$$
$$= \beta^{(t)} + \eta \sum_{i=1}^{n} \left(y_i - \mu_i(\beta^{(t)})\right) x_i.$$

- $\eta$ is a step-size parameter.
- Since the $\mu_i$ depend on $\beta$, the gradient changes as $\beta$ changes.
- Each gradient calculation takes $O(np)$ time.

# Logistic regression

- Instead of computing $\nabla_\beta \ell(\beta^{(t)})$, we compute a random approximation that is on average in the direction of $\nabla_\beta \ell(\beta^{(t)})$.

# Logistic regression

- Instead of computing $\nabla_\beta \ell(\beta^{(t)})$, we compute a random approximation that is on average in the direction of $\nabla_\beta \ell(\beta^{(t)})$.
- Here, gradient is

$$\nabla_\beta \ell(\beta) = \sum_{i=1}^{n} (y_i - \mu_i(\beta))\, x_i.$$

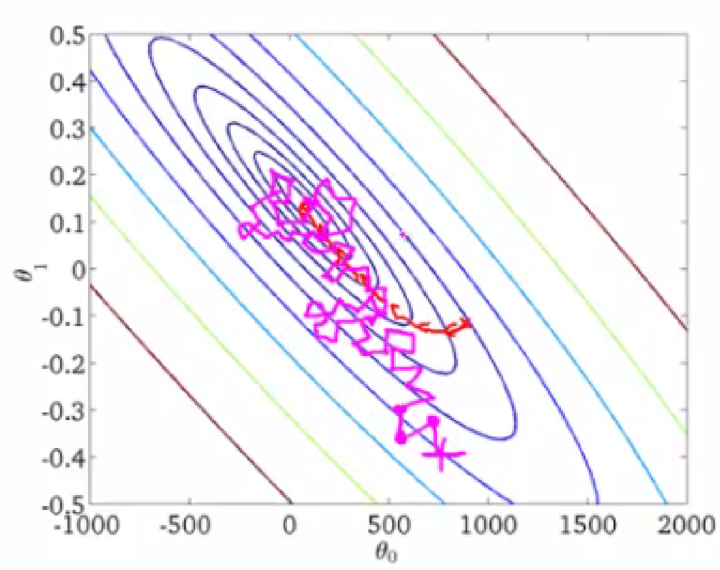
# Logistic regression

- Instead of computing $\nabla_\beta \ell(\beta^{(t)})$, we compute a random approximation that is on average in the direction of $\nabla_\beta \ell(\beta^{(t)})$.
- Here, gradient is

$$\nabla_\beta \ell(\beta) = \sum_{i=1}^{n} (y_i - \mu_i(\beta))\, x_i.$$

- We need an (easy-to-compute) quantity that is in this direction on average.

# Stochastic Gradient



# Logistic regression

- Instead of computing $\nabla_\beta \ell(\beta^{(t)})$, we compute a random approximation that is on average in the direction of $\nabla_\beta \ell(\beta^{(t)})$.

# Logistic regression

- Instead of computing $\nabla_\beta \ell(\beta^{(t)})$, we compute a random approximation that is on average in the direction of $\nabla_\beta \ell(\beta^{(t)})$.
- At step $t$, we'll choose a random index $i_t \in \{1, \ldots, n\}$, and use

$$g_{i_t} = \left(y_{i_t} - \mu_{i_t}(\beta^{(t)})\right) x_{i_t}.$$

# Logistic regression

- Instead of computing $\nabla_\beta \ell(\beta^{(t)})$, we compute a random approximation that is on average in the direction of $\nabla_\beta \ell(\beta^{(t)})$.
- At step $t$, we'll choose a random index $i_t \in \{1, \ldots, n\}$, and use

$$g_{i_t} = \left(y_{i_t} - \mu_{i_t}(\beta^{(t)})\right) x_{i_t}.$$

For instance, if we assume that the training examples are in a random order, we can cycle through the training data and interpret the choice of $(x_i, y_i)$ as uniform.

# Logistic regression

- Instead of computing $\nabla_\beta \ell(\beta^{(t)})$, we compute a random approximation that is on average in the direction of $\nabla_\beta \ell(\beta^{(t)})$.
- At step $t$, we'll choose a random index $i_t \in \{1, \dots, n\}$, and use

$$g_{i_t} = \left( y_{i_t} - \mu_{i_t}(\beta^{(t)}) \right) x_{i_t}.$$

For instance, if we assume that the training examples are in a random order, we can cycle through the training data and interpret the choice of $(x_i, y_i)$ as uniform.

- Each stochastic gradient calculation takes $O(p)$ time.

# Logistic regression

- Instead of computing $\nabla_\beta \ell(\beta^{(t)})$, we compute a random approximation that is on average in the direction of $\nabla_\beta \ell(\beta^{(t)})$.
- At step $t$, we'll choose a random index $i_t \in \{1, \ldots, n\}$, and use

$$g_{i_t} = \left(y_{i_t} - \mu_{i_t}(\beta^{(t)})\right) x_{i_t}.$$

  For instance, if we assume that the training examples are in a random order, we can cycle through the training data and interpret the choice of $(x_i, y_i)$ as uniform.
- Each stochastic gradient calculation takes $O(p)$ time.
- If $n$ is large, we might obtain a good estimate long before we have seen $n$ training examples.

# Logistic regression

- Instead of computing $\nabla_\beta \ell(\beta^{(t)})$, we compute a random approximation that is on average in the direction of $\nabla_\beta \ell(\beta^{(t)})$.
- At step $t$, we'll choose a random index $i_t \in \{1, \ldots, n\}$, and use

$$g_{i_t} = \left(y_{i_t} - \mu_{i_t}(\beta^{(t)})\right) x_{i_t}.$$

For instance, if we assume that the training examples are in a random order, we can cycle through the training data and interpret the choice of $(x_i, y_i)$ as uniform.

- Each stochastic gradient calculation takes $O(p)$ time.
- If $n$ is large, we might obtain a good estimate long before we have seen $n$ training examples.

**Stochastic gradient**

$$\beta^{(t+1)} = \beta^{(t)} + \eta \left(y_{i_t} - \mu_{i_t}(\beta^{(t)})\right) x_{i_t}.$$

# Outline

**Logistic Regression:**

- Gaussian generative to logistic discriminative models.
- Parameter estimates for logistic models.
  - Maximum likelihood: coupled non-linear equations.
  - Gradient ascent.
  - Stochastic gradient.
  - **(Detour: stochastic gradient for linear regression.)**
  - Newton's method: iteratively reweighted least squares.

# Detour: Back to linear regression

# Detour: Back to linear regression

## Residual sum of squares

# Detour: Back to linear regression

## Residual sum of squares

$$\nabla_\beta RSS(\beta) = X'X\beta - X'y$$

# Detour: Back to linear regression

## Residual sum of squares

$$\nabla_\beta RSS(\beta) = X'X\beta - X'y = \sum_{i=1}^{n} x_i(x_i'\beta - y_i).$$

# Detour: Back to linear regression

## Residual sum of squares

$$\nabla_\beta RSS(\beta) = X'X\beta - X'y = \sum_{i=1}^{n} x_i(x_i'\beta - y_i).$$

- This gave us the normal equations.

# Detour: Back to linear regression

## Residual sum of squares

$$\nabla_\beta RSS(\beta) = X'X\beta - X'y = \sum_{i=1}^{n} x_i(x_i'\beta - y_i).$$

- This gave us the normal equations.
- We could also consider a gradient descent approach.

# Detour: Back to linear regression

## Residual sum of squares

$$\nabla_\beta RSS(\beta) = X'X\beta - X'y = \sum_{i=1}^{n} x_i(x_i'\beta - y_i).$$

- This gave us the normal equations.
- We could also consider a gradient descent approach.
- Or a stochastic gradient approach:

$$\beta^{(t+1)} = \beta^{(t)} + \eta x_{i_t} \left( y_{i_t} - x_{i_t}'\beta^{(t)} \right).$$

# Detour: Back to linear regression

## Residual sum of squares

$$\nabla_\beta RSS(\beta) = X'X\beta - X'y = \sum_{i=1}^{n} x_i(x_i'\beta - y_i).$$

- This gave us the normal equations.
- We could also consider a gradient descent approach.
- Or a stochastic gradient approach:

$$\beta^{(t+1)} = \beta^{(t)} + \eta x_{i_t}\left(y_{i_t} - x_{i_t}'\beta^{(t)}\right).$$

- This might be simpler for large scale problems.

# Detour: Back to linear regression

## Residual sum of squares

$$\nabla_\beta RSS(\beta) = X'X\beta - X'y = \sum_{i=1}^{n} x_i(x_i'\beta - y_i).$$

- This gave us the normal equations.
- We could also consider a gradient descent approach.
- Or a stochastic gradient approach:

$$\beta^{(t+1)} = \beta^{(t)} + \eta x_{i_t}\left(y_{i_t} - x_{i_t}'\beta^{(t)}\right).$$

- This might be simpler for large scale problems.
  (e.g., suppose each $x$ encodes a bag-of-words representation of a document.)

# Detour: Back to linear regression

## Residual sum of squares

$$\nabla_\beta RSS(\beta) = X'X\beta - X'y = \sum_{i=1}^{n} x_i(x_i'\beta - y_i).$$

- This gave us the normal equations.
- We could also consider a gradient descent approach.
- Or a stochastic gradient approach:

$$\beta^{(t+1)} = \beta^{(t)} + \eta x_{i_t} \left( y_{i_t} - x_{i_t}' \beta^{(t)} \right).$$

- This might be simpler for large scale problems.
  (e.g., suppose each $x$ encodes a bag-of-words representation of a document.)
- Convex regularization terms (like $\|\beta\|_2^2$ and $\|\beta\|_1$) can be easily incorporated.

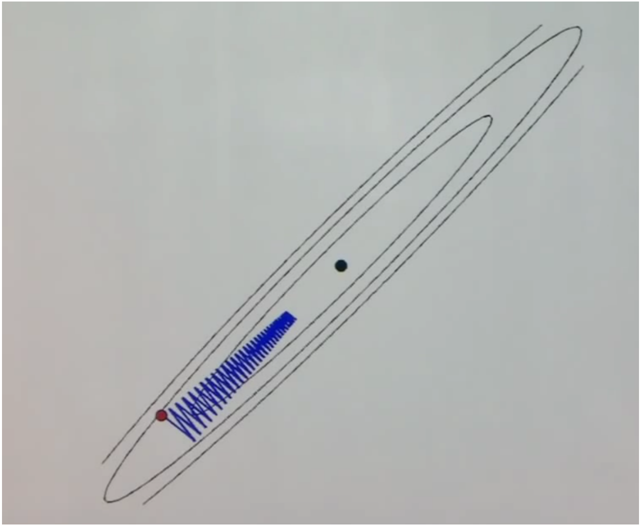
# Outline

**Logistic Regression:**

- Gaussian generative to logistic discriminative models.
- Parameter estimates for logistic models.
  - Maximum likelihood: coupled non-linear equations.
  - Gradient ascent.
  - Stochastic gradient.
  - (Detour: stochastic gradient for linear regression.)
  - **Newton's method: iteratively reweighted least squares.**

# Newton's method

- Incorporating second derivative information can significantly speed up gradient methods.



# Newton's method

## Finding roots

# Newton's method

## Finding roots

- To solve $f(x) = 0$:

## Newton's method

### Finding roots

- To solve $f(x) = 0$:
- Start at $x_0$.

# Newton's method

## Finding roots

- To solve $f(x) = 0$:
- Start at $x_0$.
- Calculate the linear (first order Taylor series) approximation of $f$ at $x_0$:

$$f(x_0) + f'(x_0)(x - x_0).$$

# Newton's method

## Finding roots

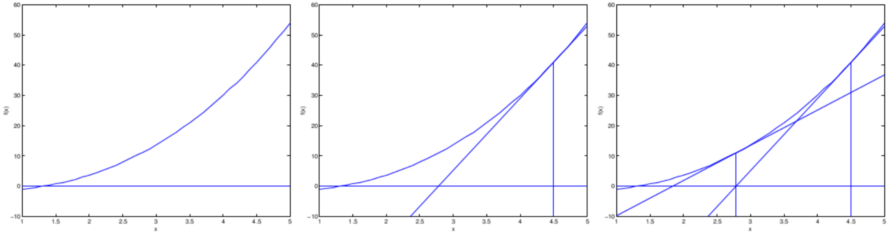
- To solve $f(x) = 0$:
- Start at $x_0$.
- Calculate the linear (first order Taylor series) approximation of $f$ at $x_0$:

$$f(x_0) + f'(x_0)(x - x_0).$$

- Solve for $f(x) = 0$:

$$x_1 = x_0 - \frac{f(x_0)}{f'(x_0)}.$$

# Newton's method



# Newton's method

## Finding maxima

# Newton's method

## Finding maxima

- To solve $\nabla f(x) = 0$:

# Newton's method

## Finding maxima

- To solve $\nabla f(x) = 0$:
- Start at $x_0$.

# Newton's method

## Finding maxima

- To solve $\nabla f(x) = 0$:
- Start at $x_0$.
- Calculate the linear (first order Taylor series) approximation of $\nabla f$ at $x_0$:

$$\nabla f(x_0) + \nabla^2 f(x_0)(x - x_0).$$

# Newton's method

## Finding maxima

- To solve $\nabla f(x) = 0$:
- Start at $x_0$.
- Calculate the linear (first order Taylor series) approximation of $\nabla f$ at $x_0$:

$$\nabla f(x_0) + \nabla^2 f(x_0)(x - x_0).$$

- Solve for $\nabla f(x) = 0$:

$$x_1 = x_0 - \left[\nabla^2 f(x_0)\right]^{-1} \nabla f(x_0).$$

# Newton's method

## Finding maxima

- To solve $\nabla f(x) = 0$:
- Start at $x_0$.
- Calculate the linear (first order Taylor series) approximation of $\nabla f$ at $x_0$:

$$\nabla f(x_0) + \nabla^2 f(x_0)(x - x_0).$$

- Solve for $\nabla f(x) = 0$:

$$x_1 = x_0 - \left[\nabla^2 f(x_0)\right]^{-1} \nabla f(x_0).$$

- **Newton-Raphson:** $x_{t+1} = x_t - \left[\nabla^2 f(x_t)\right]^{-1} \nabla f(x_t).$

# Newton-Raphson method for logistic regression

## Logistic regression

Log likelihood: $$\ell(\beta) = \sum_{i=1}^{n} y_i \log \mu_i(\beta) + (1 - y_i) \log(1 - \mu_i(\beta)),$$

# Newton-Raphson method for logistic regression

## Logistic regression

Log likelihood: $$\ell(\beta) = \sum_{i=1}^{n} y_i \log \mu_i(\beta) + (1 - y_i) \log(1 - \mu_i(\beta)),$$

First derivative: $$\nabla_\beta \ell(\beta) = \sum_{i=1}^{n} (y_i - \mu_i(\beta))\, x_i$$

# Newton-Raphson method for logistic regression

## Logistic regression

Log likelihood: $$\ell(\beta) = \sum_{i=1}^{n} y_i \log \mu_i(\beta) + (1 - y_i) \log(1 - \mu_i(\beta)),$$

First derivative: $$\nabla_\beta \ell(\beta) = \sum_{i=1}^{n} (y_i - \mu_i(\beta))\, x_i$$

Second derivative: $$\nabla_\beta^2 \ell(\beta) = \sum_{i=1}^{n} -\mu_i(\beta)(1 - \mu_i(\beta)) x_i x_i'$$

# Newton-Raphson method for logistic regression

## Logistic regression

Log likelihood: $$\ell(\beta) = \sum_{i=1}^{n} y_i \log \mu_i(\beta) + (1 - y_i) \log(1 - \mu_i(\beta)),$$

First derivative: $$\begin{aligned} \nabla_\beta \ell(\beta) &= \sum_{i=1}^{n} (y_i - \mu_i(\beta))\, x_i \\ &= X'(y - \mu). \end{aligned}$$

Second derivative: $$\nabla_\beta^2 \ell(\beta) = \sum_{i=1}^{n} -\mu_i(\beta)(1 - \mu_i(\beta)) x_i x_i'$$

# Newton-Raphson method for logistic regression

## Logistic regression

Log likelihood: $$\ell(\beta) = \sum_{i=1}^{n} y_i \log \mu_i(\beta) + (1 - y_i) \log(1 - \mu_i(\beta)),$$

First derivative: $$\nabla_\beta \ell(\beta) = \sum_{i=1}^{n} (y_i - \mu_i(\beta))\, x_i = X'(y - \mu).$$

Second derivative: $$\nabla_\beta^2 \ell(\beta) = \sum_{i=1}^{n} -\mu_i(\beta)(1 - \mu_i(\beta)) x_i x_i' = -X'\mathrm{diag}(\mu(1 - \mu))X.$$

# Newton-Raphson method for logistic regression

## Logistic regression

$$\beta^{(t+1)} = \beta^{(t)} - \left[\nabla^2 \ell(\beta^{(t)})\right]^{-1} \nabla \ell(\beta^{(t)})$$

# Newton-Raphson method for logistic regression

## Logistic regression

$$
\begin{aligned}
\beta^{(t+1)} &= \beta^{(t)} - \left[\nabla^2 \ell(\beta^{(t)})\right]^{-1} \nabla \ell(\beta^{(t)}) \\
&= \beta^{(t)} \quad \left[ \qquad\qquad\qquad\qquad \right]^{-1} \qquad\qquad .
\end{aligned}
$$

# Newton-Raphson method for logistic regression

## Logistic regression

$$\begin{aligned}\beta^{(t+1)} &= \beta^{(t)} - \left[\nabla^2 \ell(\beta^{(t)})\right]^{-1} \nabla \ell(\beta^{(t)}) \\ &= \beta^{(t)} + \left[X' \mathrm{diag}(\mu(1-\mu))X\right]^{-1} \qquad .\end{aligned}$$

# Newton-Raphson method for logistic regression

## Logistic regression

$$\begin{aligned}
\beta^{(t+1)} &= \beta^{(t)} - \left[\nabla^2 \ell(\beta^{(t)})\right]^{-1} \nabla \ell(\beta^{(t)}) \\
&= \beta^{(t)} + \left[X' \mathrm{diag}(\mu(1-\mu)) X\right]^{-1} X'(y-\mu).
\end{aligned}$$

# Newton-Raphson method for logistic regression

## Logistic regression

$$\begin{aligned}\beta^{(t+1)} &= \beta^{(t)} - \left[\nabla^2 \ell(\beta^{(t)})\right]^{-1} \nabla \ell(\beta^{(t)}) \\ &= \beta^{(t)} + \left[X' \mathrm{diag}(\mu(1-\mu))X\right]^{-1} X'(y-\mu).\end{aligned}$$

- Each iteration is like a least squares problem, except that the $x_i x_i'$ entries are weighted by $\mu_i(1-\mu_i)$.

# Newton-Raphson method for logistic regression

## Logistic regression

$$\beta^{(t+1)} = \beta^{(t)} - \left[\nabla^2 \ell(\beta^{(t)})\right]^{-1} \nabla \ell(\beta^{(t)})$$
$$= \beta^{(t)} + \left[X' \mathrm{diag}(\mu(1-\mu))X\right]^{-1} X'(y-\mu).$$

- Each iteration is like a least squares problem, except that the $x_i x_i'$ entries are weighted by $\mu_i(1-\mu_i)$.
- *Iteratively reweighted least squares*.

# Newton-Raphson method for logistic regression

## Logistic regression

$$\beta^{(t+1)} = \beta^{(t)} - \left[\nabla^2 \ell(\beta^{(t)})\right]^{-1} \nabla \ell(\beta^{(t)})$$
$$= \beta^{(t)} + \left[X' \mathrm{diag}(\mu(1-\mu))X\right]^{-1} X'(y-\mu).$$

- Each iteration is like a least squares problem, except that the $x_i x_i'$ entries are weighted by $\mu_i(1-\mu_i)$.
- *Iteratively reweighted least squares*.
- Reweighting is according to how far $\mu_i$ is from $0$ or $1$.

# Outline

**Logistic Regression:**

- Gaussian generative to logistic discriminative models.
- Parameter estimates for logistic models.
  - Maximum likelihood: coupled non-linear equations.
  - Gradient ascent.
  - Stochastic gradient.
  - (Detour: stochastic gradient for linear regression.)
  - Newton's method: iteratively reweighted least squares.