IEMS 304 Lecture 2: Simple Linear Regression

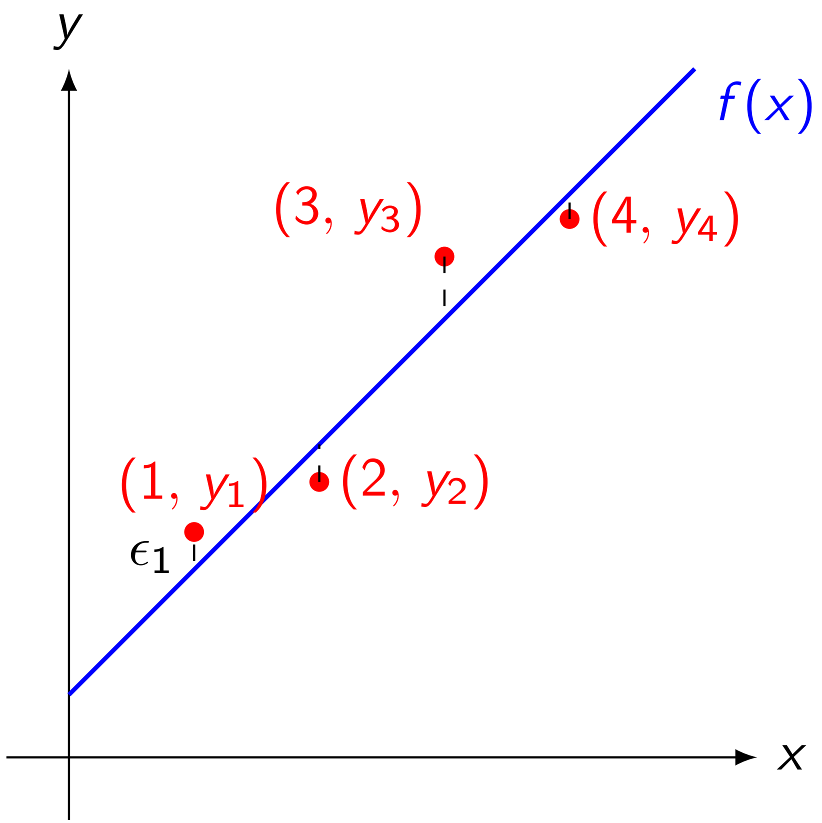
Yiping Lu
`yiping.lu@northwestern.edu`

Industrial Engineering & Management Sciences
Northwestern University

NORTHWESTERN
UNIVERSITY

Simple Linear Regression

Linear Regression



Dataset (X_1, Y_1), (X_2, Y_2), ...

real number real number

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i,$$

- X has an arbitrary distribution, possibly deterministic.
- If $X = x$, then $Y = \beta_0 + \beta_1 x + \varepsilon$, with β_0, β_1 being the *coefficients*, and ε being the *noise* variable.
- $\mathbb{E}[\varepsilon | X = x] = \overset{?}{0}$, $\mathrm{Var}(\varepsilon | X = x) = \sigma^2$.

Least Squares Estimators

One option to estimate the unknown quantities is to find the optimal fit to be precise here, minimize the mean squared error (MSE):

because I'm using L_2 loss

$$(\beta_0, \beta_1) = \arg \min_{(b_0, b_1)} \underbrace{\mathbb{E}\left[\left(Y - \overset{\text{prediction}}{(b_0 + b_1 X)}\right)^2\right]}_{\text{objective function (population)}}.$$

minimize l_2 loss for a single prediction, will return the mean

$\mathbb{E}[\varepsilon | X=x] = 0$

Variable to optimize

$\mathbb{E}[Y | X=x] = \beta_0 + \beta_1 x$

❐ How to access $\mathbb{E}$?

- The data we may consider are $\{(X_1, Y_1), \ldots, (X_n, Y_n)\}$.

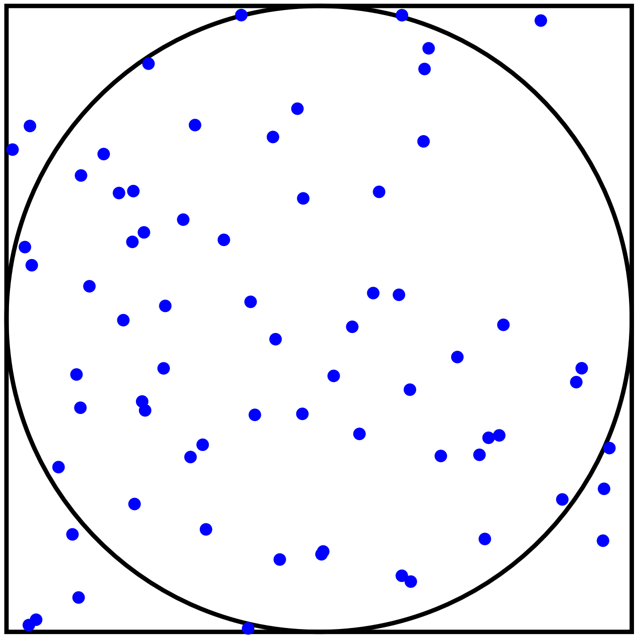
Only Thing I can compute.

$$(\hat{\beta}_0, \hat{\beta}_1) := \arg\min_{(b_0, b_1)} \underbrace{\frac{1}{n}\sum_{i=1}^{n}\left[(Y_i - (b_0 + b_1 X_i))^2\right]}_{\text{Empirical Objective Function}}$$

Monte Carlo Methods

How to Estimate π ?

- ❐ Draw a square of side length 2 (from -1 to $+1$) and inscribe a circle of radius 1.
- ❐ Randomly sample the points within the square.
- ❐ Count how many points fall inside the circle.
- ❐ The expectation of fraction of points in the circle is $\frac{\text{the circle's area}}{\text{total points' area}} \approx \frac{\pi r^2}{(2r)^2} = \frac{\pi}{4}$.
- ❐ Hence $\pi \approx 4 \times \frac{\text{points in circle}}{\text{total points}}$.



Find β_0, β_1

We minimize in-sample, empirical MSE: (mean square error)

$$(\hat{\beta}_0, \hat{\beta}_1) = \arg \min_{(b_0, b_1)} \underbrace{\frac{1}{n} \sum_{i=1}^{n} (Y_i - (b_0 + b_1 X_i))^2}_{\widehat{\mathrm{MSE}}(b_0, b_1)} .$$

Next. $\hat{\beta}_0, \hat{\beta}_1$ has closed form solution!

How ?

How to find the Minimizer of a Function

$f(x) = g_1(g_2(x))$ $\frac{\partial f}{\partial x} = \frac{\partial g_1(y)}{\partial y}\Big|_{y=g_2(x)} \cdot \frac{\partial g_2(x)}{\partial x}$

How to find the Minimizer of a function $x^* = \arg\min_x f(x)$?

Solve the equation $\nabla f(x^*) = 0$

$$f(b_0, b_1) = \frac{1}{n}\sum_{i=1}^{n}[\underbrace{Y_i - (b_0 + b_1 X_i)}_{g_2(b_0, b_1)}]^{\boxed{2}} - g_1(b_0, b_1)$$

$$\nabla_{b_0} f(b_0, b_1) = -\frac{1}{n}\sum_{i=1}^{n} \underbrace{2(Y_i - (b_0 + b_1 X_i))}_{\partial g_1} \cdot \underbrace{1}_{\partial g_2} = 0$$

$$\nabla_{b_1} f(b_0, b_1) = -\frac{1}{n}\sum_{i=1}^{n} \underbrace{2(Y_i - (b_0 + b_1 X_i))}_{\partial g_1} \cdot \underbrace{X_i}_{\partial g_2} = 0$$

linear Eq w.r.t. b_0, b_1

$$\nabla_{b_0} f = 0 \Rightarrow \frac{1}{n}\sum_{i=1}^{n} \Big(\underbrace{Y_i - (b_0 + b_1 x_i)}_{\text{recidual}} \Big) \cdot 1 = 0$$

The error of linear regression . on training data

$$\nabla_{b_1} f = 0 \Rightarrow \frac{1}{n}\sum_{i=1}^{n} \big(Y_i - (b_0 + b_1 x_i) \big) \cdot x_i = 0$$

① The residual/error on training data. is mean zero!

$\mathrm{Cov}(X \cdot Y) = \frac{1}{n}\sum_{i=1}^{n} x_i \cdot Y_i$

② The residual/error on training data. is independent to the data!

$$b_0 = \frac{1}{n}\sum_{i=1}^{n} (Y_i - b_1 x_i) = \bar{Y} - b_1 \bar{x} \qquad (\Delta)$$

$$\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i, \quad \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Plug (Δ) into $\nabla_{b_1} f = 0$

$$\frac{1}{n}\sum_{i=1}^{n} \big(Y_i - (\bar{Y} - b_1 \bar{x}) - b_1 x_i \big) x_i = 0 .$$

$$\Rightarrow \frac{1}{n}\sum_{i=1}^{n} \Big((Y_i - \bar{Y}) - b_1 (x_i - \bar{x}) \Big) x_i = 0 \qquad (\ast)$$

This is using $((x_i - \bar{x}), (Y_i - \bar{Y}))$ as dataset to fit the cimple linear regression.

Computing Eq $(\ast)$

$$\frac{1}{n}\sum_{i=1}^{n} x_i (Y_i - \bar{Y}) - \frac{1}{n}\sum_{i=1}^{n} x_i (x_i - \bar{x})\, b_1 = 0$$

$$-\bar{x} \cdot \Big(\frac{1}{n}\sum_{i=1}^{n} (Y_i - \bar{Y})\Big) = 0 \qquad -\bar{x} b_1 \Big(\frac{1}{n}\sum_{i=1}^{n} (x_i - \bar{x})\Big) = 0$$

$$\Rightarrow \frac{1}{n}\sum_{i=1}^{n} (x_i - \bar{x})(Y_i - \bar{Y}) - \frac{1}{n}\sum_{i=1}^{n} (x_i - \bar{x})(x_i - \bar{x})\, b_1 = 0$$

Find β_0, β_1

$$\hat{\beta}_1 = \frac{c_{XY}}{s_X^2}, = \frac{\text{Covariance }(X, Y)}{\text{Covariance }(X, x)}$$

where c_{XY}, s_X^2 are the sample covariance between X, Y and the sample variance of X respectively. As a reminder,

Covariance (X, Y) Var (X), Covariance(X, x)

$$c_{XY} = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{x})(Y_i - \overline{y}), s_X^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{x})^2.$$

$$0 = \overline{xy} - (\overline{y} - \hat{\beta}_1\overline{x})\overline{x} - \hat{\beta}_1\overline{x^2}$$

$$0 = c_{XY} - \hat{\beta}_1 s_X^2$$

How accurate is the Model?– Bias

$$\frac{Cov(X,Y)}{Cov(X,X)}$$

$$\|$$

only depend on the input \

$$\hat{\beta}_1 = \beta_1 + \frac{1}{ns_X^2}\sum_{i=1}^{n}(X_i - \overline{x})\boxed{\varepsilon_i}. \rightarrow \varepsilon_i \text{ is } \mathbb{E}[\varepsilon_i] = 0.$$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \bar{X})(X_i - \bar{X})}$$

$$\bar{Y} = \mathbb{E}(\beta_0 + \beta_1 X + \overset{\text{mean } 0}{\varepsilon}) = \beta_0 + \beta_1 \bar{X}$$

$$= \frac{\sum_{i=1}^{n}(X_i - \bar{X})\left((\cancel{\beta_0} + \beta_1 X_i + \varepsilon_i) - (\cancel{\beta_0} + \beta_1 \bar{X})\right)}{\sum_{i=1}^{n}(X_i - \bar{X})(X_i - \bar{X})} = \beta_1 + \frac{\sum_{i=1}^{n}(X_i - \bar{X})\varepsilon_i}{\sum_{i=1}^{n}(X_i - \bar{X})(X_i - \bar{X})}$$

Statement: $\hat{\beta}_1$ is unbiased, i.e. $\mathbb{E}[\hat{\beta}_1] = \beta_1$.

Model Fitting

❒ Find $(\hat{\beta}_0, \hat{\beta}_1)$ that minimize the least square

$$Q = \sum_{i=1}^{n} (y_i - \underbrace{(\hat{\beta}_0 + \hat{\beta}_1 x_i)}_{\hat{y}_i})^2.$$

- Denote $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ as the **fitted value**;
- Denote $e_i = y_i - \hat{y}_i$ as the **residual**.

Therefore, minimizing the least square can be understood as fitting y_i's to minimize residuals as good as possible.

How accurate is the Model?– Variance

$$\mathrm{Var}(\hat{\beta}_1) = \mathrm{Var}\left(\beta_1 + \frac{1}{ns_X^2}\sum_{i=1}^{n}(X_i - \overline{x})\varepsilon_i\right) = \frac{\sigma^2}{ns_X^2}.$$

Unconditioning on X

- ❐ Bias apply the law of total expectation:

$$\mathbb{E}[\hat{\beta}_1] = \mathbb{E}\Big[\mathbb{E}[\hat{\beta}_1 \mid X_1, \ldots, X_n]\Big] = \mathbb{E}[\beta_1] = \beta_1.$$

- ❐ Variance apply the law of total variance:

$$\begin{aligned}\mathrm{Var}(\hat{\beta}_1) &= \mathbb{E}\Big[\mathrm{Var}(\hat{\beta}_1 \mid X_1, \ldots, X_n)\Big] + \mathrm{Var}\Big(\mathbb{E}[\hat{\beta}_1 \mid X_1, \ldots, X_n]\Big) \\ &= \mathbb{E}\left[\frac{\sigma^2}{ns_X^2}\right] + \mathrm{Var}(\beta_1) = \frac{\sigma^2}{n}\mathbb{E}\left[\frac{1}{s_X^2}\right].\end{aligned}$$

Go Beyond Point Estimation

Fact. $\mathbb{E}[\hat{f}(x)] = \beta_0 + \beta_1 x$. and $\mathrm{Var}(\hat{f}(x)) = \frac{\sigma^2}{n}\left(1 + \frac{(x-\overline{x})^2}{s_X^2}\right)$.

What is the the standard error of an estimator ? $\mathrm{se}(\hat{\beta}_1) = \frac{\sigma}{\sqrt{n s_X^2}}$.

Exercise

- What happens when the noise variance, σ^2, increases?

Bias = 0   Var ↑ $\hat{\beta}_1$ is worse.

- What happens when the number of samples, n, increases? $\left(\text{Std}(\hat{\beta}_1) \propto \frac{1}{\sqrt{n}}\right)$

Bias = 0 $\text{Var} = \frac{1}{n} \times \boxed{\cdots}$ ↓ $\hat{\beta}_1$ become better.

- What influences the variance of our predictions?

- What happens when we predict at x that is very close to $\bar{x}$? How about very far?

$$\frac{\cdots}{\sum (x_i - \bar{x})(x_i - \bar{x})}$$

↑ big Var

How to Estimate σ?

Using the simple linear regression model,

$$\mathbb{E}[(Y - (\beta_0 + \beta_1 X))^2] = \sigma^2. \quad \textit{(convince yourself why.)}$$

Then, a natural estimator for σ^2 would be

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{f}(X_i))^2.$$

Notice that this is a biased estimator. Moreover $s^2 = \frac{n}{n-2}\hat{\sigma}^2$ is an unbiased estimator of σ^2. (Later)

⇓

Next Lecture.

Residual and Error

$$\text{(residual)} \quad e_i = Y_i - (\hat{\beta}_0 + \hat{\beta}_1 X_i)$$

$$\text{(noise)} \quad \varepsilon_i = Y_i - (\beta_0 + \beta_1 X_i)$$

Remark

- The sum of noise variables cannot equal zero all the time, because $\mathrm{Var}(\sum_{i=1}^{n} \varepsilon_i) = n\sigma^2$.
- The sum of residuals is *always* zero, i.e. $\sum_{i=1}^{n} e_i = 0$.
- The sample correlation between the residuals and X_i's is also 0, i.e. $\sum_{i=1}^{n}(X_i - \overline{x})e_i = 0$.

Assessing the Fit

Assessing the Fit

- ❒ As in simple regression, we calculate
  - fitted values: $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$;
  - residuals: $e_i = y_i - \hat{y}_i$;
  - error sum of squares: $\mathrm{SSE} = \sum_{i=1}^n e_i^2$;
  - total sum of squares: $\mathrm{SST} = \sum_{i=1}^n (y_i - \bar{y})^2$;
  - regression sum of squares: $\mathrm{SSR} = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2$.

$\bar{y} = \arg\min_c \sum_{i=1}^n (c - y_i)^2$ is the best constant fit of $\{y_i\}_{i=1}^n$!

- ❒ We can decompose SST as

$$\underbrace{\sum_{i=1}^n (y_i - \bar{y})^2}_{\mathrm{SST}} = \underbrace{\sum_{i=1}^n (\hat{y}_i - \bar{y})^2}_{\mathrm{SSR}} + \underbrace{\sum_{i=1}^n (y_i - \hat{y}_i)^2}_{\mathrm{SSE}}$$

R^2 Statistics and Correlation

R^2 (Coefficient of Determination):

$$R^2 = \frac{\text{SSR}}{\text{SST}}, \quad \text{where} \quad \text{SSR} = \sum (\hat{y}_i - \bar{y})^2, \quad \text{SST} = \sum (y_i - \bar{y})^2.$$

Theorem

Recall Pearson correlation coefficient: $r = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \sum(y_i - \bar{y})^2}}$, *then we have*

$$R^2 = r^2$$

Skip NOT REQUIRED

Prove $R^2 = r^2$

Since$\hat{\beta}_1 = \frac{\sum(x_i-\bar{x})(y_i-\bar{y})}{\sum(x_i-\bar{x})^2} = r\,\frac{s_y}{s_x}$, we have SSR $= \frac{(\sum(x_i-\bar{x})(y_i-\bar{y}))^2}{\sum(x_i-\bar{x})^2}$. Thus,

$$R^2 = \frac{\text{SSR}}{\text{SST}} = \frac{\left(\sum(x_i-\bar{x})(y_i-\bar{y})\right)^2}{\sum(x_i-\bar{x})^2\,\sum(y_i-\bar{y})^2} = r^2.$$

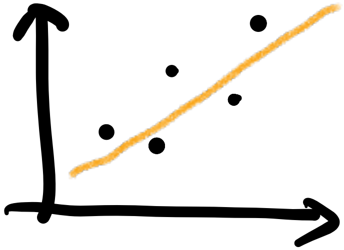
Skip NOT REQUIRED

Error

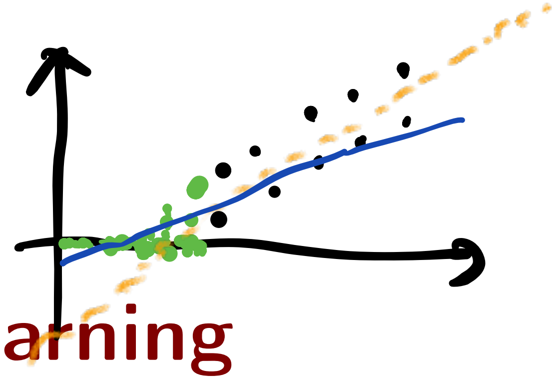
Prove: $s^2 = \frac{n}{n-2}\hat{\sigma}^2$ is an *unbiased* estimator of σ^2

SKIP NOT REQUIRED

$$Y = b_0 + b_1 X + \varepsilon$$



$$Y = \max(b_0 . + b_1 X + \varepsilon, 0)$$



Pipeline of Machine Learning

Log-Likelihood

The model looks similar,

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i,$$

with modified assumptions:

- ❒ X has an arbitrary distribution, possibly deterministic.
- ❒ If $X = x$, then $Y = \beta_0 + \beta_1 x + \varepsilon$, with β_0, β_1 being the coefficients, and ε being the noise variable.
- ❒ (stronger) $\varepsilon \sim N(0, \sigma^2)$, and is independent of X.
- ❒ (stronger) ε is *independent* across observations.

$$\mathbb{P}(\varepsilon_1 = \cdots, \varepsilon_2 = xxx) = \mathbb{P}(\varepsilon_1 = \cdots) \cdot \mathbb{P}(\varepsilon_2 = xxx)$$

Question. What is $p(Y_i | X_i; b_0, b_1, s^2)$? $Y_i = b_0 + b_1 X_i + \varepsilon_i$, $(\varepsilon_i \sim N(0, s^2))$

observes a data (X_i, Y_i) $\varepsilon_i = (\overbrace{Y_i - b_0 - b_1 X_i}^{\text{residual}})$ means $P(\varepsilon_i) = \frac{1}{\sqrt{2\pi s^2}} \exp\left\{\frac{-1}{2s^2} \cdot \underset{\text{residual}}{\varepsilon^2}\right\}$

what is the probability that Y_i is the value I observe?

Log-Likelihood

max likelihood $\iff$ minimize for $(\text{residual})^2$

Given the data, the likelihood under this set of assumption is a function of the unknown parameters, defined as

is the probability that Y_i is the value I observed

$$L(b_0, b_1, s^2) = \prod_{i=1}^{n} p(Y_i|X_i; b_0, b_1, s^2) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi s^2}} \exp\left\{-\frac{1}{2s^2}(Y_i - (b_0 + b_1 X_i))^2\right\}.$$

☆ exp(negative constant $(\text{residual})^2$)

$\exp(a)\exp(b) = \exp(a+b)$

likelihood of second data

$$\log(ab) = \log(a) + \log(b)$$

likelihood of first data

$$\log L(b_0, b_1, s^2) \overset{\text{def}}{=} \ell(b_0, b_1, s^2) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log s^2 - \frac{1}{2s^2}(Y_i - (b_0 + b_1 X_i))^2.$$

max $\sum_{\text{all data}}$ log of the likelihood of each data $\iff$ min $\sum_{\text{all data}}$ − log of likelihood

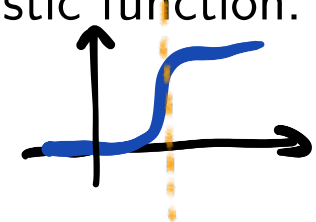
Logistic regression

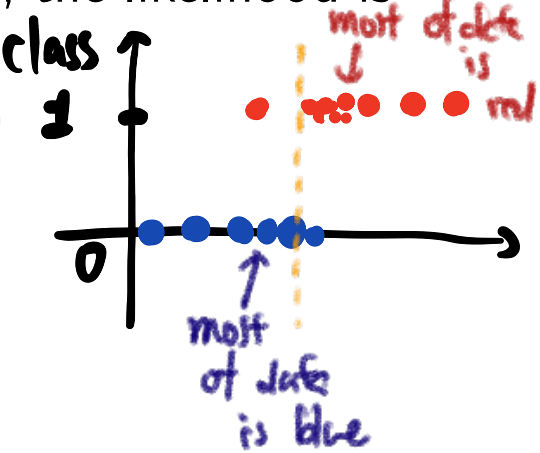
Step 1. Likelihood for a Logistic Binary Outcome:

For each observation $y_i \in \{0,1\}$ with probability p_i for $y_i = 1$, the likelihood is

$$L(y_i \mid p_i) = p_i^{y_i}(1-p_i)^{1-y_i} = \begin{cases} p_i & Y_i = 1 \\ 1-p_i & Y_i = 0 \end{cases}$$

where probability $p_i = \frac{1}{1+e^{-\beta^T x_i}}$ using the logistic function.





Step 2. Log-Likelihood: $P_i = \frac{1}{1+e^{-(b_0+b_1 x)}}$

For n independent observations, the log-likelihood function is

$$\ell(\beta) = \sum_{i=1}^{n} \left[y_i \log\left(\frac{1}{1+e^{-\beta^T x_i}}\right) + (1-y_i)\log\left(1 - \frac{1}{1+e^{-\beta^T x_i}}\right)\right].$$

Step 3. Estimation:

Maximizing $\ell(\beta)$ with respect to β gives the maximum likelihood estimates, leading to the logistic regression model.

☹ No closed-form solution.

Basic Idea: max $\underbrace{\mathbb{P}(Y|x)}_{\text{called likelihood}}$

Idea 1. $\max \mathbb{P}(Y|x) \iff \min -\log \mathbb{P}(Y|x)$

Idea 2. $\mathbb{P}(Y|x) = \prod_{i=1}^{n} \underbrace{\mathbb{P}(Y_i|X_i)}_{\text{because every data is independent /experience}}$

Fact. $\log(ab) = \log(a) + \log(b)$, $\log\left(\prod_{i=1}^{n} P_i\right) = \sum_{i=1}^{n} \log(P_i)$

Then $\max \mathbb{P}(Y|X)$

$$\iff \min -\log \mathbb{P}(Y|X) = \min -\log\left(\prod_{i=1}^{n} P(Y_i|X_i)\right)$$

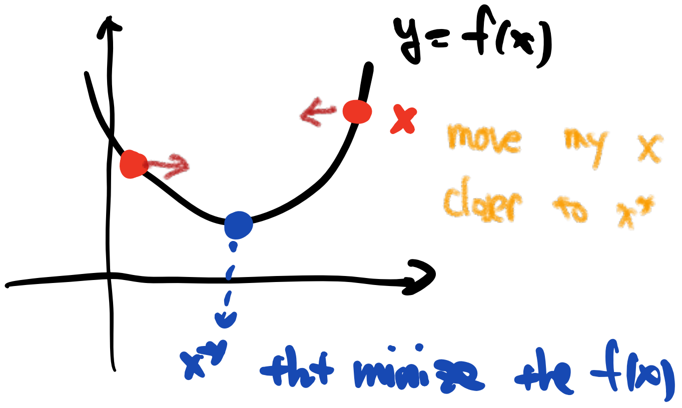
$$= \min \sum_{i=1}^{n} \underbrace{-\log(\mathbb{P}(Y_i|X_i))}_{\text{loss function for data } (X_i, Y_i)}$$

Example. Assume

$Y_i = f(X_i) + \varepsilon_i$, ε_i is Gaussian

$\Rightarrow -\log(\mathbb{P}(Y_i|X_i)) = (Y_i - f(X_i))^2$

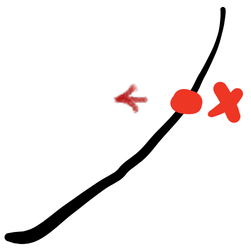
Optimization: Iterative Procedure to minimize a function



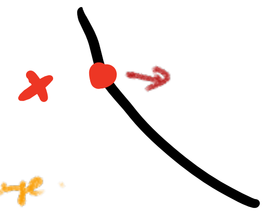
move my x closer to x*

x* that minimize the f(x)

Case1 $\nabla f(x) > 0$



Case2 $\nabla f(x) < 0$



t is the time of iterative procedure

$$x_{t+1} = x_t - \alpha \nabla f(x_t)$$

if $\alpha<0$, the function value become large

here $\alpha > 0$ is called learning rate/step size.

∇f means for two dimensional function $f(x)$, $x \in \mathbb{R}^2$

$$\nabla f(x) = \begin{bmatrix} df/dx_1 \\ df/dx_2 \end{bmatrix} \in \mathbb{R}^2$$

It means taylor expansion holds:

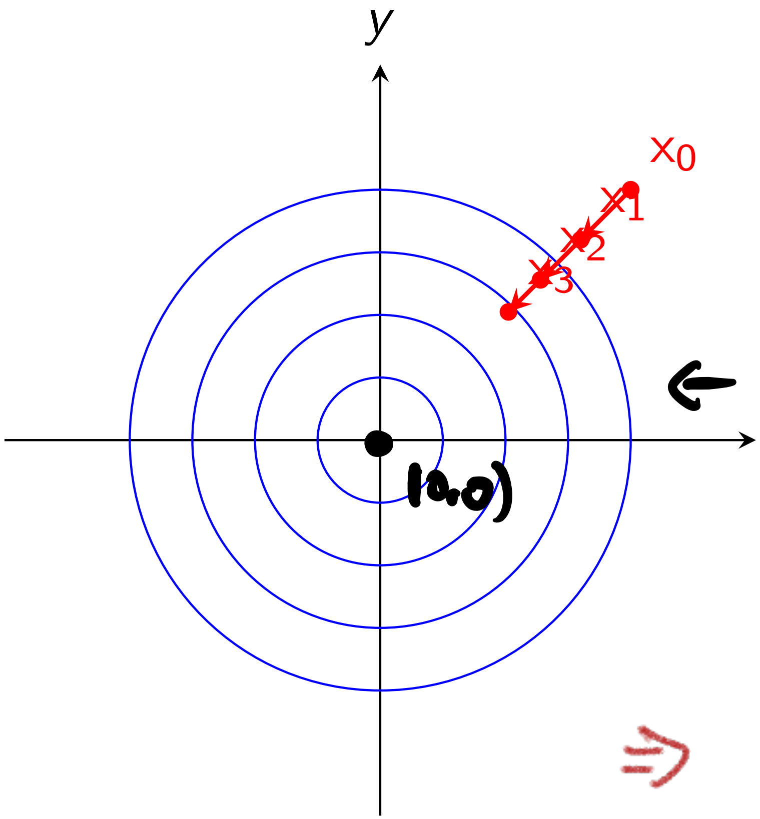
$$f(z) \approx f(y) + \nabla f(y) \cdot (z - y)$$

If f is decaying faster in x_1 direction, then $z_1 - y_1$ is larger

$$\begin{bmatrix} \frac{df}{dx_1} \\ \frac{df}{dx_2} \end{bmatrix} \begin{bmatrix} z_1 - y_1 \\ z_2 - y_2 \end{bmatrix} = \frac{df}{dx_1} \cdot \overbrace{(z_1 - y_1)}^{= -\alpha \frac{df}{dx_1}} + \frac{df}{dx_1} \underbrace{(z_2 - y_2)}_{= -\alpha \frac{df}{dx_2}}$$

Gradient Descent

- **Gradient Descent** is an iterative optimization method to find local minima of a function.
- The update rule is $x_{n+1} = x_n - \alpha \nabla f(x_n)$, where α is the learning rate.



← minimizing $x^2 + y^2$

$$\nabla f(x,y) = \begin{bmatrix} 2x \\ 2y \end{bmatrix}$$

$$\Rightarrow \quad x \rightarrow x - 2\alpha x$$

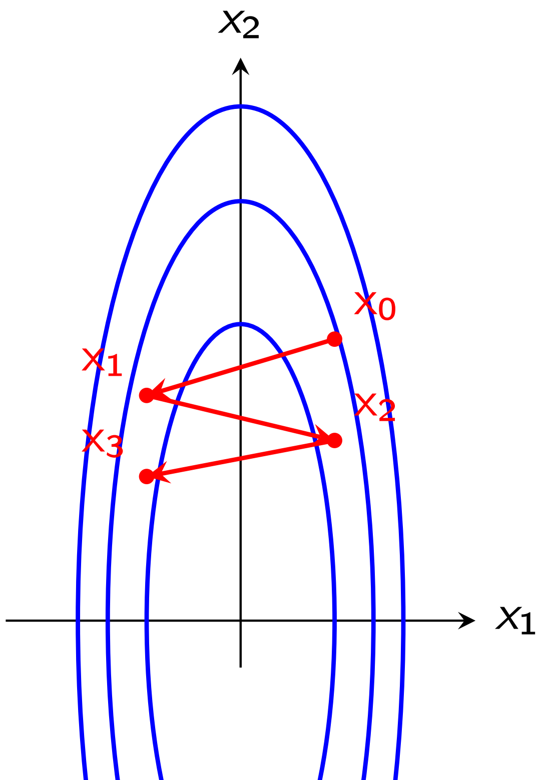
$$y \rightarrow y - 2\alpha y$$

Ill Conditioned Problems

- The function $f(x_1, x_2) = 10x_1^2 + x_2^2$ has very different curvatures along x_1 and x_2.
- Its level sets are ellipses elongated along the x_2-axis.
- With a fixed learning rate, gradient descent can overshoot in the steep x_1 direction, leading to oscillatory (zigzag) behavior.

$$H = \begin{bmatrix} 20 & 0 \\ 0 & 2 \end{bmatrix}$$

$$H^{-1} = \begin{bmatrix} \frac{1}{20} & \\ & \frac{1}{2} \end{bmatrix}$$

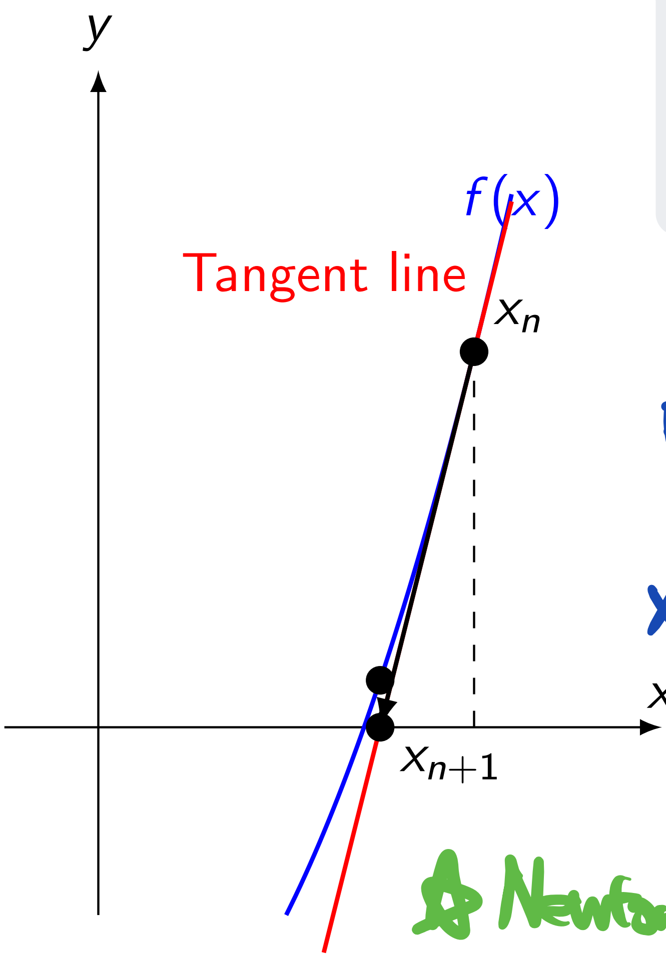


$$\nabla f(x,y) = \begin{bmatrix} 20x \\ 2y \end{bmatrix}$$

← scale is large

↑ scale is small

Newton Methods



Newton's method is an iterative technique for finding a root of a nonlinear equation $F(x) = 0$ via

$$x_{n+1} = x_n - F'(x_n)^{-1} F(x_n).$$

$$F(x) = 0$$

$$\Updownarrow$$

$$F(x_n) + F'(x_n) \cdot (x - x_n) = 0$$

$\Downarrow$ Solve linear approximation

$$x_{n+1} = x_n - F'(x_n)^{-1} F(x_n)$$

Newton's method.

In optimization

$\min f(x)$

$\Leftrightarrow$ Solve nonlinear Eq

$\nabla f(x) = 0$

$\Downarrow$

$$x_{n+1} = x_n - \underbrace{Hf^{-1}}_{\text{Hessian}^{-1}} \cdot \underbrace{\nabla f}_{\text{gradient}}$$

What happens if one optimize $f(x_1, x_2) = 10x_1^2 + x_2^2$?

☆ Newton will converge to the right solution in just one iteration!!

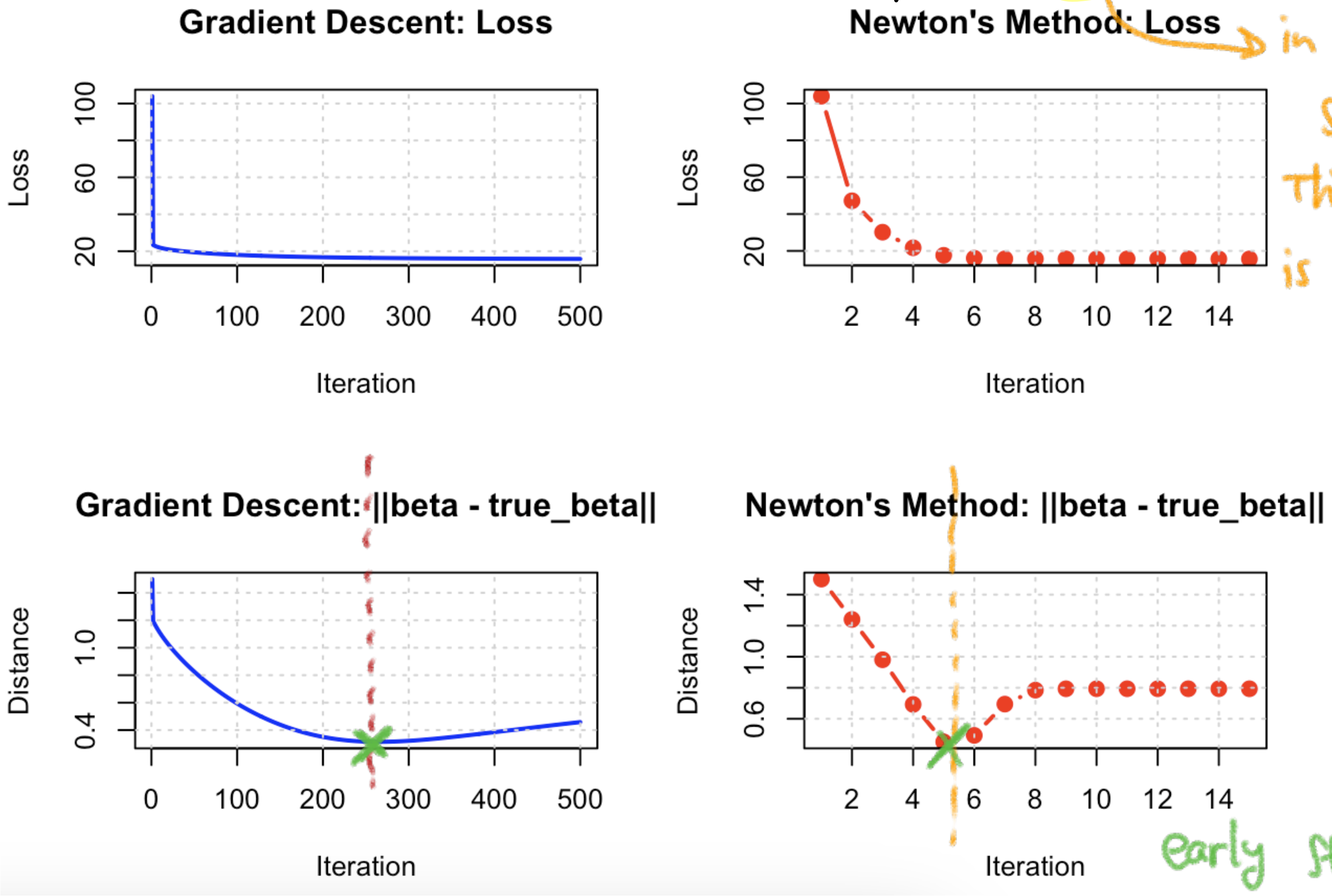
($f(x_1, x_2)$ is a quadratic function)

Homework

$x_{t+1} = x_t - H^{-1} \nabla f$

in large scale syste

This step is expensive



Gradient Descent: Loss

Newton's Method: Loss

Gradient Descent: ||beta - true_beta||

Newton's Method: ||beta - true_beta||

early stop

Pipeline of Machine Learning