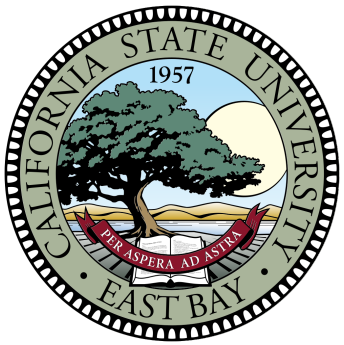


CS 6820 – Machine Learning

Lecture 6

Instructor: Eric S. Gayles, PhD.

Jan 22, 2018

Math Essentials

- Gradient is a vector
  - Each element is the slope of function along direction of one of variables
  - Each element is the partial derivative of function with respect to one of variables

$$\nabla f(\mathbf{x}) = \nabla f(x_1, x_2, \ldots, x_d) = \left[\frac{\partial f(\mathbf{x})}{\partial x_1} \quad \frac{\partial f(\mathbf{x})}{\partial x_2} \quad \cdots \quad \frac{\partial f(\mathbf{x})}{\partial x_d} \right]$$

  - Example:

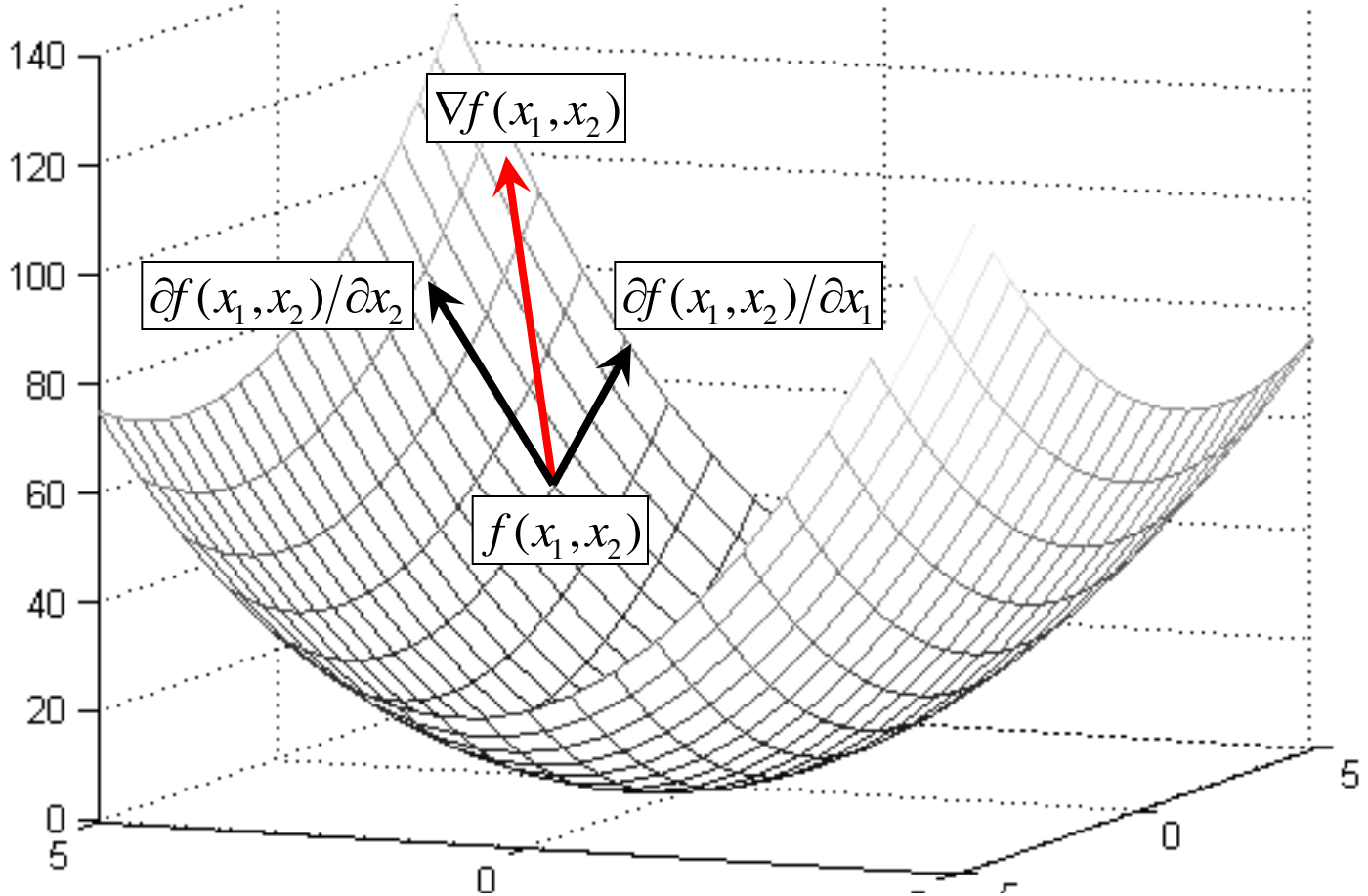
$$f(\mathbf{x}) = f(x_1, x_2) = {x_1}^2 + x_1 x_2 + 3{x_2}^2$$

$$\nabla f(\mathbf{x}) = \nabla f(x_1, x_2) = \left[\frac{\partial f(\mathbf{x})}{\partial x_1} \quad \frac{\partial f(\mathbf{x})}{\partial x_2} \right] = \left[2x_1 + x_2 \quad x_1 + 6x_2 \right]$$

* Howbert

Math Essentials

- Gradient vector points in direction of steepest ascent of function



Math Essentials

The **Hessian** matrix of $f : \mathbb{R}^d \to \mathbb{R}$ is a matrix of second-order partial derivatives:

$$\nabla^2 f = \begin{bmatrix} \frac{\partial^2 f}{\partial x_1^2} & \cdots & \frac{\partial^2 f}{\partial x_1 \partial x_d} \\ \vdots & \ddots & \vdots \\ \frac{\partial^2 f}{\partial x_d \partial x_1} & \cdots & \frac{\partial^2 f}{\partial x_d^2} \end{bmatrix} \quad \text{i.e.} \quad [\nabla^2 f]_{ij} = \frac{\partial^2 f}{\partial x_i \partial x_j}$$

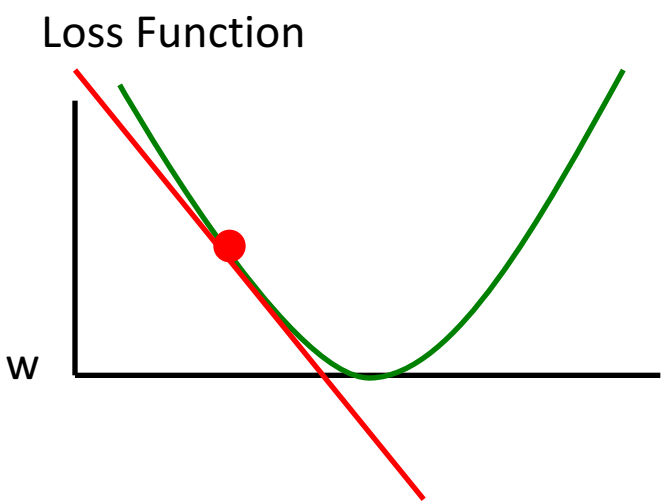
Basic Intro: Gradient Descent

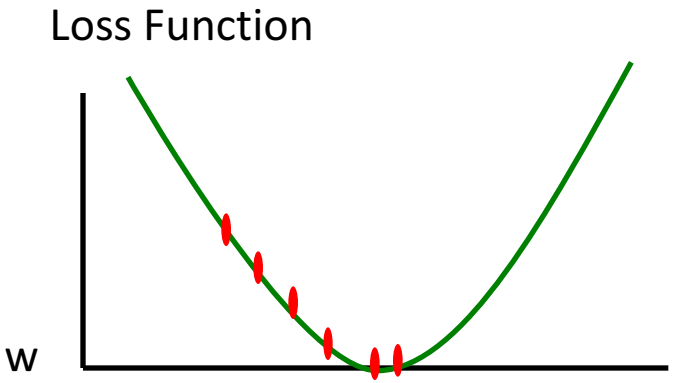
Partial derivatives provide the slope to direct the solution toward minimization of the loss function.

- Pick initial solution
- Repeat until optimized {
  - Evaluate slope and minima conditionality
  - Pick dimension(s)
  - Move a small amount (learning step) in the direction decreasing loss

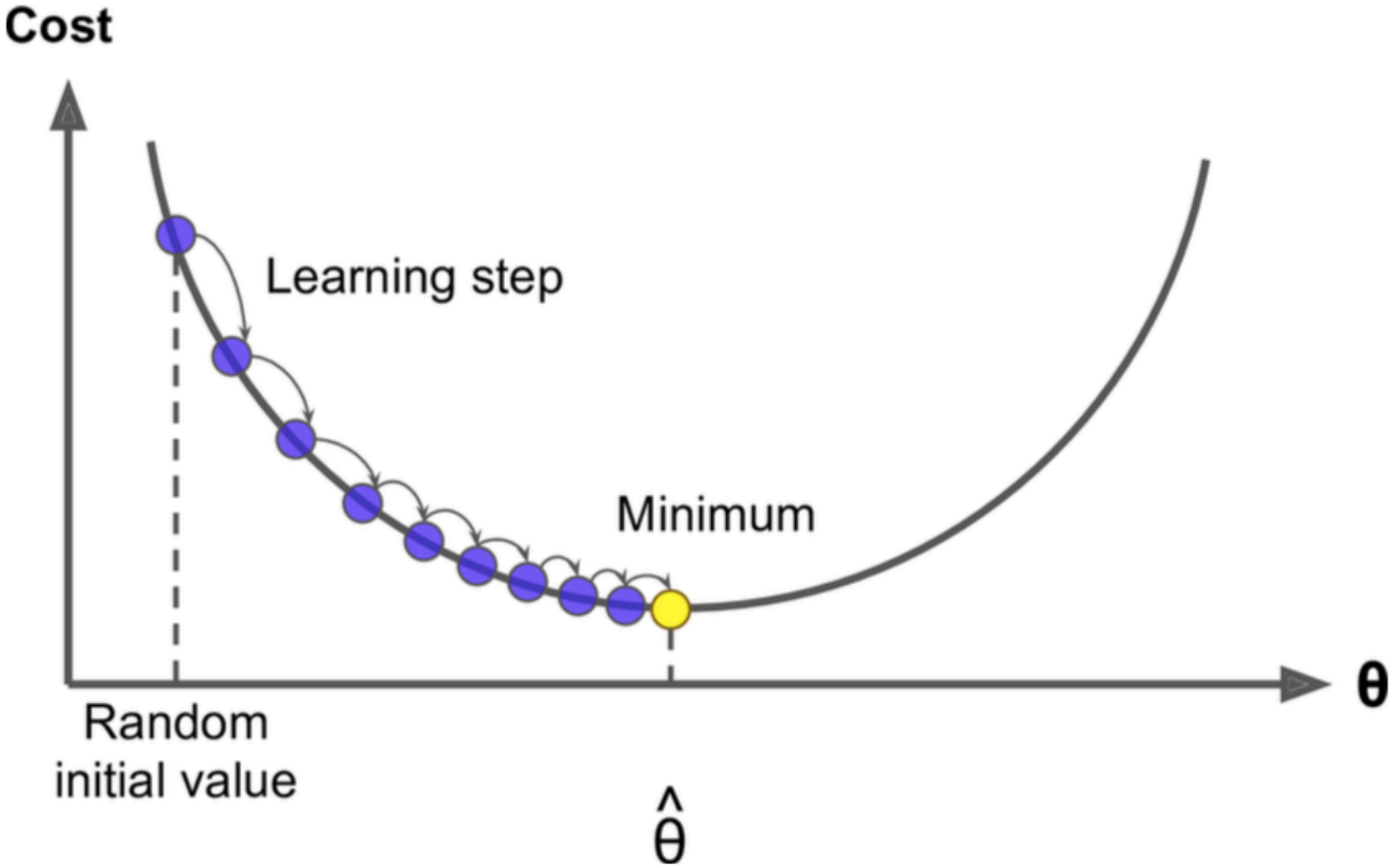
  }

$$\theta_1 := \theta_1 - \alpha \frac{\partial J}{\partial \theta_1}$$
$$\theta_2 := \theta_2 - \alpha \frac{\partial J}{\partial \theta_2}$$
$$\vdots$$
$$\theta_k := \theta_k - \alpha \frac{\partial J}{\partial \theta_k}$$

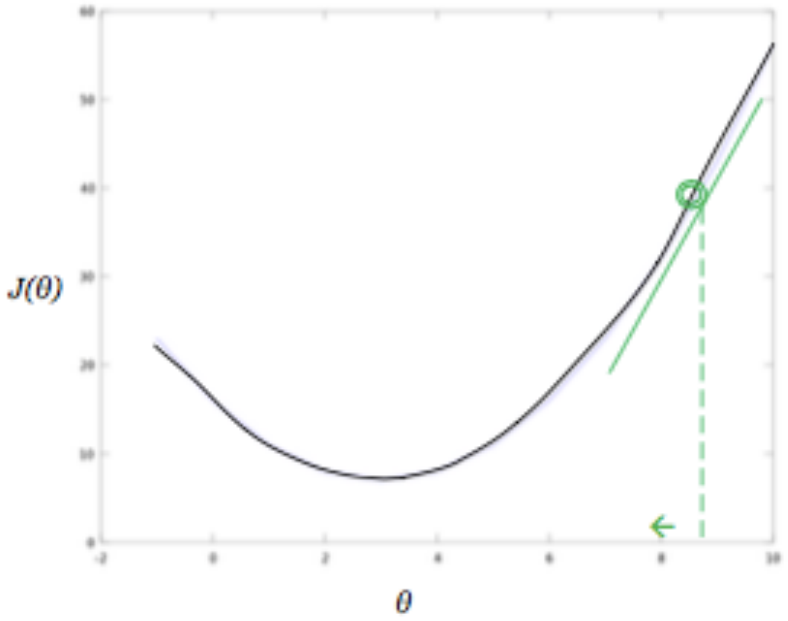


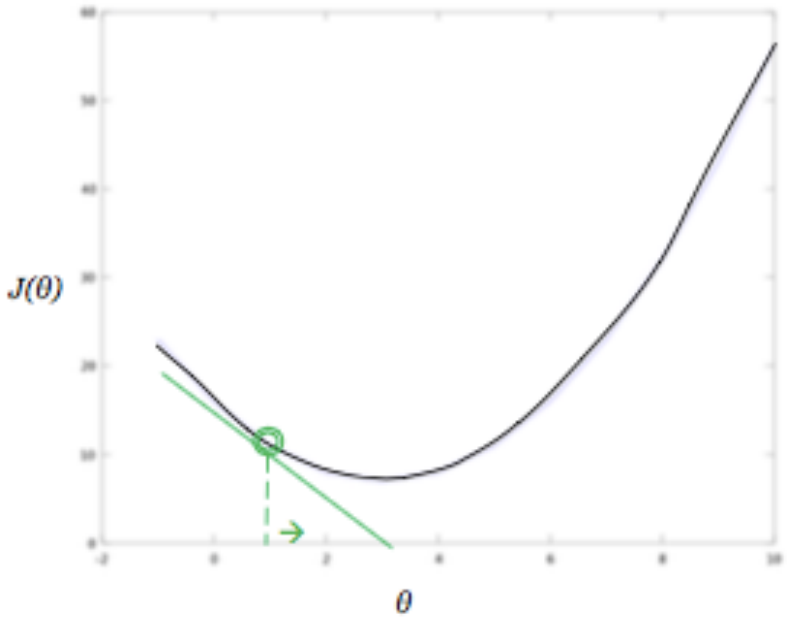


Basic Intro: Gradient Descent



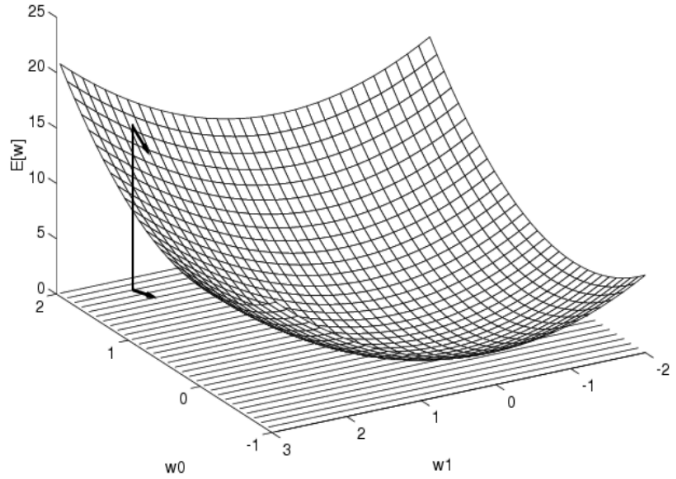
Basic Intro: Gradient Descent





* A Beginners Tutorial for Machine Learning Beginners, Hao

Basic Intro: Gradient Decent



Gradient

$$\nabla E[\vec{w}] \equiv \left[\frac{\partial E}{\partial w_0}, \frac{\partial E}{\partial w_1}, \cdots \frac{\partial E}{\partial w_n}\right]$$

Training rule:

$$\Delta \vec{w} = -\eta \nabla E[\vec{w}]$$

i.e.,

$$\Delta w_i = -\eta \frac{\partial E}{\partial w_i}$$

Basic Intro: Gradient Descent

- Simple concept: follow the gradient *downhill*
- Process:
  1. Pick a starting position: $\mathbf{x}^0 = (x_1, x_2, \ldots, x_d)$
  2. Determine the descent direction: $-\nabla f(\mathbf{x}^t)$
  3. Choose a learning rate: η
  4. Update your position: $\mathbf{x}^{t+1} = \mathbf{x}^t - \eta \cdot \nabla f(\mathbf{x}^t)$
  5. Repeat from 2) until stopping criterion is satisfied
- Typical stopping criteria
  - $\nabla f(\mathbf{x}^{t+1}) \sim 0$
  - some validation metric is optimized

* Howbert

Basic Intro: Gradient Descent

Batch gradient: use error $E_D(\mathbf{w})$ over entire training set D

Do until satisfied:

1. Compute the gradient $\nabla E_D(\mathbf{w}) = \left[\frac{\partial E_D(\mathbf{w})}{\partial w_0} \cdots \frac{\partial E_D(\mathbf{w})}{\partial w_n}\right]$
2. Update the vector of parameters: $\mathbf{w} \leftarrow \mathbf{w} - \eta \nabla E_D(\mathbf{w})$

Basic Intro: Gradient Descent

Stochastic gradient: use error $E_d(\mathbf{w})$ over single examples $d \in D$

Do until satisfied:

1. Choose (with replacement) a random training example $d \in D$
2. Compute the gradient just for d: $\nabla E_d(\mathbf{w}) = \left[\frac{\partial E_d(\mathbf{w})}{\partial w_0} \ldots \frac{\partial E_d(\mathbf{w})}{\partial w_n} \right]$
3. Update the vector of parameters: $\mathbf{w} \leftarrow \mathbf{w} - \eta \nabla E_d(\mathbf{w})$

Basic Intro: Gradient Descent

- An issue with Batch Gradient Descent is the fact that it uses the whole training set to compute the gradients at every step, which makes it very slow when the training set is large.

- An alternative is *Stochastic Gradient Descent, which* picks a random sample in the training set at every step and computes the gradients based only on that single instance.

Probability

- Probability statements describe the likelihood that particular values occur.

- The likelihood is quantified by assigning a number from the interval [0, 1] to the set of values (or a percentage from 0 to 100%).

- Higher numbers indicate that the set of values is more likely.

Math Essentials

A *probability space* is a *random process* or *experiment* with three components:

- Ω, the set of possible *outcomes O*
  - number of possible outcomes = $|\Omega| = N$
- *F*, the set of possible *events E*
  - an event comprises 0 to *N* outcomes
  - number of possible events = $|F| = 2^N$
- *P*, the *probability distribution*
  - function mapping each outcome and event to real number between 0 and 1 (the *probability* of *O* or *E*)
  - probability of an event is *sum* of probabilities of possible outcomes in event

Probability

- The probability of A or B is the sum of the individual probabilities:
  - P(A or B) = P(A) + P(B)
  - The probability of A or B equals the probability of A plus the probability of B"

Probability

- When two events (call them "A" and "B") are Mutually Exclusive it is impossible for them to happen together:
  - P(A and B) = 0
  - "The probability of A and B together equals 0 (impossible)"

Probability

- Turning left and turning right are Mutually Exclusive (you can't do both at the same time)
- Tossing a coin: Heads and Tails are Mutually Exclusive (can't happen at the same time)
- Cards: Kings and Aces are Mutually Exclusive (can’t be both)

Joint Probabilities

- Mutually Exclusive
  - Events A and B happening together is impossible:
    - P(A and B) = 0
    - A or B is the sum of A and B:
    - P(A or B) = P(A) + P(B)

- Non-Mutually Exclusive
  - A or B is the sum of A and B minus A and B:
  - P(A or B) = P(A) + P(B) – P(A and B)

Continuous Probability Distribution

- The **probability distribution** or simply **distribution** of a continuous random variable *X* is a description of the set of the probabilities associated with the possible values for *X*

$$P(a < X < b) = \int_a^b f(x)\,dx$$

The properties of the pdf are

(1) $f(x) \geq 0$

(2) $\int_{-\infty}^{\infty} f(x) = 1$

Continuous Probability Distribution

The **cumulative distribution function** (or cdf) of a continuous random variable X with probability density function $f(x)$ is

$$F(x) = P(X \leq x) = \int_{-\infty}^{x} f(u)\, du$$

for $-\infty < x < \infty$.

Continuous Probability Distribution

A random variable X with probability density function

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{\frac{-(x-\mu)^2}{2\sigma^2}} \quad \text{for} \quad -\infty < x < \infty \qquad (3\text{-}4)$$

has a **normal distribution** (and is called a **normal random variable**) with parameters μ and σ, where $-\infty < \mu < \infty$, and $\sigma > 0$. Also,

$$E(X) = \mu \quad \text{and} \quad V(X) = \sigma^2$$

Conditional Probability

- Conditional Probability contains a condition that may limit the sample space for an event.

- “the probability of event B, given event A”

Conditional Probability

The conditional probability that A occurs given that B has occurred is written as $P(A|B)$ and is defined as:

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

Conditional Probability

- Generically, if P(A | B) = P(A) then the events A and B are said to be independent.
- Conceptually: Giving B doesn't tell us anything about A.
- Mathematically: We know that if events A and B are independent, P(A and B) = P(A) x P(B).

$$P(A|B) = \frac{P(A \text{ and } B)}{P(B)} = \frac{P(A) \times P(B)}{P(B)} = P(A)$$

Bayes' Theorem

- Bayes' theorem will be key for much of the material to follow in this course.

$$P(B|A) = \frac{P(A|B) \cdot P(B)}{P(A)}.$$

Bayes' Theorem - Proof

$$P(A \cap B) = P(A)P(B|A) \tag{1}$$

On the other hand, the probability of A and B is also equal to the probability of B times the probability of A given B.

$$P(A \cap B) = P(B)P(A|B) \tag{2}$$

Equating the two yields:

$$P(B)P(A|B) = P(A)P(B|A) \tag{3}$$

and thus

$$P(A|B) = P(A)\frac{P(B|A)}{P(B)} \tag{4}$$

Math Essentials

The probability a discrete variable A takes value a is: $0 \leq P(A=a) \leq 1$ — Probability

Probabilities of alternative outcomes add: $P(A \in \{a, a'\}) = P(A=a) + P(A=a')$ — Alternatives

The probabilities of all outcomes must sum to one: $\sum_{\text{all possible } a} P(A=a) = 1$ — Normalization

$P(A=a, B=b)$ is the joint probability that both $A=a$ and $B=b$ occur. — Joint Probability

Variables can be "summed out" of joint distributions: — Marginalization

$$P(A=a) = \sum_{\text{all possible } b} P(A=a, B=b)$$

$P(A=a|B=b)$ is the probability $A=a$ occurs given the knowledge $B=b$. — Conditional Probability

$P(A=a, B=b) = P(A=a)\,P(B=b|A=a) = P(B=b)\,P(A=a|B=b)$ — Product Rule

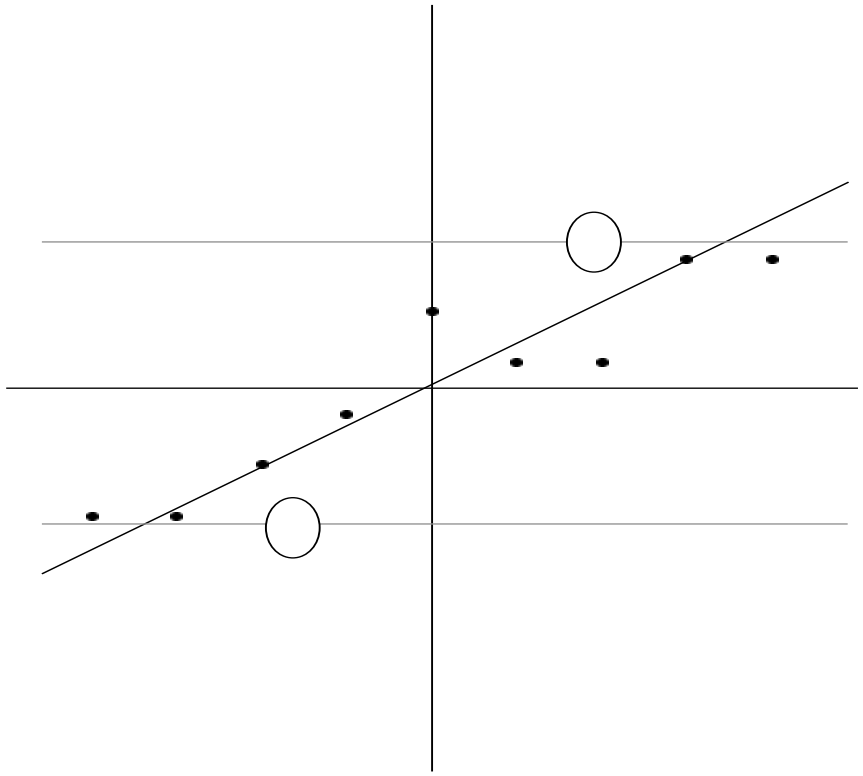
Bayes rule can be derived from the above: — Bayes Rule

$$P(A=a|B=b, \mathcal{H}) = \frac{P(B=b|A=a, \mathcal{H})\,P(A=a|\mathcal{H})}{P(B=b|\mathcal{H})} \propto P(A=a, B=b|\mathcal{H})$$

* Thanks to A Yu

What about Non Linear Phenomena?

From the graph, it appears that the dependent variable in our sample set has a non linear dependency on *x*.

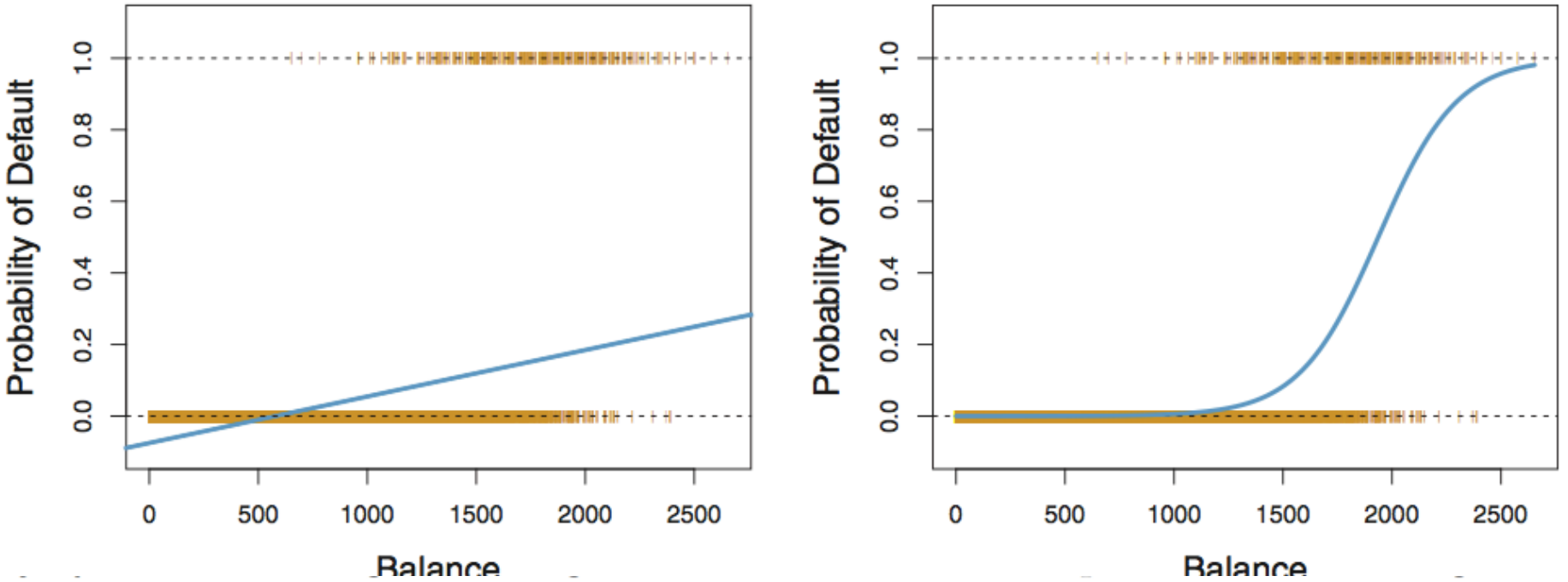


Logistic Regression

- Example from the Financial Services industry
  - Logistic regression models the probability of a credit card default.
  - For example, the probability of default given balance can be written as Pr(default = Yes|balance).
- The values of Pr(default = Yes|balance) will range between 0 and 1.
- For any given value of balance, a prediction can be made for default.
  - For example, which customers has p(balance) > 0.5.

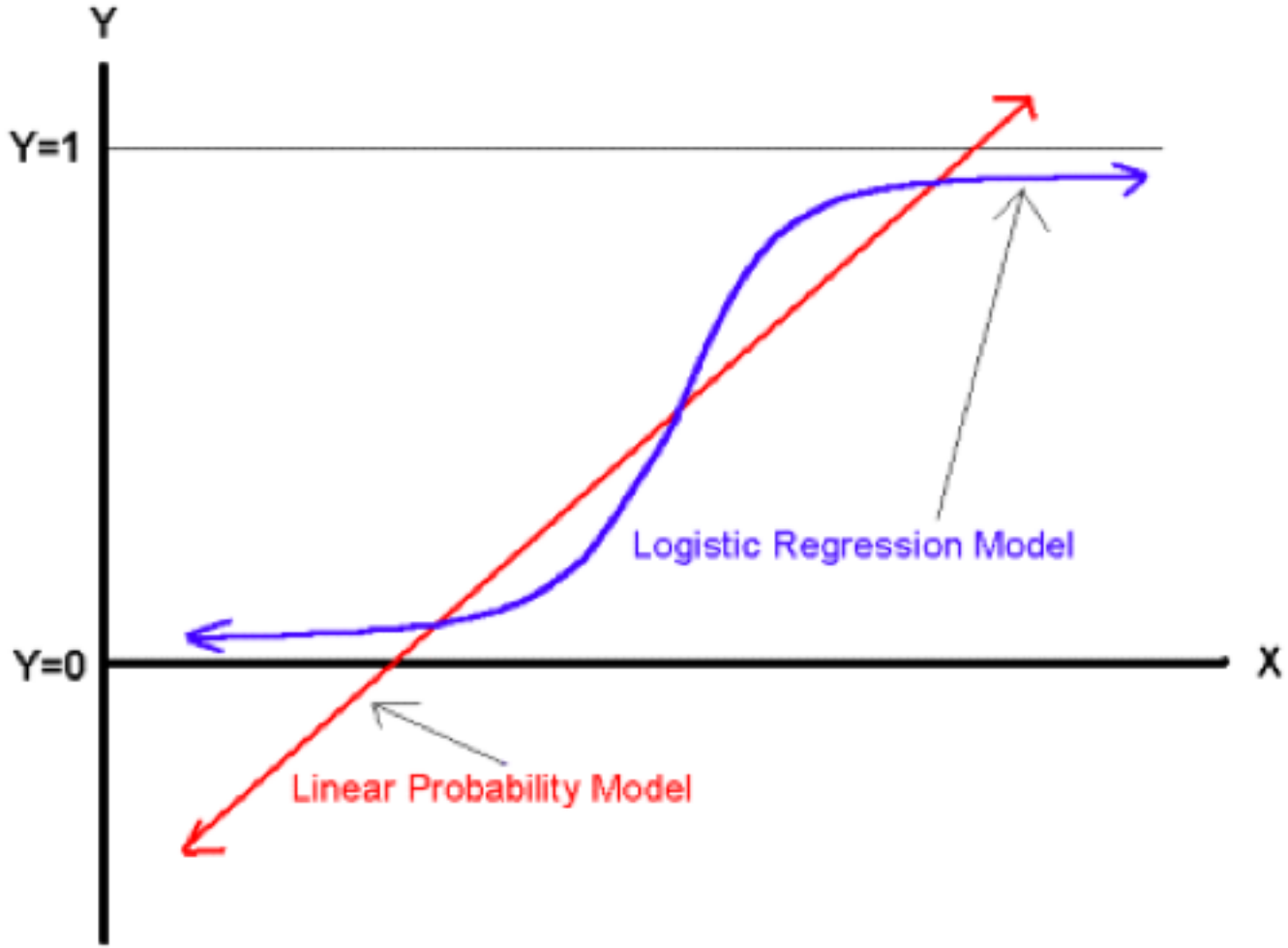
Logistic Regression

- Logistic Regression models the probability that a sample Y belongs to a particular category.



The orange ticks indicate the 0/1 values coded for `default` *(*`No` *or* `Yes`*).* Right: *Predicted probabilities of* `default` *using logistic regression. All probabilities lie between* 0 *and* 1.

Logistic Function



Logistic Function

$$p(X) = \frac{e^{\beta_0+\beta_1 X}}{1+e^{\beta_0+\beta_1 X}}$$

Logistic Regression

- Once the coefficients have been estimated, it is a simple matter to compute the probability of default for any given credit card balance.

- For example, using the following coefficient estimates we predict that the default probability for an individual with a balance of \$1, 000 is

$$\hat{p}(X) = \frac{e^{\hat{\beta}_0+\hat{\beta}_1 X}}{1+e^{\hat{\beta}_0+\hat{\beta}_1 X}} = \frac{e^{-10.6513+0.0055\times 1{,}000}}{1+e^{-10.6513+0.0055\times 1{,}000}} = 0.00576$$

Odds & Odds Ratios

- The definitions of an **odds**:

$$odds = \frac{p}{1-p}$$

- The odds has a range of 0 to ∞
- Values greater than 1 associated with an event are more likely to occur than to not occur.
- Values less than 1 associated with an event that is less likely to occur than not occur.

Logistic Regression

- For example, 1 in 5 people with an odds of 1/4 will default, since p(X) = 0.2 implies an odds of 0.2 = 1/4.

- Likewise on average nine out of every ten people with 1−0.2 an odds of 9 will default

$$\frac{p(X)}{1-p(X)} = e^{\beta_0+\beta_1 X}$$

Logistic Regression

- By taking the logarithm of both sides of we arrive at the equation below
- The left-hand side is called the log-odds or logit.
- The logistic regression model has a logit that is linear in X.
- In Linear Regression, β1 gives the average change in Y associated with a one-unit increase in X.
- In contrast, in a logistic regression model, increasing X by one unit changes the log odds by β1, or equivalently it multiplies the odds by $e^{\beta 1.}$

$$ln\left(\frac{p(X)}{1-p(X)}\right) = \beta_0 + \beta_1 X$$

Logistic Regression

- The coefficients β0 and β1 in are unknown, and must be estimated based on the available training data.
- Previously we used the least squares approach to estimate the unknown linear regression coefficients.
- We could use (non-linear) least squares to fit the model …
- However the the more general method of maximum likelihood is preferred, since it has better statistical properties.

Logistic Regression

- Basic intuition behind using maximum likelihood
- We seek estimates for β0 and β1 such that the predicted probability of default for each individual corresponds as closely as possible to the individual's observed default status.
- In other words, we try to find β_0 and β_1 such that plugging these estimates into the model for p(X), given yields a number close to one for all individuals who defaulted, and a number close to zero for all individuals who did not.
- This intuition can be formalized using a mathematical equation called a likelihood function:

Logistic Regression

- The estimates β_0 and β_1 are chosen to maximize this likelihood function.

$$\ell(\beta_0, \beta_1) = \prod_{i:y_i=1} p(x_i) \prod_{i':y_{i'}=0} (1 - p(x_{i'}))$$

Logistic Regression

- Consider the problem of predicting a binary response using multiple predictors. We can generalize as follows:

$$p(X) = \frac{e^{\beta_0+\beta_1 X_1+\cdots+\beta_p X_p}}{1+e^{\beta_0+\beta_1 X_1+\cdots+\beta_p X_p}}$$

$$ln\left(\frac{p(X)}{1-p(X)}\right) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$$