MLE AND LINEAR REGRESSION

MAXIMUM LIKELIHOOD ESTIMATION

MLE OF GAUSSIAN RANDOM VARIABLE

- Recall: Hessian method for two variables.
  How to find/determine max/min of f(x,y)?
- 1) we have two first partial derivative equations vanish
- 2) Rules for two variable Maximums and Minimums

1. Maximum

$$\begin{aligned} f_{xx} &< 0 \\ f_{yy} &< 0 \\ f_{yy}f_{xx} - f_{xy}f_{yx} &> 0 \end{aligned}$$

2. Minimum

$$\begin{aligned} f_{xx} &> 0 \\ f_{yy} &> 0 \\ f_{yy}f_{xx} - f_{xy}f_{yx} &> 0 \end{aligned}$$

3. Otherwise, we have a *Saddle Point*

MLE OF GAUSSIAN RANDOM VARIABLE

- The parameters of a Gaussian distribution are the mean (μ) and standard deviation (**σ**).
- Given observations x1, . . . , xN , the likelihood of those observations for a certain μ and **σ**
- what **λ** = (μ, **σ**) would give the best probability?

$$p(x_1, \ldots, x_N | \mu, \sigma^2) = \prod_{n=1}^{N} \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{ \frac{-(x_n - \mu)^2}{2\sigma^2} \right\}$$

- the log likelihood is

$$-\frac{1}{2} N \log(2\pi\sigma^2) - \sum_{n=1}^{N} \frac{(x_n - \mu)^2}{2\sigma^2}$$

MLE OF GAUSSIAN RANDOM VARIABLE

- Find (μ, **σ**) such that the following function is maximized

$$-\frac{1}{2}N\log(2\pi\sigma^2)-\sum_{n=1}^{N}\frac{(x_n-\mu)^2}{2\sigma^2}$$

MLE OF GAUSSIAN RANDOM VARIABLE

- Find (μ, **σ**) such that the following function is maximized

$$p(x_1, \ldots, x_N | \mu, \sigma^2) = \prod_{n=1}^{N} \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{ \frac{-(x_n - \mu)^2}{2\sigma^2} \right\}$$

- Note: : To maximize (minimize) a function of many variables you use the technique of partial differentiation, and Hessian

MLE OF GAUSSIAN RANDOM VARIABLE

- Find (μ, **σ**) such that the following function is maximized

$$p(x_1, \ldots, x_N | \mu, \sigma^2) = \prod_{n=1}^{N} \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{ \frac{-(x_n - \mu)^2}{2\sigma^2} \right\}$$

- Note: : To maximize (minimize) a function of many variables you use the technique of partial differentiation, and Hessian
- we find that the MLE of the mean is

$$\hat{\mu} = \frac{1}{N} \sum_{n=1}^{N} x_n$$

- the MLE of the variance is

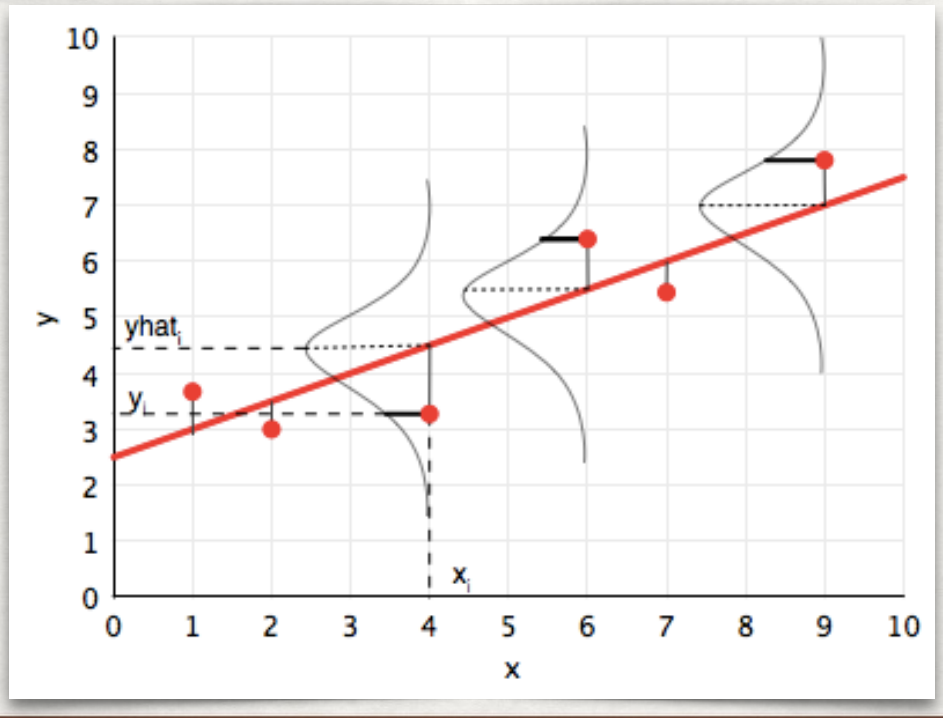
$$\hat{\sigma^2} = \frac{1}{N} \sum_{n=1}^{N} (x_n - \hat{\mu})^2$$

APPLICATION OF MLE: LINEAR REGRESSION MLE

Linear regression

- Linear regression is the workhorse algorithm that's used in many sciences, social and natural. The diagram below illustrates the probabilistic interpretation of linear regression.

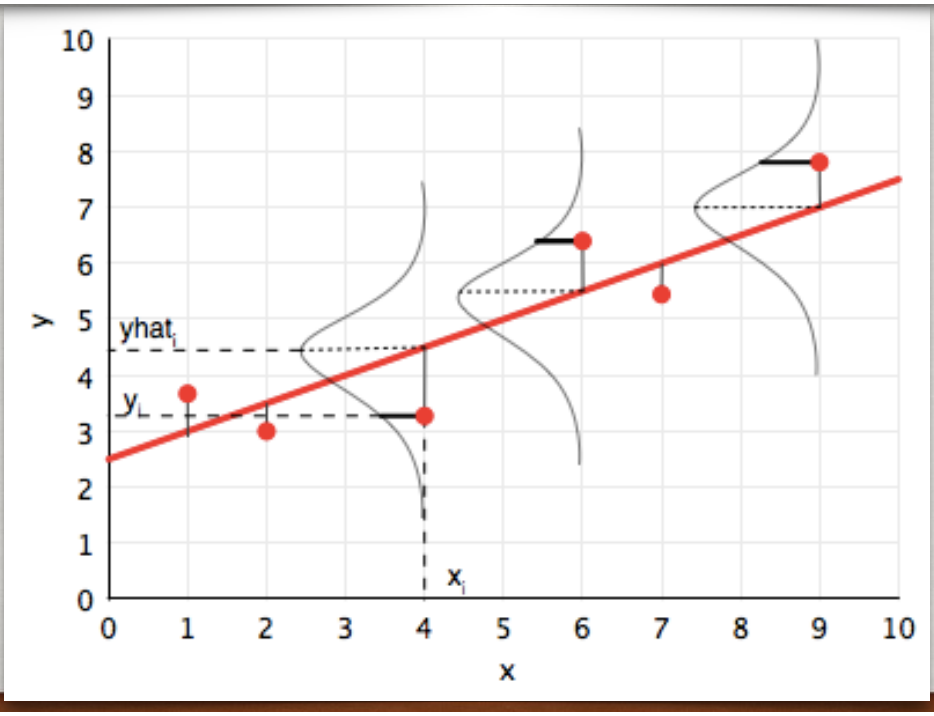
(x_i, y_i), and the corresponding prediction for x_i using the line, that is $yhat_i$ or $\hat{y}_i$.



- The fundamental assumption for the probabilistic analysis of linear regression is that
- Each xi is of m dimensional w = column vector = (w1,…, wm)

each y_i is gaussian distributed with mean $\mathbf{w} \cdot \mathbf{x}_i$ (the y predicted by the regression line so to speak) and variance σ^2:

$$y_i \sim N(\mathbf{w} \cdot \mathbf{x}_i, \sigma^2).$$



We can then write the likelihood:

$$\mathcal{L} = p(\mathbf{y}|\mathbf{x}, \mathbf{w}, \sigma) = \prod_i p(\mathbf{y}_i|\mathbf{x}_i, \mathbf{w}, \sigma)$$

Given the canonical form of the gaussian:

$$N(\mu, \sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(y-\mu)^2/2\sigma^2},$$

we can show that:

$$\mathcal{L} = (2\pi\sigma^2)^{(-n/2)} e^{\frac{-1}{2\sigma^2}\sum_i (y_i - \mathbf{w}\cdot\mathbf{x}_i)^2}.$$

The log likelihood ℓ then is given by:

$$\ell = \frac{-n}{2} log(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_i (y_i - \mathbf{w} \cdot \mathbf{x}_i)^2.$$

If you differentiate this with respect to $\mathbf{w}$ and σ, you get the MLE values of the parameter estimates:

$$\mathbf{w}_{MLE} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y},$$

where $\mathbf{X}$ is the design matrix created by stacking rows $\mathbf{x}_i$, and

$$\sigma^2_{MLE} = \frac{1}{n} \sum_i (y_i - \mathbf{w} \cdot \mathbf{x}_i)^2.$$

These are the standard results of linear regression.

Chapter 11 of Grinstead

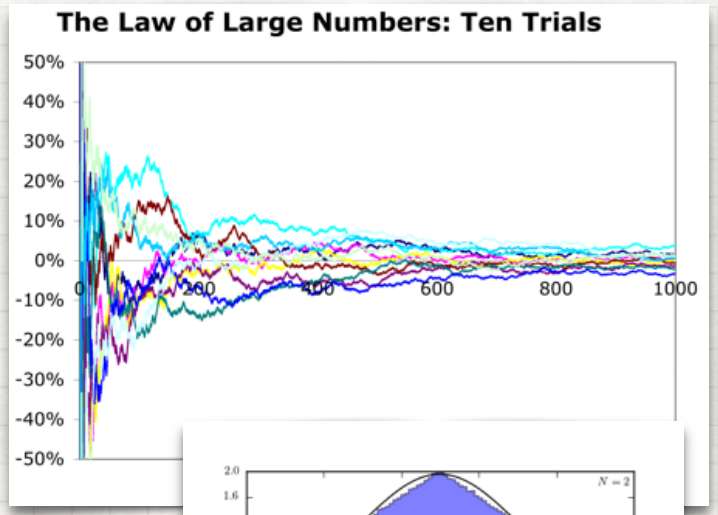
DISCRETE-TIME MARKOV CHAINS

- Topics
- State-transition matrix
- Network diagrams
- Examples: gambler's ruin
- Transient probabilities
- Steady-state probabilities

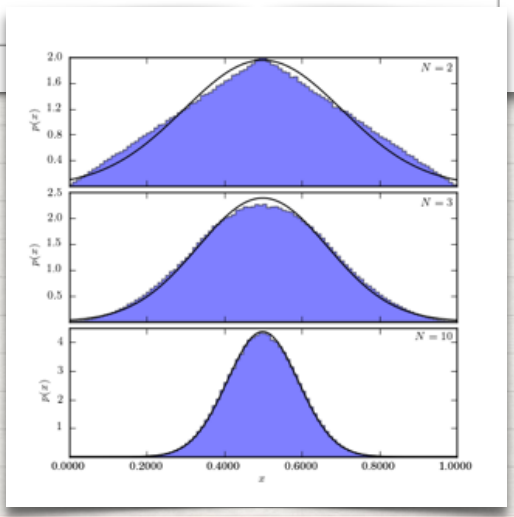
DISCRETE-TIME MARKOV CHAINS

INTRODUCTION

- Most of our study of probability has dealt with independent trials processes.
- These processes are the basis of classical probability theory
- We have discussed two of the principal theorems for these processes: the Law of Large Numbers and the Central Limit Theorem.
- **They all assume samples are IID**



central limit



DISCRETE-TIME MARKOV CHAINS

INTRODUCTION

- Central limit and Law of Large numbers work with IID:
  the previous experiments DO NOT influence our predictions for the outcomes of the next experiment. (coin flips)

- Modern probability theory studies chance processes for which the knowledge of previous outcomes DOES influence predictions for future experiments.

- In principle, when we observe a sequence of chance experiments, all of the past outcomes could influence our predictions for the next experiment. For example, this should be the case in predicting a student's grades on a sequence of exams in a course. But to allow this much generality would make it very difficult to prove general results.

DISCRETE-TIME MARKOV CHAINS

INTRODUCTION

- In principle, when we observe a sequence of chance experiments, all of the past outcomes could influence our predictions for the next experiment.
- For example, the temperature today is influenced by temperature yesterday.
- In 1907, Markov began the study of an important new type of chance process: the outcome of a given experiment can affect the outcome of the next experiment. This type of process is called a Markov chain.

DISCRETE-TIME MARKOV CHAINS

EXAMPLE

- According to weather history records for Boston, we have the following information. (Note, the record only shows nice days, snow days and rain days)
- Boston never has two nice days in a row, and we assume tomorrow's weather only depends on today's weather.
  - If they have a nice day, they are just as likely to have snow as rain the next day.
  - If they have snow or rain, they have an even chance of having the same the next day.
  - If there is change from snow or rain, only half of the time is this a change to a nice day.
- Mathematically, how to summary the above information?

DISCRETE-TIME MARKOV CHAINS

EXAMPLE

- Boston never has two nice days in a row.
  If they have a nice day, they are just as likely to have snow as rain the next day.
  If they have snow or rain, they have an even chance of having the same the next day.
  If there is change from snow or rain, only half of the time is this a change to a nice day.

$$\mathbf{P} = \begin{array}{c} \\ \mathrm{R} \\ \mathrm{N} \\ \mathrm{S} \end{array} \begin{array}{c} \begin{array}{ccc} \mathrm{R} & \mathrm{N} & \mathrm{S} \end{array} \\ \begin{pmatrix} 1/2 & 1/4 & 1/4 \\ 1/2 & 0 & 1/2 \\ 1/4 & 1/4 & 1/2 \end{pmatrix} \end{array} .$$

DISCRETE-TIME MARKOV CHAINS

EXAMPLE

- $$\mathbf{P} = \begin{array}{c} \\ \mathrm{R} \\ \mathrm{N} \\ \mathrm{S} \end{array} \begin{array}{c} \begin{array}{ccc} \mathrm{R} & \mathrm{N} & \mathrm{S} \end{array} \\ \begin{pmatrix} 1/2 & 1/4 & 1/4 \\ 1/2 & 0 & 1/2 \\ 1/4 & 1/4 & 1/2 \end{pmatrix} \end{array}.$$
- The entries in the first row of the matrix P represents the probabilities for the various kinds of weather following a rainy day.
- Similarly, the entries in the second and third rows represent the probabilities for the various kinds of weather following nice and snowy days, respectively.

DISCRETE-TIME MARKOV CHAINS

EXAMPLE

- This is an example of Discrete time Markov Chain, it consists of the following
- $$\mathbf{P} = \begin{array}{c} \\ \mathrm{R} \\ \mathrm{N} \\ \mathrm{S} \end{array} \begin{array}{c} \begin{array}{ccc} \mathrm{R} & \mathrm{N} & \mathrm{S} \end{array} \\ \begin{pmatrix} 1/2 & 1/4 & 1/4 \\ 1/2 & 0 & 1/2 \\ 1/4 & 1/4 & 1/2 \end{pmatrix} \end{array}.$$
- state space {R,N,S}
- transition among states: P
- P is called **the matrix of transition probabilities, or the transition matrix.**

DISCRETE-TIME MARKOV CHAINS

EXAMPLE

- This is an example of Discrete time Markov Chain, it consists of the following
-

$$\mathbf{P} = \begin{array}{c} \\ \mathrm{R} \\ \mathrm{N} \\ \mathrm{S} \end{array} \begin{array}{c} \begin{array}{ccc} \mathrm{R} & \mathrm{N} & \mathrm{S} \end{array} \\ \begin{pmatrix} 1/2 & 1/4 & 1/4 \\ 1/2 & 0 & 1/2 \\ 1/4 & 1/4 & 1/2 \end{pmatrix} \end{array}.$$

- The set of state (state space) is: {R,N,S}
  Time: t = {0, 1, 2, . . . day 0, day 1, day 2, ….}
- {Xn,} where n is in t ={0,1,2,3,…}— is a sequence of random variables — take values in the state space {R,N,S}, where the outcome of Xn+1 only depends on Xn,
- If day 0 rains , then the probability that day 1 will rain is: 1/2, will be nice is 1/4 and will snow is 1/4
- P(X1=R | X0 = R) = 1/2, P(X1=N | X0 = R) = 1/4…

DISCRETE-TIME MARKOV CHAINS

EXAMPLE

$$\mathbf{P} = \begin{array}{c} \\ \mathrm{R} \\ \mathrm{N} \\ \mathrm{S} \end{array} \begin{array}{c} \begin{array}{ccc} \mathrm{R} & \mathrm{N} & \mathrm{S} \end{array} \\ \begin{pmatrix} 1/2 & 1/4 & 1/4 \\ 1/2 & 0 & 1/2 \\ 1/4 & 1/4 & 1/2 \end{pmatrix} \end{array} .$$

- We say:
  P is the one-step transition matrix for a Markov chain {Xn} with states S = { 0, 1, 2 } (set R = 0, N = 1, S = 2)

$$\mathbf{P} = \begin{bmatrix} p_{00} & p_{01} & p_{02} \\ p_{10} & p_{11} & p_{12} \\ p_{20} & p_{21} & p_{22} \end{bmatrix}$$

where $p_{ij} = \Pr\{ X_1 = j \mid X_0 = i \}$

DISCRETE-TIME MARKOV CHAINS

A stochastic process $\{ X_n \}$ is called a Markov chain if

$$\Pr\{ X_{n+1} = j \mid X_0 = k_0, \ldots, X_{n-1} = k_{n-1}, X_n = i \}$$

$$= \Pr\{ X_{n+1} = j \mid X_n = i \} \quad \leftarrow \text{transition probabilities}$$

for every $i, j, k_0, \ldots, k_{n-1}$ and for every n.

- Discrete time means $n \in N = \{ 0, 1, 2, \ldots \}$.
- Markovian property means:
  The future behavior of the system depends only on the current state i and not on any of the previous states.
- eg: tomorrow's weather only depends on today, not any days before today.