CMSC 726
Lecture 3:Math Review

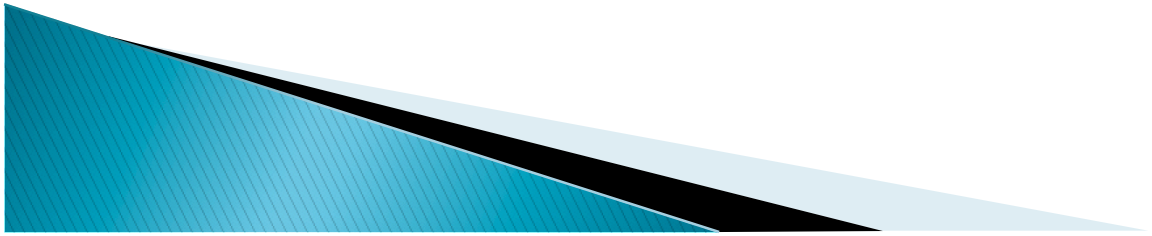
Lise Getoor

September 7, 2010

ACKNOWLEDGEMENTS: The material in this course is a synthesis of materials from many sources, including: Hal Daume III, Mark Drezde, Carlos Guestrin, Andrew Ng, Ben Taskar, **Eric Xing**, and others. I am very grateful for their generous sharing of insights and materials.

Today's Topics

- Why Math?
- Probability
- Other stuff
  - Calculus
  - Logarithms
  - Optimization
  - Linear Algebra (next time)
- Example
- Matlab Tutorial



Why do we care about math?

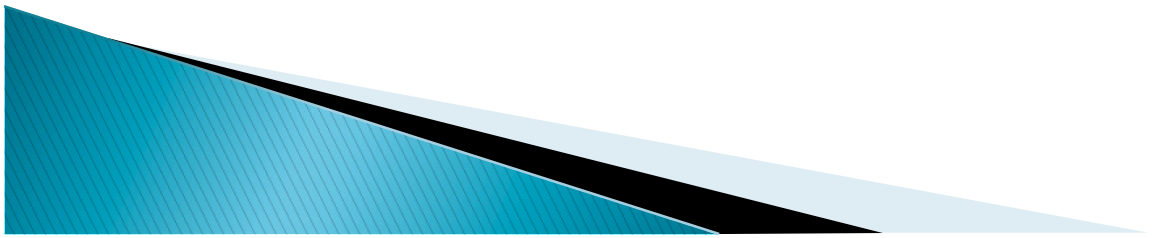
- **Probability:**
  - 1. The study of the outcome of repeated experiments
  - 2. The study of the plausibility of some event
- **Statistics:**
  - The analysis and interpretation of data
- **Calculus and linear algebra:**
  - 1. Techniques for finding maxima/minima of functions
  - 2. Convenient language for high dimensional data analysis
- **Logarithms and Exponents:**
  - Helps in dealing with small numbers (probabilities)
  - Often makes math easier (change product into sum)

Probability 101

- Probability theory assigns a numerical probability to events
- Probability of an event = *fraction of times* that event would occur if we ran an experiment many times
  - This is the *frequentist* definition of probability
- Events are drawn (they happen) from a sample space Ω (omega)
- A probability model is a function that maps any subset of Ω to a real value between 0 and 1
- Formally, $p : P(\Omega) \rightarrow [0, 1]$

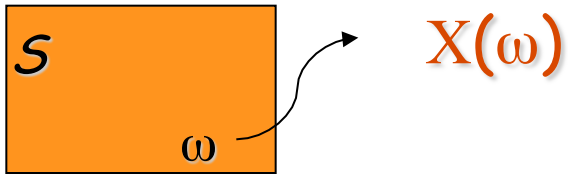
Axioms of Probability

- $p(E) \geq 0$ for all $E \subseteq \Omega$
  - Cannot have a negative event
- $p(\Omega) = 1$
  - An event must occur
- If $E_1, E_2, E_3, \cdot \cdot \cdot \subseteq \Omega$, are pairwise disjoint, then: $p(E_1 \cup E_2 \cup E_3 \cup \ldots) = p(E_1) + p(E_2) + p(E_3) + \ldots$
  - Events are additive



Random Variable

- A *random variable* is a function that associates a unique numerical value with every outcome of an experiment. (The value of the r.v. will vary from trial to trial as the experiment is repeated)
  - Discrete r.v.:
    - The outcome of a dice-roll
  - Continuous r.v.:
    - The temperature at time t



Rules of Probability

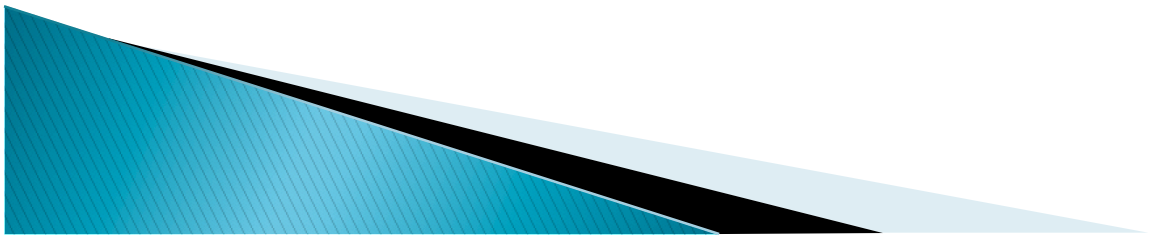
- Suppose we have two random variables X and Y, and a joint distribution p(X,Y). Then

- Sum Rule:

$$p(X) = \sum_{Y} p(X,Y)$$

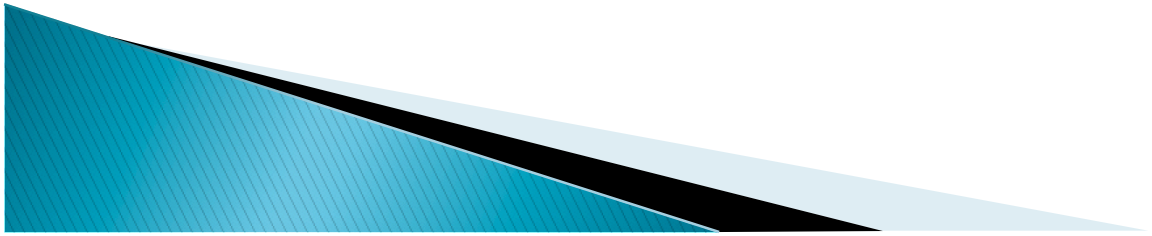
- Product Rule:

$$p(X,Y) = p(Y \mid X)\,p(X)$$



Terminology

- $p(X)$
  - Marginal probability
- $p(X,Y)$
  - Joint probability
- $p(X|Y)$
  - Conditional probability



Bayes Theorem

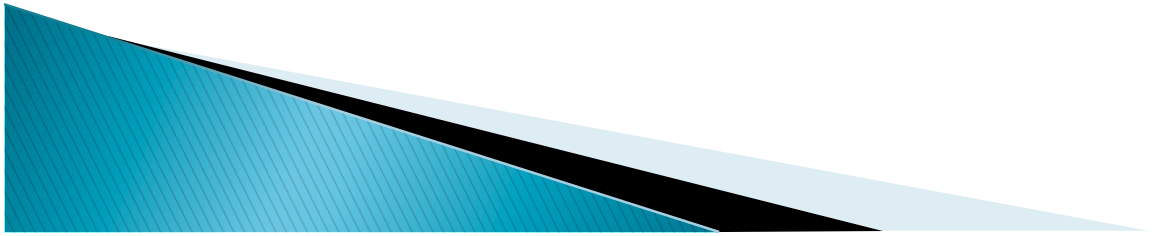
$$p(X \mid Y) = \frac{p(Y \mid X)\, p(X)}{p(Y)}$$

Bayes, Thomas (1763) An essay towards solving a problem in the doctrine of chances. *Philosophical Transactions of the Royal Society of London,* **53:370-418**

More Rules of Probability

- We often use shorthand, p(x) for p(X=x)

- Then
  - $p(x \vee y) = p(x) + p(y) - p(x,y)$
  - $p(\neg x) = 1 - p(x)$
    - Inclusion/exclusion
  - $p(\varnothing) = 0$
    - Null event

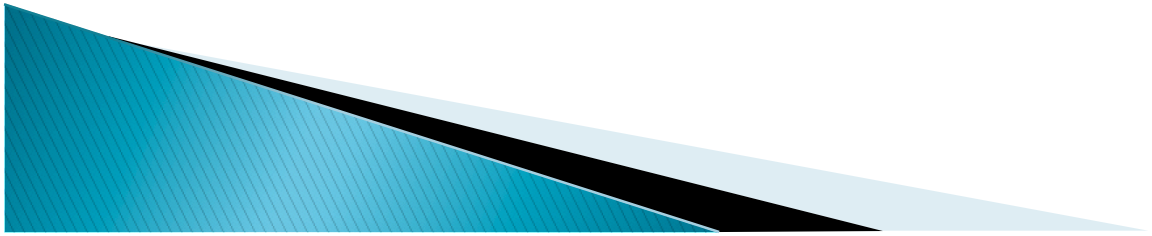


Expectations

- We want to compute the "average" result of an event
  - The "expectation" of the result
- If X is a random variable, the expectation is

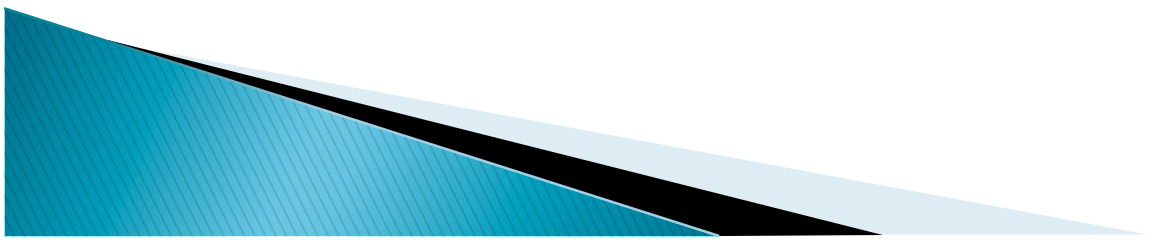
$$\mathrm{E}_p[X] = \begin{cases} \sum_{x \in X} p(X = x)x & \text{discrete} \\ \int_X \mathrm{d}x \; p(X = x)x & \text{continuous} \end{cases}$$

- Expectations are just like averages



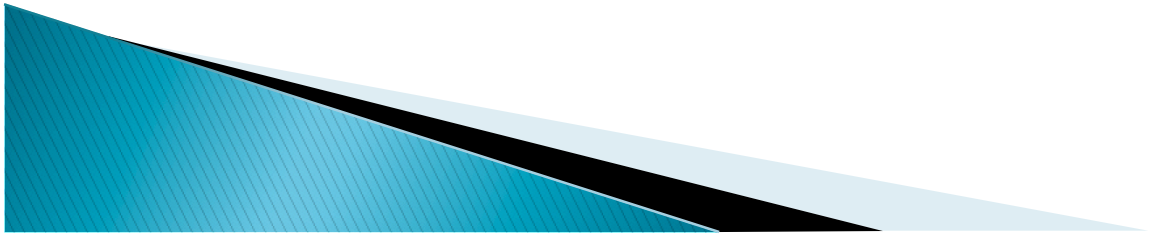
A Word on Bayesian Probabilities

- So far probabilities are frequencies of random events
- Bayesian view
  - Probabilities are quantifications of uncertainty
  - Not all events are repeatable
  - Ex. What grade will you get in this class
    - Over the semester our guess will get better
- Prior probability $p(G)$
- Posterior probability, given observations $p(G|D)$



Calculus

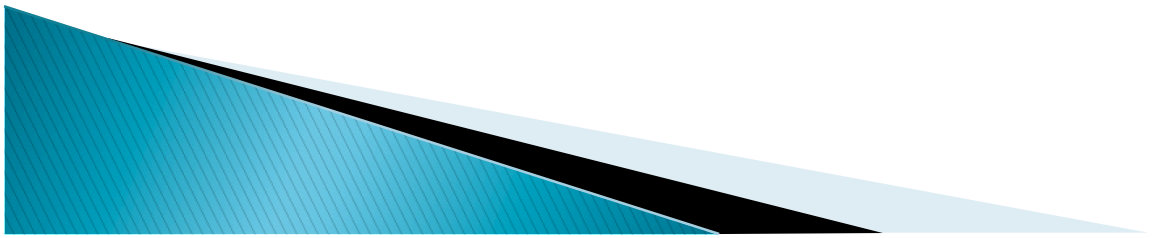
- Why?
  - Integrals – we've already seen a few, ☺
  - Derivatives: gives the rate of change of a function; for a given value x, derivative is the slope:
    - E.g:
      - $f(x) = ½ x^2$
      - What is f'(x)?
  - Sometimes need to take derivative with respect to multiple variables; we can do this separately for each variable, sequentially.



Logarithms

- Usually natural logs, occasionally base 2

- Useful properties:
  - $\log(xy) = \log(x) + \log(y)$
  - $\log \Pi f(x) = \Sigma \log f(x)$
  - $\log(x/y) = \log x - \log y$
  - $\log_e e^x = x$
  - $\log c^p = p \log c$

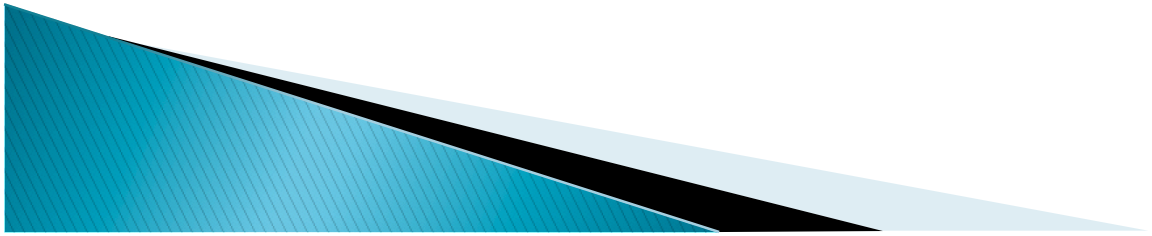


Optimization

- Assume function f(x) and we want to find a value for x such that $f(x) >= f(x')$, $\forall x' \neq x$
- Ex. $F(x) = -(x^2+3)$
- What happens if f'(x) is 0?
  - First derivative test
- What happens if f''(x) is negative/positive?
  - Second derivative test

Convexity

- Only works if the function is convex
- We'll say more about this later….

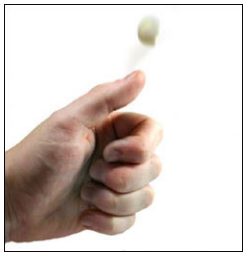


Example….

Discrete Distributions

- Bernoulli distribution: Bern(p)

$$P(x) = \begin{cases} 1-\theta & \text{if } x = 0 \\ \theta & \text{if } x = 1 \end{cases} \quad \Rightarrow \quad P(x) = \theta^x (1-\theta)^{1-x}$$



Parameter Learning from *iid* Data

- Goal: estimate distribution parameters θ from a dataset of N independent, identically distributed (*iid*), fully observed, training cases

$$D = \{x_1, \ldots, x_N\}$$

- Maximum likelihood estimation (MLE)
  1. One of the most common estimators
  2. With iid and full-observability assumption, write $L(\theta)$ as the **likelihood of the data**:

$$\begin{aligned} L(\theta) &= P(x_1, x_2, \ldots, x_N; \theta) \\ &= P(x_1;\theta)P(x_2;\theta),\ldots,P(x_N;\theta) \\ &= \prod_{i=1}^{N} P(x_i;\theta) \end{aligned}$$

  3. pick the setting of parameters most likely to have generated the data we saw:

$$\theta^* = \arg\max_\theta L(\theta) \quad = \arg\max_\theta \log L(\theta)$$

Log-likelihood of the data $l(\theta)$

Example: Bernoulli model

- Data:
  - We observed *N* ***iid*** coin tosses: *D*={1, 0, 1, …, 0}
- Representation:

Binary r.v: $x = \{0,1\}$

- Model: $P(x) = \begin{cases} 1-\theta & \text{for } x = 0 \\ \theta & \text{for } x = 1 \end{cases} \quad \Rightarrow \quad P(x) = \theta^{x}(1-\theta)^{1-x}$
- How to write the likelihood of a single observation x_i?

$$P(x_i) = \theta^{x_i}(1-\theta)^{1-x_i}$$

- The likelihood of the dataset $D=\{x_1, \ldots, x_N\}$:

$$P(x_1, x_2, \ldots, x_N \mid \theta) = \prod_{i=1}^{N} P(x_i \mid \theta) = \prod_{i=1}^{N} \left(\theta^{x_i}(1-\theta)^{1-x_i}\right) = \theta^{\sum_{i=1}^{N} x_i}(1-\theta)^{\sum_{i=1}^{N} 1-x_i} = \theta^{\#\text{head}}(1-\theta)^{\#\text{tails}}$$

Maximum Likelihood Estimation

- Objective function:

$$\boldsymbol{l}(\theta; D) = \log P(D \mid \theta) = \log \theta^{n_h} (1-\theta)^{n_t} = n_h \log \theta + (N - n_h) \log(1-\theta)$$

- We need to maximize this w.r.t. θ
- Take derivatives wrt θ

$$\frac{\partial \boldsymbol{l}}{\partial \theta} = \frac{n_h}{\theta} - \frac{N - n_h}{1-\theta} = 0 \quad \Rightarrow \quad \widehat{\theta}_{MLE} = \frac{n_h}{N} \quad \text{or} \quad \widehat{\theta}_{MLE} = \frac{1}{N} \sum_i x_i$$

Frequency as sample mean

Overfitting

- Recall that for Bernoulli Distribution, we have

$$\hat{\theta}_{ML}^{head} = \frac{n^{head}}{n^{head} + n^{tail}}$$

- What if we tossed only a few times so that we saw zero heads?
  We have $\hat{\theta}_{ML}^{head} = 0,$ and we will predict that the probability of seeing a head next is zero!!!

- The rescue: *"smoothing"*
  - Where n' is know as the pseudo- (imaginary) count

$$\hat{\theta}_{ML}^{head} = \frac{n^{head} + n'}{n^{head} + n^{tail} + \sum n'}$$

  - But can we make this more formal?

Bayesian Parameter Estimation

- Treat the distribution parameters θ also as a *random variable*
- The *a posteriori* distribution of θ after seeing the data is:

$$p(\theta \mid D) = \frac{p(D \mid \theta)p(\theta)}{p(D)} = \frac{p(D \mid \theta)p(\theta)}{\int p(D \mid \theta)p(\theta)d\theta}$$

uses Bayes Rule

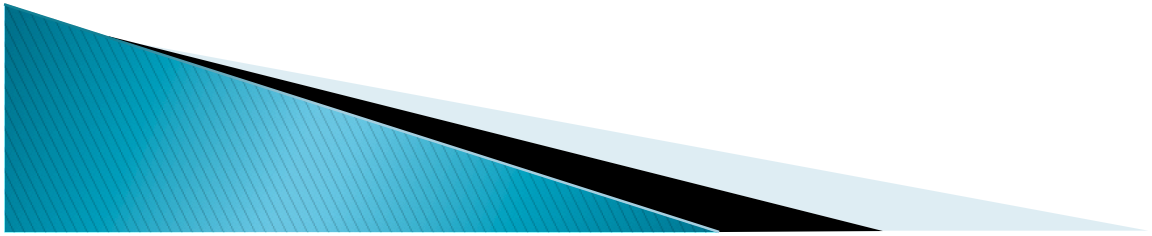
$$\text{posterior} = \frac{\text{likelihood} \times \text{prior}}{\text{marginal likelihood}}$$

The prior p(.) encodes our prior knowledge about the domain

Frequentist Parameter Estimation

Two people with different priors $p(\theta)$ will end up with different estimates $p(\theta|D)$.

- Frequentists dislike this “subjectivity”.
- Frequentists think of the parameter as a fixed, unknown constant, not a random variable.
- Hence they have to come up with different "objective" **estimators** (ways of computing from data), instead of using Bayes’ rule.
  - These estimators have different properties, such as being “unbiased”, “minimum variance”, etc.
  - The maximum likelihood estimator, is one such estimator.



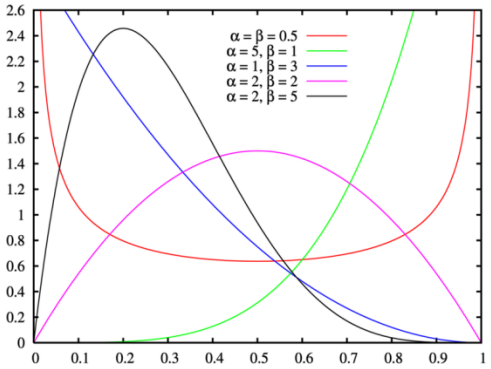
Discussion

θ or $p(\theta)$, this is the problem!

Bayesian estimation for Bernoulli

- Beta distribution:

$$P(\theta;\alpha,\beta)=\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1}=B(\alpha,\beta)\theta^{\alpha-1}(1-\theta)^{\beta-1}$$



  - When x is discrete $\Gamma(x+1)=x\Gamma(x)=x!$

- Posterior distribution of θ :

$$P(\theta\mid x_1,\ldots,x_N)=\frac{p(x_1,\ldots,x_N\mid\theta)p(\theta)}{p(x_1,\ldots,x_N)}\propto\theta^{n_h}(1-\theta)^{n_t}\times\theta^{\alpha-1}(1-\theta)^{\beta-1}=\theta^{n_h+\alpha-1}(1-\theta)^{n_t+\beta-1}$$

  - Notice the isomorphism of the posterior to the prior,
  - such a prior is called a **conjugate prior**
  - α *and* β are hyperparameters (parameters of the prior) and correspond to the number of "virtual" heads/tails (pseudo counts)

Bayesian estimation for Bernoulli, con'd

- Posterior distribution of θ :

$$P(\theta \mid x_1,\ldots,x_N) = \frac{p(x_1,\ldots,x_N \mid \theta)p(\theta)}{p(x_1,\ldots,x_N)} \propto \theta^{n_h}(1-\theta)^{n_t} \times \theta^{\alpha-1}(1-\theta)^{\beta-1} = \theta^{n_h+\alpha-1}(1-\theta)^{n_t+\beta-1}$$

- Maximum *a posteriori* (MAP) estimation:

$$\theta_{MAP} = \arg\max_{\theta} \log P(\theta \mid x_1,\ldots,x_N)$$

- :

$$\theta_{MAP} = \frac{n_h + \alpha - 1}{N + \alpha + \beta - 2}$$

Bata parameters can be understood as pseudo-counts

Prior strength: A=$\alpha+\beta$

- A can be interoperated as the size of an imaginary data set from which we obtain the **pseudo-counts**

Effect of Prior Strength

- Suppose we have a uniform prior, and we observe $\vec{n} = (n_h = 2, n_t = 8)$
- Weak prior A $= 2$, ($\alpha = \beta = 1$). Posterior prediction:

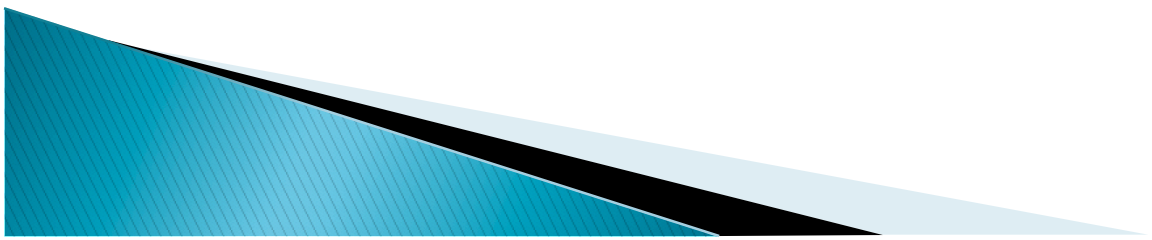
$$p(x = h \mid n_h = 2, n_t = 8, \alpha = 1) = \frac{2 + 1 - 1}{10 + 2 - 2} = 0.2$$

- Strong prior A $= 20$, ($\alpha = \beta = 10$). Posterior prediction:

$$p(x = h \mid n_h = 2, n_t = 8, \alpha = 10, \beta = 10) = \frac{2 + 10 - 1}{10 + 20 - 2} = 0.39$$

- However, if we have enough data, it washes away the prior. e.g. $\vec{n} = (n_h = 200, n_t = 800)$.
- Then the estimates under weak and strong prior are $\frac{200+1-1}{1000+2-2}$ and $\frac{200+10-1}{1000+20-2}$, respectively, both of which are close to 0.2

Matlab Tutorial



Next Time….

- Reading: decision tree handout, available on course site.