Biological Sequence Analysis
(Ch 11. Background on probability)

Biointelligence Laboratory
School of Computer Sci. & Eng.
Seoul National University
Seoul 151-742, Korea

This slide file is available online at
http://bi.snu.ac.kr/

List of Contents

- Introduction
- Random variables
- Probability density, distributions
- Transformation
- Plots of statistical distributions
- Relative entropy and mutual information
- Many random variables
- One DNA sequence
- Maximum likelihood
- Sampling
- Metropolis sampling
- EM algorithm

Introduction

- Sequence similarity test:
  g g a g a c t g t a g a c a g c t a a t g c t a a
g a a c g c c t a g c c a c g a g c c t t a t e
  P (more than 10 matches) = ? (0.04)

- Parameter, data, hypothesis, random variable

One DNA Sequence

- Shotgun sequencing: Find long DNA sequence by many overlapping short sequences (500 bases).
One DNA Sequence (2)

1. What is the mean proportion of the genome covered by contigs?
2. What is the mean number of contigs?
3. What is the mean contig size?

(Discrete) Random Variables

A discrete numerical quantity corresponding to an observed outcome of experiment

(E.g.) experiment: rolling two six-sided dice
outcome: the two numbers on the dice
discrete random variables: the sum of the two numbers, the difference of the two numbers, the smaller number of the two,

Random variables are usually represented by uppercase symbols (X, Y, …) and the realized values of the random variables are represented by lowercase symbols (x, y, …)

Probability Distributions

For a finite set X: the probability distribution is simply an assignment of a probability $p_x$ to each outcome $x$ in X.
(e.g.) the probability of outcomes of rolling a dice {1/12, 1/12, 1/12, 1/6, 1/4, 1/3}

For a continuous set X: the probability density $f(x)$

$p(x - \Delta x / 2 \leq x \leq x + \Delta x / 2) = f(x) \Delta x$

$p(x_0 \leq x \leq x_1) = \int_{x_0}^{x_1} f(x) dx$

$f(x) \geq 0$

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

Note: $f(x)$ can be greater than 0

Probability density, distribution (1/2)

Relation of distribution and density

$F(t) = P(X \leq t) = \sum_{x \leq t} P(X = x)$

Expectation (mean)

$E(X) = \sum_{x} x P(X = x)$

For continuous X:

$P(X = x) = 0, \quad f(x) \text{ can be } > 1$
Probability density, distribution (2/2)

- Relation of distribution and density
  \[ P(X = x) = \sum_{t \geq x} P(X = t) - \sum_{t < x} P(X = t) \]
  \[ f(x) = \frac{d}{dt} F(t) \bigg|_{t=x} \]

- Conditions for the probability density:
  1. \( f(x) \geq 0 \)
  2. \( \int f(x)dx = 1 \)

- Variance: \( E(X - \mu)^2 = EX^2 - \mu^2 \)

Transformation

- One random variable:
  \[ X_2 = g(X_1) \text{ for monotone } g \]
  \[ F_2(t) = P(X_2 < t) = P(g(X_1) < t) = P(X_1 < g^{-1}(t)) = F_1(g^{-1}(t)) \]
  \[ f_2(x) = f_1(g^{-1})(g^{-1})' \]

- Many random variable: For 1-1 relations between \( X \) and \( U \)
  \[ \bar{X} = (X_1, \ldots, X_n) \]
  \[ \bar{U} = (U_1(X_1, \ldots, X_n), \ldots, U_n(X_1, \ldots, X_n)) \]
  \[ f_U(u_1, \ldots, u_n) = f_X(x_1, \ldots, x_n) |J| = f_X(x_1, \ldots, x_n) |J| \]

Statistical Independence

- Two events are independent if the outcome of one event does not affect the outcome of the other event.
- Discrete random variables are independent if the value of one does not affect the probabilities associated with the values of another random variable.
- (E.g.) Experiment: rolling a fair die
  1. \( A \): the number is even, \( B \): the number is greater than or equal to 3.
  \[ P(A \mid B) = P(A), \ P(B \mid A) = P(B) : A \text{ and } B \text{ are independent} \]
  2. \( C \): the number is greater than 3
  \[ P(C) = 1/2 , \ P(C \mid A) = 2/3 , \ P(A \mid C) = 1/3 , \ P(A) = 1/2 : A \text{ and } C \text{ are not independent} \]

Uniform Distribution

- Discrete uniform: each outcome can occur equally likely.
- For \( N \) outcomes: \( P(x) = 1/N \)
- Continuous uniform: \( f(x) = 1/(b-a) \text{ or } 1/\text{area}(A) \)
Bernoulli Distribution

- Bernoulli trial is a single trial with two possible outcomes (success / failure).
- Bernoulli random variable $Y$: number of successes in this trial.

$$p(y \text{ successes in a trial}) = p^y (1-p)^{1-y}, \quad y = 0,1$$

Binomial Distribution

- Defined on a finite set of all the possible results of $N$ trials with a binary outcome (‘0’ or ‘1’).
- The random variable is the number of success in the fixed $N$ trials.

$$p(k \text{ successes out of } N) = \binom{N}{k} p^k (1-p)^{N-k}$$

$$m = \sum k p(k) = \sum_{k=1}^{N} \binom{N}{k} p^k (1-p)^{N-k} = Np$$

$$\sigma^2 = \sum (k-m)^2 p(k) = \sum_{k=1}^{N} (k-m)^2 \binom{N}{k} p^k (1-p)^{N-k} = Np(1-p)$$

Multinomial Distribution

- Extend the binomial to $K$ independent outcomes with probabilities $\theta_i, \ (i = 1, \cdots, K)$.

$$p(n, i = 1, \cdots, K | \theta) = \binom{n}{n_1 \cdots n_K} \prod_{i=1}^{K} \theta_i^{n_i}$$

(e.g.) Rolling a fair dice $N$ times:

$P \text{ (rolling 12 times and getting each number twice)}$

$= 12! / 2!^6 (1/6)^{12} = 3.4 \times 10^{-3}$

Geometric Distribution

- Random variable is the number of trials before the first failure (the length of a success run).
- Test of the significance of the long run in the sequences.

$$P(y) = (1-p)^y, \quad (y = 0, 1, 2, \cdots)$$

$$F(y) = P(Y \leq y) = 1 - p^{y+1}$$

- Geometric-like random variables in BLAST theory:

$$1 - F(y-1) = P(Y \geq y) \sim C p^y, \quad (0 < C < 1)$$
Negative Binomial Distribution

- Random variable is the number of trials in the fixed number of successes
- \( P(N=n) = P(\text{the first } (n-1) \text{ trials result in exactly } (m-1) \text{ successes and } (n-m) \text{ failures and trial } n \text{ results in success}) \)
- \( P(n) = \binom{n-1}{m-1} p^m (1-p)^{n-m}, \ (n = m, m+1, m+2, \cdots) \)

Generalized Geometric Distribution

- Random variable is the number of trials before the \((k+1)\text{-th}\) failure.
- \( P(y) = \binom{y}{k} p^{y-k} (1-p)^{k+1}, \ (y = k, k+1, k+2, \cdots) \)

Poisson Distribution

- A limiting form of the binomial distribution (as \( n \) becomes large, \( p \) becomes smaller) with moderate \( np = \lambda \)
- \( P(y) = \frac{e^{-\lambda} \lambda^y}{y!}, \ (y = 0, 1, 2, \cdots) \)

Exponential Distribution

- Exponential distribution is a continuous analogue of the geometric distribution
- \( f(x) = \lambda e^{-\lambda x}, \ x \geq 0 \)
- \( F(x) = 1 - e^{-\lambda x}, \ x \geq 0 \)
- \( E(X) = 1/\lambda \)
- \( Var(X) = 1/\lambda^2 \)
- \( Med(X) = \frac{\log 2}{\lambda} = 0.301 E(X) \) (right skewed)
Relation with Geometric Distribution

- Suppose $X$ has an exponential distribution and $Y$ is the integer part of $X$
  
  $P(Y = y) = P(y \leq X < y + 1) = (1 - e^{-\lambda}) e^{-\lambda y} = (1 - p) p^y$

  ($p = e^{-\lambda}$)

- The density function of the fractal part $D = X - Y$:

  $f(d) = \frac{\lambda e^{-\lambda d}}{1 - e^{-\lambda}}$

Gaussian Distribution

- As $N$ gets larger, mean and variance of the binomial distribution increase linearly $\Rightarrow$ rescale $k$ by

  $u = \frac{k - m}{\sigma} = \frac{k - Np}{\sqrt{Np(1 - p)}}$

  then the binomial becomes a Gaussian with the density as

  $f(u) = \frac{1}{\sqrt{2\pi}} \exp(-u^2 / 2)$

Dirichlet Distribution (1/3)

- The Dirichlet distribution is a conjugate distribution for the multinomial distribution. (The Dirichlet prior wrt. multinomial distribution gives Dirichlet posterior again.)

  $D(\theta_1, \cdots, \theta_K | \alpha) = Z^{-1}(\alpha) \prod_{i=1}^{K} \theta_i^{\alpha_i - 1} \delta(\sum_{i=1}^{K} \theta_i - 1) \quad (0 \leq \theta_i \leq 1)$

  $Z(\alpha) = \prod_{i=1}^{K} \Gamma(\alpha_i) / \Gamma(\sum_{i=1}^{K} \alpha_i) \quad (\Gamma(x + 1) = x\Gamma(x))$

- $\theta_i$'s corresponds to the parameters of the multinomial distribution.

Figure 11.1. The limit for large $N$ of a binomial tends to a Gaussian. In this case $N = 40$ and $p = 1/4$ in (11.1).
The mean of Dirichlet is equal to the normalized parameters. \( E[\theta_i] = \frac{\alpha_i}{\sum_i \alpha_i} \)

(e.g.) The three distributions all have the same mean (1/8, 2/8, 5/8) though with different values of \(\alpha\)
(1,2,5); (10,20,50); (.1, .2, .5)

For \(K=2\), Dirichlet reduces to the Beta distribution

(e.g.) The dice factory: sampling \(\theta_1,\ldots,\theta_6\) from Dirichlet parameterized by \(\alpha_1,\ldots,\alpha_6\)

Factory A: all six parameters set to 10
Factory B: all six parameters set to 2

On average both factories produce fair dice (average = 1/6)

But if we find a loaded dice with (.1 .1 .1 .1 .1 .6), it is more likely from factory B:

The variance of Dirichlet is inversely proportional to the sum of parameters.

The gamma distribution is conjugate to the Poisson:
the probability of seeing \(n\) events over some interval when there is a probability \(p\) of an individual event occurring in that interval. \(f(n) = \frac{e^{-p} p^n}{n!}\)

The gamma distribution is used to model the probabilities of rate.

The gamma distribution is used to model the rate of evolution at different sites in DNA sequences.
Extreme Value Distribution (1/2)

- For $N$ samples from $g(x)$, the probability that the largest of them are less than $x$ is $G(x)^N$
- $G(x) = \int_{-\infty}^{x} g(u) \, du$
- $h(x) = N g(x) G(x)^{N-1}$
- Extreme value density (EVD) for $g(x)$ is the limit of $h(x)$
- EVD is used to model the breaking point of a chain, to assessing the significance of the maximum score from a set of alignments

Extreme Value Distribution (2/2)

- EVD for exponential density $g(x) = \alpha e^{-\alpha x}$
- $h(x) = \alpha e^{-\alpha x} \left(1 - \alpha e^{-\alpha x} / N\right)^{N-1} \rightarrow \alpha e^{-\alpha x} \exp(-e^{-\alpha z})$ for $N \geq 10$
- (Gumbel distribution)
- (e.g.) For $N = 1, 2, 10, 100$, $N \geq 10$ gives a good approx. to the EVD

Statistical Distribution (1/7)

- Binomial Distribution

Statistical Distribution (2/7)

- Poisson Distribution

Fifty dots distributed randomly over 30 events (discrete), $m = 1$.
Example 1: $m = 1$.
Example 2: $m = 2, 3, 5$ and $10$. 
Statistical Distribution

- **Negative-Binomial Distribution**

- **Geometric Distribution**

Statistical Distribution

- **Exponential Distribution**

- **Normal Distribution**
Lognormal Distribution

A measure of average uncertainty of an outcome
Shannon entropy: \( H(X) = -\sum_i P(x_i) \log P(x_i) \)
Entropy is maximized when all the \( P(x_i) \) are equal and the maximum is then \( \log K \). If we are certain of the outcome, then the entropy is zero.

Information: \( I(X) = H_{\text{before}} - H_{\text{after}} \)

Entropy of equi-probable DNA symbol \( (A,C,G,T) \) is 2 bits.
Information content of a conserved position: A particular position is always an A or a G with \( p(A)=0.7 \) and \( p(G)=0.3 \). Thus \( H_{\text{before}} - H_{\text{after}} = 2 - .88 = 1.12 \) bits
(The more conserved the position, the higher the information content.)

Relative entropy and mutual information (1/4)

Relative entropy (KL divergence) \( P \) wrt. \( Q \):
\[
H(P \| Q) = \sum_i P(x_i) \log \frac{P(x_i)}{Q(x_i)}
\]
Information content and relative entropy are the same if \( Q \) represents the initial state
\( H(P \| Q) \neq H(Q \| P) \) (Not a metric)
Positive of relative entropy
\[ \log(x) \leq x - 1 \]

\[ -H(P \parallel Q) = \sum_i P(x_i) \log(Q(x_i)/P(x_i)) \leq \sum_i P(x_i)(Q(x_i)/P(x_i) - 1) = 0 \]

Independency can be measured by the relative entropy between \( P(X,Y) \) and \( P(X)P(Y) \).

\[ M(X;Y) = \sum_{i,j} P(x_i, y_j) \log \frac{P(x_i, y_j)}{P(x_i)P(y_j)} \]

\( M(X, Y) \) can be interpreted as the amount of information that we acquire about outcome \( X \) when we are told outcome \( Y \).

(E.g.) Acceptor sites: 757 acceptor sites from a database with human genes. 30 bases upstream, 20 bases downstream are extracted from each acceptor.

Relative entropy:
\[ \sum_a p_i(a) \log[p_i(a)/q_a] \]

where \( q_a \) is the overall distribution of the four nucleotides in the sequences.

Mutual information:
\[ \sum_{a,b} p_i(a,b) \log[p_i(a,b)/p_i(a)p_{i+1}(b)] \]
Many random variables (1/2)

- Marginal probability
  \[ P(Y_i = y_i, \cdots, Y_k = y_k) = \sum_{y_{i+1}, \cdots, y_k} P(Y_i = y_i, \cdots, Y_k = y_k) \]
  \[ f(x_1, \cdots, x_k) = \cdots \int f(x_1, \cdots, x_k) dx_1 \cdots dx_k \]

- Conditional probability
  \[ P(Y_i = y_i, \cdots, Y_k = y_k | Y_1 = y_1, \cdots, Y_i = y_i) \]
  \[ P(Y_i = y_i, \cdots, Y_k = y_k) \]
  \[ f(x_1, \cdots, x_k | x_1, \cdots, x_i) = \frac{f(x_1, \cdots, x_k)}{f(x_1, \cdots, x_i)} \]

Many random variables (2/2)

- Covariance
  \[ \sigma_{y_1y_2} = \sum_{y_1,y_2} (y_1 - \mu_1)(y_2 - \mu_2) P(Y_1 = y_1, Y_2 = y_2) \]
  \[ \sigma_{x_1x_2} = \int \int (x_1 - \mu_1)(x_2 - \mu_2) f(x_1, x_2) dx_1 dx_2 \]

- Correlation:
  \[ \rho_{12} = \frac{\sigma_{y_1y_2}}{\sigma_1 \sigma_2} \]

One DNA Sequence (1/10)

- Shotgun sequencing: Find long DNA sequence by many overlapping short sequences.

One DNA Sequence (2/10)

- 1. What is the mean proportion of the genome covered by contigs?
- 2. What is the mean number of contigs?
- 3. What is the mean contig size?
There are \( N \) fragments, each of length \( L \), the whole DNA is of length \( G \).

The position of the left hand end of any fragment is uniformly distributed in \((0, G)\).

The number of fragments falling in the interval length \( h \) becomes \( Np = Nh/G \) from binomial distribution.

For large \( N \) and small interval \( h \), this distribution becomes Poisson with mean \( a = Nh/G \).

Mean proportion of genome covered by one or more fragment

\[
P = P \text{ (a point chosen at random is covered at least by one fragment)}
\]

\[
= P \text{ (the left end of at least one fragment is in the interval } L \text{ starting from that point)}, \text{ (by poisson)}
\]

\[
= 1 - e^{-a} = 1 - e^{-a/NL/G} \quad (a = NL/G)
\]

The length of fragments to cover the whole genome:

\[
P = .99 \Rightarrow a = NL/G = 4.6 \quad (a \text{: coverage})
\]

\[
P = .999 \Rightarrow a = 6.9 \text{ (about 7 times the genome length)}
\]

Human genome is \(3 \times 10^9\) nucleotides : 3,000,000 of them are missing with \( P = .999\).

What is the mean number of contigs?

- success number: a unique right most fragment
- trial number: \( N \)
- \( p = P \) (a fragment is the right-most member of a contig)
- \( = P \) (no other fragment has its left-end point on the fragment in question)
- \( = e^a \)
- Mean number of contigs

\[
Ne^{-a} = Ne^{-NL/G}
\]

For \( G = 100,000 \ L = 500 \):

\[
\begin{array}{cccccccccc}
\text{a} & 0.5 & 0.75 & 1 & 1.5 & 2 & 3 & 4 & 5 & 6 & 7 \\
\hline
\text{Mean number of contigs} & 60.7 & 70.8 & 73.6 & 66.9 & 54.1 & 29.9 & 14.7 & 6.7 & 3.0 & 1.3
\end{array}
\]

If there is a small number of fragments, there must be a small number of contigs.

A large number of fragments tend to form a small number of large contigs.
Mean contig size ?

- Expectation of a sum of random variables where \( N \) is random
- Moment generating function of a random variable
- Probability generating function of a random variable
- Geometric distribution
- Exponential distribution

Moment generating function (mgf) of a random variable \( Y \):

\[
m(\theta) = E(e^{\theta Y}) = \int e^{\theta y} f(y) dy
\]

Finding mean and variance from mgf:

\[
\mu = \left( \frac{dm(\theta)}{d\theta} \right)_{\theta=0}, \quad \sigma^2 = \left( \frac{d^2m(\theta)}{d\theta^2} \right)_{\theta=0} - \mu^2
\]

or

\[
\mu = \left( \frac{d \log m(\theta)}{d\theta} \right)_{\theta=0}, \quad \sigma^2 = \left( \frac{d^2 \log m(\theta)}{d\theta^2} \right)_{\theta=0}
\]

Probability generating function (pgf) of a (discrete) random variable \( Y \):

\[
p(t) = E(t^Y) = \sum_y P(y) t^y
\]

\[
\mu = \left( \frac{d}{dt} p(t) \right)_{t=1}, \quad \sigma^2 = \left( \frac{d^2}{dt^2} p(t) \right)_{t=1} + \mu - \mu^2
\]

Expectation of a sum of random variables \( S = X_1 + \cdots + X_n \) when \( N \) is random:

\[
P(S = y) = \sum_n P_n P(S = y | n)
\]

pgf of \( N \):

\[
p(t) = \sum_n P_n t^n
\]

pgf of \( S | n \):

\[
(q(t))^n P(S = y) = \text{coeff. of } t^y \text{ in } \sum_n P_n (q(t))^n
\]

thus pgf of \( S \):

\[
p(q(t)), \quad \text{and by chain rule } E(S) = E(N)E(X)
\]
Contig length = sum of n successful left end segments of fragments overlapped in the contig + L

The distance between two successful left-hand points ~ $\text{Exp}(\lambda)$, $\lambda = N/G$ (known from Poisson process)

The number of overlapping fragments ~ Geo ($p$)

$$p = \int_0^L \lambda e^{-\lambda x} dx = 1 - e^{-\lambda L}$$

Mean of the random distance (any left-left end) ~ conditional exponential ($0 < X < L$)

$$f(x|0 < x < L) = \frac{f(x)}{\int_0^L f(x)dx}$$

$$f(x) = \frac{\lambda e^{-\lambda x}}{1 - e^{-\lambda L}}$$

Mean Contig length:

$$E(N)E(X | 0 < X < L) + L = (e^\mu - 1) \left( \frac{1}{\lambda} - \frac{L}{e^\mu - 1} \right) + L$$

(e.g.) For $L = 500$, $a = 2$, $G = 100,000$ : the mean contig size = 1600

Maximum likelihood

$$\theta^{ML} = \arg \max_{\theta} P(D|\theta,M)$$

Consistent : parameter value used to generate the data set will also be the value that maximizes the likelihood in the limit.

(Eg.) For $K$ observable outcomes of the model, the frequency of occurrence $w$ will tend to $P(\omega|\theta_0,M)$ and log-likelihood for parameter $\theta$: $\sum (n_l \sum n_i) \log P(\omega_l|\theta,M)$ tends to $\sum P(\omega_l|\theta_0,M) \log P(\omega_l|\theta,M)$.

By the positivity of the relevant entropy implies that for all $\theta$

$$\sum P(\omega_l|\theta_0,M) \log P(\omega_l|\theta_0,M) \geq \sum P(\omega_l|\theta,M) \log P(\omega_l|\theta,M)$$

Thus the likelihood is maximized by $\theta_0$

Drawbacks: can give poor estimate for scanty data

Posterior probability distribution

Bayes theorem

$$P(\theta|D,M) = \frac{P(D|\theta,M)P(\theta|M)}{P(D|M)}$$

Use of posteriors: MAP (maximum a posteriori probability) estimate

$$\theta^{MAP} = \arg \max_{\theta} P(D|\theta,M)P(\theta|M)$$

PME (posterior mean estimator) : parameters weighted by the posterior

$$\theta^{PME} = \int \theta P(\theta|n)d\theta$$
Change of variables

- Given a density of $x: f(x)$ and a change of variable $x = \phi(y)$, the density of $y: g(y)$ becomes $g(y) = f(\phi(y)) |\phi'(y)|$
- The maximum of $g(y)$, MAP or PME may shift from that of $f(x)$.

Sampling

- Given a finite set with probabilities $P(x)$, to sample from this set means to pick elements $x$ randomly with probability $P(x)$.
- Sampling by random number generator in $[0, 1]$:
  $P(\text{selecting } x_i) = P(p(x_1) + \cdots + p(x_{i-1}) < \text{rand} [0, 1] < p(x_i) + \cdots + p(x_{i-1}) + p(x_i))$
  $= P(x_i)$

Sampling by transformation from a uniform distribution

- For a uniform density $f(x)$ and a map $x = \phi(y)$,
  $g(y) = f(\phi(y))\phi'(y) = \phi'(y), \phi(y) = \int_0^y g(u)du$
  $y = \phi^{-1}(x)$
- (E.g.) Sampling from a Gaussian:
  $g(y), x \text{ is given. Find } y$:
  $\text{Define } \phi(y) = \int_0^y e^{-u^2/2} / \sqrt{2\pi} \ du$
  $\text{And let } y = \phi^{-1}(x)$

Sampling with the Metropolis algorithm (1/3)

- For sampling when the analytic methods are not available
- Generate a sequence $\{y_i\}$ to approximate $P$ as close as we like from $\tau(y \mid y_{i-1})$ under a condition:
  (detailed balance) $P(x) \tau(y \mid x) = P(y) \tau(x \mid y)$
- Detailed balance implies:
  $\lim_{N \to \infty} \frac{1}{N} \#(y_i = x) = P(x)$
- The sequence under this transition process will sample $P$ correctly
Sampling with the Metropolis algorithm (2/3)

- Metropolis algorithm:
  - Symmetric proposal mechanism: Given a point \( x \), this selects a point \( y \) with probability \( F(y|x) \) and symmetric.
  - Acceptance mechanism: accepts proposed \( y \) with probability \( \min (1, P(y)/P(x)) \).
    (A point \( y \) with larger posterior probability than the current \( x \) is always accepted, and one with lower probability is accepted randomly with probability \( P(y)/P(x) \)).

Sampling with the Metropolis algorithm (3/3)

- Metropolis algorithm (balance):
  \[
P(x)\pi(y|x) = P(x)F(y|x)\min(1, P(y)/P(x))
  = F(y|x)\min(P(x), P(y))
  = F(x|y)\min(P(y), P(x))
  = P(y)\pi(x|y)
\]

Gibbs sampling

- Sampling from conditional distributions
  \( P(x_i | x_1, \cdots, x_{i-1}, x_{i+1}, \cdots, x_N) \) for each \( i \).
- Proposal distribution is the conditional distribution: Always accept the sample.

Estimation of probabilities from counts (1/2)

- \( \theta_i^{ML} = n_i / N \)
- \( P(n | \theta^{ML}) > P(n | \theta) \) for any \( \theta \neq \theta^{ML} \) and an observation \( n \):
  \[
  \log \frac{P(n | \theta^{ML})}{P(n | \theta)} = \log \frac{\prod_i (\theta_i^{ML})^{n_i}}{\prod_i \theta_i^{n_i}}
  = \sum_i n_i \log \frac{\theta_i^{ML}}{\theta_i}
  = N \sum_i \theta_i^{ML} \log \frac{\theta_i^{ML}}{\theta_i} > 0
  \]
Estimation of probabilities from counts (2/2)

- For scarce data: use prior
  \[ P(n | \theta) = M^{-1}(n) \prod_{i=1}^{k} \theta_i^{n_i} \]
  \[ D(\theta | \alpha) = \frac{1}{Z(\alpha)} \prod_{i=1}^{k} \theta_i^{n_i-1} \]
  \[ P(\theta | n) = \frac{P(n | \theta)D(\theta | \alpha)}{P(n)} = \frac{1}{P(n)M(n)Z(\alpha)} \prod_{i} \theta_i^{n_i+\alpha_i-1} \]
  \[ \theta_{i}^{PME} = \frac{n_i + \alpha_i}{N + A} \]
  \[ P(\theta | n) = D(\theta | n + \alpha) \]

Mixtures of Dirichlets

\[ P(\theta | \alpha^1, \ldots, \alpha^m) = \sum_k q_k D(\theta | \alpha^k), \quad q_k = P(\alpha^k) \]
\[ P(\theta | n) = \sum_k P(\theta | \alpha^k, n) P(\alpha^k | n) = \sum_k P(\alpha^k | n) D(\theta | n + \alpha^k) \]
\[ \theta_{i}^{PME} = \sum_k P(\alpha^k | n) \frac{n_i + \alpha_i^k}{N + A} \]

EM algorithm (1/3)

- A general algorithm for ML estimation with missing data
- Baum-Welch algorithm for estimating hidden Markov model probabilities is a special case of the EM algorithm.

  \[ \text{Maximize:} \]
  \[ \log P(x | \theta) = \log P(x, y | \theta) - \log P(y | x, \theta) \]
  \[ = \sum_y P(y | x, \theta') \log P(x, y | \theta) - \sum_y P(y | x, \theta) \log P(y | x, \theta) \]

  \[ Q(\theta | \theta') = \sum_y P(y | x, \theta') \log P(x, y | \theta) \]

EM algorithm (2/3)

\[ \log P(x | \theta) - \log P(x | \theta') = Q(\theta | \theta') - Q(\theta' | \theta') + \sum_y P(y | x, \theta') \log \frac{P(y | x, \theta)}{P(y | x, \theta')} \]

Thus,
\[ \log P(x | \theta) - \log P(x | \theta') \geq Q(\theta | \theta') - Q(\theta' | \theta') \]
EM algorithm (3/3)

- Choosing \( \theta^{t+1} = \arg \max_{\theta} Q(\theta | \theta^t) \) will always make the difference positive and thus the likelihood of the new model is larger than the likelihood of the old one.
- EM Algorithm:
  - (E-step) Calculate Q function
  - (M-step) Maximize \( Q(\theta | \theta^t) \) wrt. \( \theta \)
- The likelihood increases in each iteration
- Instead of maximizing in the (M-step), algorithms that increase Q are called generalized EM (GEM).