Definitions

- Probabilistic models
  - A model means a system that simulates the object under consideration
  - A probabilistic model is one that produces different outcomes with different probabilities (BSA)

Why probabilistic models

- The biological system being analyzed is stochastic
- Or noisy
- Or completely deterministic, but because a number of hidden variables effecting its behavior are unknown, the observed data might be best explained with a probabilistic model

Probability

- Experiment: a procedure involving chance that leads to different results
- Outcome: the result of a single trial of an experiment
- Event: one or more outcomes of an experiment
- Probability: the measure of how likely an event is
  - Between 0 (will not occur) and 1 (will occur)

Example: a fair 6-sided dice

- Outcome: The possible outcomes of this experiment are 1, 2, 3, 4, 5 and 6
- Events: 1; 6; even
- Probability: outcomes are equally likely to occur
  - \( P(A) = \frac{\text{The Number Of Ways Event A Can Occur}}{\text{The Total Number Of Possible Outcomes}} \)
  - \( P(1) = P(6) = \frac{1}{6}; \ P(\text{even}) = \frac{3}{6} = \frac{1}{2} \)

Random variable

- Random variables are functions that assign a unique number to each possible outcome of an experiment
- An example
  - Experiment: tossing a coin
  - Outcome space: \{heads, tails\}
  - \( X = \begin{cases} 1 & \text{if heads} \\ 0 & \text{if tails} \end{cases} \)
  - More exactly, \( X \) is a discrete random variable
    - \( P(X=1) = \frac{1}{2}, P(X=0) = \frac{1}{2} \)
Probability distribution

- Probability distribution: the assignment of a probability \( P(x) \) to each outcome \( x \).
- A fair dice: outcomes are equally likely to occur \( \Rightarrow \) the probability distribution over the all six outcomes \( P(x)=\frac{1}{6}, x=1,2,3,4,5 \text{ or } 6 \).
- A loaded dice: outcomes are unequally likely to occur \( \Rightarrow \) the probability distribution over the all six outcomes \( P(x)=f(x), x=1,2,3,4,5 \text{ or } 6, \text{ but } \sum f(x)=1 \).

Probability density function (pdf)

- Probability density functions (pdf) are for continuous rather than discrete random variables; \( f(x) \).
- A pdf must be integrated over an interval to yield a probability, since \( P(X=x)=0 \)

\[
P(a \leq X \leq b) = \int_a^b f(x)\,dx
\]

- Cumulative distribution function (cdf)

\[
P(X \leq x) = \int_{-\infty}^x f(t)\,dt
\]

Joint probability

- Two experiments (random variables) \( X \) and \( Y \)
  - \( P(X,Y) \) \( \Rightarrow \) joint probability (distribution) of \( X \) and \( Y \)
  - \( P(X,Y)=P(X|Y)P(Y)=P(Y|X)P(X) \)
  - \( P(X|Y)=P(X), X \text{ and } Y \text{ are independent} \)

Example: experiment 1 (selecting a dice), experiment 2 (rolling the selected dice)
  - \( P(y): y=D_1 \text{ or } D_2 \)
  - \( P(i,D_j)=P(i\,|\,D_j)P(D_j) \)
  - \( P(i\,|\,D_2)=P(i\,|\,D_2), \text{ independent events} \)

Marginal probability

- The distribution of the marginal variables (the marginal distribution) is obtained by marginalizing over the distribution of the variables being discarded (so the discarded variables are marginalized out)

\[
P(X)=\sum_y P(X|Y)P(Y)
\]

Example: experiment 1 (selecting a dice), experiment 2 (rolling the selected dice)
  - \( P(y|D_1 \text{ or } D_2) \)
  - \( P(i\,|\,D_1|P(D_1)+P(i\,|\,D_2|P(D_2) \)
  - \( P(i\,|\,D_1)=P(i\,|\,D_1), \text{ independent events} \)
  - \( P(i\,|\,D_2)=P(i\,|\,D_2) \text, \text{ independent events} \)
  - \( P(i\,|\,D_1\,P(D_1)+P(i\,|\,D_2)=P(i\,|\,D_1) \)

The probability of a DNA sequence

- Event: Observing a DNA sequence \( S=s_1s_2...s_n \), \( s_i \in \{A,C,G,T\} \).
- Random sequence model (or Independent and identically-distributed, i.i.d. model): \( s_i \) occurs at random with the probability \( P(s_i) \), independent of all other residues in the sequence;

\[
P(S)=\prod_{i=1}^n P(s_i)
\]

- This model will be used as a background model (or called a null hypothesis).
1. Common Distributions

- Bernoulli trials (where the result of each Bernoulli trial is true with probability of \( p \) and false with probability of \( 1-p \)).

- The binomial distribution gives the discrete probability distribution of obtaining \( n \) successes out of \( N \) trials.

\[ P(X=k) = \binom{N}{k} p^k (1-p)^{N-k} \]

- The Poisson distribution gives the discrete probability distribution of a given number of events during a fixed interval of time (or space, distance, area et al), if these events are independent, and the average rate of occurrence \( \lambda \) is known.

\[ P(X=k) = \frac{e^{-\lambda} \lambda^k}{k!} \]

- The Gaussian (or normal) distribution is a continuous probability distribution that is often used to model real-world phenomena.

\[ f(x; \mu, \sigma^2) = \frac{1}{\sigma \sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \]

\( \mu \): mean (expectation); \( \sigma^2 \): variance (\( \sigma \): the standard deviation).

- If we define a new variable \( u=(x-\mu)/\sigma \),

\[ f(x) \sim \frac{1}{\sqrt{2\pi}} e^{-u^2/2} \]

2. Conditional probability

- Conditioning the joint distribution on a particular observation.

- Conditional probability \( P(X|Y) \): the measure of how likely an event \( X \) happens under the condition \( Y \).

\[ P(x|y) = \frac{P(x,y)}{P(y)} \]

- Example: two dices \( D_1, D_2 \)
  - \( P(D_1) \rightarrow \) probability for picking \( i \) using dicer \( D_1 \)
  - \( P(D_2) \rightarrow \) probability for picking \( i \) using dicer \( D_2 \)

3. Probability models

- A system that produces different outcomes with different probabilities.

- It can simulate a class of objects (events), assigning each an associated probability.

- Simple objects (processes) \( \rightarrow \) probability distributions

4. Typical probability distributions

- Binomial distribution
- Gaussian distribution
- Multinomial distribution
- Poisson distribution
- Dirichlet distribution

5. Binomial distribution

- An experiment with binary outcomes: 0 or 1;

- Probability distribution of a single experiment: \( P('1')=p \) and \( P('0')=1-p \);

- Probability distribution of \( N \) tries of the same experiment

\[ Bi(k \text{ '1' s out of } N \text{ tries}) \sim \binom{N}{k} p^k (1-p)^{N-k} \]

6. Gaussian distribution

- When \( N \rightarrow \infty \), \( Bi \rightarrow \text{Gaussian distribution} \)

- The Gaussian (normal) distribution is a continuous probability distribution with probability density function defined as:

\[ f(x; \mu, \sigma^2) = \frac{1}{\sigma \sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \]

- \( \mu \): mean (expectation); \( \sigma^2 \): variance (\( \sigma \): the standard deviation)

- If we define a new variable \( u=(x-\mu)/\sigma \)

\[ f(x) \sim \frac{1}{\sqrt{2\pi}} e^{-u^2/2} \]

7. Gaussian distribution

- The standard normal distribution when \( \mu = 0 \) and \( \sigma^2 = 1 \)
Multinomial distribution

- An experiment with $K$ independent outcomes with probabilities $\theta_i$, $i=1,\ldots,K$, $\sum \theta_i = 1$.
- Probability distribution of $N$ tries of the same experiment, getting $n_i$ occurrences of outcome $i$, $\sum n_i = N$ ($n=\{n_i\}$).

\[
P(n|\theta) = \frac{M(n)}{\prod_i \theta_i^{n_i}}
\]

\[
M(n) = \frac{n_1!n_2!\cdots n_K!}{(\sum n_k)!}
\]

Example: a fair dice

- Probability: outcomes (1,2,..,6) are equally likely to occur
- Probability of rolling 1 dozen times (12) and getting each outcome twice:
  \[
  \frac{12!}{(6!)^2} \times 1.87 \times 10^{-4}
  \]

Example: a loaded dice

- Probability: outcomes (1,2,..,6) are unequally likely to occur: $P(6)=0.5$, $P(1)=P(2)=\ldots=P(5)=0.1$
- Probability of rolling 1 dozen times (12) and getting each outcome twice:
  \[
  \frac{12!}{(0.5)^2 \times 0.1} \times 1.87 \times 10^{-4}
  \]

Poisson distribution

- Poisson gives the probability of seeing $n$ events over some interval, when there is a probability $p$ of an individual event occurring in that period.

Poisson distribution for sequencing coverage modeling

Assuming uniform distribution of reads:
- Length of genomic segment: $L$
- Number of reads: $n$
- Coverage $\lambda = nL/L$

How much coverage is enough (or what is sufficient oversampling)?
- Lander-Waterman model: $P(x) = (\lambda^x e^{-\lambda}) / x!$
  $P(x=0) = e^{-\lambda}$

where $\lambda$ is coverage
Dirichlet distribution

- Outcomes: \( \theta = (\theta_1, \theta_2, \ldots, \theta_K) \)
- Density: \( D(\theta|\alpha) = Z^{-1}(\alpha) \prod_{i=1}^{K} \theta_i^{\alpha_i-1} \delta(\sum_{i=1}^{K} \theta_i - 1) \)
  \( Z(\alpha) = \int \prod_{i=1}^{K} \theta_i^{\alpha_i-1} \delta(\sum_{i=1}^{K} \theta_i - 1) \, d\theta = \prod_{i=1}^{K} \Gamma(\alpha_i) / \Gamma(\sum_{i=1}^{K} \alpha_i) \)
- \( (\alpha_1, \alpha_2, \ldots, \alpha_K) \) are constants \( \rightarrow \) different \( \alpha \) gives different probability distribution over \( \theta \).
- \( K=2 \rightarrow \) Beta distribution

Example: dice factories

- Dice factories produce all kinds of dices: \( \theta(1), \theta(2), \ldots, \theta(6) \)
- A dice factory distinguish itself from the others by parameters \( \alpha(\alpha_1, \alpha_2, \alpha_3, \alpha_4, \alpha_5, \alpha_6) \)
- The probability of producing a dice \( i \) in the factory \( \alpha \) is determined by \( \theta(i|\alpha) \)

Probabilistic model

- Selecting a model
  - A model can be anything from a simple distribution to a complex stochastic grammar with many implicit probability distributions
  - Probabilistic distributions (Gaussian, binomial, etc)
  - Probabilistic graphical models
    - Markov models
    - Hidden Markov models (HMM)
    - Bayesian models
    - Stochastic grammars
- Data \( \rightarrow \) model (learning)
  - The parameters of the model have to be inferred from the data
  - MLE (maximum likelihood estimation) & MAP (maximum a posterior probability)
- Model \( \rightarrow \) data (inference/sampling)

Example: a loaded dice

- Loaded dice: to estimate parameters \( \theta_1, \theta_2, \ldots, \theta_6 \) based on N observations \( D=d_1, d_2, \ldots, d_N \)
- \( \theta_i=n_i/N \), where \( n_i \) is the occurrence of \( i \) outcome (observed frequencies), is the maximum likelihood solution (BSA 11.5)
  \( P(n|\theta_{MLE}) > P(n|\theta) \) for any \( \theta \neq \theta_{MLE} \)
- Learning from counts

MLE

- Estimating the model parameters (learning): from large sets of trusted examples
- Given a set of data \( D \) (training set), find a model with parameters \( \theta \) with the maximal likelihood \( P(D|\theta) \)
  \( \hat{\theta}_{MLE} = \arg \max_{\theta} P(D|\theta) \)

When to use MLE

- A drawback of MLE is that it can give poor estimations when the data are scarce
  - E.g., if you flip coin twice, you may only get heads, then \( P(\text{tail}) = 0 \)
- It may be wiser to apply prior knowledge (e.g., we assume \( P(\text{tail}) \) is close to 0.5)
  - Use MAP instead
Parameters. Now we have,

\[ P(\theta|D) = \frac{P(D|\theta)P(\theta)}{P(D)} \]

\[ P(\theta) \rightarrow \text{prior probability} \]

\[ P(\theta|D) \rightarrow \text{posterior probability} \]

MAP

- Bayesian statistics
  \[ P(\theta|D) = \frac{P(D|\theta)P(\theta)}{P(D)} \]
  \[ = \frac{\int P(D|\theta)P(\theta) \, d\theta}{\int P(D|\theta) \, d\theta} \]

- MAP
  \[ \hat{	heta}_{MAP} = \arg \max_{\theta} P(\theta|D) \]
  \[ = \arg \max_{\theta} \frac{P(D|\theta)P(\theta)}{P(D)} \]
  \[ = \arg \max_{\theta} P(D|\theta)P(\theta) \]

Example: two dices

- Prior probabilities: fair dice 0.99; loaded dice: 0.01;
- Loaded dice: P(6)=0.5, P(1)=...=P(5)=0.1
- Data: 3 consecutive ‘6’s:
  - P(loaded|3 6’s)=P(loaded)*[P(3 6’s|loaded)*P(3 6’s)]=0.01*(0.5 / 6)
  - P(fair|3 6’s)=P(fair)*[P(3 6’s|fair)*P(3 6’s)]=0.99 * ((1/6)^3 / 6)
- Model comparison by using likelihood ratio: P(loaded|3 6’s)/P(fair|3 6’s) < 1
- So fair dice is more likely to generate the observation.

Learning from counts: including prior

- Use prior knowledge when the data is scarce
- Use Dirichlet distribution as prior for the multinomial distribution:
  - Posterior \[ \frac{P(\theta|n)}{P(n)} = \frac{P(n|\theta)P(\theta)}{P(n)} \]
  - Posterior mean estimator
    \[ \hat{\theta}_{ME} = \int \theta D(\theta|n + \alpha) \, d\theta = \int \theta \frac{1}{\Gamma(\alpha + n)} \theta^{n + \alpha - 1} \, d\theta \]
  - Equivalent to add \( \alpha \) as pseudo-counts to the observation \( n_i \) (BSA 11.5)
  - We can forget about statistics and use pseudo-counts in the parameter estimation!

Entropy

- Probabilities distributions \( P(x) \) over \( K \) events
  \[ H(x) = - \sum P(x) \log P(x) \]
  - Maximized for uniform distribution \( P(x)=1/K \)
  - A measure of average uncertainty

Sampling

- Probabilistic model with parameter \( \theta \rightarrow P(x|\theta) \) for event \( x \)
- Sampling: generate a large set of events \( x \) with probability \( P(x|\theta) \)
- Random number generator (function \( rand() \)) picks a number randomly from the interval \([0,1]\) with the uniform density;
- Sampling from a probabilistic model \( P(x|\theta) \) to a uniform distribution
  - For a finite set \( x \in X \), find s.t. \( P(x_1)+...+P(x_{s-1}) < \text{rand}(0,1) < P(x_1)+...+P(x_s) + P(x) \)

Mutual information

- Measure of independence of two random variable \( X \) and \( Y \)
  \[ P(X|Y)=P(X) \text{ and } Y \text{ are independent} \rightarrow P(X|Y)/P(X|Y)=1 \]
  \[ M(X;Y)=\sum_{x,y} P(x,y)\log[P(x,y)/P(x)P(y)] \]
  - 0 \( \rightarrow \) independent