Ch 7 Probability Theory and Stochastic Simulation:

Frequentist statistics:

- Probability of observing E: $p(E) \approx \frac{N_E}{N}$
- Joint Probability: $p(E_1 \cap E_2) = p(E_1)p(E_2 \mid E_1)$
- Expectation: $E(W) = \langle W \rangle \approx \frac{1}{N_{\exp}} \sum_{\nu=1}^{N_{\exp}} W_\nu$

Bayes' Theorem: $p(E_1)p(E_2 \mid E_1) = p(E_2)p(E_1 \mid E_2)$

- Bayes' Theorem is general.

Definitions:

- variance: $\mathrm{var}(W) = E\left[(W - E(W))^2\right] = E(W^2) - \left[E(W)^2\right]$
- (X, Y independent, var(X+Y) = var(X) + var(Y))
- standard deviation: $\sigma = \sqrt{\mathrm{var}(W)}$
- covariance $\mathrm{cov}(X,Y) = E\{[X - E(X)][Y - E(Y)]\}$, for two random variables X and Y
- covariance matrix

Important Probability Distributions Definitions:

- Discrete random variable
  - For $X_i = \{X_1, X_2, ..., X_m\}$
  - $N(X_i) =$ number of observations of Xi
  - T is the total number of observations $\sum_{i=1}^{M} N(X_j) = T$
  - Probability is definied by: $P(X_i) = \frac{N(X_i)}{T}$
  - Normalization is defined by $\sum_{i=1}^{M} P(X_j) = 1$
- Continuous random variable
  - This is just the continuous version of the above, defined by integrals instead of limits, differentials instead of increments
  - Normalization condition: $\int_{x_{lo}}^{x_{hi}} p(x)dx = 1$
  - Expectation $E(x) = \langle x \rangle = \int_{x_{lo}}^{x_{hi}} xp(x)dx$
- Cumulative Probability distribution
  - Basis of **RAND** in matlab

- $F(X_M) = \int_{x_{lo}}^{x} p(x')dx' = u$
- u is defined as uniformly distributed $0 \le u \le 1$

Bernoulli trials

- Concept that observed error is the net sum of many small random errors

Random Walk Problem

- key point: independence of coin tosses
- Main results: $\langle x \rangle = 0$ $\langle x^2 \rangle = nl^2$

Binomial Distribution

- probability distribution: $P(n, n_H) = \binom{n}{n_H} p_H^{\,n_H} (1 - p_H)^{(n - n_H)}$
- binomial coefficient: $\binom{n}{n_H} = \dfrac{n!}{n_H!(n - n_H)!}$
- **BINORND** Matlab to generate random number distributed using binomial distribution

Gaussian (Normal) Distribution

- Take binomial distribution, change into probability of observing net displacement after n steps of length l
  - $p(x; n, l) = \dfrac{n!}{\left[\frac{(n + x/l)}{2}\right]!\left[\frac{(n - x/l)}{2}\right]!}\left(\frac{1}{2}\right)^n$
- Evaluate in limit that n→ ∞, take natural log, and use Stirling's approximation
- Algebra, and taylor expand around the ln terms
- Taking the exponential and normalizing such that: $\int_{-\infty}^{\infty} P(x; n, l)dx = 1$
- $P(x; \sigma) = \dfrac{1}{\sigma\sqrt{2\pi}} \exp\left[-\dfrac{x^2}{2\sigma^2}\right]$ $\sigma^2 = nl^2$
- Binomial Distribution of random walk reduces to Gaussian Distribution as n→ ∞
- Central Limit Theorem: sequence of random variables, which are not distributed normally, the statistic
  - $S_n = \dfrac{1}{\sqrt{n}} \sum_{j=1}^{N} \dfrac{\xi_j - \mu_{j.}}{\sigma_j}$

- random variable: ξ_j with mean $\mu_{j.}$ and variance $\sigma_j^{\ 2}$

- is normally distributed in the limit that n→ ∞, with variance = 1
  - $P(S_n) = \frac{1}{\sqrt{2\pi}} \exp\left[-\frac{S_n^{\ 2}}{2}\right]$
- Non-zero Mean (basis of **randn**)
  - $N(\mu, \sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right]$

- Multivariate Gaussian Distribution (use of covariance matrix)
  - Covariance Matrix: $[\mathrm{cov}(\underline{\nu})]_{ij} = E\{[\nu_i - E(\nu_i)][\nu_j - E(\nu_j)]\}$
  - Covariance Matrix is always symmetric and positive definite
  - For independent components: $\mathrm{cov}(\underline{\nu}) = \sigma^2 I$
  - $\mathrm{cov}(\underline{\nu}) = \mathrm{cov}(A\underline{x}) = A[\mathrm{cov}(\underline{x})]A^T$
  - if v is a random vector and c is a constant vector:
  - $\mathrm{var}(\underline{c}\cdot\underline{\nu}) = \mathrm{var}(\underline{c}^T\underline{\nu}) = \underline{c}^T[\mathrm{cov}(\underline{\nu})]\underline{c} = \underline{c}\cdot[\mathrm{cov}(\underline{\nu})]\underline{c}$
  - $P(\underline{\nu}; \underline{\mu}, \Sigma) = \frac{1}{(2\pi)^{N/2}\sqrt{|\Sigma|}} \exp\left\{-\frac{1}{2}(\underline{\nu}-\underline{\mu})^T \Sigma^{-1}(\underline{\nu}-\underline{\mu})\right\}$

Poisson Distribution

- Poisson distribution can be used to determine probability of success if there are n trials, derived in the limit as n→ ∞
- Total number of successes in trial is a random variable, which d
- Another limiting case of binomial distribution
- $P(\xi; n, p) = \frac{(pn)^{\xi}}{\xi!} e^{-pn}$
- p = probability of individual success
- n = number of trials
- ξ = result if success or failure, typically {1,0} with different probabilities

Boltzmann/Maxwell Distributions

- $P(\underline{q}) = \frac{1}{Q} \exp\left[-\frac{E(q)}{kT}\right]$
- Q is the normalization constant
- Replacing E(q) for kinetic energy we arrive at Maxwell Distribution
  - $P(\underline{\nu}) \propto \exp\left[-\frac{m|\underline{\nu}|^2}{2kT}\right]$

Brownian Dynamics and Stochastic Differential Equations

- velocity autocorrelation function
    - $C_{V_x}(t \geq 0) \approx C_{V_x}(0)e^{-t/\tau_{V_x}} \qquad \tau_{v_x} = \frac{2\rho R^2}{9\mu}$
    - $\langle V_x(t)V_x(0)\rangle = 2D\delta(t)$
- Dirac Delta Function
    - $\delta(t) = \lim_{\sigma \to 0} \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{t^2}{2\sigma^2}\right]$
    - $\int_{-\infty}^{\infty} f(t)\delta(t)dt = f(0)$
- Langevin equation
- Wiener process
- Stochastic Differential equations
    - Explicit Euler SDE method
    - $x(t+\Delta t) - x(t) = -\frac{1}{\zeta}\left(\left.\frac{dU}{dx}\right|_{x(t)}\right)(\Delta t) + [2D]^{1/2}(\Delta W_t)$
- Ito's Stochastic Calculus
    - Example: Black-Scholes
    - Fokker-Planck
    - Einstein Relation
    - Brownian motion in multiple dimensions
- MCMC
    - Stat Mech example
    - Metropolis recipe (pg497)
    - Example: Ising Lattice
    - Field theory
    - Monte Carlo Integration
    - Simulated annealings
    - Genetic Programming

<u>Bayesian Statistics and Parameter Estimation</u>

Goal of this material is to draw conclusions from data ("statistical inference") and estimate parameters. Basic definitions

- Predictor variables: $\underline{x} = [x_1\ x_2\ x_3 \dots\ x_M]$
- Response variable: $\underline{y}^{(R)} = [y_1\ y_2\ y_3\ \dots\ y_L]$
- $\underline{\theta}$: model parameters

Main goal: match model prediction to that of the observed response by selecting $\underline{\theta}$.

Single-Response Linear Regression

- set of predictor variables, known a priori: $\underline{x}^{[k]} = [x_1^{[k]}\ x_2^{[k]}\ x_3^{[k]} \ldots x_M^{[k]}]$, for the kth experiment
- measurement $y^{[k]}$
- assume a linear model: $y^{[k]} = \beta_0 + \beta_1 x_1^{[k]} + \beta_2 x_2^{[k]} + \ldots + \beta_M x_M^{[k]} + \varepsilon^{[k]}$
- the error in $\varepsilon^{[k]}$ is responsible for the difference between model and observed
- define $\underline{\theta} = [\beta_0\ \beta_1\ \beta_2 \ldots \beta_M]^T$
- response is: $y^{[k]} = \underline{x}^{[k]} \cdot \underline{\theta}^{(true)} + \varepsilon^{[k]}$
- model prediction is: $\hat{y}^{[k]} = \underline{x}^{[k]} \cdot \underline{\theta}^{(true)}$
- define design matrix X, which contains all information about every experiment (with different predictor variables)

$$X = \begin{bmatrix} ---\underline{x}^{[1]}--- \\ ---\underline{x}^{[2]}--- \\ \vdots \\ ---\underline{x}^{[N]}--- \end{bmatrix}$$

- vector of predicted responses:

$$\underline{\hat{y}}(\underline{\theta}) = \begin{bmatrix} \underline{\hat{y}}^{[1]}(\underline{\theta}) \\ \underline{\hat{y}}^{[2]}(\underline{\theta}) \\ \vdots \\ \underline{\hat{y}}^{[N]}(\underline{\theta}) \end{bmatrix} = X\underline{\theta}$$

Linear Least Squares Regression

- minimize sum of squared errors: $S(\underline{\theta}) = \sum_{k=1}^{N} \left[y^{[k]} - \hat{y}^{[k]}(\underline{\theta}) \right]^2$
- First derivative = 0, 2nd derivative is > 0, using these conditions with above equation you can derive a linear system
- $(X^T X)\underline{\theta}_{LS} = X^T \underline{y} \rightarrow \underline{\theta}_{LS} = (X^T X)^{-1} X^T \underline{y}$ **(review point?)**
- $X^T X$, contains information about experimental design to probe the parameter values
- $X^T X$ is a real, symmetric matrix that is positive-semidefinite
- Solving this is through standard linear solving, or QR decomposition or some other method
- All this estimates parameters, but does not give us accuracy of our estimates

Bayesian view of statistical inference

- Statement of belief (especially in random number generators)

Bayesian view of single-response regression

- Begin with $y^{[k]} = \underline{x}^{[k]} \cdot \underline{\theta}^{(true)} + \varepsilon^{[k]}$
- When we repeat this experiment multiple times, we get a vector $\underline{\varepsilon}$
- With Gauss-Markov Conditions: $E\left(\varepsilon^{[k]}\right) = 0 \quad \operatorname{cov}\left(\varepsilon^{[k]}, \varepsilon^{[j]}\right) = \delta_{kj}\sigma^2$
- We also assume that our error is normally distributed
- Probability of observing some response $\underline{y}$
  - $p\left(\underline{y} \mid \underline{\theta}, \sigma\right) = \left(\frac{1}{\sqrt{2\pi}}\right)^N \sigma^{-N} \exp\left[-\frac{1}{2\sigma^2} S\left(\underline{\theta}\right)\right]$
- We use Bayes' Theorem to get probability of $\underline{\theta}$ and σ
- Posterior density: $p\left(\underline{\theta}, \sigma \mid \underline{y}\right) = \frac{p\left(\underline{y} \mid \underline{\theta}, \sigma\right) p\left(\underline{\theta}, \sigma\right)}{p\left(\underline{y}\right)}$
- p($\underline{y}$) is a normalizing factor
- we redefine p($\underline{y}$ | $\underline{\theta}$, σ) to l($\underline{\theta}$, σ| $\underline{y}$)
- in the Bayesian framework we want to maximize posterior density
- Non-informative priors: p($\underline{\theta}$,σ)=p($\underline{\theta}$)p(σ) $\quad p\left(\underline{\theta}\right) \sim c \quad p\left(\underline{\theta}\right) \propto \sigma^{-1}$

Nonlinear least squares

- the treatment via least squares still works, we just use numerical optimization, utilizing a cost function, to get there: (**review point?**)
- $F_{cost}\left(\underline{\theta}\right) = \frac{1}{2} S\left(\underline{\theta}\right) = \frac{1}{2} \sum_{k=1}^{N} \left[y^{[k]} - f\left(\underline{x}^{[k]}; \underline{\theta}\right)\right]^2$
- use of linearized design matrix
- Hessians (first order approximation to get to X^TX). Remember to get convergence, approximate Hessian needs to be positive-definite.
- Levenberg-Marquardt method: ill-conditioned systems

Generating Confidence Intervals

- t-statistic
  - $t \equiv \frac{\bar{y} - \theta}{\left(s / \sqrt{N}\right)}$
  - $p\left(t \mid \nu\right) \propto \left[1 + \frac{t^2}{\nu}\right]^{-\frac{(\nu+1)}{2}}$
  - in the limit that ν approaches infinity, t-distribution reduces to Normal distribution
- confidence intervals for model parameters

- $\theta_j = \theta_{M,j} \pm T_{\nu,\alpha/2} s \left\{ \left[X^T X \big|_{\theta_M} \right]^{-1}_{jj} \right\}^{1/2}$
- $\nu = N - \dim(\theta)$

MCMC in Bayesian Analysis