Prep: bring ID, water, pen and this set of notes; don't be afraid of wasting paper!
You got this!

Erdős-Rényi Random Graphs $\mathcal{G}(n,p), p \in [0,1]$

Basic Results:

- $\mathbb{E}[|E|] = \binom{n}{2}p$
- Degree of a node $D \sim \text{Binomial}(n-1,p)$. $\mathbb{E}[D] = (n-1)p$
- If $p(n) = \frac{\lambda}{n}$ then $D \approx \text{Poisson}(\lambda)$ as $n \to \infty$
- $\mathbb{P}[\text{a specific vertex isolated}] = (1-p)^{n-1}$

Connectivity Theorems: Let $p(n) := \lambda \frac{\ln n}{n}$

- If $\lambda < 1$, $\lim_{n\to\infty} \mathbb{P}\left[\mathcal{G}(n,p(n)) \text{ connected}\right] = 0$
  - Almost surely disconnected
  - Bound X_n (# of disconnected nodes)
- If $\lambda > 1$, $\lim_{n\to\infty} \mathbb{P}\left[\mathcal{G}(n,p(n)) \text{ connected}\right] = 1$
  - Almost surely connected
- If $p(n) = \frac{c+\ln n}{n}$, with constant $c \in \mathbb{R}$
  - $\lim_{n\to\infty} \mathbb{P}\left[\mathcal{G}(n,p(n)) \text{ connected}\right] = e^{-e^{-c}}$
- If $np < 1$, then $\mathcal{G}(n,p)$ have no connected component of size $\geq O(\log N)$ almost surely
- If $np = 1$, then $\mathcal{G}(n,p)$ have a largest component on the order of $n^{\frac{2}{3}}$ almost surely
- If $np \to c > 1$, then $\mathcal{G}(n,p)$ have a unique giant component with no other component having $\geq O(\log N)$ vertices almost surely.

MAP and MLE

Let X be causes and Y be observations

Maximum A Posteriori:

- $MAP[X|Y=y] = \arg\max_{x\in\mathcal{X}} \mathbb{P}[X=x]\mathbb{P}[Y=y|X=x]$
  - Find $MAP[X|Y]$ by pattern on $MAP[X|Y=y]$

Maximum Likelihood Estimation:

- Special case of MAP where $\mathbb{P}[X=x] = c\ \forall x$
- $MLE[X|Y=y] = \arg\max_{x\in\mathcal{X}} \mathbb{P}[Y=y|X=x]$

Four Versions of Bayes

- X continuous; Y continuous
$$f_{X|Y}(x|y) = \frac{f_X(x)f_{Y|X}(y|x)}{\int_{-\infty}^{\infty} f_X(t)f_{Y|X}(y|t)\,\mathrm{d}t}$$
- X continuous; Y discrete
$$f_{X|Y}(x|y) = \frac{f_X(x)\mathbb{P}_{Y|X}[y|x]}{\int_{-\infty}^{\infty} f_X(t)\mathbb{P}_{Y|X}[y|t]\,\mathrm{d}t}$$
- X discrete; Y discrete
$$\mathbb{P}_{X|Y}(x|y) = \frac{\mathbb{P}_X[x]\mathbb{P}_{Y|X}[y|x]}{\sum_{t=-\infty}^{\infty} \mathbb{P}_X[t]\mathbb{P}_{Y|X}[y|t]}$$
- X discrete; Y continuous
$$\mathbb{P}_{X|Y}(x|y) = \frac{\mathbb{P}_X[x]f_{Y|X}(y|x)}{\sum_{t=-\infty}^{\infty} \mathbb{P}_X[t]f_{Y|X}(y|t)}$$

Hypothesis Testing

- $X = 0$: null hypothesis
- $X = 1$: alternate hypothesis
- Y: data
- $\hat{X}: Y \to \{0,1\}$: decision rule
- Probability of False Alarm: $\mathbb{P}\left[\hat{X} = 1|X=0\right]$
- Probability of Correct Detection: $\mathbb{P}\left[\hat{X} = 1|X=1\right]$
- Type II Error: $\mathbb{P}\left[\hat{X} = 0|X=1\right] = 1 - PCD$

Optimization Problem:

$$\max_{\hat{X}} \mathbb{P}[\hat{X}=1|X=1] \quad \text{s.t. } \mathbb{P}\left[\hat{X}=1|X=0\right] \leq \beta$$

Equivalent Terms

	$X = 0$	$X = 1$
$\hat{X}(Y) = 0$		• False Negative • Type II Error
$\hat{X}(Y) = 1$	• False Positive • Significance Level • Type I Error	• Power

Neyman-Pearson Lemma

Define likelihood function:

$$L(Y=y) = \frac{\mathbb{P}[Y=y|X=1]}{\mathbb{P}[Y=y|X=0]} = \frac{f(Y=y|X=1)}{f(Y=y|X=0)}$$

NP states optimal decision rule is in the form:

$$\hat{X}(y) = \begin{cases} 1, & L(Y) > \lambda \\ \text{Bernoulli}(\gamma), & L(Y) = \lambda \\ 0, & L(Y) < \lambda \end{cases}$$

where $\gamma \in [0,1]$ chosen such that:

- $\mathbb{P}\left[\hat{X}(Y) = 1|X=0\right] = \beta$
- $\mathbb{P}[L(Y) > \lambda|X=0] + \gamma\mathbb{P}[L(Y) = \gamma|X=0] = \beta$

- Find equivalent conditions for $L(x) > c$, like $\log L(x)$
- Exploit monotonicity of Y with respect to $L(Y)$ (can be increasing or decreasing)

Hilbert Space $\mathcal{H} \coloneqq \{X: \mathbb{E}[X^2] < \infty\}$

- $\mathcal{H}$: complete inner product space; any Cauchy sequence converges in the space
- $\langle X, Y\rangle \coloneqq \mathbb{E}[XY]$, $\|X\|^2 = \mathbb{E}[X^2]$
- $\cos\theta = \frac{\langle X,Y\rangle}{\|X\|\|Y\|}$
- If X, Y zero mean, $\mathrm{Var}[X] = \|X\|^2$, $\cos\theta = \mathrm{corr}(X, Y)$
- $\|X - \mathbb{E}[X]\| = \sqrt{\mathbb{E}[(X-\mathbb{E}[X])^2]} = \sqrt{\mathrm{Var}[X]}$
- $|\mathbb{E}[XY]| = |\langle X, Y\rangle| \le \|X\|\|Y\| = \sqrt{\mathbb{E}[X^2]\mathbb{E}[Y^2]}$

Orthogonality $\langle X, Y\rangle = 0$

- If either X, Y zero mean, $\mathbb{E}[XY] = 0 \Rightarrow \mathrm{Cov}[X,Y] = 0$ (i.e. uncorrelated)
- $\hat{Y} = \Pi_U(Y) = \arg\min_{Z\in U}\|Y - Z\|^2$ unique
- $\mathbb{E}\left[\left(Y - \Pi_U(Y)\right)Z\right] = \mathbb{E}\left[\left(Y - \hat{Y}\right)Z\right] = 0$
- If $\langle X, Y\rangle = 0$, $\|X+Y\|^2 = \|X\|^2 + \|Y\|^2$

Linear Least Squares Estimation (LLSE)

- $\mathbb{L}(X) = \{a + bX : a, b \in \mathbb{R}\} = \mathrm{Span}\{1, X\}$
- $\mathbb{L}[Y|X] = \Pi_{\mathrm{Span}\{1,X\}}(Y) = \Pi_{\mathrm{Span}\left\{1, \frac{X-\mathbb{E}[X]}{\sqrt{\mathrm{Var}[X]}}\right\}}(Y)$
- $\mathbb{L}[Y|X] = \mathbb{E}[Y] + \frac{\mathrm{Cov}[X,Y]}{\mathrm{Var}[X]}(X - \mathbb{E}[X])$
- [Error] $\|Y - L[Y|X]\|^2 = \mathrm{Var}[Y] - \frac{\mathrm{Cov}[X,Y]^2}{\mathrm{Var}[X]}$
- [Unbiased] $\mathbb{E}\left[Y - \mathbb{L}[Y|X]\right] = 0$
- [Uncorrelated] $\mathbb{E}[X(Y - \mathbb{L}[Y|X])] = 0$

Orthogonal Updates:

- If X, Y, Z zero mean, $\mathbb{L}[Y|X,Z] = \mathbb{L}[Y|X] + \mathbb{L}\left[Y\middle|\tilde{Z}\right]$ where $\tilde{Z} \coloneqq Z - \mathbb{L}[Z|X]$
- If $\langle Z, X\rangle = 0$, $\mathbb{L}[Y|X,Z] = \mathbb{L}[Y|X] + \mathbb{L}[Y|Z]$

Random Vectors:

- $\mathbb{L}[Y|X] = \mathbb{E}[X] + \mathrm{Cov}[X,Y]\Sigma_Y^{-1}(Y - \mathbb{E}[Y])$
- $\mathbb{E}[\|Y - \mathbb{L}[Y|X]\|^2] = \mathrm{tr}(\Sigma_X - \mathrm{Cov}[Y,X]\Sigma_Y^{-1}\mathrm{Cov}[X,Y])$

Minimum Mean Square Estimation (MMSE)

- $\mathbb{E}[Y|X] \coloneqq \arg\min_{\phi}\mathbb{E}\left[\left(Y - \phi(X)\right)^2\right]$
- Equivalently, $\mathbb{E}[(Y - \mathbb{E}[Y|X])\phi(X)] = 0\ \forall\phi$
- If Φ satisfies $\mathbb{E}\left[\left(Y - \Phi(X)\right)\phi(X)\right] = 0\ \forall\phi$, then $\mathbb{E}[Y|X] \equiv \Phi$
- $\forall\phi\ \mathbb{E}\left[\left(Y - \phi(X)\right)^2\right] \ge \mathbb{E}[(Y - \mathbb{E}[Y|X])^2]$

Kalman Filter (Vector)

Properties of Random Vectors (revisited)

- $\mathrm{Var}[AX] = A\mathrm{Var}[X]A^T$
- $\mathrm{Cov}[AX, BY] = A\mathrm{Cov}[X,Y]B^T$, bilinear
- $\mathrm{Cov}[X,Y] = \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])^T] = \mathbb{E}[XY^T] - \mathbb{E}[X]\mathbb{E}[Y]^T$
- Assuming Y zero mean, $\Pi_Y(X) = \mathrm{Cov}[X,Y]\mathrm{Var}[Y]^{-1}Y = \mathbb{E}[XY^T](\mathbb{E}[YY^T])^{-1}Y$
- $tr(\mathbb{E}[ab^T]) = \mathbb{E}[b^Ta]$

Modelling Variables

- $X_n \in \mathbb{R}^d$: state of dynamical system
- X_0: starting state
- $A \in \mathbb{R}^{d\times d}$: transition model
- $V_n \in \mathbb{R}^d \sim N(0, \Sigma_V)$: process noise, i.i.d.
- $Y_n \in \mathbb{R}^e$: observations
- $C \in \mathbb{R}^{e\times d}$: observation model
- $W_n \in \mathbb{R}^e \sim N(0, \Sigma_W)$: observation noise, i.i.d.

Transition Equations:

- $X_n = AX_{n-1} + V_n \quad n \ge 1$
- $Y_n = CX_n + W_n \quad n \ge 1$

Jointly Gaussian Random Variables

Definition: X is <u>jointly Gaussian</u> if $X = AZ + \mu$ where $Z_i \sim N(0,1)$ i.i.d., $A \in \mathbb{R}^{n\times l}$, $\mu \in \mathbb{R}^n$

Properties of Random Vectors and Covariance

- $\mathbb{E}[X] = \mu$, $\Sigma = \mathbb{E}[(X-\mu)(X-\mu)^T] = AA^T$
- $\Sigma_{ij} = \mathrm{Cov}\left[X_i, X_j\right]$, $\mathbb{E}[ZZ^T] = \mathbb{I}_l$
- $\Sigma \succcurlyeq 0$ is equivalent to:
  - $\Sigma = AA^T$ (Cholesky Decomposition)
  - $\forall x$, $x^T\Sigma x \ge 0$
  - Σ has real, nonnegative eigenvalues
- $\hat{X} = X - \mu$ is the centered version of X
- $\mathrm{Var}\left[u^T\hat{X}\right] = u^T\Sigma u$ (if u unit vector, interpret as variance of projection of $\hat{X}$ along u)
- $\Sigma = U\Lambda U^T \Rightarrow A = U\Lambda^{\frac{1}{2}}U^T$

Properties of Jointly Gaussian $X \sim N(\mu, \Sigma)$

- $f_X(x) = \frac{1}{\sqrt{(2\pi)^n\det\Sigma}}e^{-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)}$ assuming $\Sigma \succ 0$
- Independent if and only if uncorrelated (i.e. Σ diagonal)
- Linear combinations of jointly Gaussian RV are jointly Gaussian (use matrix)
- If any linear combination of $X_1, \dots, X_n$ i.e. u^TX for $u \in \mathbb{R}^n$ follows a normal distribution, then X jointly Gaussian.
- If X, Y jointly Gaussian, $\mathbb{E}[X|Y] \equiv \mathbb{L}[X|Y]$
- Level curves of jointly Gaussian RVs are ellipse: any slice is a normal distribution

$$\mathbb{E}[(Y - \mathbb{L}[X|Y])X] = 0 \Rightarrow \mathrm{Cov}[Y - \mathbb{L}[X|Y], X] = 0$$

General Modelling Equations:

- $X_n = A^n X_0 + \sum_{i=1}^{n} A^{n-i} V_i$
- $\mathbb{E}[X_n] = A^n \mathbb{E}[X_0]$
- $\text{Var}[X_n] = A^n \text{Var}[X_0](A^T)^n + \sum_{i=0}^{n-1} A^i \Sigma_V (A^T)^i$
- $\lim_{n\to\infty} \text{Var}[X_n] = \sum_{i=0}^{\infty} A^i \Sigma_V (A^T)^i$ (when $\|A\| < 1$)
- $Y_n = C(A^n X_0 + \sum_{i=1}^{n} A^{n-i} V_i) + W_n$

Prediction Variables

- $\hat{X}_{n|k} \in \mathbb{R}^d \coloneqq \mathbb{L}[X_n | Y_1, \dots, Y_k]$: estimate X_n given observations $Y_1, \dots, Y_k$
- $\hat{X}_{0|0} = X_0$ (know initial state)
- $\Sigma_{n|n} \coloneqq \text{Var}\left[X_n - \hat{X}_{n|n}\right]$: estimation variance
- $\Sigma_{n|k} \coloneqq \text{Var}\left[X_n - \hat{X}_{n|k}\right]$: prediction variance
- $\mathbb{E}\left[\left\|X_n - \hat{X}_{n|n}\right\|^2\right] = \text{tr}(\Sigma_{n|n})$: estimation error
- $\mathbb{E}\left[\left\|X_n - \hat{X}_{n|k}\right\|^2\right]$: prediction error
- $\tilde{Y}_n$: innovation at time n (the orthogonal component added by Y_n to $\text{Span}\{1, Y_1, \dots, Y_{n-1}\}$)
- K_n: Kalman gain at time n (the projection of X_n onto the span of $\tilde{Y}_n$)

Properties of Prediction Variables:

- $\hat{X}_{n|k} = A^{n-k} \hat{X}_{k|k}$
- $\Sigma_{n|k} = \mathbb{E}\left[(X_n - \hat{X}_{n|k})(X_n - \hat{X}_{n|k})^T\right]$ (since $X_n - \hat{X}_{n|k}$ is zero mean)
- $\hat{X}_{n|n} = \mathbb{L}[X_n | Y_1, \dots, Y_n] = \hat{X}_{n|n-1} + K_n \tilde{Y}_n = A\hat{X}_{n-1|n-1} + K_n \tilde{Y}_n$
- $\tilde{Y}_n = Y_n - \Pi_{\text{Span}\{1, Y_1, \dots, Y_{n-1}\}}(Y_n) = Y_n - CA\hat{X}_{n-1|n-1}$
- $K_n = \langle X_n, \tilde{Y}_n \rangle = \text{Cov}[X_n, \tilde{Y}_n]\text{Var}[\tilde{Y}_n]^{-1} = \Sigma_{n|n-1} C^T \left(C\Sigma_{n|n-1} C^T + \Sigma_W\right)^{-1}$
- $\hat{X}_{n|n} = A\hat{X}_{n-1|n-1} + K_n \tilde{Y}_n = (\mathbb{I} - K_n C)A\hat{X}_{n-1|n-1} + K_n Y_n$ (i.e. optimal estimate of X_n lies between past prediction and present observation)

Derived Equations:

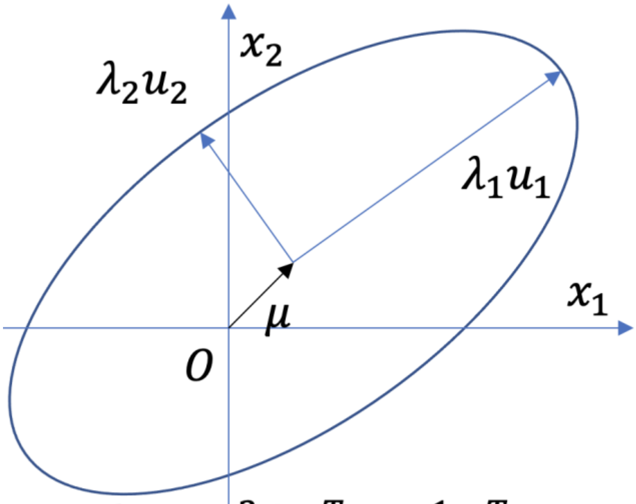
- Let $B_j = (\mathbb{I} - K_j C)A$; $\hat{X}_{n|n} = B_n B_{n-1} \dots B_1 \hat{X}_{0|0} + \sum_{i=1}^{n} B_n B_{n-1} \dots B_{i+1} K_i Y_i$
- $\mathbb{E}\left[\hat{X}_{n|n}\right] = B_n B_{n-1} \dots B_1 \mathbb{E}[X_0]$

$\Rightarrow Y - \mathbb{L}[X|Y], X$ independent $\Rightarrow \mathbb{E}[\phi(X)(Y - \mathbb{L}[X|Y])] = 0 \; \forall \phi \Rightarrow \mathbb{E}[X|Y] \equiv \mathbb{L}[X|Y]$

Famous example: $X \sim N(0,1)$, $Y = WX$, $W = \begin{cases} 1, & \text{w.p. } 0.5 \\ -1, & \text{w.p. } 0.5 \end{cases}$ independent of X. X, Y normal distribution, uncorrelated, but not independent.

Density Level Curves:

$g(x) = (x-\mu)^T \Sigma^{-1} (x - \mu)$	$\Sigma = U\Lambda U^T$
$g(x) = x^T U \Lambda^{-1} U^T x = \sum_{i=1}^{n} \frac{1}{\lambda_i} (U^T x)_i^2$	



$\{x \in \mathbb{R}^2 \mid x^T U \Lambda^{-1} U^T x = 1\}$

Kalman Filter Algorithm

Prediction Phase (after time step $n-1$)

- Given: $\left(\hat{X}_{n-1|n-1}, \Sigma_{n-1|n-1}\right)$
- $\hat{X}_{n|n-1} \leftarrow A\hat{X}_{n-1|n-1}$
- $\Sigma_{n|n-1} \leftarrow \text{Var}\left(X_n - A\hat{X}_{n-1|n-1}\right) = A\Sigma_{n-1|n-1}A^T + \Sigma_V$
- $K_n \leftarrow \Sigma_{n|n-1} C^T \left(C\Sigma_{n|n-1} C^T + \Sigma_W\right)^{-1}$ (i.e. can already find Kalman gain here)

Update Phase (at time step n)

- Know: $\left(\hat{X}_{n|n-1}, \Sigma_{n|n-1}\right)$, Y_n
- $\tilde{Y}_n \leftarrow Y_n - C\hat{X}_{n|n-1}$
- $\hat{X}_{n|n} \leftarrow \hat{X}_{n|n-1} + K_n \tilde{Y}_n$
- $\Sigma_{n|n} \leftarrow \text{Var}\left[X_n - \left((I - K_n C)\hat{X}_{n|n-1} + K_n Y_n\right)\right] = (I - K_n C)\Sigma_{n|n-1}$

Kalman Filter Summary

Modelling:

- $X_n = AX_{n-1} + V_n$
- $Y_n = CX_n + W_n$

Kalman Filter (Scalar)

- $X_n = AX_{n-1} + V_n$
- $Y_n = X_n + W_n$
- $\Sigma_{n|n-1} = A^2 \Sigma_{n-1|n-1} + \Sigma_V$
- $K_n = \frac{\Sigma_{n|n-1}}{\Sigma_{n|n-1} + \Sigma_W}$

Algorithm:	Last Resorts
• Initialize $\left(\hat{X}_{0\|0}, \Sigma_{0\|0}\right) \leftarrow (X_0, \mathrm{Var}[X_0])$ • Offline compute estimation variances and Kalman gains: ○ $\Sigma_{n\|n-1} = A\Sigma_{n-1\|n-1}A^T + \Sigma_V$ (prediction) ○ $K_n = \Sigma_{n\|n-1}C^T\left(\left(C\Sigma_{n\|n-1}C^T + \Sigma_W\right)^{-1}\right)$ (gain) ○ $\Sigma_{n\|n} = (I - K_nC)\Sigma_{n\|n-1}$ • Online compute state estimate as new observations arrive: ○ $\hat{X}_{n\|n-1} = A\hat{X}_{n-1\|n-1}$ (prediction) ○ $\hat{Y}_n = Y_n - C\hat{X}_{n\|n-1}$ (innovation) ○ $\hat{X}_{n\|n} = \hat{X}_{n\|n-1} + K_n\tilde{Y}_n$ (update)	• Trust the process • Model the problem, apply relevant treatments • Make life easier, possible to apply the same arguments to $\log\cdot$?