UC Berkeley
Department of Electrical Engineering and Computer Sciences

EECS 126: Probability and Random Processes

Homework 11
Fall 2023

1. **Connected Random Graph**

We start with the empty graph on n vertices. Iteratively, we add an undirected edge, chosen uniformly at random from the edges that are not yet present in the graph, until the graph is connected.

Hint: Recall the coupon collector's problem.

a. Suppose that there are currently k connected components in the graph. Let X_k be the number of edges we need to add until there are $k-1$ connected components. Show that $\mathbb{E}(X_k) \leq \frac{n-1}{k-1}$.

b. Let X be the total number of edges in the final connected graph. Show that $\mathbb{E}(X) \leq Cn \log n$ for some constant C.

Solution:

a. Let p_k be the probability that the first edge we add brings us to $k-1$ components. As we continue to add edges, the probability that each new edge added will produce $k-1$ components from k components is increasing, starting from p_k. If $Y_k \sim \text{Geometric}(p_k)$, then $\mathbb{E}(X_k) \leq \mathbb{E}(Y_k) = \frac{1}{p_k}$. (*See remark.*)
To bound p_k, suppose there are currently k components and u is one endpoint of the edge that we are currently adding. There are at most $n-1$ vertices to which we can connect u, and at least $k-1$ of these will reduce the number of components, so $p_k \geq \frac{k-1}{n-1}$.

b. Observe that $X = \sum_{k=2}^{n} X_k$. Using part a,

$$\mathbb{E}(X) = \sum_{k=2}^{n} \mathbb{E}(X_k) \leq \sum_{k=2}^{n} \frac{n-1}{k-1} = (n-1)H_{n-1} \leq n \log n.$$

Remark. Y_k is intuitively "larger" than X_k, but it is difficult to explain the precise meaning of this in the context of randomness. We say that Y_k *stochastically dominates* X_k if $\mathbb{P}(Y_k \geq x) \geq \mathbb{P}(X_k \geq x)$ for all x, which holds here and implies that $\mathbb{E}(Y_k) \geq \mathbb{E}(X_k)$.

To explain why, we could use the *tail sum* formula or consider a **coupling** argument: suppose that each time we add an edge, we flip a coin of probability p_k. If the coin comes up heads, we add an edge that connects two components; otherwise, we still have some additional chance of connecting two components. In this case, the number of edges until we have $k-1$ components is at most the number of flips until we see heads, or $X_k \leq Y_k$, so $\mathbb{E}(X_k) \leq \mathbb{E}(Y_k)$.

2. **Isolated Vertices**

Consider an Erdős–Rényi random graph $\mathcal{G}(n, p(n))$, where n is the number of vertices and $p(n)$ is the probability that any specific edge appears in the graph. Let X_n be the number of isolated vertices in $\mathcal{G}(n, p(n))$.

a. Show that $\mathbb{E}(X_n) \to \exp(-c)$ as $n \to \infty$ when $p(n) = \frac{(\ln n)+c}{n}$ for some constant c.

b. Conclude that $\mathbb{E}(X_n) \to \infty$ when $p(n) \ll \frac{\ln n}{n}$.

c. Conclude that $\mathbb{E}(X_n) \to 0$, and $X_n \to 0$ in probability, when $p(n) \gg \frac{\ln n}{n}$.

The asymptotic notation $f(n) \ll g(n)$ means that $\frac{f(n)}{g(n)} \to 0$ as $n \to \infty$.

Hint: From Taylor series expansion, $\ln(1 + x) \approx x$ when x is small.

Solution:

a. The probability that any specific vertex is isolated is $(1 - p(n))^{n-1}$, so

$$\mathbb{E}(X_n) = n(1 - p(n))^{n-1}.$$

When $p(n) = \frac{(\ln n)+c}{n}$, using the approximation $\ln(1 + x) \approx x$ for small x,

$$\ln \mathbb{E}(X_n) = \ln n + (n-1) \ln\left(1 - \frac{(\ln n) + c}{n}\right) \sim \ln n - \frac{(n-1)((\ln n) + c)}{n} \to -c.$$

Thus $\mathbb{E}(X_n) \to \exp(-c)$.

b. When $p(n) \ll \frac{\ln n}{n}$, we have $p(n) < \frac{(\ln n)+c}{n}$ for all c, which means that the limit of $\mathbb{E}(X_n)$ is lower bounded by $\exp(-c)$ for all c, or $\mathbb{E}(X_n) \to \infty$.

c. When $p(n) \gg \frac{\ln n}{n}$, then by the same reasoning, the limit of $\mathbb{E}(X_n)$ is upper bounded by $\exp(-c)$ for all c, or $\mathbb{E}(X_n) \to 0$. Now, by Markov's inequality,

$$\mathbb{P}(X_n > 0) = \mathbb{P}(X_n \geq 1) \leq \mathbb{E}(X_n) \to 0,$$

so $X_n \to 0$ in probability as $n \to \infty$.

Remark. We have shown that $\frac{\ln n}{n}$ is a *threshold* for the expected number of isolated vertices, but it is also a threshold for connectivity: if $p(n) = \lambda \frac{\ln n}{n}$, then the probability that the graph is connected tends to 1 when $\lambda > 1$ and tends to 0 when $\lambda < 1$.

$p(n) = \lambda \frac{\ln n}{n}$ is called a *coarse* parameterization and $p(n) = \frac{(\ln n)+c}{n}$ a *fine* parameterization. If a finer parameterization is used, then we can observe subtler, "smoother" transitions instead of "sharp" thresholds. In fact, it is known that for $p(n) = \frac{(\ln n)+c}{n}$, $X_n \to \text{Poisson}(\exp(-c))$ in distribution as $n \to \infty$.

3. **Community Detection Using MAP**

It may be helpful to work on this problem in conjunction with the relevant lab. The *stochastic block model* (SBM) defines the random graph $\mathcal{G}(n, p, q)$ consisting of two communities of size $\frac{n}{2}$ each, such that the probability an edge exists between two nodes of the same community is p, and the probability an edge exists between two nodes in different communities is $q < p$. The goal of the problem is to exactly determine the two communities, given only the graph.

Show that the MAP estimate of the two communities is equivalent to finding the *min-bisection* or *balanced min-cut* of the graph, the split of G into two groups of size $\frac{n}{2}$ that has the minimum edge weight across the partition. Assume that any assignment of the communities is a priori equally likely.

Solution: Let $G \sim \mathcal{G}(n, p, q)$, and let A be a random variable representing the assignment or labelling of the two communities. We are interested in

$$\text{MAP}(A \mid G) = \operatorname*{argmax}_A \mathbb{P}(G \mid A) \cdot \mathbb{P}(A) = \operatorname*{argmax}_A \mathbb{P}(G \mid A).$$

Note that the MAP rule is equivalent to the MLE as each assignment of labels is equally likely. Let k be the number of edges across the partition in assignment A, and let m be the number of edges in G. Then

$$\begin{aligned}\mathbb{P}(G \mid A) &= q^k (1-q)^{\binom{\frac{n}{2}}{2} - k} \cdot p^{m-k} (1-p)^{2\binom{n/2}{2} - (m-k)} \\ &= \left(\frac{q}{1-q} \cdot \frac{1-p}{p}\right)^k \cdot \left(\frac{p}{1-p}\right)^m \cdot (1-p)^{2\binom{n/2}{2}} \cdot (1-q)^{n^2/4}.\end{aligned}$$

Now, the last three terms do not depend on the assignment of labels, and thus do not affect the likelihood function. We also see that

$$p > q \implies \left(\frac{q}{1-q} \cdot \frac{1-p}{p}\right) < 1,$$

so increasing k corresponds to decreasing the likelihood. Therefore, the MAP rule is to select the partition with the smallest number of edges across it, which is exactly the min-bisection of the graph.

4. **Bayesian Estimation of Exponential Distribution**

We have seen the MLE (non-Bayesian perspective) and MAP estimation (Bayesian perspective). In this problem, we will introduce the fully Bayesian approach to statistical estimation.

Suppose that X is Exponential with unknown rate Λ. As a Bayesian practitioner, you have a prior belief that the random variable Λ is equally likely to be λ_1 or λ_2.

Now, you collect one sample X_1 from X.

a. Find the posterior distribution $\mathbb{P}(\Lambda = \lambda_1 \mid X_1 = x_1)$.

b. If we were using the MLE or MAP rule, we would choose a single value λ for Λ, sometimes called a *point estimate*. This amounts to saying X has Exponential distribution with rate λ. In the Bayesian approach, we will instead keep the full information of the posterior distribution of Λ, and we compute the distribution of X as

$$f_X(x) = \sum_{\lambda \in \{\lambda_1, \lambda_2\}} f_{X|\Lambda}(x \mid \lambda) \cdot \mathbb{P}(\Lambda = \lambda \mid X_1 = x_1).$$

Note that we do not necessarily have an Exponential distribution for X anymore. Compute $f_X(x)$ in closed form.

c. From the previous part, you may have guessed that the fully Bayesian approach is often computationally intractable, which is one of the main reasons why point estimates are common in practice. Supposing that $\lambda_1 > \lambda_2$, compute the MAP estimate for Λ, and calculate $f_X(x)$ again using the MAP rule.

Solution:

a. The prior distribution is $\mathbb{P}(\Lambda = \lambda_1) = \mathbb{P}(\Lambda = \lambda_2) = \frac{1}{2}$, and the likelihood of the data is

$$f_{X_1|\Lambda}(x_1 \mid \lambda) = \lambda e^{-\lambda x_1},$$

so by Bayes' rule, the posterior distribution is

$$\mathbb{P}(\Lambda = \lambda_1 \mid X_1 = x_1) = \frac{\frac{1}{2}\lambda_1 e^{-\lambda_1 x_1}}{\frac{1}{2}\lambda_1 e^{-\lambda_1 x_1} + \frac{1}{2}\lambda_2 e^{-\lambda_2 x_1}} = \frac{\lambda_1 e^{-\lambda_1 x_1}}{\lambda_1 e^{-\lambda_1 x_1} + \lambda_2 e^{-\lambda_2 x_1}}.$$

b. By the previous part, we have

$$f_X(x) = \frac{\lambda_1^2 e^{-\lambda_1(x+x_1)} + \lambda_2^2 e^{-\lambda_2(x+x_1)}}{\lambda_1 e^{-\lambda_1 x_1} + \lambda_2 e^{-\lambda_2 x_1}}.$$

c. The MAP rule says to choose the value of λ that maximizes the posterior probability $\mathbb{P}(\Lambda = \lambda \mid X_1 = x_1)$, i.e. choose λ_1 if $\lambda_1 e^{-\lambda_1 x_1} > \lambda_2 e^{-\lambda_2 x_1}$, in which case X is Exponential with rate λ_1. When $\lambda_1 > \lambda_2$,

$$\text{MAP}(\Lambda \mid X_1) = \lambda_1 \, \mathbb{1}\left\{X_1 < \frac{\ln(\lambda_1/\lambda_2)}{\lambda_1 - \lambda_2}\right\} + \lambda_2 \, \mathbb{1}\left\{X_1 > \frac{\ln(\lambda_1/\lambda_2)}{\lambda_1 - \lambda_2}\right\}.$$

$$f_X(x) = \begin{cases} \lambda_1 e^{-\lambda_1 x} & \text{if } X_1 < \frac{\ln(\lambda_1/\lambda_2)}{\lambda_1 - \lambda_2} \\ \lambda_2 e^{-\lambda_2 x} & \text{if } X_1 > \frac{\ln(\lambda_1/\lambda_2)}{\lambda_1 - \lambda_2}. \end{cases}$$

5. **Linear Regression, MLE, and MAP**

Suppose you draw n i.i.d. data points $(x_1, y_1), \ldots, (x_n, y_n)$, where the true relationship is given by $Y = WX + \varepsilon$ for $\varepsilon \sim \mathcal{N}(0, \sigma^2)$. In other words, Y has a linear dependence on X with additive Gaussian noise.

a. Show that finding the MLE of W given the data points $\{(x_i, y_i)\}_{i=1}^n$ is equivalent to minimizing mean squared error, or minimizing the cost function

$$J(w) = \sum_{i=1}^{n} (y_i - wx_i)^2.$$

b. Now suppose that W has a *Laplace* prior distribution,

$$f_W(w) = \frac{1}{2\beta} e^{-|w|/\beta}.$$

Show that finding the MAP estimate of W given the data points $\{(x_i, y_i)\}_{i=1}^n$ is equivalent to minimizing the cost function

$$J(w) = \sum_{i=1}^{n} (y_i - wx_i)^2 + \lambda |w|.$$

(You should determine what λ is.) This is interpreted as a one-dimensional ℓ^1-regularized least-squares criterion, also known as LASSO.

Solution:

a. The likelihood of the data is

$$L((x_1, y_1), \ldots, (x_n, y_n) \mid W = w) = \prod_{i=1}^{n} L((x_i, y_i) \mid W = w)$$

as the data points are conditionally independent given W;

$$= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} e^{-(y_i - wx_i)^2/(2\sigma^2)}$$

as the likelihood of (x_i, y_i) given $W = w$ is the density of ε_i evaluated at $y_i - wx_i$;

$$\propto \prod_{i=1}^{n} e^{-(y_i - wx_i)^2/(2\sigma^2)},$$

discarding constant factors that do not depend on the data points or w. We now find it more convenient to work with the log-likelihood

$$\ell((x_1, y_1), \ldots, (x_n, y_n) \mid W = w) = -\frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - wx_i)^2.$$

We wish to maximize the log-likelihood with respect to w, which is equivalent to *minimizing* the cost function

$$J(w) = \sum_{i=1}^{n} (y_i - wx_i)^2.$$

b. The likelihood of W given the data points is

$$\begin{aligned} L(w \mid (x_1, y_1), \ldots, (x_n, y_n)) &\propto L((x_1, y_1), \ldots, (x_n, y_n) \mid W = w) \cdot f_W(w) \\ &= f_W(w) \prod_{i=1}^{n} L((x_i, y_i) \mid W = w) \\ &= \frac{1}{2\beta} e^{-|w|/\beta} \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} e^{-(y_i - wx_i)^2/(2\sigma^2)} \\ &\propto e^{-|w|/\beta} \prod_{i=1}^{n} e^{-(y_i - wx_i)^2/(2\sigma^2)}. \end{aligned}$$

Again, we find it more convenient to work with the log-likelihood

$$\ell(w \mid (x_1, y_1), \ldots, (x_n, y_n)) = -\frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - wx_i)^2 - \frac{1}{\beta}|w|.$$

Maximizing the log-likelihood is equivalent to minimizing the cost function

$$J(w) = \sum_{i=1}^{n} (y_i - wx_i)^2 + \lambda |w|$$

with $\lambda = 2\sigma^2/\beta$.

6. **Minimum-Error Property of MAP**

a. Let $X \in \{0,1\}$, and suppose we have the prior $\mathbb{P}(X=0)=\pi_0$ and $\mathbb{P}(X=1)=\pi_1$. Let $\hat{X}_{\text{MAP}}$ be the MAP estimate of X given the random variable Y, and let $\hat{X}$ be any other estimate of X given Y. Show that

$$\mathbb{P}(X \neq \hat{X}_{\text{MAP}}) \leq \mathbb{P}(X \neq \hat{X}).$$

b. Now, also suppose that type I errors (declaring $\hat{X}=1$ when $X=0$) incur a cost of $c_1 \geq 0$ and type II errors (declaring $\hat{X}=0$ when $X=1$) a cost of $c_2 \geq 0$. Derive the decision rule $\hat{X}$ that minimizes the total cost

$$c_1\,\mathbb{P}(\hat{X}=1, X=0) + c_2\,\mathbb{P}(\hat{X}=0, X=1).$$

Solution:

a. We write $\hat{X}_{\text{MAP}} = r^*(Y)$, where

$$r^*(y) = \operatorname*{argmax}_x \mathbb{P}(X=x, Y=y) = \operatorname*{argmin}_x \mathbb{P}(X \neq x, Y=y).$$

Now, the error probability for a general estimate $\hat{X}$ is

$$\begin{aligned}
\mathbb{P}(X \neq \hat{X}) &= \sum_y \mathbb{P}(X \neq \hat{X}, Y=y) \\
&= \sum_y \sum_z \mathbb{P}(X \neq z, Y=y) \cdot \mathbb{P}(\hat{X}=z \mid Y=y) \\
&\geq \sum_y \sum_z \mathbb{P}(X \neq r^*(y), Y=y) \cdot \mathbb{P}(\hat{X}=z \mid Y=y) \\
&= \sum_y \mathbb{P}(X \neq r^*(y), Y=y) \\
&= \mathbb{P}(X \neq r^*(Y)).
\end{aligned}$$

Remark. $\hat{X}$ being an estimate of X given Y means that it is conditionally independent of X given Y; that is, $X \to Y \to \hat{X}$ forms a Markov chain, as we saw in HW 06 Q4 and HW 07 Q1. This allowed us to drop the conditioning on X in the term $\mathbb{P}(\hat{X}=z \mid Y=y)$.

Remark. The error probability $\mathbb{P}(X \neq \hat{X}) = \mathbb{E}(\mathbb{1}\{X \neq \hat{X}\})$ is also known as the *Bayes risk* of $\hat{X}$ under the 0–1 loss function. We have shown that $\hat{X}_{\text{MAP}}$ minimizes $\mathbb{E}(\mathbb{1}\{X \neq \hat{X}\})$, i.e. MAP is the *Bayes-optimal* decision rule for estimating $X \in \{0,1\}$ under 0–1 loss.

Alternate solution. As $X \in \{0,1\}$, the MAP estimate is the threshold decision rule

$$\hat{X}_{\text{MAP}} = \mathbb{1}\{p_{Y|X}(Y \mid 1) \cdot \pi_1 \geq p_{Y|X}(Y \mid 0) \cdot \pi_0\} = \mathbb{1}\{L(Y) \geq \tfrac{\pi_0}{\pi_1}\},$$

$\pi_1 > 0$ without loss of generality. We can rewrite the error probability for $\hat{X}$ as

$$\begin{aligned}
\mathbb{P}(X \neq \hat{X}) &= \pi_0\,\mathbb{P}(\hat{X}=1 \mid X=0) + \pi_1\,\mathbb{P}(\hat{X}=0 \mid X=1) \\
&= \pi_0\,\mathbb{E}(\hat{X} \mid X=0) + \pi_1(1 - \mathbb{E}(\hat{X} \mid X=1))
\end{aligned}$$

$$= \pi_1 \,\mathbb{E}(\tfrac{\pi_0}{\pi_1}\hat{X} \mid X = 0) + \pi_1 - \pi_1 \,\mathbb{E}(L(Y)\hat{X} \mid X = 0)$$

$$= \pi_1 - \pi_1 \,\mathbb{E}((L(Y) - \tfrac{\pi_0}{\pi_1})\hat{X} \mid X = 0).$$

Observe that $(L(Y) - \frac{\pi_0}{\pi_1})\hat{X}_{\mathrm{MAP}} \geq (L(Y) - \frac{\pi_0}{\pi_1})\hat{X}$ by the definition of $\hat{X}_{\mathrm{MAP}}$. Thus the error probability of the MAP estimate is minimal.

b. Suppose $c_1 + c_2 > 0$ without loss of generality, and let $c \coloneqq c_1\pi_0 + c_2\pi_1$. The total cost of $\hat{X}$ is precisely

$$c_1\pi_0 \,\mathbb{P}(\hat{X} = 1 \mid X = 0) + c_2\pi_1 \,\mathbb{P}(\hat{X} = 0 \mid X = 1) = c \,\mathbb{P}(X \neq \hat{X}),$$

c times the error probability of $\hat{X}$ for the prior $\mathbb{P}(X = 0) = \frac{c_1\pi_0}{c}$ and $\mathbb{P}(X = 1) = \frac{c_2\pi_1}{c}$. By part a, the total cost is minimized by the MAP estimate under this reweighted prior.