Probability and Computation: Problem sheet 2 Solutions

You are encouraged to submit your solutions at student reception or by emailing them to nr454 by 2pm Friday 07th of February

Algorithms

Question 1. *In this question we consider the following* NP*-complete problem called* `VERTEX-COVER`

Instance: *A graph* $G = (V, E)$.   ***Output:*** *Subset* $S \subseteq V$ *such that each edge has an end in* S *and* $\|S\|$ *is minimized.*

(i) Can you come up with a simple randomised algorithm for `VERTEX-COVER`*?*

(ii) Is this algorithm correct? (That is, does it always output a valid vertex-cover S*)*

(iii) Can you show that, for any graph G*, the expected size of the vertex cover* S *produced by your algorithm is at most twice the size of an optimal one?*

Solution: Our algorithm **Rand-VC** is as follows:

For each $e = uv \in E$*, if* e *is not already covered, add* u *to* S *with probability* $1/2$*, otherwise add* v.

To see part (ii), each edge has at least one of its endpoints in S and so every edge is covered.

Part (iii) is the following result:

Proposition. *Let* S *be the vertex cover output by* ***Rand-VC*** *and* C *be an optimal vertex cover. Then*

$$\mathbf{E}[\,|S|\,] \leq 2|C|.$$

Proof. To begin we partition the edges by assigning them to vertices in C, an optimal vertex cover. Let $e_1, \dots, e_k$ be the edges assigned to some $v \in C$. For each edge e_i examined by the algorithm, at most one vertex was added to the vertex cover S output by **Rand-VC**. Let S_v denote the set of such vertices added to S by the edges $e_1, \dots, e_{k_v}$ assigned to $v \in C$.

Claim. *For all* $v \in C$, $\mathbf{E}[\,|S_v|\,] \leq 2$.

Proof of claim. We prove the claim by induction on the number of edges $k = k_v$ associated with v. If $k_v = 0$ the claim holds vacuously since no edge is adjacent to v and so cannot contribute any vertices to S_v. Assume the claim holds for any $n \leq k-1$ edges then if v is associated with k edges and we start by considering $e_k = uv$ then

$$\begin{aligned}\mathbf{E}[\,S_v\,] &= \mathbf{E}[\,|S_v| \mid v\text{added to } S_v\,] \cdot \mathbf{P}[\,v\text{added to } S_v\,] + \mathbf{E}[\,|S_v| \mid u\text{added to } S_v\,] \cdot \mathbf{P}[\,u\text{added to } S_v\,] \\ &\leq 1 \cdot (1/2) + (1 + \mathbf{E}[\,S'\,]) \cdot (1/2),\end{aligned}$$

where S' is the set of vertices added **Rand-VC** when it considers the edges $e_1, \dots, e_{k-1}$. Thus by the inductive hypothesis $\mathbf{E}[\,|S'|\,] \leq 2$ and so

$$\mathbf{E}[\,S_v\,] \leq 1/2 + (1+2) \cdot (1/2) \leq 2,$$

as claimed. Alternatively one could see that, if the number of edges covered by v in the optimal cover was infinite, then $|S_v| \sim \text{Geo}(1/2)$ would have geometric distribution with parameter $1/2$. This is since each time we terminate with probability $1/2$ if we pick v and add it to S_v, or we add the other endpoint to S_v and keep going. ◊

Now to prove the proposition, by the Claim, we have

$$\mathbf{E}[\,|S|\,] = \sum_{v\in C} \mathbf{E}[\,|S|\,] \le \sum_{v\in C} 2 \le 2|C|,$$

since the total number of vertices in the cover S found by **Rand-VC** is the sum of the vertices added by the edges associated with each $v \in C$. □

Conditional Expectation

Question 2. *Show properties* $2-6$ *of slide* 7 *of Lecture* 5.

Solution: Property 4: Let a be a possible value of X. Since X is independent of Y we have $\mathbf{P}[\,Y=y|X=a\,] = \mathbf{P}[\,Y=y\,]$ then

$$\mathbf{E}[\,Y|X=a\,] = \sum_y y\mathbf{P}[\,Y=y|X=a\,] \overset{indep}{=} \sum_y y\mathbf{P}[\,Y=y\,] = \mathbf{E}[\,Y\,]\,,$$

deducing that $\mathbf{E}[\,Y|X\,] = \mathbf{E}[\,Y\,]$.

Property 5: Let a be a possible value of X, then

$$\begin{aligned}\mathbf{E}[\,YZ|X=a\,] = \mathbf{E}[\,F(X)Z|X=a\,] &= \sum_x\sum_z F(x)z\mathbf{P}[\,X=x, Z=z|X=a\,]\\ &= F(a)\sum_z z\mathbf{P}[\,Z=z|X=a\,] = F(a)\mathbf{E}[\,Z|X=a\,]\,,\end{aligned}$$

deducing that $\mathbf{E}[\,YZ|X\,] = Y\mathbf{E}[\,Z|X\,]$.

Question 3. *Let* $X_1,\ldots,X_n$ *be independent discrete random variables and let* $Z = f(X_1,\ldots,X_n)$ *for some function* f. *Prove that*

$$\mathbf{E}[\,Z|X_1,\ldots,X_i\,] = \sum_{x_{i+1},\ldots,x_n} f(X_1,\ldots,X_i,x_{i+1},\ldots,x_n)\mathbf{P}[\,X_i=x_{i+1},\ldots,X_n=x_n\,]$$

Solution: Let $(a_1,\ldots,a_n)$ be a possible value of $(X_1,\ldots,X_n)$, then

$$\begin{aligned}\mathbf{E}[\,Z|X_1=a_1,\ldots,X_n=a_n\,] &= \sum_{x_1}\sum_{x_2}\cdots\sum_{x_n} f(x_1,\ldots,x_n)\mathbf{P}[\,X_1=x_1,\ldots,X_n=x_n|X_1=a_1,\ldots,X_i=a_i\,]\\ &= \sum_{x_{i+1}}\cdots\sum_{x_n} f(a_1,\ldots,a_i,x_{i+1},\ldots,x_n)\mathbf{P}[\,X_{i+1}=x_{i+1},\ldots,X_n=x_n\,]\,.\end{aligned}$$

Question 4. *Conditional Variance. Define the conditional variance of* Y *given* X *as*

$$\mathbf{Var}\,[\,Y|X\,] = \mathbf{E}\big[\,(Y-\mathbf{E}[\,Y|X\,])^2|X\,\big]\,.$$

(i) Prove that $\mathbf{Var}\,[\,Y\,] = \mathbf{E}[\,\mathbf{Var}\,[\,Y|X\,]\,] + \mathbf{Var}\,[\,\mathbf{E}[\,Y|X\,]\,]$

(ii) Consider n *bins and a random number* M *of balls, where* $\mathbf{E}[\,M\,]=\mu$ *and* $\mathbf{Var}\,[\,M\,]=\sigma^2$. *Compute the variance of the number of balls that are assigned to the first bin.*

Solution: Remember that $\mathbf{Var}[Y] = \mathbf{E}\big[(Y - \mathbf{E}[Y])^2\big]$, and, equivalently, $\mathbf{Var}[Y] = \mathbf{E}\big[Y^2\big] - \mathbf{E}[Y]^2$

By definition we have that

$$\mathbf{E}[\mathbf{Var}[Y|X]] = \mathbf{E}\big[\mathbf{E}\big[(Y - \mathbf{E}[Y|X])^2|X\big]\big] \stackrel{p1}{=} \mathbf{E}\big[(Y - \mathbf{E}[Y|X])^2)\big] = \mathbf{E}\Big[Y^2 - 2Y\mathbf{E}[Y|X] + \mathbf{E}[Y|X]^2\Big]$$

($p1$ refers to the properties of Lecture 6, slide 19). By $p1$ we get

$$\mathbf{E}[Y\mathbf{E}[Y|X]] \stackrel{p1}{=} \mathbf{E}[\mathbf{E}[Y\mathbf{E}[Y|X]\,|X]] \stackrel{p5}{=} \mathbf{E}\Big[\mathbf{E}[Y|X]^2\Big]$$

by linearity of conditional expectation (p3) we get

$$\mathbf{E}[\mathbf{Var}[Y|X]] = \mathbf{E}\big[Y^2\big] - \mathbf{E}\Big[\mathbf{E}[Y|X]^2\Big] \tag{1}$$

Also, note that $\mathbf{E}[\mathbf{E}[Y|X]] = \mathbf{E}[Y]$ then

$$\mathbf{Var}[\mathbf{E}[Y|X]] = \mathbf{E}\Big[\mathbf{E}[Y|X]^2\Big] - \mathbf{E}[Y]^2 \tag{2}$$

By adding equations (1) and (2) we get the result.

For the second part, let X be the number of balls that are assigned to the first bin. We compute $\mathbf{Var}[X] = \mathbf{E}[\mathbf{Var}[X|M]] + \mathbf{Var}[\mathbf{E}[X|M]]$. In lecture 6 slide 24 we computed $\mathbf{E}[X|M] = \sum_{i=1}^{\infty} \mathbf{1}_{\{i \le M\}} = M/n$. Moreover, by definition of conditional variance, we get

$$\begin{aligned}\mathbf{Var}[X|M] = \mathbf{E}\Big[(X - M/n)^2\Big|M\Big] &\stackrel{p3,p5}{=} \mathbf{E}\big[X^2|M\big] - 2(M/n)\mathbf{E}[X|M] + (M/n)^2 \\ &= \mathbf{E}\big[X^2|M\big] - (M/n)^2\end{aligned}$$

We just need to compute $\mathbf{E}\big[X^2|M\big]$. Let X_i be 1 if ball number i is assigned to the first bin, otherwise X_i is 0. Then $X = \sum_{i=1}^{\infty} X_i \mathbf{1}_{\{i \le M\}}$ and therefore

$$X^2 = \sum_{i=1}^{\infty}\sum_{j=1}^{\infty} X_i X_j \mathbf{1}_{\{i,j \le M\}}$$

Note that $X_i^2 = X_i$, then

$$\mathbf{E}\big[X^2|M\big] = \sum_{i=1}^{\infty} \mathbf{E}[X_i|M] + 2 \sum_{1=i<j<\infty} \mathbf{E}[X_i X_j|M]\,\mathbf{1}_{\{i,j \le M\}}$$

Finally, use that the location of a ball is independent of how many balls we assigned in total. Therefore $\mathbf{E}[X_i|M] = 1/n$ and $\mathbf{E}[X_i X_j|M] = 1/n^2$ for $i \neq j$. We conclude that

$$\mathbf{Var}[X|M] = M/n + M(M-1)/n^2 - (M/n)^2 = M/n - M/n^2$$

and $\mathbf{E}[\mathbf{Var}[X|M]] = (\mu/n)\left(1 - \frac{1}{n}\right)$

On the other hand, remember that $\mathbf{E}[X|M] = M/n$. Then

$$\mathbf{Var}[\mathbf{E}[X|M]] = \mathbf{Var}[M/n] = \frac{1}{n^2}\mathbf{Var}[M] = \sigma^2/n^2.$$

By adding $\mathbf{Var}[\mathbf{E}[X|M]]$ and $\mathbf{E}[\mathbf{Var}[X|M]]$ we get the result.

Question 5. *Consider a coin that shows head with probability p. What is the expected number of flips required to observe a run of n consecutive heads?*

Solution: To get a run of n heads, we first need to get a run of $n-1$ head. After that we either get another head, or we get a tail and restart the process. Denote by T_n the number of coins needed to get a run of n heads, and T_n' be the number of coins needed to get a run of n heads after restarting. Denote by X_n the indicator variable that tells us if we have to restart after T_{n-1} or not. Then

$$T_n = T_{n-1} + 1 + X_n T_n'$$

Note that X_n is independent of T_n' since T_n' does not dependent on the previous coins. Then $\mathbf{E}[T_n] = \mathbf{E}[T_{n-1}] + 1 + \mathbf{E}[X_n]\mathbf{E}[T_n'] = \mathbf{E}[T_{n-1}] + 1 + (1-p)\mathbf{E}[T_n']$ Using that $\mathbf{E}[T_n'] = ET_n$ we get the recursion

$$\mathbf{E}[T_n] = \frac{1}{p}\mathbf{E}[T_{n-1}] + \frac{1}{p}.$$

By iterating the previous recursion, and by using that $\mathbf{E}[T_1] = \frac{1}{p}$ we conclude that

$$\mathbf{E}[T_n] = \frac{1}{p^n} + \ldots + \frac{1}{p}.$$

Hint.

Q5: *Recall how we deduce the expectation of a geometric random variable in class.*