Foundations of Data Mining - Homework 2

Idan Shabat, 203511597

December 2, 2017

This homework was done with the collaboration of Mor Geva, 200831618.

1 Question 1

(a) Given $s = (s_0, s_1, s_2)^T = Mb$, where b is a non-negative vector of length n:

- If $s_0 = 0$, return **False**.
- Else, If $(\frac{s_1}{s_0})^2 \neq \frac{s_2}{s_0}$, return **False**.
- Else, return s_0.

We now prove that b contains exactly one non-zero entry **iff** this algorithm doesn't return **False**, and in this case, it returns the non-zero entry.

Suppose that b contains exactly one non-zero entry, say $b = (0, ..., 0, b_i, 0, ..., 0)^T$ where $b_i > 0$ placed in the i'th coordinate. Then by M definition, $s_0 = b_i$, $s_1 = ib_i$ and $s_2 = i^2 b_i$, and indeed, $b_i \neq 0$ and $(\frac{s_1}{s_0})^2 = i^2 = \frac{s_2}{s_0}$, so the algorithm doesn't return **False**, but instead return $s_0 = b_i$.

Now suppose that the algorithm doesn't return **False**, i.e. $s_0 > 0$ and $(\frac{s_1}{s_0})^2 = \frac{s_2}{s_0}$. If $b = (b_1, ..., b_n)$, then $s_0 = \sum_{i=1}^n b_i$, $s_1 = \sum_{i=1}^n ib_i$, $s_2 = \sum_{i=1}^n i^2 b_i$. We define $\lambda_i = \frac{b_i}{\sum_{i=1}^n b_i} = \frac{b_i}{s_0}$, so for each i, $\lambda_i \geq 0$ and $\sum_{i=1}^n \lambda_i = 1$. We now apply Jensen inequality for the strictly convex function $f(x) = x^2$, with the points $1, 2, ..., n$ and the coefficients $\lambda_1, \lambda_2, ..., \lambda_n$:

$$(\frac{s_1}{s_0})^2 = (\frac{\sum_{i=1}^n ib_i}{s_0})^2 = (\sum_{i=1}^n \lambda_i i)^2 \leq \sum_{i=1}^n \lambda_i i^2 = \frac{s_2}{s_0}$$

where equality holds only if all λ_i except one is 0 (since f is **strictly** convex: $f'' \equiv 2 > 0$). Assuming equality does hold, we get that exactly one of the b_i is non-zero.

(b) We use the linear sketch $s = Mb$, where M is a $3\log(n) \times n$ matrix, defined as follows:

For each $j \in \{0, 1, 2, ..., \log(n)\}$, let $h_j : \{1, ..., n\} \to \{0, 1\}$ be a hash function s.t. for each $i \neq k$, $h_j(i)$ and $h_j(k)$ are independent, and $P(h_j(i) = 1) = P(h_j(k) = 1) = \frac{2^j}{n}$ (for that we can use a uniform hash function $\hat{h} : \{1, ..., n\} \to (0, 1)$ and let $h_j(i) = 1 \iff \hat{h}(i) < \frac{2^j}{n}$). Now define:

$$M_j = \begin{pmatrix} 1 \cdot h_j(1) & 1 \cdot h_j(2) & 1 \cdot h_j(3) & \dots & 1 \cdot h_j(n) \\ 1 \cdot h_j(1) & 2 \cdot h_j(2) & 3 \cdot h_j(3) & \dots & n \cdot h_j(n) \\ 1 \cdot h_j(1) & 4 \cdot h_j(2) & 9 \cdot h_j(3) & \dots & n^2 \cdot h_j(n) \end{pmatrix}$$

and:

$$M = \begin{pmatrix} M_0 \\ M_1 \\ \vdots \\ M_{\log(n)} \end{pmatrix}$$

The query algorithm: Given $s = \left(s_0^0, s_1^0, s_2^0, s_0^1, s_1^1, s_2^1, \ldots, s_0^{\log(n)}, s_1^{\log(n)}, s_2^{\log(n)}\right)^T$, we apply the algorithm in (a) for each triplet (s_0^j, s_1^j, s_2^j) for $j \in \{0, 1, 2, ..., \log(n)\}$. Let j_s be the first j such that the algorithm from (a) doesn't return **False** (if exists). $s_0^{j_s}$ will be the returned element from the query. If j_s doesn't exist, the query fails.

For the rest of this question, we use the following notation: Given independent hash functions $h_0, ..., h_{\log(n)}$ as above, Let

$$B_j = \{k \in \{1, ..., n\} \mid h_j(k) = 1 \text{ and } b_k \neq 0\}$$

I.e. B_j is the set of the elements (coordinates) from b that the function h_j had chosen, and that are non-zero.

Claim: Given a vector b, each $b_i \neq 0$ has the same probability to be chosen by the query algorithm.

Proof: With a similar proof to the one in (a), we can show that $|B_j| = 1 \iff$ (a)'s algorithm doesn't return **False** on (s_0^j, s_1^j, s_2^j). Let $b_i \neq 0$. Since

$$P(b_i \text{ is chosen}) = P(b_i = s_0^{j_s}) = \sum_{j=0}^{\log(n)} P(b_i = s_0^{j_s} \mid j_s = j) P(j_s = j) + 0 \cdot P(j_s \text{ doesn't exist})$$

it suffices to show that for each j, $P(b_i = s_0^{j_s} \mid j_s = j)$ is the same for all $b_i \neq 0$. Indeed:

$$P(b_i = s_0^{j_s} \mid j_s = j) = P(B_j = \{b_i\}) = P(b_i \in B_j \text{ and } \forall_{\text{positive } b_k \text{ with } k \neq i} b_k \notin B_j) = \frac{2^j}{n}(1 - \frac{2^j}{n})^{m-1}$$

where m is the number of positive b_k's in b. The last expression doesn't depent on i, so the claim is proved.

Probability of a successful query: Again, let m be the number of positive b_k's in b. For a query to fail, (a)'s algorithm should return **False** for each triplet (s_0^j, s_1^j, s_2^j). Fix a $j \in \{0, 1, ..., \log(n)\}$. We know that (a)'s algorithm fails iff $|B_j| \neq 1$. Note that $|B_j|$'s distribution is exactly $Bin(m, \frac{2^j}{n})$, so the probability that the algorithm fails on (s_0^j, s_1^j, s_2^j) is

$$P(|B_j| \neq 1) = 1 - P(|B_j| = 1) = 1 - m\frac{2^j}{n}(1 - \frac{2^j}{n})^{m-1}$$

Since the hash functions are independent, we get that the probability of a successful query, i.e. that (a)'s algorithm doesn't return **False** for at least one j is

$$P(\exists_j |B_j| = 1) = 1 - P(\forall_j |B_j| \neq 1) = 1 - \prod_{j=0}^{\log(n)} P(|B_j| \neq 1) =$$

$$1 - \prod_{j=0}^{\log(n)} (1 - m\frac{2^j}{n}(1 - \frac{2^j}{n})^{m-1}) \geq 1 - \prod_{j=0}^{\log(n)} \exp(-m\frac{2^j}{n}(1 - \frac{2^j}{n})^{m-1}) = 1 - e^{-\frac{m}{n}\sum_{j=0}^{\log(n)} 2^j (1 - \frac{2^j}{n})^{m-1}}$$

(using the fact that for every $x \geq 0$, $1 - x \leq e^{-x}$).

Size of the sketch: The length of the sketch $s = \left(s_0^0, s_1^0, s_2^0, s_0^1, s_1^1, s_2^1, \ldots, s_0^{\log(n)}, s_1^{\log(n)}, s_2^{\log(n)}\right)^T$ is $3\log(n)$, and each entry is at most $n^2 \cdot ||b||_1$, so it takes $2\log(n) + \log(||b||_1)$ bits. Therefore, the full sketch size is

$$O\left(\log^2(n) + \log(n)\log(||b||_1)\right)$$

2 Question 2

(a) First, for each $j = 1, ..., d$ we define a matrix M_j of size $w \times n$ (n is the size of the vector b):

$$[M_j]_{i,k} = \begin{cases} 1 & h_j(k) = i \\ 0 & otherwise \end{cases}$$

We now define the matrix M of size $dw \times n$:

$$M = \begin{pmatrix} M_1 \\ M_2 \\ \vdots \\ M_d \end{pmatrix}$$

If $s = Mb$, we write $s = (s^1, s^2, ..., s^d)^T$ where s^j is a vector of length w, so $s^j = M_j b$. Note that the k'th column of M_j indicates the mapping of k by h_j - there's 1 in the $h_j(k)$ coordinate, and 0 elsewhere. That means that after an update (k, x), x will be added to the $h_j(k)$ coordinate of s^j, and 0 to the others. Hence, the array of counters is simply another arrangement of s:

$$\begin{pmatrix} (s^1)^T \\ (s^2)^T \\ \vdots \\ (s^d)^T \end{pmatrix}_{d \times w}$$

The length of s is dw.

(b) For every $k \in \{1, ..., n\}$ and $j \in \{1, ..., d\}$, we add x to $count[j, h_j(k)]$ after each update (k, x). Therefore, the final value of $count[j, h_j(k)]$ is at least the sum of all such x's, which is exactly b_k. We get $b_k \leq count[j, h_j(k)]$ for each j, hence

$$b_k \leq \hat{b_k} = \min_{j=1,...,d} count[j, h_j(k)]$$

For the other inequality, notice that since $h_1, ..., h_d$ are independent:

$$P\big(\min_{j=1,...,d} count[j, h_j(k)] = \hat{b_k} > b_k + \epsilon||b||_1\big) = P(\forall_j count[j, h_j(k)] > b_k + \epsilon||b||_1) =$$

$$= \prod_{j=1}^{d} P(count[j, h_j(k)] > b_k + \epsilon||b||_1)$$

Fix a k, and let X_i be an indicator random variable s.t. $X_i = 1 \iff h_j(i) = h_j(k)$. Since the value of $count[j, h_j(k)]$ rises by x after every update of the form (i, x), where $X_i = 1$, we get that:

$$count[j, h_j(k)] = \sum_{i=1}^{n} b_i X_i$$

So

$$E[count[j, h_j(k)]] = \sum_{i=1}^{n} b_i E[X_i] = \sum_{i=1}^{n} b_i P(h_j(i) = h_j(k)) = b_k + \sum_{i \neq k} b_i \frac{1}{w} =$$

$$= b_k + \frac{1}{w}||b||_1 - \frac{1}{w} b_k \leq b_k + \frac{1}{w}||b||_1 \leq b_k + \frac{\epsilon}{e}||b||_1$$

By Markov inequality (we saw that $b_k \leq count[j, h_j(k)]$, therefore $count[j, h_j(k)] - b_k$ is a positive variable):

$$P(count[j, h_j(k)] - b_k > \epsilon||b||_1) \leq \frac{E[count[j, h_j(k)] - b_k]}{\epsilon||b||_1} \leq \frac{b_k + \frac{\epsilon}{e}||b||_1 - b_k}{\epsilon||b||_1} = e^{-1}$$

Finally

$$P(\hat{b_k} > b_k + \epsilon||b||_1) = \prod_{j=1}^{d} P(count[j, h_j(k)] > b_k + \epsilon||b||_1) \leq$$

$$\leq \prod_{j=1}^{d} e^{-1} = e^{-d} \leq e^{-\ln(\frac{1}{\delta})} = \delta$$

$$P(\hat{b_k} \leq b_k + \epsilon||b||_1) \geq 1 - \delta$$

(c) As before, for each $j \in \{1, ..., d'\}$ and $i, k \in \{1, ..., n\}$, we define the indicator X_i to be 1 iff $h_j(i) = h_j(k)$. Again, we get $count[j, h_j(k)] = \sum_{i=1}^{n} b_i X_i$, or

$$count[j, h_j(k)] - b_k = \sum_{i \neq k} b_i X_i \quad \Rightarrow$$

$$|count[j, h_j(k)] - b_k| = |\sum_{i \neq k} b_i X_i| \leq \sum_{i \neq k} |b_i| X_i \leq \sum_{i=1}^{n} |b_i| X_i \quad \Rightarrow$$

$$E[|count[j, h_j(k)] - b_k|] \leq \sum_{i=1}^{n} |b_i| E[X_i] = \frac{1}{w}||b||_1 \leq \frac{\epsilon||b||_1}{e}$$

By Markov inequality (with the non-negative variable $|count[j, h_j(k)] - b_k|$):

$$P\big(|count[j, h_j(k)] - b_k| > 2\epsilon||b||_1\big) \leq \frac{E[|count[j, h_j(k)] - b_k|]}{2\epsilon||b||_1} \leq$$

$$\leq \frac{\frac{\epsilon||b||_1}{e}}{2\epsilon||b||_1} = \frac{1}{2e}$$

Fix a $k \in \{1, ..., n\}$, and let Y_j be an indicator random variable such that $Y_j = 1 \iff |count[j, h_j(k)] - b_k| > 2\epsilon||b||_1$. Let $Y = \sum_{j=1}^{d'} Y_j$.

Note that if $median_{j=1,...,d'} count[j, h_j(k)] \notin \big[b_k - 2\epsilon||b||_1, b_k + 2\epsilon||b||_1\big]$, that means that at least half of the $count[j, h_j(k)]$ aren't in this interval (since there are half of them bigger than the median and half of them smaller than the median). In other words, in that case: $Y \leq \frac{d'}{2}$, hence,

$$P(\hat{b_k} \notin \big[b_k - 2\epsilon||b||_1, b_k + 2\epsilon||b||_1\big]) \leq P(Y \leq \frac{d'}{2})$$

We now use Chernoff bound with $Y = \sum_{j=1}^{d'} Y_j$. By Markov inequality, we showed that $P(Y_j = 0) \leq \frac{1}{2e}$, so

$$E[Y] = \sum_{j=1}^{d'} E[Y_j] \geq d'(1 - \frac{1}{2e})$$

Choose $c = \frac{e-1}{2e-1} \in [0, 1]$ and:

$$P\big(Y \leq \frac{d'}{2}\big) = P\big(Y \leq \frac{e}{2e-1}(1-\frac{1}{2e})d'\big) \leq P\big(Y \leq \frac{e}{2e-1}E[Y]\big) =$$

$$= P\big(Y \leq (1-c)E[Y]\big) \leq e^{-\frac{1}{2}c^2 E[Y]} \leq e^{-\frac{(e-1)^2(2e-1)}{2(2e-1)^2 2e}d'} \leq e^{-0.05d'}$$

Taking $d' = 20d$, we get that

$$P(\hat{b_k} \notin \big[b_k - 2\epsilon||b||_1, b_k + 2\epsilon||b||_1\big]) \leq e^{-0.05d'} = e^{-d} \leq e^{-\ln(\frac{1}{\delta})} = \delta \quad \Rightarrow$$

$$P(b_k - 2\epsilon||b||_1 \leq \hat{b_k} \leq b_k + 2\epsilon||b||_1) \geq 1 - \delta$$

3 Estimating weighted Jaccard similarity

(a) Suppose that the samples $S(U), S(V)$ are subsets of $\{1, ..., n\}$ - those are the sampled coordinates from u, v. We use the following notations (note that $\forall_{i \in S(U)} u_i > 0$ and $\forall_{i \in S(V)} v_i > 0$):

$$\tau_u = \max_{i \in S(U)} \frac{h(i)}{u_i}$$
$$\tau_v = \max_{i \in S(V)} \frac{h(i)}{v_i}$$
$$X_i = \begin{cases} 1 & i \in S(U) \cap S(V) \\ 0 & otherwise \end{cases}$$
$$p_i = P(X_i = 1)$$

For calculating p_i, we use conditioned probabilty. If the values of h are given for all $j \neq i$, then i gets to $S(U)$ iff (i, u_i) is within the bottom-k elements, with respect to $\frac{h(i)}{u_i}$. Since τ_u is the maximum of these bottom-k elements, we require that $\frac{h(i)}{u_i} < \tau_u$ i.e. $h(i) < \tau_u u_i$. Similarly, $i \in S(V)$ iff $h(i) < \tau_v v_i$. Therefore $i \in S(U) \cap S(V) \iff h(i) < \min\{\tau_u u_i, \tau_v v_i\}$, which happens in probability

$$p_i = \min\{1, \min\{\tau_u u_i, \tau_v v_i\}\} = \min\{1, \tau_u u_i, \tau_v v_i\}$$

The estimator for $||\min\{U, V\}||_1$ is:

$$\hat{e_1} = \sum_{i \in S(U) \cap S(V)} \frac{\min\{u_i, v_i\}}{p_i}$$

The estimator is unbiased: Note that we can write $\hat{e_1} = \sum_{i=1}^{n} \frac{\min\{u_i, v_i\}}{p_i} X_i$, so in expectation:

$$E[\hat{e_1}] = \sum_{i=1}^{n} \frac{\min\{u_i, v_i\}}{p_i} E[X_i] = \sum_{i=1}^{n} \frac{\min\{u_i, v_i\}}{p_i} p_i = \sum_{i=1}^{n} \min\{u_i, v_i\} = ||\min\{U, V\}||_1$$

(b) With the same notations as in (a), the estimator for $||\max\{U, V\}||_1$ is:

$$\hat{e_2} = \sum_{i \in S(U) \cap S(V)} \frac{\max\{u_i, v_i\}}{p_i}$$

The estimator is unbiased: Note that we can write $\hat{e_2} = \sum_{i=1}^{n} \frac{\max\{u_i, v_i\}}{p_i} X_i$, so in expectation:

$$E[\hat{e_2}] = \sum_{i=1}^{n} \frac{\max\{u_i, v_i\}}{p_i} E[X_i] = \sum_{i=1}^{n} \frac{\max\{u_i, v_i\}}{p_i} p_i = \sum_{i=1}^{n} \max\{u_i, v_i\} = ||\max\{U, V\}||_1$$

(c) Given a weighted sampling of the vector A, we saw in class that the probability of (i, a_i) to get into the sample is proportional to its weight a_i: αa_i. If the sampled set is of size k, then:

$$k = E[\text{number of sampled elements}] = \sum_{i=1}^{n} \alpha a_i = \alpha ||A||_1 \Rightarrow$$
$$\alpha = \frac{k}{||A||_1}$$

By symmetry, the probability of (i, a_i) to get into the sampling we saw in (b) is also proportional to a_i, but the set of sampled items (which is $S(U) \cap S(V)$) is at most k. So here we get $P(i \in S(U) \cap S(V)) \leq \frac{k}{||A||_1} a_i$. Therefore, in a direct sampling from A, more elements are sampled (in expectation), so the estimate is better (i.e. lower variance).

4 Difference estimation from samples

(a) For every $i \in \{1, ..., n\}$, let $p_i = \min\{1, \alpha w_i\} = P((i, w_i) \in S)$. We define an estimator for $w_1 - w_2$ as follows:

$$\hat{D} = \begin{cases} \frac{w_1 - w_2}{p_1 p_2} & (1, w_1), (2, w_2) \in S \\ 0 & otherwise \end{cases}$$

The estimator is well-defined and non-negative, since $p_1, p_2 > 0$ (because $1, \alpha, w_1, w_2 > 0$), and we know that $w_1 - w_2 \geq 0$.

The probability that $(1, w_1), (2, w_2) \in S$ is exactly $p_1 p_2$, since S elements are chosen independently. Therefore,

$$E[\hat{D}] = p_1 p_2 \cdot \frac{w_1 - w_2}{p_1 p_2} + (1 - p_1 p_2) \cdot 0 = w_1 - w_2$$

Hence the estimator is unbiased.

(b)

$$E[\hat{D}^2] = p_1 p_2 \cdot (\frac{w_1 - w_2}{p_1 p_2})^2 + (1 - p_1 p_2) \cdot 0^2 = \frac{(w_1 - w_2)^2}{\min\{1, \alpha w_1\} \min\{1, \alpha w_2\}} = \frac{(w_1 - w_2)^2}{\min\{1, \alpha w_2, \alpha^2 w_1 w_2\}}$$

The last equation is by considering all the cases: $\alpha w_1, \alpha w_2 \geq 1$, $\alpha w_1, \alpha w_2 \leq 1$ or $\alpha w_1 \geq 1 \geq \alpha w_2$. Other cases aren't possible, since $w_1 \geq w_2$. We have

$$Var[\hat{D}] = E[\hat{D}^2] - E[\hat{D}]^2 = (w_1 - w_2)^2 (\frac{1}{\min\{1, \alpha w_2, \alpha^2 w_1 w_2\}} - 1)$$

(c) Since (i, w_i) is not sampled, we conclude that the probability of sampling it was < 1. We know that the probability was $\min\{1, \alpha w_i\}$, therefore $\alpha w_i < 1 \Rightarrow w_i < \frac{1}{\alpha}$. On the other hand, it is given that $w_i \geq 0$, so we finally get:

$$0 \leq w_i < \frac{1}{\alpha}$$

(d) The following estimator can be used:

$$\hat{F} = \begin{cases} \frac{w_1}{p_1} - \frac{w_2}{p_2} & (1, w_1) \in S \text{ and } (2, w_2) \in S \\ \frac{w_1}{p_1} & (1, w_1) \in S \text{ and } (2, w_2) \notin S \\ -\frac{w_2}{p_2} & (1, w_1) \notin S \text{ and } (2, w_2) \in S \\ 0 & (1, w_1) \notin S \text{ and } (2, w_2) \notin S \end{cases}$$

Notice that when $w_i = 0$, the probability that $(i, w_i) \in S$ is $\min\{1, \alpha \cdot 0\} = 0$, and then $(i, w_i) \notin S$ for sure. Hence, if $(i, w_i) \in S$, it is known that $w_i > 0$ and therefore $p_i > 0$, so $\hat{F}$ is well-defined.

For computing the expectation, we first suppose that $w_1, w_2 > 0$ (therefore $p_1, p_2 > 0$):

$$E[\hat{F}] = p_1 p_2 (\frac{w_1}{p_1} - \frac{w_2}{p_2}) + p_1 (1 - p_2) \frac{w_1}{p_1} - (1 - p_1) p_2 \frac{w_2}{p_2} + (1 - p_1)(1 - p_2) \cdot 0 =$$
$$= p_2 w_1 - p_1 w_2 + w_1 - p_2 w_1 - w_2 + p_1 w_2 = w_1 - w_2$$

For $w_1 > 0, w_2 = 0$, we know that $p_2 = 0$, so $(2, w_2) \notin S$, and we get:

$$E[\hat{F}] = p_1 \frac{w_1}{p_1} + (1 - p_1) \cdot 0 = w_1 = w_1 - w_2$$

Finally, for $w_1 = w_2 = 0$, necessarily $(1, w_1), (2, w_2) \notin S$, so $\hat{F} \equiv 0 \Rightarrow E[\hat{F}] = 0 = w_1 - w_2$.

We got that anyway, $E[\hat{F}] = w_1 - w_2$, so the estimator is unbiased.

We now prove that there is no non-negative unbiased estimator in this case: By negation, suppose we have such estimator $\hat{C}$. We can see $\hat{C}$ as a function of the sketch S, that was picked from a vector w, and it's unbiased iff for every non-negative w such that $w_1 \geq w_2$:

$$\forall_w \; E_S[\hat{C}(S)] = w_1 - w_2$$

Let w be a vector such that $w_1 = w_2 = 1$, and let w' be the exact same vector except that $w'_2 = 0$. We know that for a sketch S of w: $E[\hat{C}(S)] = 1 - 1 = 0$ and also $\forall_S \hat{C}(S) \geq 0$. Combining the two, we get that $\forall_S \hat{C}(S) = 0$ (if the expectation of $X \geq 0$ is 0, it means that $X = 0$ almost surely).

Now let S be a sketch of w'. Since $w'_2 = 0$, the probability that $(2, w_2) \in S$ is 0, therefore S doesn't contain $(2, w_2)$ - **so it's also a sketch of** w. Hence $\hat{C}(S) = 0$ for every sketch S of w', and then we got:

$$1 = w_1 - 0 = w'_1 - w'_2 = E[\hat{C}(S)] = E[0] = 0$$

Contradiction.

5 Difference estimation from hash-based samples

(a) By the sample definition, if (i, w_i) wasn't sampled, it means that $h(i) > \alpha w_i$, i.e. $w_i < \frac{h(i)}{\alpha}$. On the other hand, it is given that $w_i \geq 0$, so we finally get:

$$0 \leq w_i < \frac{h(i)}{\alpha}$$

(b) If $(1, w_1) \in S$ and $(2, w_2) \in S$, then of course we can tell the actual value of $w_1 - w_2$.

If $(1, w_1) \in S$ and $(2, w_2) \notin S$, then we know that $w_1 - w_2 \leq w_1$ (since $w_2 \geq 0$). From (a), we know that $w_1 - w_2 > w_1 - \frac{h(2)}{\alpha}$, and it is given also that $w_1 - w_2 \geq 0$, so we have $w_1 - w_2 \geq \max\{0, w_1 - \frac{h(2)}{\alpha}\}$.

If $(1, w_1) \notin S$ and $(2, w_2) \in S$, then from (a) we get $w_1 - w_2 < \frac{h(1)}{\alpha} - w_2$, and it's given that $w_1 - w_2 \geq 0$.

If $(1, w_1) \notin S$ and $(2, w_2) \notin S$, then from (a) and from the fact that $w_i \geq 0$, we get $w_1 - w_2 < \frac{h(1)}{\alpha} - 0 = \frac{h(1)}{\alpha}$. On the other hand, it is given that $w_1 - w_2 \geq 0$.

Concluding all the results:

$$w_1 - w_2 \in \begin{cases} \{w_1 - w_2\} & (1, w_1) \in S \text{ and } (2, w_2) \in S \\ [\max\{0, w_1 - \frac{h(2)}{\alpha}\}, w_1] & (1, w_1) \in S \text{ and } (2, w_2) \notin S \\ [0, \frac{h(1)}{\alpha} - w_2) & (1, w_1) \notin S \text{ and } (2, w_2) \in S \\ [0, \frac{h(1)}{\alpha}) & (1, w_1) \notin S \text{ and } (2, w_2) \notin S \end{cases}$$

(c) Define:

$$\hat{G} = \begin{cases} \frac{(w_1 - w_2)(1 - \frac{p_1 - p_2}{2})}{p_1 p_2} & (1, w_1), (2, w_2) \in S \\ \frac{1}{p_1} \max\{w_1 - \frac{h(2)}{\alpha}, 0\} & (1, w_1) \in S, (2, w_2) \notin S \\ 0 & otherwise \end{cases}$$

where $p_i = P((i, w_i) \in S) = \min\{1, \alpha w_i\} = \alpha w_i$ (it's given that $\alpha w_2 \leq \alpha w_1 \leq 1$).

The estimator is non-negative: $\frac{p_1 - p_2}{2} \leq p_1 - p_2 \leq p_1 \leq 1 \Rightarrow 1 - \frac{p_1 - p_2}{\alpha} \geq 0$, and we are given that $w_1 - w_2 \geq 0$.

$\hat{G}$ is also unbiased: We know from (a) that $(i, w_i) \notin S \iff w_i < \frac{h(i)}{\alpha} \iff p_i = \alpha w_i < h(i)$, so we have:

$$E[\hat{G} \mid (1, w_1) \in S, (2, w_2) \notin S] = E[\hat{G} \mid h(1) \leq p_1, h(2) > p_2] =$$
$$= P(h(2) > p_1 \mid h(1) \leq p_1, h(2) > p_2) \cdot E[\hat{G} \mid h(1) \leq p_1, h(2) > p_1] +$$
$$+ P(h(2) \leq p_1 \mid h(1) \leq p_1, h(2) > p_2) \cdot E[\hat{G} \mid h(1) \leq p_1, p_2 < h(2) \leq p_1] =$$
$$= P(h(2) > p_1 \mid h(1) \leq p_1, h(2) > p_2) \cdot 0 +$$
$$+ \frac{p_1 - p_2}{1 - p_2} \cdot \frac{1}{p_1}(w_1 - \frac{1}{\alpha} E[h(2) \mid h(1) \leq p_1, p_2 < h(2) \leq p_1]) = \frac{p_1 - p_2}{p_1(1 - p_2)}(w_1 - \frac{1}{\alpha} \cdot \frac{p_1 + p_2}{2}) =$$
$$= \frac{p_1 - p_2}{p_1(1 - p_2)}(w_1 - \frac{w_1 + w_2}{2}) = \frac{p_1 - p_2}{2p_1(1 - p_2)}(w_1 - w_2)$$

So we get:

$$E[\hat{G}] = P((1, w_1), (2, w_2) \in S) E[\hat{G} \mid (1, w_1), (2, w_2) \in S] +$$
$$+ P((1, w_1) \in S, (2, w_2) \notin S) E[\hat{G} \mid (1, w_1) \in S, (2, w_2) \notin S] =$$
$$= p_1 p_2 \cdot \frac{(w_1 - w_2)(1 - \frac{p_1 - p_2}{2})}{p_1 p_2} + P(h(1) \leq p_1, p_2 < h(2)) \frac{p_1 - p_2}{2p_1(1 - p_2)}(w_1 - w_2) =$$

$$= p_1p_2 \cdot \frac{(w_1-w_2)(1-\frac{p_1-p_2}{2})}{p_1p_2} + p_1(1-p_2)\frac{p_1-p_2}{2p_1(1-p_2)}(w_1-w_2) = (w_1-w_2)(1-\frac{p_1-p_2}{2}) + (w_1-w_2)\frac{p_1-p_2}{2} =$$

$$= (w_1-w_2)(1-\frac{p_1-p_2}{2}+\frac{p_1-p_2}{2}) = w_1-w_2$$

For comparing the variances of $\hat{D}$ (from question 4(a)) and of $\hat{G}$, we consider the following arguments:

- If $(1,w_1),(2,w_2) \in S$, then $E \le \hat{G} \le \hat{D}$, where $E = w_1 - w_2 = E[\hat{D}] = E[\hat{G}]$.
  proof: First, $w_1 \ge w_2 \Rightarrow \frac{p_1-p_2}{2} = \frac{1}{2}\alpha(w_1-w_2) \ge 0 \Rightarrow 1-\frac{p_1-p_2}{2} \le 1$, so $\hat{G} = \frac{(w_1-w_2)(1-\frac{p_1-p_2}{2})}{p_1p_2} \le \frac{w_1-w_2}{p_1p_2} = \hat{D}$.
  Second, when p_2 is fixed and we look at the function $f(p_1) = 1-\frac{p_1-p_2}{2}-p_1p_2$, the derivative is $f'(p_1) = -\frac{1}{2}-p_2 < 0$, so f is monotonically decreasing, hence, for every $0 \le p_1 \le 1$: $1-\frac{p_1-p_2}{2}-p_1p_2 \ge 1-\frac{1-p_2}{2}-1\cdot p_2 = \frac{1-p_2}{2} \ge 0$. Therefore, $1-\frac{p_1-p_2}{2} \ge p_1p_2 \Rightarrow \hat{G} = \frac{(w_1-w_2)(1-\frac{p_1-p_2}{2})}{p_1p_2} \ge w_1-w_2 = E$.
  corrollary: If $(1,w_1),(2,w_2) \in S$: $0 \le \hat{G}-E \le \hat{D}-E$, and therefore $(\hat{G}-E)^2 \le (\hat{D}-E)^2$.
- If $(1,w_1) \in S, (2,w_2) \notin S$, then $\hat{D} \le \hat{G} \le E$.
  proof: Notice that in this case, $\hat{D} = 0$, and as we saw $\hat{G}$ is non-negative, so: $\hat{D} \le \hat{G}$. Also, we saw that in this case: $\hat{G} = \frac{p_1-p_2}{2p_1(1-p_2)}(w_1-w_2) \le w_1-w_2 = E$.
  corrollary: If $(1,w_1) \in S, (2,w_2) \notin S, h(2) < p_1$: $\hat{D}-E \le \hat{G}-E \le 0$, and therefore $(\hat{G}-E)^2 \le (\hat{D}-E)^2$.
- The probability that none of the two cases happens is equal to the probability that $(1,w_2) \notin S$: $1-p_1$. In this case, both $\hat{D}$ and $\hat{G}$ are 0, so $E[(\hat{D}-E)^2] = E[(\hat{G}-E)^2] = E^2$.
- By variance deinition:

$$Var(\hat{G}) = E[(\hat{G}-E)^2] =$$

$$= p_1p_2E\big[(\hat{G}-E)^2 \mid (1,w_1),(2,w_2) \in S\big] +$$

$$+p_1(1-p_2)E\big[(\hat{G}-E)^2 \mid (1,w_1) \in S,(2,w_2) \notin S\big] + (1-p_1)E^2 \le$$

$$\le p_1p_2E\big[(\hat{D}-E)^2 \mid (1,w_1),(2,w_2) \in S\big] +$$

$$+p_1(1-p_2)E\big[(\hat{D}-E)^2 \mid (1,w_1) \in S,(2,w_2) \notin S\big] + (1-p_1)E^2 =$$

$$= E[(\hat{D}-E)^2] = Var(\hat{D})$$

  In conclusion, we have an estimator $\hat{G}$, that is non-negative, unbiased and with lower variance than $\hat{D}$.