Variance Reduction in Frequency Estimators via Control Variates Method (Supplementary Materials)

Rameshwar Pratap[1] **Raghav Kulkarni**[2]

[1]Indian Institute of Technology (IIT), Mandi H.P., India.
[2]Chennai Mathematical Institute (CMI) Chennai, India.

A MISSING PROOFS:

PROOF OF THEOREM 5:

Proof. We restate the the random variable X is as follows: $X = \sum_{j\in[n]} f_j Y_j$, where Y_j denotes an indicator random variable of the event "$h(j) = h(a)$" for $j \in [n]$. By 2-universality of the family from which h is drawn we have $\mathbb{E}[Y_j] = 1/k$. Thus, by linearity of expectation we have

$$\mathbb{E}[X] = \mathbb{E}\left[f_a + \sum_{j\in[n]/\{a\}} f_j Y_j\right].$$

$$= f_a + \sum_{j\in[n]/\{a\}} \frac{f_j}{k} = f_a + \frac{||\mathbf{f}||_1 - f_a}{k}, \tag{1}$$

where $||\mathbf{f}||_1 = \sum_{i\in[n]} f_i$. We now calculate the variance of the random variable X.

$$\mathrm{Var}[X] = \mathrm{Var}\left(f_a + \sum_{j\in[n]/\{a\}} f_j Y_j\right).$$

$$= \mathrm{Var}\left(\sum_{j\in[n]/\{a\}} f_j Y_j\right). \tag{2}$$

$$= \sum_{j\in[n]/\{a\}} \mathrm{Var}[f_j Y_j] + \sum_{i\neq j, i,j\in[n]/\{a\}} \mathrm{Cov}[f_i Y_i, f_j Y_j]. \tag{3}$$

$$= \sum_{j\in[n]/\{a\}} \left(\mathbb{E}[f_j^2 Y_j^2] - \mathbb{E}[f_j Y_j]^2\right) + \sum_{i\neq j, i,j\in[n]/\{a\}} \left(\mathbb{E}[f_i Y_i f_j Y_j] - \mathbb{E}[f_i Y_i]\mathbb{E}[f_j Y_j]\right).$$

$$= \sum_{j\in[n]/\{a\}} f_j^2 \left(\mathbb{E}[Y_j] - \mathbb{E}[Y_j]^2\right) + \sum_{i\neq j, i,j\in[n]/\{a\}} f_i f_j \left(\mathbb{E}[Y_i Y_j] - \mathbb{E}[Y_i]\mathbb{E}[Y_j]\right).$$

$$= \sum_{j\in[n]/\{a\}} f_j^2 \left(\frac{1}{k} - \frac{1}{k^2}\right) + \sum_{i\neq j, i,j\in[n]/\{a\}} f_i f_j \left(\frac{1}{k^2} - \frac{1}{k^2}\right). \tag{4}$$

$$= \left(\frac{1}{k} - \frac{1}{k^2}\right) \sum_{j\in[n]/\{a\}} f_j^2 + 0.$$

$$= \frac{||\mathbf{f}||_2^2 - f_a^2}{k}\left(1 - \frac{1}{k}\right). \tag{5}$$

Equations (2), and (3) hold due to Fact 4. Equation (4) holds as $h(.)$ is 2-universal hash function, which gives $\mathbb{E}[Y_i Y_j] = \mathbb{E}[Y_i]\mathbb{E}[Y_j] = 1/k^2$. Equations (1), and 5 complete a proof of the theorem. □

PROOF OF THEOREM 6**:**

Proof. We recall our random variable for our estimate as follows:

$$X = g(a)\sum_{j=1}^{n} f_j g(j) Y_j.$$

$$= g(a)^2 f_a Y_a + \sum_{j\in[n]/\{a\}} f_j g(a) g(j) Y_j. \tag{6}$$

$$= f_a + g(a) \sum_{j\in[n]/\{a\}} f_j g(j) Y_j. \tag{7}$$

For each $j \in [n]/\{a\}$ we have the following two equalities, which we will repeatedly use.

$$\mathbb{E}[g(j)] = 0,$$

$$\mathbb{E}[Y_j^2] = \mathbb{E}[Y_j] = \Pr[h(j) = h(a)] = 1/k. \tag{8}$$

Equation (8) holds as $g(.)$ is from 2-universal family and can take sign between $\{-1, +1\}$ each with probability $1/2$. Equation (8) holds since g and h are independent. Thus, we have

$$\mathbb{E}[g(j)Y_j] = \mathbb{E}[g(j)]\mathbb{E}[Y_j] = 0 \times \mathbb{E}[Y_j] = 0. \tag{9}$$

Due to Equations (7),(9), we have

$$\mathbb{E}[X] = f_a + g(a) \sum_{j\in[n]/\{a\}} f_j \mathbb{E}[g(j)Y_j] = f_a. \tag{10}$$

Thus, the output $X = \hat{f_a}$ is an unbiased estimator for the desired frequency f_a. We now give a variance analysis on the estimate.

$$\mathrm{Var}[X] = \mathrm{Var}\left[f_a + \sum_{j\in[n]/\{a\}} f_j g(a) g(j) Y_j\right].$$

$$= \mathrm{Var}\left[\sum_{j\in[n]/\{a\}} f_j g(a) g(j) Y_j\right]. \tag{11}$$

$$= g(a)^2 \mathrm{Var}\left[\sum_{j\in[n]/\{a\}} f_j g(j) Y_j\right].$$

$$= \mathrm{Var}\left[\sum_{j\in[n]/\{a\}} f_j g(j) Y_j\right]. \tag{12}$$

$$= \mathbb{E}\left[\left(\sum_{j\in[n]/\{a\}} f_j g(j) Y_j\right)^2\right] - \mathbb{E}\left[\sum_{j\in[n]/\{a\}} f_j g(j) Y_j\right]^2.$$

$$= \mathbb{E}\left[\left(\sum_{j\in[n]/\{a\}} f_j g(j) Y_j\right)^2\right]$$

$$= \mathbb{E}\left[\sum_{j\in[n]/\{a\}} f_j^2 g(j)^2 Y_j^2 + \sum_{j\neq l} f_j f_l g(j) g(l) Y_j Y_l\right]. \tag{13}$$

$$= \mathbb{E}\left[\sum_{j\in[n]/\{a\}} f_j^2 Y_j\right] = \sum_{j\in[n]/\{a\}} \frac{f_j^2}{k} = \frac{||\mathbf{f}||_2^2 - f_a^2}{k}. \quad (14)$$

Equation (11) and (12) hold due to Fact 4, and $g(a)^2 = 1$. Equation (13) hold due to Equation (8). Equations (10) and (14) completes a proof of the theorem. □

PROOF OF COROLLARY 7:

Proof. The random variable X mentioned in Theorem 5 captures the estimated frequency (an overestimate indeed). Due to Theorem 5, we have $\mathbb{E}[X] = f_a + \frac{||\mathbf{f}||_1 - f_a}{k}$, and $\mathrm{Var}[X] = \frac{||\mathbf{f}||_2^2 - f_a^2}{k}\left(1 - \frac{1}{k}\right)$. For a random variable R with mean $\mathbb{E}[R]$ and variance $\mathrm{Var}[R]$ satisfies the following concentration guarantee

$$\Pr\left[|R - \mathbb{E}[R]| \geq \epsilon' \sqrt{\mathrm{Var}[R]}\right] \leq \frac{1}{\epsilon'^2}.$$

We obtain the following by putting R as our random variance X, and $\epsilon' = \frac{\varepsilon\sqrt{||\mathbf{f}||_2^2 - f_a^2}}{\sqrt{\mathrm{Var}[X]}}$ in the above equation.

$$\Pr\left[\left|\hat{f}_a - \left(f_a + \frac{||\mathbf{f}||_1 - f_a}{k}\right)\right| \geq \varepsilon\sqrt{||\mathbf{f}||_2^2 - f_a^2}\right] \leq \frac{k-1}{\varepsilon^2 k^2}.$$
$$\leq \frac{1}{\varepsilon^2 k} = \frac{1}{3}.$$

The last equality holds due to our choice of the parameter k. Due to Theorem 1, the variance of our CV estimator is given as follows:

$$\mathrm{Var}(X + \hat{c}(Z - \mathbb{E}[Z])) = \mathrm{Var}(X) - \frac{(||\mathbf{f}||_1 - f_a)^2}{(n-1)k}\left(1 - \frac{1}{k}\right). \quad (15)$$

$$= \left(\frac{||\mathbf{f}||_2^2 - f_a^2}{k} - \frac{(||\mathbf{f}||_1 - f_a)^2}{(n-1)k}\right) \cdot \left(1 - \frac{1}{k}\right). \quad (16)$$

$$= \left(\frac{(n-1)(||\mathbf{f}||_2^2 - f_a^2) - (||\mathbf{f}||_1 - f_a)^2}{(n-1)k}\right) \cdot \left(1 - \frac{1}{k}\right). \quad (17)$$

Further, due to Chebyshev's inequality, for the query item a its estimated frequency $\tilde{f}_a$ outputted by our CV estimate satisfies the following:

$$\Pr\left[\left|\tilde{f}_a - \left(f_a + \frac{||\mathbf{f}||_1 - f_a}{k}\right)\right| \geq \varepsilon\sqrt{||\mathbf{f}||_2^2 - f_a^2}\right] \leq \frac{\mathrm{Var}(X + \hat{c}(Z - \mathbb{E}[Z]))}{\varepsilon^2(||\mathbf{f}||_2^2 - f_a^2)}. \quad (18)$$

$$= \left(\frac{(n-1)(||\mathbf{f}||_2^2 - f_a^2) - (||\mathbf{f}||_1 - f_a)^2}{\varepsilon^2 k(n-1)(||\mathbf{f}||_2^2 - f_a^2)}\right) \cdot \left(1 - \frac{1}{k}\right). \quad (19)$$

$$\leq \frac{(n-1)(||\mathbf{f}||_2^2 - f_a^2) - (||\mathbf{f}||_1 - f_a)^2}{\varepsilon^2 k(n-1)(||\mathbf{f}||_2^2 - f_a^2)}. \quad (20)$$

$$= \frac{1}{3}. \quad (21)$$

The last equality follows by putting

$$k = \frac{3}{\varepsilon^2} \cdot \left(\frac{(n-1)(||\mathbf{f}||_2^2 - f_a^2) - (||\mathbf{f}||_1 - f_a)^2}{(n-1)\left(||\mathbf{f}||_2^2 - f_a^2\right)}\right),$$

in Equation (21). □

PROOF OF COROLLARY 8**:**

Proof. The random variable X mentioned in Theorem 6 captures the estimated frequency. Due to Theorem 6, we have

$$\mathbb{E}[X] = f_a, \text{ and } \mathrm{Var}[X] = \frac{||\mathbf{f}||_2^2 - f_a^2}{k}.$$

We now apply Chebyshev's inequality on the above expression which gives us the desired concentration guarantee

$$\begin{aligned}
\Pr\left[|\hat{f}_a - f_a| \geq \varepsilon\sqrt{||\mathbf{f}||_2^2 - f_a^2}\right] &= \Pr\left[|X - \mathbb{E}[X]| \geq \varepsilon\sqrt{||\mathbf{f}||_2^2 - f_a^2}\right]. \\
&\leq \frac{\mathrm{Var}[X]}{\varepsilon^2(||\mathbf{f}||_2^2 - f_a^2)}. \\
&= \frac{1}{k\varepsilon^2} = \frac{1}{3}.
\end{aligned}$$

The last equality holds due to our choice of the parameter k. Due to Theorem 2, the variance of our CV estimator is given as follows:

$$\begin{aligned}
\mathrm{Var}(X + \hat{c}(Z - \mathbb{E}[Z])) &= \mathrm{Var}(X) - \frac{(||\mathbf{f}||_1 - f_a)^2}{(n-1)k}. && (22) \\
&= \frac{||\mathbf{f}||_2^2 - f_a^2}{k} - \frac{(||\mathbf{f}||_1 - f_a)^2}{(n-1)k}. && (23) \\
&= \frac{(n-1)(||\mathbf{f}||_2^2 - f_a^2) - (||\mathbf{f}||_1 - f_a)^2}{(n-1)k}. && (24)
\end{aligned}$$

The equality follows after some simple algebraic calculations. Further, due to Chebyshev's inequality, for the query item a its estimated frequency $\tilde{f}_a$ outputted by CV satisfies the following:

$$\begin{aligned}
\Pr\left[|\tilde{f}_a - f_a| \geq \varepsilon\sqrt{||\mathbf{f}||_2^2 - f_a^2}\right] &\leq \frac{\mathrm{Var}(X + \hat{c}(Z - \mathbb{E}[Z]))}{\varepsilon^2(||\mathbf{f}||_2^2 - f_a^2)}. && (25) \\
&= \frac{(n-1)(||\mathbf{f}||_2^2 - f_a^2) - (||\mathbf{f}||_1 - f_a)^2}{\varepsilon^2 k(n-1)(||\mathbf{f}||_2^2 - f_a^2)}. && (26) \\
&= \frac{1}{3}. && (27)
\end{aligned}$$

The last equality follows by putting

$$k = \frac{3}{\varepsilon^2} \cdot \left(\frac{(n-1)(||\mathbf{f}||_2^2 - f_a^2) - (||\mathbf{f}||_1 - f_a)^2}{(n-1)\left(||\mathbf{f}||_2^2 - f_a^2\right)}\right).$$

in Equation (27). □