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

Rameshwar Pratap
Raghav Kulkarni

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

A MISSING PROOFS:

PROOF OF THEOREM 5:

Proof. We restate 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

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

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

(1)

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

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

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

(2)

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

(3)

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

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

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

(4)

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

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

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

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

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

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.

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

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

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

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

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

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

\[
= \mathbb{E} \left[ \sum_{j \in [n]/\{a\}} f_j^2 g(j)^2 Y_j^2 + \sum_{j \neq l} f_j f_l g(j) g(l) Y_j Y_l \right]. \tag{13}
\]
\[
= \mathbb{E} \left[ \sum_{j \in [n]/\{a\}} f_j^2 Y_j \right] = \sum_{j \in [n]/\{a\}} \frac{f_j^2}{k} = \frac{||f||_2^2 - f_a^2}{k}. \tag{14}
\]

Equation (11) and (12) hold due to Fact 4, and \(g(a)^2 = 1\). Equation (13) hold due to Equation (5). 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 \(\mathbb{E}[X] = f_a + \frac{||f||_1 - f_a}{k} \), and \(\text{Var}[X] = \frac{||f||_2^2 - f_a^2}{k} (1 - \frac{1}{k})\). For a random variable \(R\) with mean \(\mathbb{E}[R]\) and variance \(\text{Var}[R]\) satisfies the following concentration guarantee

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

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

\[
\Pr \left[ |\hat{f}_a - \left( f_a + \frac{||f||_1 - f_a}{k} \right) | \geq \epsilon \sqrt{||f||_2^2 - f_a^2} \right] \leq \frac{k - 1}{\epsilon'^2}. \tag{17}
\]

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:

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

\[
= \left( \frac{||f||_2^2 - f_a^2}{k} - \frac{(||f||_1 - f_a)^2}{(n - 1)k} \right) \cdot \left( 1 - \frac{1}{k} \right). \tag{16}
\]

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

\[
\Pr \left[ |\hat{f}_a - \left( f_a + \frac{||f||_1 - f_a}{k} \right) | \geq \epsilon \sqrt{||f||_2^2 - f_a^2} \right] \leq \frac{\text{Var}(X + \hat{c}(Z - \mathbb{E}[Z]))}{\epsilon'^2} \tag{18}
\]

\[
= \left( \frac{(n - 1)(||f||_2^2 - f_a^2) - (||f||_1 - f_a)^2}{\epsilon'^2 k(n - 1)(||f||_2^2 - f_a^2)} \right) \cdot \left( 1 - \frac{1}{k} \right). \tag{19}
\]

\[
\leq \frac{(n - 1)(||f||_2^2 - f_a^2) - (||f||_1 - f_a)^2}{\epsilon'^2 k(n - 1)(||f||_2^2 - f_a^2)}. \tag{20}
\]

\[
= \frac{1}{3}. \tag{21}
\]

The last equality follows by putting

\[
k = \frac{3}{\epsilon'^2} \cdot \left( \frac{(n - 1)(||f||_2^2 - f_a^2) - (||f||_1 - f_a)^2}{(n - 1)(||f||_2^2 - f_a^2)} \right),
\]

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

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

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

$$\Pr \left[ |\hat{f}_a - f_a| \geq \varepsilon \sqrt{||f||_2^2 - f_a^2} \right] = \Pr \left[ |X - E[X]| \geq \varepsilon \sqrt{||f||_2^2 - f_a^2} \right].$$

\[\leq \frac{\text{Var}[X]}{\varepsilon^2 (||f||_2^2 - f_a^2)} = \frac{1}{k\varepsilon^2} = \frac{1}{3}.\]

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:

$$\text{Var}(X + \hat{c}(Z - E[Z])) = \text{Var}(X) - \frac{(||f||_1 - f_a)^2}{(n-1)k}. \quad (22)$$

$$= \frac{||f||_2^2 - f_a^2}{k} - \frac{(||f||_1 - f_a)^2}{(n-1)k}. \quad (23)$$

$$= \frac{(n-1)(||f||_2^2 - f_a^2) - (||f||_1 - f_a)^2}{(n-1)k}. \quad (24)$$

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

$$\Pr \left[ |\hat{f}_a - f_a| \geq \varepsilon \sqrt{||f||_2^2 - f_a^2} \right] \leq \frac{\text{Var}(X + \hat{c}(Z - E[Z]))}{\varepsilon^2 (||f||_2^2 - f_a^2)}. \quad (25)$$

$$= \frac{(n-1)(||f||_2^2 - f_a^2) - (||f||_1 - f_a)^2}{\varepsilon^2 k(n-1)(||f||_2^2 - f_a^2)}. \quad (26)$$

$$= \frac{1}{3}. \quad (27)$$

The last equality follows by putting

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

in Equation (27). \qed