Supplementary Material

1. Improving the First Class of Hashing Algorithms: Maximum Likelihood Estimators

The main goal is to prove **Theorem 1** stated in the main paper. We restate the procedure of maximum likelihood estimation here in this context.

Suppose we have k hashes, and define v_{ik} (equivalently $v_{jk}, v_{\boldsymbol{w}k}$) to be the respective values under the k^{th} hash. Consider the k^{th} triple given by $(v_{ik}, v_{jk}, v_{\boldsymbol{w}k})$. There are only five possible sets of triples: a) all elements are equal, b) two elements are equal and the third is different, and c) all elements are distinct. Table 1 shows the different types of triples.

	A	B	C	D	E
v_{is}	equal	different	equal	equal	distinct
v_{js}	equal	equal	different	equal	distinct
$v_{\boldsymbol{w}s}$	equal	equal	equal	different	distinct

Table 1: Table of possible triples.

Suppose we count the triples in each set and denote this as n_l, where $l \in \{A, B, C, D, E\}$, and p_l the probability of observing a triple falling in the set l. We note that $n_A + n_B + n_C + n_D + n_E = k$.

Since we have pre-computed $d(\boldsymbol{x}_i, \boldsymbol{w})$ and $d(\boldsymbol{x}_j, \boldsymbol{w})$, and $f^{(1)}$ is linear, we can invert this function to find the corresponding $\rho_h(\boldsymbol{x}_i, \boldsymbol{w}), \rho_h(\boldsymbol{x}_j, \boldsymbol{w})$. Moreover, we can write $p_A + p_C = \rho_h(\boldsymbol{x}_i, \boldsymbol{w})$, $p_A + p_B = \rho_h(\boldsymbol{x}_j, \boldsymbol{w})$, $p_A + p_D = \rho_h(\boldsymbol{x}_i, \boldsymbol{x}_j)$, and $p_A + p_B + p_C + p_D + p_E = 1$. Finally, since $d(\boldsymbol{x}_i, \boldsymbol{x}_j)$ is given by $p_A + p_D$, we write the log likelihood function $\ell(p_A, p_B, p_C, p_D, p_E)$ in terms of p_A and p_D, and get

$$\begin{aligned}\ell(p_A, p_D) = n_A \log(p_A) + n_B \log(\rho_h(\boldsymbol{x}_j, \boldsymbol{w}) - p_A) + n_C \log(\rho_h(\boldsymbol{x}_i, \boldsymbol{w}) - p_A) \\ + n_D \log(p_D) + n_E \log(1 - \rho_h(\boldsymbol{x}_i, \boldsymbol{w}) - \rho_h(\boldsymbol{x}_j, \boldsymbol{w}) + p_A - p_D)\end{aligned} \tag{1}$$

as we want to compute the maximum likelihood estimate of $\hat{p}_A + \hat{p}_D$ to give an estimate of $\rho_h(\boldsymbol{x}_i, \boldsymbol{x}_j)$.

Theorem 1 *Suppose we have a hashing algorithm where the estimate of interest is given by*

$$\rho_h(\boldsymbol{x}_i, \boldsymbol{x}_j) = \mathbb{E}[f(Y^{(1)})] \tag{2}$$

where $Y^{(1)}$ is a Bernoulli random variable, f is a linear function, and the output of the hashing algorithm takes discrete values. Suppose we add a weighted vector $\boldsymbol{w}$, and compute

the maximum likelihood estimate via (1). Then: a) this estimator is unbiased, and b) the asymptotic variance of this estimator is always lower than or equal to the variance of the estimate without using the MLE.

Proof In part a), the estimator is given by $\hat{p}_A + \hat{p}_D$. We can see that observe from Table 1 that sets A and D correspond to when the hashed values $v_{is} = v_{js}$ for $1 \leq s \leq k$, and by definition, we have $\rho_h(\boldsymbol{x}_i, \boldsymbol{x}_j) = \mathbb{P}[v_i = v_j]$, hence by the law of large numbers, $\sum_{s=1}^{k} 1_{\{v_{is}=v_{js}\}}/k \approx \rho_h(\boldsymbol{x}_i, \boldsymbol{x}_j)$ and hence this estimator is unbiased.

We now prove part b).

Recall that for random variables $Y_1, Y_2, \ldots, Y_k$, we have that

$$\mathrm{Var}[\frac{Y_1 + \ldots + Y_k}{k}] = \frac{1}{k^2}\left(\sum_{s=1}^{k} \mathrm{Var}[Y_s] + 2\sum_{s \geq t} \mathrm{Cov}(Y_s, Y_t)\right) \tag{3}$$

In the original case without any modifications, we have $Y_s := 1_{\{v_{is}=v_{js}\}}$ for $1 \leq s \leq k$, and each Y_s are i.i.d. from a Bernoulli distribution.

Hence with k hashes, we must have

$$\begin{aligned}
\mathrm{Var}[\frac{Y_1 + \ldots + Y_k}{k}] &= \frac{1}{k^2}\left(\sum_{s=1}^{k} \mathrm{Var}[Y_s]\right) && (4)\\
&= \frac{\mathrm{Var}[1_{\{v_i=v_j\}}]}{k} && (5)\\
&= \frac{(p_A + p_D)(p_B + p_C + p_E)}{k} && (6)
\end{aligned}$$

Suppose we now look at our new estimator, where we make use of the pre-computed values. From Equation (1), we can compute the partial derivatives to be

$$\frac{\partial \ell}{\partial p_A} = \frac{n_A}{p_A} + \frac{n_B}{p_A - \rho_h(\boldsymbol{x}_j, \boldsymbol{w})} + \frac{n_C}{p_A - \rho_h(\boldsymbol{x}_i, \boldsymbol{w})} + \frac{n_E}{1 + p_A - p_D - \rho_h(\boldsymbol{x}_i, \boldsymbol{w}) - \rho_h(\boldsymbol{x}_j, \boldsymbol{w})} \tag{7}$$

$$\frac{\partial \ell}{\partial p_D} = \frac{n_D}{p_D} - \frac{n_E}{1 + p_A - p_D - \rho_h(\boldsymbol{x}_i, \boldsymbol{w}) - \rho_h(\boldsymbol{x}_j, \boldsymbol{w})} \tag{8}$$

and Hessian matrix of second partial derivatives as

$$\begin{aligned}
H &= \begin{pmatrix} -\frac{n_A}{p_A^2} - \frac{n_B}{(\rho_h(\boldsymbol{x}_j,\boldsymbol{w}) - p_A)^2} - \frac{n_C}{(\rho_h(\boldsymbol{x}_i,\boldsymbol{w}) - p_A)^2} - \frac{n_E}{(1 - \rho_h(\boldsymbol{x}_i,\boldsymbol{w}) - \rho_h(\boldsymbol{x}_j,\boldsymbol{w}) + p_A - p_D)^2} & -\frac{n_E}{(1 - \rho_h(\boldsymbol{x}_i,\boldsymbol{w}) - \rho_h(\boldsymbol{x}_j,\boldsymbol{w}) + p_A - p_D)^2} \\ -\frac{n_E}{(1 - \rho_h(\boldsymbol{x}_i,\boldsymbol{w}) - \rho_h(\boldsymbol{x}_j,\boldsymbol{w}) + p_A - p_D)^2} & -\frac{n_E}{(1 - \rho_h(\boldsymbol{x}_i,\boldsymbol{w}) - \rho_h(\boldsymbol{x}_j,\boldsymbol{w}) + p_A - p_D)^2} - \frac{n_D}{p_D^2} \end{pmatrix} \\
&= \begin{pmatrix} -\frac{n_A}{p_A^2} - \frac{n_B}{p_B^2} - \frac{n_C}{p_C^2} - \frac{n_E}{p_E^2} & \frac{n_E}{p_E^2} \\ \frac{n_E}{p_E^2} & -\frac{n_D}{p_D^2} - \frac{n_E}{p_E^2} \end{pmatrix} && (9)
\end{aligned}$$

From the Hessian, we can compute the expected Fisher Information $\mathcal{I}$ of p_A and p_D given by

$$\mathcal{I} = -\mathbb{E}\left[H\right] \tag{10}$$

$$= \begin{pmatrix} \frac{\mathbb{E}[n_A]}{p_A^2} + \frac{\mathbb{E}[n_B]}{p_B^2} + \frac{\mathbb{E}[n_C]}{p_C^2} + \frac{\mathbb{E}[n_E]}{p_E^2} & -\frac{\mathbb{E}[n_E]}{p_E^2} \\ -\frac{\mathbb{E}[n_E]}{p_E^2} & \frac{\mathbb{E}[n_D]}{p_D^2} + \frac{\mathbb{E}[n_E]}{p_E^2} \end{pmatrix} \tag{11}$$

$$= \begin{pmatrix} \frac{kp_A}{p_A^2} + \frac{kp_B}{p_B^2} + \frac{kp_C}{p_C^2} + \frac{kp_E}{p_E^2} & -\frac{kp_E}{p_E^2} \\ -\frac{kp_E}{p_E^2} & \frac{kp_D}{p_D^2} + \frac{kp_E}{p_E^2} \end{pmatrix} \tag{12}$$

$$= k \begin{pmatrix} \frac{1}{p_A} + \frac{1}{p_B} + \frac{1}{p_C} + \frac{1}{p_E} & -\frac{1}{p_E} \\ -\frac{1}{p_E} & \frac{1}{p_D} + \frac{1}{p_E} \end{pmatrix} \tag{13}$$

Now, $\begin{pmatrix} \hat{p}_A \\ \hat{p}_D \end{pmatrix}$ converges in distribution to a bivariate normal random variable $N\left(\begin{pmatrix} p_A \\ p_D \end{pmatrix}, \mathcal{I}^{-1}\right)$, where

$$\mathcal{I}^{-1} = \frac{1}{k\left[\left(\frac{1}{p_A} + \frac{1}{p_B} + \frac{1}{p_C} + \frac{1}{p_E}\right)\left(\frac{1}{p_D} + \frac{1}{p_E}\right) - \frac{1}{p_E^2}\right]} \begin{pmatrix} \frac{1}{p_D} + \frac{1}{p_E} & \frac{1}{p_E} \\ \frac{1}{p_E} & \frac{1}{p_A} + \frac{1}{p_B} + \frac{1}{p_C} + \frac{1}{p_E} \end{pmatrix} \tag{14}$$

We can now write the variance of our new estimator as

$$\mathrm{Var}[\hat{p}_A + \hat{p}_D] = \mathrm{Var}[\hat{p}_A] + \mathrm{Var}[\hat{p}_D] + 2\mathrm{Cov}[\hat{p}_A, \hat{p}_D] \tag{15}$$

$$= \frac{\left(\left(\frac{1}{p_D} + \frac{1}{p_E}\right) + \left(\frac{1}{p_A} + \frac{1}{p_B} + \frac{1}{p_C} + \frac{1}{p_E}\right) + \frac{2}{p_E}\right)}{k\left[\left(\frac{1}{p_A} + \frac{1}{p_B} + \frac{1}{p_C} + \frac{1}{p_E}\right)\left(\frac{1}{p_D} + \frac{1}{p_E}\right) - \frac{1}{p_E^2}\right]} \tag{16}$$

$$= \frac{\frac{1}{p_A} + \frac{1}{p_B} + \frac{1}{p_C} + \frac{1}{p_D} + \frac{4}{p_E}}{k\left[\left(\frac{1}{p_A} + \frac{1}{p_B} + \frac{1}{p_C} + \frac{1}{p_E}\right)\left(\frac{1}{p_D} + \frac{1}{p_E}\right) - \frac{1}{p_E^2}\right]} \tag{17}$$

Finally, we want to show that the asymptotic variance of this estimator is always lower than or equal to the variance of the original estimator, which is equivalent to showing that

$$(p_A + p_D)(p_B + p_C + p_E) - \frac{\frac{1}{p_A} + \frac{1}{p_B} + \frac{1}{p_C} + \frac{1}{p_D} + \frac{4}{p_E}}{\left(\frac{1}{p_A} + \frac{1}{p_B} + \frac{1}{p_C} + \frac{1}{p_E}\right)\left(\frac{1}{p_D} + \frac{1}{p_E}\right) - \frac{1}{p_E^2}} \geq 0 \tag{18}$$

Suppose we let

$$A := (p_A + p_D)(p_B + p_C + p_E)\left[\left(\frac{1}{p_A} + \frac{1}{p_B} + \frac{1}{p_C} + \frac{1}{p_E}\right)\left(\frac{1}{p_D} + \frac{1}{p_E}\right) - \frac{1}{p_E^2}\right] \tag{19}$$

$$B := \frac{1}{p_A} + \frac{1}{p_B} + \frac{1}{p_C} + \frac{1}{p_D} + \frac{4}{p_E} \tag{20}$$

We then want to show that $A \geq B$ for all values of p_A, p_B, p_C, p_D, p_E.

The key idea here is to replace the “1" terms by $p_A + p_B + p_C + p_D + p_E$ whenever they occur.

Note that A remains unchanged, because we can write

$$A := \frac{(p_A + p_D)}{1} \frac{(p_B + p_C + p_E)}{1} \left[\left(\frac{1}{p_A} + \frac{1}{p_B} + \frac{1}{p_C} + \frac{1}{p_E} \right) \left(\frac{1}{p_D} + \frac{1}{p_E} \right) - \frac{1}{p_E^2} \right] \tag{21}$$

and the $p_A + p_B + p_C + p_D + p_E$ terms cancel.

However, B now becomes

$$\begin{aligned} B := & \frac{p_A + p_B + p_C + p_D + p_E}{p_A} + \frac{p_A + p_B + p_C + p_D + p_E}{p_B} + \frac{p_A + p_B + p_C + p_D + p_E}{p_C} \\ & + \frac{p_A + p_B + p_C + p_D + p_E}{p_D} + \frac{4(p_A + p_B + p_C + p_D + p_E)}{p_E} \end{aligned} \tag{22}$$

Now, we need to show that $A - B \geq 0$ for all $0 \leq p_A, p_B, p_C, p_D, p_E \leq 1$. We first write

$$A - B = \frac{C}{p_A p_B p_C p_D p_E} \tag{23}$$

where

$$\begin{aligned} C = & (p_A - p_D)^2 (p_B + p_C) p_B p_C + (p_B - p_C)^2 (p_A + p_D) p_A p_D \\ & + p_A p_E (p_A (p_B + p_C)(p_B + p_C + p_E) - 4 p_B p_C p_D) \qquad (24) \\ = & \alpha p_D^2 + \beta p_D + \gamma \qquad (25) \end{aligned}$$

as a quadratic in p_D by replacing p_E by $1 - p_A - p_B - p_C - p_D$, with

$$\alpha = (p_A + p_B)(p_A + p_C)(p_B + p_C) \tag{26}$$

$$\beta = 2p_A(p_B p_C (p_B + p_C - 2) + p_A^2 (p_B + p_C) + p_A (p_B + p_C - 1)(p_B + p_C)) \tag{27}$$

$$\gamma = p_A^2 (p_A + p_B - 1)(p_A + p_C - 1)(p_B + p_C) \tag{28}$$

We can complete the square in our quadratic at (25) and get

$$C = \alpha \left(p_D + \frac{\beta}{2\alpha} \right)^2 - \frac{\beta^2}{4\alpha} + \gamma \tag{29}$$

and now we just need to show that that

$$-\frac{\beta^2}{4\alpha} + \gamma \geq 0 \tag{30}$$

which is a term in p_A, p_B and p_C. The left hand side of (30) gives

$$\frac{p_A^2 (p_B - p_C)^2 (p_A p_B (1 - p_A - p_B) + p_A p_C (1 - p_A - p_C) + p_B p_C (1 - p_B - p_C) - 2 p_A p_B p_C)}{(p_A + p_B)(p_A + p_C)(p_B + p_C)} \tag{31}$$

but we note that since $p_A + p_B + p_C + p_D + p_E = 1$, we can write the third factor of the numerator of (31) as

$$p_A p_B (p_C + p_D + p_E) + p_A p_C (p_B + p_D + p_E) + p_B p_C (p_A + p_D + p_E) - 2 p_A p_B p_C \tag{32}$$

and hence note we have a common factor of $p_A p_B p_C$ which cancels out when we expand (32)

$$(p_A p_B + p_A p_C + p_B p_C)(p_D + p_E) + p_A p_B p_C \tag{33}$$

which is always non-negative. We can hence write (31) as

$$\frac{p_A^2 (p_B - p_C)^2 ((p_A p_B + p_A p_C + p_B p_C)(p_D + p_E) + p_A p_B p_C)}{(p_A + p_B)(p_A + p_C)(p_B + p_C)} \geq 0 \tag{34}$$

which implies that (25) is always non-negative, and thus

$$(p_A + p_D)(p_B + p_C + p_E) - \frac{\frac{1}{p_A} + \frac{1}{p_B} + \frac{1}{p_C} + \frac{1}{p_D} + \frac{4}{p_E}}{\left(\frac{1}{p_A} + \frac{1}{p_B} + \frac{1}{p_C} + \frac{1}{p_E}\right)\left(\frac{1}{p_D} + \frac{1}{p_E}\right) - \frac{1}{p_E^2}} \geq 0 \tag{35}$$

with equality when $p_B = p_C$ and $p_D = p_A(1 - p_A - p_B)/(p_A + p_B)$.

This completes the proof. ■

1.1. Random Projections and Very Sparse Random Projections

We prove the following theorem

Theorem 2 *Suppose $\boldsymbol{w}_1, \ldots, \boldsymbol{w}_S$ are chosen to be pairwise orthogonal. With the estimator $v_i v_j$, we can write the control variate*

$$Z = \sum_{s,t} \alpha_{s,t} v_{\boldsymbol{w}_s} v_{\boldsymbol{w}_t} \tag{36}$$

with $\alpha_{s,t} = d(\boldsymbol{x}_i, \boldsymbol{w}_s) d(\boldsymbol{x}_j, \boldsymbol{w}_t) + d(\boldsymbol{x}_i, \boldsymbol{w}_t) d(\boldsymbol{x}_j, \boldsymbol{w}_s)$, with control variate correction $\hat{c} = -\frac{1}{2}$.

Proof For each hash, our estimate of the inner product $\hat{\theta}$ can be written as

$$\hat{\theta} = v_i v_j + \hat{c} \sum_{s,t} \alpha_{s,t} \left(v_{\boldsymbol{w}_s} v_{\boldsymbol{w}_t} - \mathbb{E}[v_{\boldsymbol{w}_s} v_{\boldsymbol{w}_t}] \right) \tag{37}$$

The variance of our estimates can hence be written as

$$\begin{aligned} \operatorname{Var}[\hat{\theta}] &= \operatorname{Var}[v_i v_j + \hat{c} \sum_{s,t} \alpha_{s,t} v_{\boldsymbol{w}_s} v_{\boldsymbol{w}_t}] && (38) \\ &= \operatorname{Var}[v_i v_j] + \hat{c}^2 \sum_{s,t} \alpha_{s,t}^2 \operatorname{Var}[v_{\boldsymbol{w}_s} v_{\boldsymbol{w}_t}] + 2\hat{c} \alpha_{s,t} \sum_{s,t} \operatorname{Cov}(v_i v_j, v_{\boldsymbol{w}_s} v_{\boldsymbol{w}_t}) \\ &\quad + 2\hat{c} \sum_{s,t,m,n} \alpha_{s,t} \alpha_{m,n} \operatorname{Cov}(v_{\boldsymbol{w}_m} v_{\boldsymbol{w}_n}, v_{\boldsymbol{w}_s} v_{\boldsymbol{w}_t}) && (39) \end{aligned}$$

Suppose we bring in $\hat{c}$ into the αs, i.e. set $\tilde{\alpha} = \hat{c}\alpha$, and we map each of the tuples (s, t) to $k \in \{1, 2, \ldots, S(S-1)/2\}$. Then we can rewrite the above equation to get

$$\begin{aligned} \operatorname{Var}[\hat{\theta}] &= \operatorname{Var}[v_i v_j + \sum_{k} \alpha_k v_{\boldsymbol{w}_s} v_{\boldsymbol{w}_t}] && (40) \\ &= \operatorname{Var}[v_i v_j] + \sum_{k} \alpha_k^2 \operatorname{Var}[v_{\boldsymbol{w}_s} v_{\boldsymbol{w}_t}] + 2\alpha_k \sum_{k} \operatorname{Cov}(v_i v_j, v_{\boldsymbol{w}_s} v_{\boldsymbol{w}_t}) \\ &\quad + 2 \sum_{k,k'} \alpha_k \alpha_{k'} \operatorname{Cov}(v_{\boldsymbol{w}_m} v_{\boldsymbol{w}_n}, v_{\boldsymbol{w}_s} v_{\boldsymbol{w}_t}) && (41) \end{aligned}$$

Then if we take partial derivatives with respect to αs, we have for some arbitrary α_k term

$$\begin{aligned}\frac{\partial \operatorname{Var}[\hat{\theta}]}{\partial \alpha_k} &= 2\alpha_k \operatorname{Var}[v_{\boldsymbol{w}_s} v_{\boldsymbol{w}_t}] + 2\operatorname{Cov}(v_i v_j, v_{\boldsymbol{w}_s} v_{\boldsymbol{w}_t}) \\ &\quad + 2\sum_{k' \neq k} \alpha_{k'} \operatorname{Cov}(v_{\boldsymbol{w}_m} v_{\boldsymbol{w}_n}, v_{\boldsymbol{w}_s} v_{\boldsymbol{w}_t}) \end{aligned} \tag{42}$$

Since we know the weighted vectors are chosen to be orthogonal to each other, then $\operatorname{Cov}(v_{\boldsymbol{w}_m} v_{\boldsymbol{w}_n}, v_{\boldsymbol{w}_s} v_{\boldsymbol{w}_t}) = 0$ if $(m,n) \neq (s,t)$ (or $(m,n) \neq (t,s)$). Hence (42) simplifies to

$$\frac{\partial \operatorname{Var}[\hat{\theta}]}{\partial \alpha_k} = 2\alpha_k \operatorname{Var}[v_{\boldsymbol{w}_s} v_{\boldsymbol{w}_t}] + 2\operatorname{Cov}(v_i v_j, v_{\boldsymbol{w}_s} v_{\boldsymbol{w}_t}) \tag{43}$$

and equating this to zero, we must have that

$$\alpha_k = -\operatorname{Cov}(v_i, v_j, v_{\boldsymbol{w}_s} v_{\boldsymbol{w}_t}) / \operatorname{Var}[v_{\boldsymbol{w}_s} v_{\boldsymbol{w}_t}] \tag{44}$$

Since $v_i s$ are distributed multivariate normal, and $\boldsymbol{x}_i$s and the weighted vectors are normalized to unit length 1, then any

$$\operatorname{Cov}(v_i v_j, v_k v_l) = d(\boldsymbol{x}_j, \boldsymbol{x}_k) d(\boldsymbol{x}_i, \boldsymbol{x}_l) + d(\boldsymbol{x}_j, \boldsymbol{x}_l) d(\boldsymbol{x}_i, \boldsymbol{x}_k) \tag{45}$$

and in general we get

$$\alpha_k = -\frac{1}{2} d(\boldsymbol{x}_j, \boldsymbol{x}_k) d(\boldsymbol{x}_i, \boldsymbol{x}_l) + d(\boldsymbol{x}_j, \boldsymbol{x}_l) d(\boldsymbol{x}_i, \boldsymbol{x}_k) \tag{46}$$

Since we brought $\hat{c}$ into the αs, we thus have the optimal control variate coefficient and terms to be

$$\hat{c} = \frac{1}{2} \qquad \alpha_{i,j} = d(\boldsymbol{x}_j, \boldsymbol{x}_k) d(\boldsymbol{x}_i, \boldsymbol{x}_l) + d(\boldsymbol{x}_j, \boldsymbol{x}_l) d(\boldsymbol{x}_i, \boldsymbol{x}_k) \tag{47}$$

or in general, any form of

$$\hat{c} = \frac{1}{2}\beta \qquad \alpha_{i,j} = \frac{1}{\beta} d(\boldsymbol{x}_j, \boldsymbol{x}_k) d(\boldsymbol{x}_i, \boldsymbol{x}_l) + d(\boldsymbol{x}_j, \boldsymbol{x}_l) d(\boldsymbol{x}_i, \boldsymbol{x}_k) \tag{48}$$

for any $\beta \in \mathbb{R}$.

■