Technical Report: Accelerating Approximate Analytical Join Queries over Unstructured Data with Error Guarantees

Yuxuan Zhu, Tengjun Jin, Kaimeng Zhu, Siheng Pan, Chenghao Mo and Daniel Kang
Siebel School of Computing and Data Science
University of Illinois Urbana-Champaign
Email: {yxx404,tengjun2,kaimeng2,span14,cmo8,ddkang}@illinois.edu

Abstract

BAS accelerates approximate analytical join queries over unstructured data and provides statistical guarantees via confidence intervals. Given an expensive Oracle method that can process the analytical join query with high accuracy, BAS integrates blocking and sampling algorithms to reduce the Oracle cost. Given an Oracle budget $b = b_1 + b_2$, BAS automatically allocate data tuples to sampling or blocking and minimizes the mean squared error of the estimation. In this document, we theoretically analyze that BAS converges to the optimal allocation at the rate of $\mathcal{O}(1/\sqrt{b_1})$ and asymptotically outperforms or matches the standalone sampling algorithm. Furthermore, we show that BAS be extended to accelerate approximate selection join queries.

Index Terms

theoretical analysis, join, analytics, unstructured data, approximate query processing

Symbol	Description
D	stratified dataset of data tuples in the cross product of join tables
W	normalized similarity scores of tuples in D
O	Oracle
g	attribute where the aggregate is computed
b, b_1, b_2	Oracle budget of the overall procedure, pilot sampling, and sampling+blocking execution
p	confidence
α	maximum blocking ratio
K	number of strata
$\widehat{\mathsf{COUNT}}, \widehat{\mathsf{SUM}}, \widehat{\mathsf{AVG}}, \widehat{\mathsf{AGG}}$	estimated aggregate for the entire dataset, sampling regime (with a subscript s), and blocking regime (with a subscript b)
l, u	lower, upper bound of the confidence interval
$\mathcal{O}(\cdot)$	Bachmann-Landau big-O notation
$\beta, \beta^*, \hat{\beta}^*$	the allocated strata to blocking that is given, optimal, and estimated
$n_i^{(1)}, n_i$	assigned Oracle budget for stratum i during pilot sampling and blocking+sampling execution
$\sigma_i^2, \hat{\sigma}_i^2$	(estimated) sampling variance of stratum i
$\mu, \hat{\mu}$	(estimated) expectation of stratum i
$S_i^{(1)}, S_i$	sampled data tuples from stratum i during pilot sampling and blocking+sampling
t_j	t-statistic of the j-th resampling iteration
$\tilde{D}, \tilde{D}^{(s)}$	the set of matching tuples of the entire dataset or the sampling regime

TABLE I: Summary of Notation

I. Setup

In this section, we describe BAS in detail as a setup for the theoretical analysis. We will go over the entire procedure of BAS, including stratification, pilot sampling, blocking, sampling, and resampling.

Inputs and Outputs. BAS takes as inputs the cross product of tables D, the similarity scores W, the Oracle O, the overall Oracle budget b divided for pilot sampling (b_1) and bloking+sampling execution (b_2), the confidence p, the maximum blocking ratio α, and the number of strata K. BAS outputs an estimated aggregate $\hat{\mu}$ and its confidence interval $[l, u]$.

Stratification. We divide D into a maximum blocking regime and a minimum sampling regime (D_0). The maximum blocking regime contains tuples with the top $\alpha \cdot b$ similarity scores, where α is a parameter between 0 and 1 to control the size of the maximum blocking regime. Next, we stratify the maximum blocking regime into K strata $(D_1, \ldots, D_K)$ with equal sizes. The number of strata K is automatically determined to ensure that each stratum has an Oracle budget of at least 1,000.

Pilot Sampling. For each stratum, we execute WWJ to obtain a pilot sample $S_i^{(1)}$. The Oracle budget for the stratum i in the pilot sampling is calculated as

$$n_i^{(1)} = b_1 \cdot \frac{\sum_{s \in D_i} W(s)}{\sum_{s \in D} W(s)} \tag{1}$$

We can estimate the sampling variance as

$$\hat{\sigma}_i^2 = \frac{1}{n^{(1)} - 1} \sum_{s \in S_i^{(1)}} \left(\frac{g(s)O(s)}{W(s)|D_i|} - \left(\frac{1}{n_i^{(1)}} \sum_{s' \in S_i^{(1)}} \frac{g(s')O(s')}{W(s')|D_i|} \right) \right)^2 W(s)$$

Given an allocation β, we can then estimate the MSE of BAS for a SUM aggregate as follows

$$\widehat{MSE}_{\mathsf{SUM}}(D, \beta, W, b_2) = \sum_{0 \leq i \leq k, i \notin \beta} \frac{|D_i|^2}{n_i} \cdot \hat{\sigma}_i^2$$

where n_i is the assigned Oracle budget for stratum i in the stage of blocking+sampling execution, calculated as

$$n_i = \begin{cases} (b_2 - \sum_{j \in \beta} |D_i|) \cdot \frac{\sum_{s \in D_i} W(s)}{\sum_{1 \leq j \leq K, j \notin \beta} \sum_{s \in D_j} W(s)} & i \notin \beta \\ |D_i| & i \in \beta \end{cases}$$

Next, we obtain the estimated optimal allocation by solving the following optimization problem using iterative methods.

$$\hat{\beta}^* = \underset{\beta \subset \{1, \ldots, K\}}{\arg\min} \widehat{MSE}_{\mathsf{SUM}}(D, \beta, W, b_2)$$

Blocking+Sampling. Given the optimal allocation $\hat{\beta}^*$, we use the Oracle budget b_1 to execute the Oracle on the strata that are allocated to the blocking regime. On the remaining strata, we execute WWJ to obtain a sample using the remaining Oracle budget. We can obtain the sample S_i on the stratum i that is allocated to sampling. Then, we merge the result of all strata and the results of pilot sampling to estimate the aggregate. Specifically, we estimate the aggregate as follows:

$$\widehat{\mathsf{COUNT}}_s = \frac{1}{\sum_{i \notin \hat{\beta}^*} n_i} \sum_{i \notin \hat{\beta}^*} \sum_{s \in S_i} \frac{O(s)}{W(s)|D_i|}, \quad \mathsf{COUNT}_b = \sum_{i \in \hat{\beta}^*} \sum_{s \in D_i} O(s)$$

$$\widehat{\mathsf{COUNT}}_s = \frac{1}{\sum_{i \notin \hat{\beta}^*} n_i} \sum_{i \notin \hat{\beta}^*} \sum_{s \in S_i} \frac{g(s)O(s)}{W(s)|D_i|}, \quad \mathsf{COUNT}_b = \sum_{i \in \hat{\beta}^*} \sum_{s \in D_i} g(s)O(s)$$

$$\widehat{\mathsf{COUNT}} = \widehat{\mathsf{COUNT}}_s + \mathsf{COUNT}_b, \quad \widehat{\mathsf{SUM}} = \widehat{\mathsf{SUM}}_s + \mathsf{SUM}_b$$

$$\widehat{\mathsf{AVG}} = \left(\widehat{\mathsf{SUM}}_s + \mathsf{SUM}_b \right) \Big/ \left(\widehat{\mathsf{COUNT}}_s + \mathsf{COUNT}_b \right)$$

Resampling. We apply the bootstrap-t resampling scheme to calculate the CI [1]. The bootstrap-t scheme estimates the standard error $\frac{\hat{\mu} - \mu}{\sigma}$ (i.e., t-statistic) of the underlying distribution by resampling existing samples. To process aggregation queries, we first calculate the mean and standard deviation of the estimator. Next, we use sampling with replacement to resample from all existing samples $(S^{(1)} \cup S)$. We calculate the t-statistic of the j-th iteration as follows.

$$t_j = \frac{\widehat{\mathsf{AGG}}_j - \widehat{\mathsf{AGG}}}{\hat{\sigma}_j}$$

where $\widehat{\mathsf{AGG}}_j$ and $\hat{\sigma}_j$ are the estimated aggreagte and standard deviation on the j-th resample. To achieve statistical guarantees, we repeat the resampling a sufficient number of times (e.g., 1000). Finally, we use the percentiles of resampled t-statistics to construct the CI, that is

$$l = \widehat{\mathsf{AGG}} - \text{Percentile}\left(t, \frac{1-p}{2}\right), \quad u = \widehat{\mathsf{AGG}} - \text{Percentile}\left(t, 1 - \frac{1-p}{2}\right)$$

II. BaS Converges to the Optimal Allocation

Theorem 1. *The MSE with the estimated minimizer $\hat{\beta}^*$ converges to that with the true minimizer β^* with the rate $\mathcal{O}\left(1/\sqrt{b_1}\right)$:*

$$\frac{MSE(D,\widehat{\beta}^*,W,b_2)-MSE(D,\beta^*,W,b_2)}{MSE(D,\beta^*,W,b_2)}=\mathcal{O}\left(\frac{1}{\sqrt{b_1}}\right)$$

Proof: We first show the estimated MSE converges to the true MSE. To achieve that, we establish the convergence rate of the sample mean $\hat{\mu}_i$ and variance $\hat{\sigma}_i^2$ for stratum i because MSE estimations are calculated with summations, multiplications, and divisions of basic estimators $\hat{\mu}_i, \hat{\sigma}_i^2$.

We construct the concentration inequality for the sample mean $\hat{\mu}_i$ and sample variance $\hat{\sigma}_i^2$ of each stratum $D_i, i=0,\dots,K$. Given the i.i.d. sample, we can bound the relative errors of $\hat{\mu}_i$ and $\hat{\sigma}_i^2$ with Chebeyshev's Inequality, where the variance of an i.i.d. sample variance is calculated as $Var\left[\hat{\sigma}_i^2\right]=\frac{2\sigma_i^4}{n_i-1}$ [2].

$$\mathbb{P}\left[\left|\frac{\mu_i-\hat{\mu}_i}{\mu_i}\right|\leq\frac{\sigma_i}{\mu_i\sqrt{\delta n_i^{(1)}}}\right]\geq 1-\frac{\delta}{2} \tag{2}$$

$$\mathbb{P}\left[\left|\frac{\sigma_i^2-\hat{\sigma}_i^2}{\sigma_i^2}\right|\leq\sqrt{\frac{2}{\delta(n_i^{(1)}-1)}}\right]\geq 1-\frac{\delta}{2} \tag{3}$$

where δ is a small probability.

Since $n_i^{(1)}$ is linearly related to the Oracle budget b_1 (Eq. 1), we have the following big-$\mathcal{O}$ notations showing the convergence rate that holds with high probability.

$$\frac{\mu_i-\hat{\mu}_i}{\mu_i}=\mathcal{O}\left(b_1^{-1/2}\right),\quad \frac{\sigma_i^2-\hat{\sigma}_i^2}{\sigma_i^2}=\mathcal{O}\left(b_1^{-1/2}\right) \tag{4}$$

Next, we show the convergence rate of relative error is preserved after summations, multiplications, and divisions. Namely, for unbiased estimators μ_1 and μ_2 with relative error converging to 0 with a rate of $\mathcal{O}\left(b_1^{-1}\right)$ $(t>0)$, we have Equations 5 - 7.

$$\begin{aligned}(\mu_1+\mu_2)-(\hat{\mu}_1+\hat{\mu}_2)&=\mu_1\cdot\mathcal{O}\left(b_1^{-1}\right)+\mu_2\cdot\mathcal{O}\left(b_1^{-1}\right)\\&=(\mu_1+\mu_2)\cdot\mathcal{O}\left(b_1^{-1}\right) && (5)\\ \mu_1\mu_2-\hat{\mu}_1\hat{\mu}_2&=\mu_1\mu_2-\mu_1\left(1+\mathcal{O}\left(b_1^{-1}\right)\right)\mu_2\left(1+\mathcal{O}\left(b_1^{-1}\right)\right)\\&=\mu_1\mu_2\left(\mathcal{O}\left(b_1^{-1}\right)+\mathcal{O}\left(b_1^{-2t}\right)\right)\\&=\mu_1\mu_2\mathcal{O}\left(b_1^{-1}\right) && (6)\\ \mu^{-1}-\hat{\mu}^{-1}&=\mu^{-1}\hat{\mu}^{-1}(\hat{\mu}-\mu)\\&=\mu^{-2}\left(1+\mathcal{O}\left(b_1^{-1}\right)\right)^{-1}\mu\mathcal{O}\left(b_1^{-1}\right)\\&=\mu^{-1}\mathcal{O}\left(b_1^{-1}\right) && (7)\end{aligned}$$

Therefore, given the convergence rate of basic estimators (Eq. 4) and propagation rules (Eq. 5-7), the relative error of our optimization objective converges to 0 with the same rate $\mathcal{O}\left(b_1^{-1/2}\right)$, with high probability. Namely, the following probabilistic bound holds for any $\beta\subset\{1,\dots,K\}$ with a high probability of $1-\delta/2$

$$\mathbb{P}\left[\left|\frac{MSE_{\mathsf{AGG}}(D,\beta,W,b_2)-\widehat{MSE}_{\mathsf{AGG}}(D,\beta,W,b_2)}{MSE_{\mathsf{AGG}}(D,\beta,W,b_2)}\right|\leq\frac{C}{\sqrt{b_1}}\right]\geq 1-\frac{\delta}{2} \tag{8}$$

where C is a constant independent of n. Based on Equation 8, we can derive the following bound for MSE that holds with a high probability of $1-\delta/2$.

$$\frac{1}{1+\frac{C}{\sqrt{n}}}\widehat{MSE}_{\mathsf{AGG}}(\beta)\leq MSE_{\mathsf{AGG}}(\beta)\leq\frac{1}{1-\frac{C}{\sqrt{n}}}\widehat{MSE}_{\mathsf{AGG}}(\beta)$$

where we omit the parameters D, W, b_2 for simplicity.

We then derive the upper bound of the difference between the MSE of BAS and the optimal MSE. Given the estimated minimizer $\hat{\beta}^*$ and the true minimizer β^*, we can establish the following upper bound of the difference between $MSE_{\mathsf{AGG}}(\hat{\beta}^*)$ and $MSE_{\mathsf{AGG}}(\beta^*)$.

$$\begin{aligned} MSE_{\mathsf{AGG}}(\hat{\beta}^*) - MSE_{\mathsf{AGG}}(\beta^*) &\leq \frac{1}{1-\frac{C_1}{\sqrt{n}}}\widehat{MSE}_{\mathsf{AGG}}(\hat{\beta}^*) - MSE_{\mathsf{AGG}}(\beta^*) && (9)\\ &\leq \frac{1}{1-\frac{C_1}{\sqrt{n}}}\widehat{MSE}_{\mathsf{AGG}}(\beta^*) - MSE_{\mathsf{AGG}}(\beta^*) && (10)\\ &\leq \frac{1}{1-\frac{C_1}{\sqrt{n}}}\left(1+\frac{C_2}{\sqrt{n}}\right) MSE_{\mathsf{AGG}}(\beta^*) - MSE_{\mathsf{AGG}}(\beta^*) && (11)\\ &= \frac{C_1+C_2}{\sqrt{n}-C_1} MSE_{\mathsf{AGG}}(\beta^*) && (12) \end{aligned}$$

where inequalities 9 and 11 apply the probabilistic inequalities of $MSE_{\mathsf{AGG}}(\hat{\beta}^*)$ and $MSE_{\mathsf{AGG}}(\beta^*)$ respectively, inequality 10 is due to the definition of estimated minimizer $\hat{\beta}^*$, and C_1, C_2 are constants from the probabilistic inequalities.

Finally, we derive the upper bound of the relative error of the optimization objective that holds with high probability.

$$\frac{MSE_{\mathsf{AGG}}(\hat{\beta}^*) - MSE_{\mathsf{AGG}}(\beta^*)}{MSE_{\mathsf{AGG}}(\beta^*)} \leq \frac{C_1+C_2}{\sqrt{n}-C_1}$$

This upper bound shows that the relative error between the MSE with estimated minimizer $\hat{\beta}^*$ and the MSE with actual minimizer β^* converges to 0 at the rate $\mathcal{O}\left(b_1^{-1/2}\right)$ with high probability. ■

III. BAS OUTPERFORMS OR MATCHES WWJ

Theorem 2. *If there exists an allocation β such that the following two conditions hold*

$$\mathbb{E}_{s\in\tilde{D}^{(s)}}\left[\frac{1/|\tilde{D}^{(s)}|}{W(s)}\right] \leq \mathbb{E}_{s\in D}\left[\frac{1/|D|}{W(s)}\right] \tag{13}$$

$$\frac{|\tilde{D}^{(s)}|^2}{b_2^{(s)}} \leq \frac{|D|^2}{b} \tag{14}$$

BAS *outperforms* WWJ *asymptotically, i.e.,*

$$MSE_{\mathsf{SUM}} = C \cdot MSE_{\mathsf{SUM}}^{(w)} + \mathcal{O}\left(b_1^{-1}b_2^{-1/2}\right)$$

where C is a coefficient less than 1:

$$C < \frac{|\tilde{D}^{(s)}|^2/b_2^{(s)}}{|\tilde{D}|^2/b}\frac{\mathbb{E}_{s\in\tilde{D}^{(s)}}\left[\frac{1/|\tilde{D}^{(s)}|}{W(s)}\right]}{\mathbb{E}_{s\in\tilde{D}}\left[\frac{1/|\tilde{D}|}{W(s)}\right]} \leq 1$$

Otherwise, BAS *matches* WWJ *asymptotically, i.e.,*

$$MSE_{\mathsf{SUM}} \leq MSE_{\mathsf{SUM}}^{(w)} + \mathcal{O}\left(b_1^{-1}b_2^{-1/2}\right)$$

Proof: We first derive the ratio between the MSE of BAS and importance sampling, given a deterministic allocation β. The MSe of BAS for a SUM aggregate can be calculated as follows.

$$\begin{aligned} MSE_{\mathsf{SUM}} &= \sum_{i\notin\beta}\frac{1}{n_i}\left(\mathbb{E}_{s\sim W, s\in D_i}\left[\left(\frac{g(s)O(s)}{W(s)}\right)^2\right] - \left(\mathbb{E}_{s\sim W, s\in D_i}\left[\frac{g(s)O(s)}{W(s)}\right]\right)^2\right)\\ &= \frac{1}{b_2^{(s)}}\sum_{i\notin\beta}\left(|D_i|\mathbb{E}\left[\frac{g(s)^2O(s)}{r_iW(s)}\right] - \frac{1}{r_i}\left(|D_i|\mathbb{E}\left[g(s)O(s)\right]\right)^2\right) \end{aligned}$$

where

$$b_2^{(s)} = b_2 - \sum_{i\in\beta}|D_i|, \quad r_i = \frac{\sum_{s\in D_i}W(s)}{\sum_{j\notin\beta}\sum_{s\in D_j}W(s)}$$

Assuming the independence of the SUM column $g(\cdot)$ and the oracle results $O(\cdot)$, we can further simplify the expression of MSE as follows.

$$MSE_{\mathsf{SUM}} = \frac{1}{b_2^{(s)}} \left(\mathbb{E}\left[g(s)^2\right] \sum_{i \notin \beta} \left(|D_i| \mathbb{E}\left[\frac{O(s)}{r_i W(s)}\right] \right) - \frac{\mathbb{E}[g(s)]^2}{r_i} \sum_{i \notin \beta} \left(|D_i| \mathbb{E}\left[O(s)\right]\right)^2 \right) \tag{15}$$

We notice that r_i unweights the proxy scores over a stratum into the proxy scores over the whole sampling regions. In this case, we merge $r_i W(s)$ as $W(s)$. Furthermore, we can simplify the sum of the expectations of strata into the expectation of the whole sampling region.

$$MSE_{\mathsf{SUM}} = \frac{1}{b_2^{(s)}} \left(|D| \cdot \mathbb{E}\left[g(s)^2\right] \cdot \mathbb{E}\left[\frac{O(s)}{W(s)}\right] - \frac{\mathbb{E}[g(s)]^2}{r_i} \sum_{i \notin \beta} \left(|D_i| \mathbb{E}\left[O(s)\right]\right)^2 \right)$$

Next, we rewrite the expectations over all tuples with expectations over matching tuples by evaluating the oracle results $O(\cdot)$.

$$\begin{aligned} MSE_{\mathsf{SUM}} &= \frac{1}{b_2^{(s)}} \left(\mathbb{E}\left[g(s)^2\right] \cdot \left|\tilde{D}^{(s)}\right| \cdot \mathbb{E}_{s \in \tilde{D}^{(s)}} \left[\frac{1}{W(t)}\right] - \mathbb{E}[g(s)]^2 \sum_{i \notin \beta} \frac{|\tilde{D}_i|^2}{r_i} \right) \\ &= \frac{\left|\tilde{D}^{(s)}\right|}{b_2^{(s)}} \left(\mathbb{E}\left[g(s)^2\right] \mathbb{E}_{s \in \tilde{D}^{(s)}} \left[\frac{1}{W(t)}\right] - \mathbb{E}[g(s)]^2 \sum_{i \notin \beta} q_i \left|\tilde{D}_i^{(s)}\right| \right) \end{aligned}$$

where

$$q_i = \frac{\left|\tilde{D}_i^{(s)}\right|}{\sum_{j \notin \beta} \left|\tilde{D}_j^{(s)}\right|} \Big/ r_i$$

We rewrite the MSE of WWJ for a SUM aggregate.

$$\begin{aligned} MSE_{\mathsf{SUM}}^{(w)} &= \frac{1}{b} \left(\mathbb{E}_{s \sim W} \left[\left(\frac{g(s) O(s)}{W(s)} \right)^2 \right] - \left(\mathbb{E}_{s \sim W} \left[\frac{g(s) O(s)}{W(s)} \right] \right)^2 \right) \\ &= \frac{1}{b} \left(|D| \cdot \mathbb{E}\left[\frac{g(s)^2 O(s)}{W(s)}\right] - \left(|D| \cdot \mathbb{E}\left[g(s) O(s)\right]\right)^2 \right) \\ &= \frac{1}{b} \left(\mathbb{E}\left[g(s)^2\right] \cdot |D| \cdot \mathbb{E}\left[\frac{O(s)}{W(s)}\right] - |\tilde{D}|^2 \cdot \mathbb{E}[g(s)]^2 \right) \\ &= \frac{|\tilde{D}|}{b} \left(\mathbb{E}\left[g(s)^2\right] \mathbb{E}_{s \in \tilde{D}} \left[\frac{1}{W(t)}\right] - \mathbb{E}[g(s)]^2 |\tilde{D}| \right) \end{aligned}$$

We take the ratio between the MSE of BAS and that of WWJ. We assume the variance of SUM-column in the sampling region is the same as that for all tuples.

$$\begin{aligned} \frac{MSE_{\mathsf{SUM}}}{MSE_{\mathsf{SUM}}^{(w)}} &= \frac{|\tilde{D}^{(s)}| / b_2^{(s)}}{|\tilde{D}| / b} \frac{\mathbb{E}\left[g(s)^2\right] \mathbb{E}_{s \in \tilde{D}^{(s)}} \left[\frac{1}{W(s)}\right] - \mathbb{E}[g(s)]^2 \sum_{i \notin \beta} q_i \left|\tilde{D}_i^{(s)}\right|}{\mathbb{E}\left[g(s)^2\right] \mathbb{E}_{s \in \tilde{D}} \left[\frac{1}{W(s)}\right] - \mathbb{E}[g(s)]^2 |\tilde{D}|} \\ &= \frac{|\tilde{D}^{(s)}| / b_2^{(s)}}{|\tilde{D}| / b} \frac{\mathbb{E}_{s \in \tilde{D}^{(s)}} \left[\frac{1}{W(s)}\right] - \frac{\mathbb{E}[g(s)]^2}{\mathbb{E}[g(s)^2]} \sum_{i \notin \beta} q_i \left|\tilde{D}_i^{(s)}\right|}{\mathbb{E}_{s \in \tilde{D}} \left[\frac{1}{P(s)}\right] - \frac{\mathbb{E}[g(s)]^2}{\mathbb{E}[g(s)^2]} |\tilde{D}|} \\ &= \frac{|\tilde{D}^{(s)}| / b_2^{(s)}}{|\tilde{D}| / b} \frac{\mathbb{E}_{s \in \tilde{D}^{(s)}} \left[\frac{1}{W(t)}\right] - \frac{1}{1+CV_g} \sum_{i \notin \beta} q_i \left|\tilde{D}_i^{(s)}\right|}{\mathbb{E}_{s \in \tilde{D}} \left[\frac{1}{W(s)}\right] - \frac{1}{1+CV_g} |\tilde{D}|} \end{aligned}$$

where CV_g is the coefficient of variation of the SUM-column $g(\cdot)$,

$$CV_g = \frac{Var[g(s)]}{E[g(s)]^2}$$

We apply Holder's Inequality to obtain an upper bound to the effect of stratification on the MSE as follows.

$$\sum_{i\notin\beta} q_i \left|\tilde{D}_i^{(s)}\right| = \frac{1}{|\tilde{D}^{(s)}|}\sum_{i\notin\beta}\frac{\left|\tilde{D}_i^{(s)}\right|^2}{r_i} = \frac{1}{|\tilde{D}^{(s)}|}\left(\sum_{i\notin\beta}\frac{\left|\tilde{D}_i^{(s)}\right|^2}{r_i}\right)\left(\sum_{i\notin\beta} r_i\right) \geq \frac{1}{|\tilde{D}^{(s)}|}\left(\sum_{i\notin\beta}\frac{\left|\tilde{D}_i^{(s)}\right|}{\sqrt{r_i}}\sqrt{r_i}\right)^2 = \left|\tilde{D}^{(s)}\right|$$

Therefore, the ratio has the following upper bound.

$$\frac{MSE_{\mathsf{SUM}}}{MSE_{\mathsf{SUM}}^{(w)}} \leq \frac{|\tilde{D}^{(s)}|/b_2^{(s)}}{|\tilde{D}|/b}\,\frac{\mathbb{E}_{s\in\tilde{D}^{(s)}}\left[\frac{1}{W(s)}\right] - \frac{1}{1+CV_g}|\tilde{D}^{(s)}|}{\mathbb{E}_{s\in\tilde{D}}\left[\frac{1}{W(s)}\right] - \frac{1}{1+CV_g}|\tilde{D}|} = \frac{|\tilde{D}^{(s)}|^2/b_2^{(s)}}{|\tilde{D}|^2/b}\,\frac{\mathbb{E}_{s\in\tilde{D}^{(s)}}\left[\frac{1/|\tilde{D}^{(s)}|}{W(s)}\right] - \frac{1}{1+CV_g}}{\mathbb{E}_{s\in\tilde{D}}\left[\frac{1/|\tilde{D}|}{W(s)}\right] - \frac{1}{1+CV_g}}$$

If the following condition holds

$$\mathbb{E}_{s\in\tilde{D}^{(s)}}\left[\frac{1/|\tilde{D}^{(s)}|}{W(s)}\right] \leq \mathbb{E}_{s\in\tilde{D}}\left[\frac{1/|\tilde{D}|}{W(s)}\right], \tag{16}$$

we have the following upper bound for the ratio

$$\frac{MSE_{\mathsf{SUM}}}{MSE_{\mathsf{SUM}}^{(IS)}} < \frac{|\tilde{D}^{(s)}|^2/b_2^{(s)}}{|\tilde{D}|^2/b}\,\frac{\mathbb{E}_{s\in\tilde{D}^{(s)}}\left[\frac{1/|\tilde{D}^{(s)}|}{W(s)}\right]}{\mathbb{E}_{s\in\tilde{D}}\left[\frac{1/|\tilde{D}|}{W(s)}\right]}$$

Furthermore, if the matching tuples are sparse in the sampling region, i.e.,

$$|\tilde{D}^{(s)}|^2/b_2^{(s)} < |\tilde{D}|^2/b \tag{17}$$

Then,

$$MSE_{\mathsf{SUM}} = C \cdot MSE_{\mathsf{SUM}}^{(IS)}$$

where C is a coefficient less than 1,

$$C < \frac{\mathbb{E}_{s\in\tilde{D}^{(s)}}\left[\frac{1/|\tilde{D}^{(s)}|}{W(s)}\right]}{\mathbb{E}_{s\in\tilde{D}}\left[\frac{1/|\tilde{D}|}{W(s)}\right]} \leq 1$$

We then apply the Theorem 1 to replace the MSE of BaS with deterministic allocation with the MSE of BaS with pilot sampling. Namely, we have the following approximation:

$$MSE_{\mathsf{SUM}}(\hat{\beta}^*) = MSE_{\mathsf{SUM}}(\beta^*) \cdot \left(1 + \mathcal{O}\left(b_1^{-1/2}\right)\right) = C \cdot MSE_{\mathsf{SUM}}^{(w)}\left(1 + \mathcal{O}\left(b_1^{-1/2}\right)\right)$$

Since $MSE_{\mathsf{SUM}}^{(w)}$ converges to 0 at the rate $\mathcal{O}\left(b^{-1}\right)$. Therefore, we conclude that if the conditions 16 and 17 hold, JoinML outperforms WWJ asymptotically. Namely,

$$MSE_{\mathsf{SUM}}(\hat{\beta}^*) = C \cdot MSE_{\mathsf{SUM}}^{(IS)} + \mathcal{O}\left(b_1^{-1/2} b^{-1}\right) \tag{18}$$

On the other hand, if the conditions 16 and 17 do not hold, we can set $\beta = \emptyset$, which will make the sampling region become the entire data tuples. In this case, the upper bound of the ratio will be 1. Namely,

$$MSE_{\mathsf{SUM}} \leq MSE_{\mathsf{SUM}}^{(IS)}$$

Taking Theorem 1 into account, we will have the following upper bound for MSE of BaS.

$$MSE_{\mathsf{SUM}} \leq MSE_{\mathsf{SUM}}\left(1 + \mathcal{O}\left(b_1^{-1/2}\right)\right) \leq MSE_{\mathsf{SUM}}^{(w)}\left(1 + \mathcal{O}\left(b_1^{-1/2}\right)\right) = MSE_{\mathsf{SUM}}^{(w)} + \mathcal{O}\left(b_1^{-1/2} b^{-1}\right) \tag{19}$$

To conclude, we have shown that the MSE of BaS either outperforms (Eq. 18) or matches (Eq. 19) that of WWJ asymptotically.

■

IV. BaS for Selection Join Queries

Lemma 3. *With a probability higher than p, we can achieve the overall recall target γ if γ_s satisfies*

$$\gamma_s \geq \gamma - (1-\gamma)\frac{\mathsf{COUNT}_b}{\mathrm{UB}(\mathsf{COUNT}_s, Var[\mathsf{COUNT}_s], b, p)}$$

where

$$\mathrm{UB}(\mu, \sigma^2, b, p) = \mu + \frac{\sigma}{\sqrt{b}}\sqrt{2\log\frac{2}{1-p}}$$

Proof: We first show the upper bound of the number of matching tuples in the sampling region $\widehat{\mathsf{COUNT}}_s$. Then, the required recall target γ_s of the sampling region follows automatically.

Given an i.i.d. sample of size n drawn from a population with mean μ and finite and non-zero variance σ^2, the upper bound of the sample mean can be estimated with normal approximation [3], [4]. Namely,

$$\mathbb{P}\left[\hat{\mu} \geq \mu + \frac{\sigma}{\sqrt{n}}\sqrt{2\log\frac{2}{1-p}}\right] \leq \frac{1-p}{2}$$

where p is the confidence.

We rewrite the recall target of the sampling region γ_s' with the overall recall target γ specified by the user.

$$\gamma_s = \frac{\gamma\left(\widehat{\mathsf{COUNT}}_s + \widehat{\mathsf{COUNT}}_b\right) - \widehat{\mathsf{COUNT}}_b}{\widehat{\mathsf{COUNT}}_s} = \gamma - (1-\gamma)\frac{\widehat{\mathsf{COUNT}}_b}{\widehat{\mathsf{COUNT}}_s}$$

We observe that γ_s is monotonically increasing with respect to $\widehat{\mathsf{COUNT}}_s$. Therefore, we use the upper bound of $\widehat{\mathsf{COUNT}}_s$ to estimate the required γ_s such that the user-specified overall recall target can be guaranteed with high probability. ■

References

[1] P. Hall, "Theoretical comparison of bootstrap confidence intervals," *The Annals of Statistics*, 1988.

[2] R. L. Berger and G. Casella, *Statistical inference*. Duxbury, 2001.

[3] D. Kang, E. Gan, P. Bailis, T. Hashimoto, and M. Zaharia, "Approximate selection with guarantees using proxies," *PVLDB*, 2020.

[4] S. Agarwal, H. Milner, A. Kleiner, A. Talwalkar, M. Jordan, S. Madden, B. Mozafari, and I. Stoica, "Knowing when you're wrong: building fast and reliable approximate query processing systems," in *SIGMOD*, 2014.