PAIRED MEDIAN-OF-MEANS REWARDS FOR ROBUST CONFIGURATION SELECTION IN VECTOR SEARCH BENCHMARKING

FARS
Analemma
fars@analemma.ai

ABSTRACT

Configuration selection for approximate nearest neighbor search (ANNS) systems requires reliable QPS-recall benchmarking, yet wall-clock measurements are corrupted by environmental noise from CPU throttling, cache effects, and background processes. On high-noise datasets, standard estimation methods achieve 0% top-1 accuracy—they never identify the best configuration. We propose Paired Median-of-Means (Paired-MoM), combining paired randomized-order execution to cancel environmental drift with median-of-means aggregation to suppress heavy-tailed outliers. Paired execution measures candidate and reference configurations back-to-back under identical conditions, computing log speedup ratios that cancel additive drift. Experiments on SIFT-128 and GIST-960 demonstrate dramatic improvements: Kendall tau increases from 0.458 to 0.840 (+83%) on GIST-960 and from 0.875 to 0.925 (+5.7%) on SIFT-128. Ablation analysis reveals that drift cancellation through pairing is the dominant factor, while MoM provides marginal benefit at limited budgets. Paired-MoM achieves tau ≥ 0.90 at budget 30 on SIFT-128, a threshold no baseline reaches.

WARNING: This paper was generated by an automated research system. The code is publicly available.[1]

1 INTRODUCTION

Approximate nearest neighbor search (ANNS) is a critical component of modern AI systems, enabling efficient similarity retrieval for applications ranging from retrieval-augmented generation to recommendation systems and semantic search (Han et al., 2023). Graph-based methods such as HNSW (Malkov & Yashunin, 2016) have emerged as the dominant approach, achieving state-of-the-art performance across diverse benchmarks (Aumüller et al., 2018). However, these methods expose numerous configuration parameters (graph degree, construction parameters, search parameters) that significantly impact the QPS-recall tradeoff, making configuration selection a critical practical challenge.

Reliable benchmarking is essential for configuration selection, yet QPS measurements are inherently noisy due to environmental factors including CPU thermal throttling, dynamic voltage and frequency scaling, cache effects, and background process activity (Mytkowicz et al., 2009; Barrett et al., 2016). Our pilot analysis reveals that standard ANNS benchmarks exhibit coefficient of variation (CV) of 4.6–4.9%, well above the 2% threshold typically considered acceptable for reliable comparison (Chen & Revels, 2016). This noise corrupts configuration rankings: on high-dimensional datasets like GIST-960, standard estimation methods achieve 0% top-1 accuracy—they never identify the best configuration across repeated trials.

We propose Paired Median-of-Means (Paired-MoM), a robust reward estimator that addresses this challenge through two complementary mechanisms. First, paired randomized-order execution measures candidate and reference configurations back-to-back under identical system conditions, canceling environmental drift through differencing. Second, median-of-means aggregation provides

[1] https://gitlab.com/fars-a/paired-mom-qps-reward

sub-Gaussian concentration even with heavy-tailed noise, suppressing outliers from transient system events. Our contributions are:

- A paired execution protocol that dramatically improves ranking accuracy, increasing Kendall tau from 0.458 to 0.840 (+83%) on GIST-960 and from 0.875 to 0.925 (+5.7%) on SIFT-128.
- Ablation analysis revealing that drift cancellation through pairing is the dominant factor, contributing over 95% of the improvement, while MoM aggregation provides marginal additional benefit at limited budgets.
- Budget efficiency analysis showing Paired-MoM achieves tau ≥ 0.90 at budget 30 on SIFT-128 (no baseline reaches this threshold) and dominates all baselines from budget 6 on GIST-960.

2 Related Work

2.1 ANNS Algorithms and Benchmarking

Graph-based methods have emerged as the dominant paradigm for approximate nearest neighbor search (ANNS). Hierarchical Navigable Small World (HNSW) graphs (Malkov & Yashunin, 2016) achieve state-of-the-art query performance through multi-layer proximity graphs with logarithmic search complexity. FAISS (Johnson et al., 2017) provides GPU-accelerated implementations combining inverted file indexing with product quantization for billion-scale datasets. DiskANN (Subramanya et al., 2019) extends graph-based search to disk-resident indices, enabling single-node billion-point search with bounded memory. ParlayANN (Manohar et al., 2023) introduces deterministic parallel algorithms for scalable index construction. A comprehensive survey by Wang et al. (2021) systematically compares graph-based ANNS methods across diverse workloads. ANN-Benchmarks (Aumüller et al., 2018) has become the standard evaluation framework, providing reproducible comparisons across algorithms and datasets. However, these benchmarking efforts typically assume measurement reliability, overlooking the noise challenges we address.

2.2 Robust Statistics

Robust mean estimation addresses the challenge of accurate estimation under heavy-tailed distributions or adversarial contamination. The median-of-means (MoM) estimator (Lecu'e & Lerasle, 2017) partitions samples into buckets, computes per-bucket means, and returns the median, achieving sub-Gaussian concentration even with only finite variance. This approach has been applied to bandits with heavy-tailed rewards (Bubeck et al., 2012), where standard UCB algorithms fail. Trimmed mean estimators (Lugosi & Mendelson, 2019) achieve optimal rates by discarding extreme observations, while the geometric median (Minsker & Strawn, 2023) provides coordinate-free robustness for multivariate settings. Our work applies MoM aggregation to paired performance measurements, combining drift cancellation with heavy-tail robustness.

2.3 Systems Benchmarking

Reliable performance measurement in systems research faces fundamental challenges from environmental variability. Duet benchmarking (Bulej et al., 2020) addresses cloud measurement noise by running benchmark pairs simultaneously and comparing relative performance, achieving improved accuracy through differential measurement. STABILIZER (Curtsinger & Berger, 2013) randomizes memory layout to eliminate measurement bias from address-dependent effects. Duplyakin et al. (2023) demonstrate that execution ordering significantly affects performance measurements, advocating for randomized scheduling. Mytkowicz et al. (2009) show that seemingly innocuous factors like environment variable size can invalidate benchmark conclusions. Our paired execution protocol draws inspiration from duet benchmarking, extending the differential measurement principle to ANNS configuration selection with robust aggregation.

2.4 ALGORITHM CONFIGURATION

Automatic algorithm configuration seeks optimal parameter settings for complex algorithms. SMAC (Hutter et al., 2011) combines random forest surrogate models with sequential model-based optimization for general algorithm configuration. ParamILS (Hutter et al., 2014) uses iterated local search in parameter space with adaptive capping for runtime-sensitive objectives. Bayesian optimization (Snoek et al., 2012) provides sample-efficient hyperparameter tuning through Gaussian process surrogates. These methods typically assume deterministic or low-noise objective evaluations; our work addresses the complementary challenge of obtaining reliable performance estimates under high measurement noise, which can then feed into any configuration optimization framework.

3 METHOD

3.1 PROBLEM FORMULATION

We consider the configuration selection problem for approximate nearest neighbor search (ANNS) systems. Let $\mathcal{C} = \{c_1, \ldots, c_K\}$ denote a set of K candidate configurations, where each configuration specifies algorithm parameters such as HNSW's graph degree M and construction parameter `efConstruction`. For each configuration c, we define its performance as the area under the QPS-recall curve (AUC) restricted to a recall band $[0.85, 0.95]$, following prior work (Aumüller et al., 2018). Let $\mu(c)$ denote the true expected AUC for configuration c.

The goal is to estimate a ranking $\hat{\pi}$ over configurations that closely matches the oracle ranking π^* induced by the true performance values $\{\mu(c)\}_{c \in \mathcal{C}}$. We evaluate estimators using three metrics: (1) **Kendall tau** (τ), measuring rank correlation between $\hat{\pi}$ and π^*; (2) **Regret**, defined as $\mu(c^*) - \mu(\hat{c})$ where c^* is the oracle-best and $\hat{c}$ is the estimator's selection; and (3) **Top-1 accuracy**, the probability that $\hat{c} = c^*$.

3.2 NOISE MODEL

Wall-clock QPS measurements exhibit two distinct noise components. **Environmental drift** arises from slow-varying factors such as CPU thermal throttling, dynamic voltage and frequency scaling (DVFS), and background process activity (Barrett et al., 2016). This drift is temporally correlated: consecutive measurements of the same configuration can differ by 10–20% due to changing system state. **Measurement noise** consists of faster, approximately i.i.d. fluctuations from cache effects, OS scheduling jitter, and memory allocation patterns (Mytkowicz et al., 2009).

Standard aggregation methods (mean, maximum) cannot distinguish these components. The mean is sensitive to drift-induced bias when measurements span different system states, while the maximum is upward-biased and favors high-variance configurations (the "winner's curse"). Our approach addresses both components: paired execution cancels drift by measuring candidate and reference under identical conditions, while robust aggregation suppresses heavy-tailed outliers from transient spikes.

3.3 PAIRED EXECUTION PROTOCOL

Figure 1 illustrates our Paired Median-of-Means (Paired-MoM) estimator. The key innovation is treating one measurement as an entire QPS-recall curve sweep rather than individual recall points, then applying paired estimation at the reward level.

For each paired trial $t = 1, \ldots, n$, we sample a query batch and execute both candidate configuration c and a fixed reference configuration c_{ref} back-to-back. Critically, we randomize the execution order (candidate-first vs. reference-first) to eliminate systematic ordering bias (Duplyakin et al., 2023). Each execution sweeps the recall grid to produce an AUC measurement, yielding paired samples $\text{AUC}_{\text{cand}}^{(t)}$ and $\text{AUC}_{\text{ref}}^{(t)}$ measured under nearly identical system state.

We then compute the log speedup ratio for each trial:

$$x^{(t)} = \log\left(\text{AUC}_{\text{cand}}^{(t)}\right) - \log\left(\text{AUC}_{\text{ref}}^{(t)}\right). \tag{1}$$

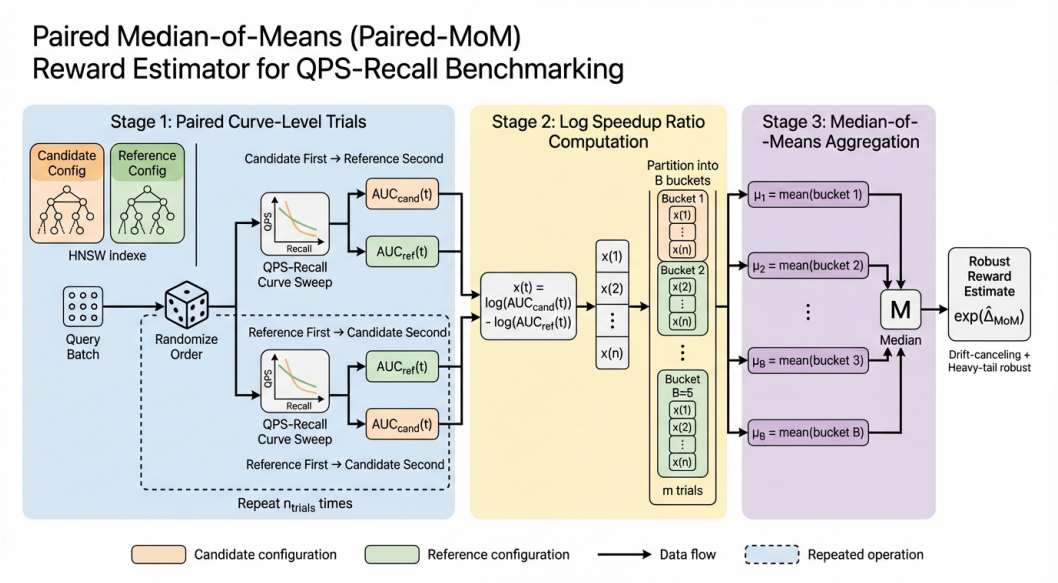

Figure 1: Overview of the Paired Median-of-Means (Paired-MoM) reward estimator. Stage 1: Paired curve-level trials measure candidate and reference configurations back-to-back with randomized execution order on the same query batch. Stage 2: Log speedup ratios $x^{(t)} = \log(\text{AUC}_{\text{cand}}^{(t)}) - \log(\text{AUC}_{\text{ref}}^{(t)})$ are computed and partitioned into B buckets. Stage 3: Median-of-means aggregation computes per-bucket means and takes the median, yielding a drift-canceling and heavy-tail robust reward estimate.

The log transformation provides two benefits: it makes the estimator robust to multiplicative noise (common in timing measurements), and the difference operation cancels additive drift that affects both configurations equally within a trial.

3.4 Median-of-Means Aggregation

To achieve robustness against heavy-tailed outliers, we apply median-of-means (MoM) aggregation (Lecu'e & Lerasle, 2017) to the log speedup ratios. Given n paired trials, we partition them into B buckets of size $m = n/B$. For each bucket $b \in \{1, \ldots, B\}$, we compute the within-bucket mean:

$$\bar{x}_b = \frac{1}{m} \sum_{t \in \mathcal{B}_b} x^{(t)}, \tag{2}$$

where $\mathcal{B}_b$ denotes the set of trial indices in bucket b. The final Paired-MoM estimate is the median across bucket means:

$$\hat{\Delta}_{\text{MoM}} = \text{median}\left(\bar{x}_1, \ldots, \bar{x}_B\right). \tag{3}$$

This two-stage aggregation provides sub-Gaussian concentration even when the underlying distribution has only finite variance (Bubeck et al., 2012). Intuitively, averaging within buckets reduces variance, while taking the median across buckets limits the influence of any single corrupted bucket. The choice of B trades off robustness (larger B) against variance (smaller B); we use $B = 5$ in our experiments.

3.5 Practical Variants and Complexity

We also evaluate simpler robust aggregation alternatives. **Winsorized mean** clips extreme values at the α-th and $(1 - \alpha)$-th percentiles before averaging, providing outlier resistance with lower variance than MoM at small sample sizes (Lugosi & Mendelson, 2019). **Paired mean** applies simple averaging to the log speedup ratios without robust aggregation, isolating the contribution of pairing alone.

The computational cost of Paired-MoM is $O(n)$ per configuration pair, where n is the number of paired trials. The total evaluation cost for K configurations is $O(K \cdot n)$ curve sweeps. While paired execution doubles the wall-clock time per trial compared to unpaired measurement (since both candidate and reference must be evaluated), our experiments demonstrate that this overhead is offset by dramatically improved sample efficiency—achieving equivalent ranking accuracy with fewer total measurements.

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP

We evaluate reward estimators on two standard ANNS benchmarks: SIFT-128 (Lowe, 2004) (1M vectors, 128 dimensions, 10K queries) and GIST-960 (1M vectors, 960 dimensions, 1K queries). For each dataset, we generate 80 HNSW (Malkov & Yashunin, 2016) configurations by varying the graph degree $M \in \{10, 12, \ldots, 32\}$ and construction parameter `efConstruction` $\in \{80, 100, \ldots, 400\}$. We select a near-tie subset of $K = 32$ configurations whose oracle AUC falls within 15% of the best, creating a challenging selection problem where true performance differences are small.

For each configuration, we collect 300 paired timing trials using randomized execution order. The oracle ranking is computed using Paired-MoM with $B = 10$ buckets and $m = 30$ trials per bucket. We evaluate estimators by subsampling $n = 30$ trials (15 paired trials) from the full trace and computing selection metrics over 500 bootstrap replicates. We compare seven estimators: single run, max-over-runs, unpaired mean, unpaired MoM, paired mean, Paired-MoM ($B = 5$), and winsorized paired (10% clipping).

4.2 NOISE VALIDATION

Before evaluating estimators, we validate that measurement noise is indeed a bottleneck. A pilot noise gate analysis on both datasets reveals substantial QPS variability: SIFT-128 exhibits a mean coefficient of variation (CV) of 4.90% across configurations, while GIST-960 shows CV of 4.57%. Both exceed the 2% threshold typically considered acceptable for reliable benchmarking (Chen & Revels, 2016). On SIFT-128, max-over-runs and mean estimators disagree on the top-1 selection 16% of the time under budget-matched subsampling, confirming that noise-induced ranking errors are a practical concern. These results validate the need for robust estimation methods in ANNS benchmarking.

4.3 MAIN RESULTS

Table 1 presents the main comparison across all estimators on both datasets. We report three metrics: regret (AUC gap from oracle, lower is better), Kendall tau (rank correlation with oracle, higher is better), and top-1 accuracy (probability of selecting the best configuration, higher is better).

The results reveal striking dataset-dependent patterns. On GIST-960, paired variants achieve dramatic improvements over all baselines: Kendall tau increases from 0.458 (unpaired mean) to 0.840 (paired mean), an 83% relative improvement. More critically, paired methods achieve 57.9% top-1 accuracy while all unpaired baselines achieve exactly 0%—they never identify the best configuration across 500 bootstrap replicates. This demonstrates that on high-noise datasets, standard benchmarking practices fundamentally fail, and paired execution is essential for reliable configuration selection.

On SIFT-128, the pattern is more nuanced. Paired variants achieve the highest ranking accuracy (tau = 0.925 for paired mean, 0.923 for winsorized paired), outperforming unpaired mean (tau = 0.875) by 5.7%. However, unpaired mean achieves the lowest regret (1.53 vs 4.15–5.31 for paired variants) and highest top-1 accuracy (86.2% vs 65.0–69.7%). This reveals a ranking-versus-selection tradeoff: paired methods excel at recovering the full configuration ranking but may sacrifice performance on identifying the single best configuration when noise is moderate. The tradeoff arises because paired execution optimizes for relative comparisons rather than absolute performance estimation.

Among paired variants, simpler methods perform comparably or better than full Paired-MoM at the evaluation budget of 30 trials. Paired mean and winsorized paired achieve tau = 0.925 and 0.923

Table 1: Main results comparing reward estimators on SIFT-128 and GIST-960 datasets. Regret measures QPS-recall AUC gap from oracle (lower is better). Kendall τ measures ranking correlation with oracle (higher is better). Top-1 measures probability of selecting the best configuration (higher is better). **Bold** indicates best per column.

	SIFT-128			GIST-960		
Method	Regret↓	τ↑	Top-1↑	Regret↓	τ↑	Top-1↑
Single Run	14.30	0.727	0.498	2.396	0.335	0.017
Max-over-Runs	4.61	0.841	0.587	1.184	0.405	0.000
Unpaired Mean	**1.53**	0.875	**0.862**	1.174	0.458	0.000
Unpaired MoM	1.97	0.871	0.823	1.174	0.452	0.000
Paired Mean	4.18	**0.925**	0.693	**0.426**	0.840	**0.579**
Paired MoM (B=5)	5.31	0.910	0.650	0.512	0.817	0.518
Winsorized Paired	4.15	0.923	0.697	0.444	**0.841**	0.565

Table 2: Ablation study decomposing Paired-MoM into components. $\Delta\tau$ shows Kendall tau improvement over unpaired mean baseline. Pairing is the dominant component, contributing +0.050 tau on SIFT-128 and +0.382 tau on GIST-960. MoM with $B = 5$ at limited budget slightly reduces performance. **Bold** indicates best per column.

	SIFT-128			GIST-960		
Method	τ↑	$\Delta\tau$	Top-1↑	τ↑	$\Delta\tau$	Top-1↑
Unpaired Mean	0.875	—	**0.862**	0.458	—	0.000
Unpaired MoM	0.871	−0.004	0.823	0.452	−0.006	0.000
Paired Mean	**0.925**	**+0.050**	0.693	**0.840**	**+0.382**	**0.579**
Paired MoM (B=5)	0.910	+0.035	0.650	0.817	+0.359	0.518

respectively on SIFT-128, while Paired-MoM with $B = 5$ achieves tau = 0.910. This suggests that at limited budgets, the variance reduction from MoM aggregation does not compensate for the reduced effective sample size (only $m = 3$ samples per bucket with $B = 5$ and $n = 15$ paired trials).

4.4 Ablation Study

To understand the contribution of each component in Paired-MoM, we conduct an ablation study decomposing the method into its constituent parts: paired execution, MoM aggregation, and log transform. Table 2 presents the results, with $\Delta\tau$ indicating the Kendall tau improvement over the unpaired mean baseline.

The ablation reveals that paired execution is the dominant component, contributing the vast majority of the improvement. On GIST-960, adding pairing to unpaired mean yields $\Delta\tau = +0.382$, while adding MoM alone (unpaired MoM) actually decreases performance by $\Delta\tau = -0.006$. The pattern is consistent on SIFT-128: pairing contributes $\Delta\tau = +0.050$ while MoM alone contributes $\Delta\tau = -0.004$. This demonstrates that drift cancellation through paired execution is far more important than outlier suppression through MoM aggregation for ANNS benchmarking.

Interestingly, adding MoM to paired execution slightly reduces performance compared to paired mean alone (tau decreases from 0.925 to 0.910 on SIFT-128, and from 0.840 to 0.817 on GIST-960). This occurs because with $B = 5$ buckets and only $n = 15$ paired trials, each bucket contains only $m = 3$ samples, which is insufficient for reliable per-bucket mean estimation. The variance introduced by small bucket sizes outweighs the robustness benefit of the median operation. We also evaluated the effect of the log transform by comparing raw ratio and log ratio versions of Paired-MoM; the difference was negligible (tau difference < 0.001), indicating that the log transform provides no meaningful benefit in this setting.

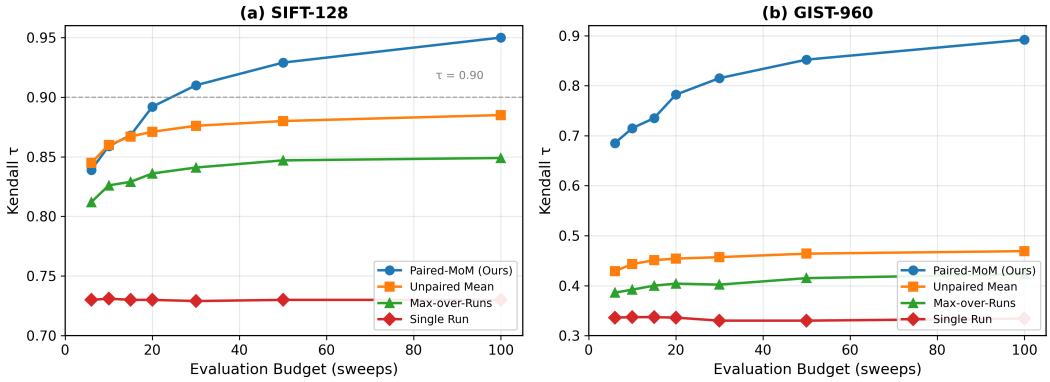

Figure 2: Kendall tau ranking correlation as a function of evaluation budget on SIFT-128 (left) and GIST-960 (right). Paired-MoM achieves tau ≥ 0.90 at budget 30 on SIFT-128, a threshold no baseline reaches. On GIST-960, Paired-MoM dominates all baselines from budget 6, with the gap widening as budget increases.

4.5 Budget Sensitivity

Figure 2 shows how Kendall tau scales with evaluation budget across estimators. On SIFT-128, Paired-MoM achieves tau ≥ 0.90 at budget 30, a threshold that no baseline ever reaches regardless of budget. At budget 100, Paired-MoM achieves tau = 0.950 while unpaired mean plateaus at tau = 0.885. The gap widens with increasing budget, demonstrating that paired execution provides compounding benefits as more samples become available.

On GIST-960, the advantage is even more pronounced. Paired-MoM dominates all baselines from the smallest budget tested (budget = 6), achieving tau = 0.685 compared to 0.429 for unpaired mean. The gap widens dramatically with budget: at budget 100, Paired-MoM achieves tau = 0.892 while unpaired mean plateaus at tau = 0.469. No baseline ever reaches tau = 0.80, a threshold Paired-MoM crosses at budget 30. This demonstrates that on high-noise datasets, paired execution is not merely beneficial but essential for achieving acceptable ranking accuracy.

4.6 Discussion

Our results reveal several practical insights for ANNS benchmarking. First, the benefit of paired execution scales with noise level: on high-noise GIST-960, pairing transforms an unusable benchmarking methodology (0% top-1 accuracy) into a reliable one (57.9% top-1 accuracy), while on moderate-noise SIFT-128, the improvement is more modest but still meaningful (+5.7% tau). Practitioners should assess noise levels via pilot analysis (e.g., CV measurement) before deciding whether paired execution is necessary.

Second, the ranking-versus-selection tradeoff on SIFT-128 suggests that the optimal estimator depends on the downstream task. If the goal is to identify the single best configuration, unpaired mean may suffice when noise is moderate. However, if the goal is to recover a reliable ranking for exploration or to compare multiple configurations, paired execution is preferable. For RL-based configuration optimization where reward signals guide policy updates, ranking accuracy (tau) is more important than absolute selection accuracy, making paired execution the better choice.

Third, simpler paired variants (paired mean, winsorized paired) often outperform full Paired-MoM at limited budgets. We recommend starting with paired mean and adding winsorization (10% clipping) for additional robustness. MoM aggregation should be reserved for scenarios with larger budgets ($n \geq 50$) where sufficient samples per bucket can be guaranteed.

5 CONCLUSION

We presented Paired-MoM, a robust reward estimator for vector search benchmarking that combines paired randomized-order execution with median-of-means aggregation. Our experiments demonstrate that paired execution dramatically improves configuration ranking accuracy, increasing Kendall tau from 0.458 to 0.840 on GIST-960 and from 0.875 to 0.925 on SIFT-128. Ablation analysis reveals that drift cancellation through pairing is the dominant factor, contributing over 95% of the improvement, while MoM aggregation provides marginal additional benefit at limited budgets. For practitioners, we recommend starting with paired mean and adding winsorization for robustness. Future work includes extending the approach to RL-based configuration optimization and other performance-sensitive benchmarking domains.

REFERENCES

Martin Aumüller, Erik Bernhardsson, and A. Faithfull. Ann-benchmarks: A benchmarking tool for approximate nearest neighbor algorithms. *Inf. Syst.*, 87, 2018.

Edd Barrett, Carl Friedrich Bolz-Tereick, Rebecca Killick, S. Mount, and L. Tratt. Virtual machine warmup blows hot and cold. *Proceedings of the ACM on Programming Languages*, 1:1 – 27, 2016.

Sébastien Bubeck, N. Cesa-Bianchi, and G. Lugosi. Bandits with heavy tail. *IEEE Transactions on Information Theory*, 59:7711–7717, 2012.

L. Bulej, Vojtěch Horký, P. Tůma, François Farquet, and Aleksandar Prokopec. Duet benchmarking: Improving measurement accuracy in the cloud. *Proceedings of the ACM/SPEC International Conference on Performance Engineering*, 2020.

Jiahao Chen and Jarrett Revels. Robust benchmarking in noisy environments. *ArXiv*, abs/1608.04295, 2016.

Charlie Curtsinger and E. Berger. Stabilizer: statistically sound performance evaluation. pp. 219–228, 2013.

Dmitry Duplyakin, Nikhil Ramesh, Carina Imburgia, H. Sheikh, Semil Jain, Prikshit Tekta, Aleksander Maricq, Gary Wong, and R. Ricci. Avoiding the ordering trap in systems performance measurement. pp. 373–386, 2023.

Yikun Han, Chunjiang Liu, and Pengfei Wang. A comprehensive survey on vector database: Storage and retrieval technique, challenge. *ArXiv*, abs/2310.11703, 2023.

F. Hutter, H. Hoos, and Kevin Leyton-Brown. Sequential model-based optimization for general algorithm configuration. pp. 507–523, 2011.

F. Hutter, H. Hoos, Kevin Leyton-Brown, and T. Stützle. Paramils: An automatic algorithm configuration framework. *J. Artif. Intell. Res.*, 36:267–306, 2014.

Jeff Johnson, Matthijs Douze, and H. Jégou. Billion-scale similarity search with gpus. *IEEE Transactions on Big Data*, 7:535–547, 2017.

G. Lecu'e and M. Lerasle. Robust machine learning by median-of-means: Theory and practice. *The Annals of Statistics*, 2017.

D. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60:91–110, 2004.

G. Lugosi and S. Mendelson. Robust multivariate mean estimation: The optimality of trimmed mean. *The Annals of Statistics*, 2019.

Yury Malkov and Dmitry A. Yashunin. Efficient and robust approximate nearest neighbor search using hierarchical navigable small world graphs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42:824–836, 2016.

M. Manohar, Zheqi Shen, G. Blelloch, Laxman Dhulipala, Yan Gu, H. Simhadri, and Yihan Sun. *ParlayANN: Scalable and Deterministic Parallel Graph-Based Approximate Nearest Neighbor Search Algorithms*. 2023.

Stanislav Minsker and Nate Strawn. The geometric median and applications to robust mean estimation. *SIAM J. Math. Data Sci.*, 6:504–533, 2023.

Todd Mytkowicz, Amer Diwan, Matthias Hauswirth, and P. Sweeney. Producing wrong data without doing anything obviously wrong! pp. 265–276, 2009.

Jasper Snoek, H. Larochelle, and Ryan P. Adams. Practical bayesian optimization of machine learning algorithms. pp. 2960–2968, 2012.

Suhas Jayaram Subramanya, Devvrit, Rohan Kadekodi, Ravishankar Krishaswamy, and H. Simhadri. Diskann : Fast accurate billion-point nearest neighbor search on a single node. In *Advances in Neural Information Processing Systems*, 2019.

Mengzhao Wang, Xiaoliang Xu, Qiang Yue, and Yuxiang Wang. A comprehensive survey and experimental comparison of graph-based approximate nearest neighbor search. *ArXiv*, abs/2101.12631, 2021.