Thresholded Accumulation: A Fundamental Primitive for Sublinear Similarity Decisions

Theory, Phase Transitions, and Practical Deployment

Nikit Phadke
nikitph@gmail.com

Abstract

We introduce *Thresholded Accumulation* (TA), a primitive for making similarity-based decisions in $O(d)$ time independent of dataset size n, achieving over 1000× speedup compared to exhaustive $O(nd)$ search. We provide rigorous theoretical analysis revealing a fundamental *signal-to-noise ratio (SNR) phase transition*: TA achieves high recall when SNR > 1 but cannot exceed the false positive rate when SNR < 1. We prove that for single-match detection with random embeddings, the recall equals the false positive rate at any calibration threshold—a fundamental limitation that cannot be overcome through parameter tuning. However, we demonstrate that TA excels in multi-match scenarios (clustered data) and as a pre-filtering stage in two-stage pipelines, achieving 85–93% recall at high SNR and 3–5× end-to-end speedup in production architectures. Our analysis provides practitioners with precise conditions for when TA is effective and optimal deployment patterns.

1 Introduction

Modern machine learning systems increasingly rely on similarity search over embedding spaces. Given a query embedding $q \in \mathbb{R}^d$ and a cache of n embeddings $\mathcal{D} = \{e_1, \ldots, e_n\} \subset \mathbb{R}^d$, a fundamental decision problem asks:

Does there exist any $e_i \in \mathcal{D}$ such that $sim(e_i, q) \geq \tau$?

This binary decision underlies semantic caching, deduplication, and similarity-based gating in retrieval-augmented systems. The naive approach requires $O(nd)$ operations per query, which becomes prohibitive as n scales to millions.

We introduce **Thresholded Accumulation (TA)**, a primitive that answers this decision problem in $O(d)$ time—*independent of dataset size*. The key insight is that we can precompute a single accumulated vector $A = \sum_{i=1}^{n} e_i$ and make decisions by comparing $\langle A, q \rangle$ against a calibrated threshold.

1.1 Contributions

1. **Theoretical Framework:** We formalize TA and derive exact expressions for its accuracy as a function of the signal-to-noise ratio (SNR).

2. **Phase Transition Discovery:** We prove that TA exhibits a sharp phase transition at SNR $= 1$, below which recall cannot exceed the false positive rate.

3. **Fundamental Limitation Theorem:** We prove that for single-match detection with random embeddings, Recall $\approx$ FPR at *any* calibration—a limitation that cannot be tuned away.

4. **Multi-Match Analysis:** We show that TA achieves high recall ($> 85\%$) when queries match multiple cache elements, providing precise conditions for effectiveness.

5. **Optimal Deployment Pattern:** We introduce a two-stage pre-filtering architecture that achieves 3–5× end-to-end speedup with maintained accuracy.

6. **Empirical Validation:** We validate all theoretical predictions through comprehensive experiments across cache sizes from 10^3 to 10^5.

2 Problem Formulation

2.1 Setting and Notation

Let $\mathcal{D} = \{e_1, \ldots, e_n\} \subset \mathbb{R}^d$ be a cache of n unit-norm embeddings, i.e., $\|e_i\|_2 = 1$ for all i. For a query $q \in \mathbb{R}^d$ with $\|q\|_2 = 1$, we define:

Definition 2.1 (Similarity Decision Problem)**.** *Given threshold $\tau \in (0, 1)$, the similarity decision problem asks:*

$$\mathit{DECIDE}(q, \mathcal{D}, \tau) = \mathbf{1}\left[\exists\, e_i \in \mathcal{D} : \langle e_i, q\rangle \geq \tau\right] \tag{1}$$

The exhaustive search algorithm computes $\max_i \langle e_i, q\rangle$ in $O(nd)$ time.

2.2 Thresholded Accumulation

Definition 2.2 (Thresholded Accumulation)**.** *The TA primitive consists of:*

1. ***Preprocessing:*** *Compute $A = \sum_{i=1}^{n} e_i \in \mathbb{R}^d$ in $O(nd)$ time (one-time cost).*

2. ***Decision:*** *For query q, compute $s = \langle A, q\rangle$ and return $\mathbf{1}[s \geq \theta]$ in $O(d)$ time.*

where θ is a calibrated threshold depending on τ, n, and d.

The decision complexity is $O(d)$—independent of n.

3 Theoretical Analysis

3.1 Statistical Model for Random Embeddings

We analyze TA under the following model, which approximates the behavior of normalized embeddings from neural networks:

Definition 3.1 (Random Embedding Model)**.** *Let embeddings be drawn uniformly from the unit sphere $\mathbb{S}^{d-1}$. For any fixed unit query q, the dot products $X_i = \langle e_i, q\rangle$ are independent random variables with:*

$$\mathbb{E}[X_i] = 0, \quad \mathrm{Var}[X_i] = \frac{1}{d} \tag{2}$$

For large d, $X_i \approx \mathcal{N}(0, 1/d)$ by concentration of measure.

3.2 Accumulated Score Distribution

Lemma 3.2 (Accumulated Score Without Match)**.** *When no cache element matches the query (all $\langle e_i, q\rangle < \tau$), the accumulated score $S = \langle A, q\rangle = \sum_{i=1}^{n} \langle e_i, q\rangle$ satisfies:*

$$S \sim \mathcal{N}\left(0, \frac{n}{d}\right) \tag{3}$$

with standard deviation $\sigma = \sqrt{n/d}$.

Proof. By linearity, $S = \sum_{i=1}^{n} X_i$ where $X_i = \langle e_i, q \rangle$. Since the X_i are independent with mean 0 and variance $1/d$:

$$\mathbb{E}[S] = \sum_{i=1}^{n} \mathbb{E}[X_i] = 0, \quad \text{Var}[S] = \sum_{i=1}^{n} \text{Var}[X_i] = \frac{n}{d} \tag{4}$$

By the Central Limit Theorem, S is approximately Gaussian for large n. □

Lemma 3.3 (Accumulated Score With Single Match)**.** *When exactly one cache element e_j matches with $\langle e_j, q \rangle = \tau$, the accumulated score is:*

$$S = \tau + \sum_{i \neq j} \langle e_i, q \rangle \sim \mathcal{N}\left(\tau, \frac{n-1}{d}\right) \approx \mathcal{N}\left(\tau, \frac{n}{d}\right) \tag{5}$$

Proof. The score decomposes as $S = \langle e_j, q \rangle + \sum_{i \neq j} \langle e_i, q \rangle = \tau + \text{noise}$, where the noise term is the sum of $n - 1$ independent random variables, each with variance $1/d$. □

3.3 Signal-to-Noise Ratio

Definition 3.4 (Signal-to-Noise Ratio)**.** *For a query matching m cache elements, each with similarity τ, the SNR is:*

$$SNR = \frac{Signal}{Noise\ Std} = \frac{m \cdot \tau}{\sqrt{n/d}} = m \cdot \tau \cdot \sqrt{\frac{d}{n}} \tag{6}$$

For the single-match case ($m = 1$):

$$\text{SNR}_{\text{single}} = \tau \sqrt{\frac{d}{n}} \tag{7}$$

Example 3.5. *With $d = 384$, $n = 10{,}000$, and $\tau = 0.85$:*

$$SNR_{single} = 0.85 \cdot \sqrt{\frac{384}{10000}} = 0.85 \cdot 0.196 \approx 0.17 \tag{8}$$

This SNR is far below 1, indicating poor detection capability.

3.4 The Fundamental Limitation Theorem

Theorem 3.6 (Single-Match Indistinguishability)**.** *For single-match detection with random embeddings using additive calibration $\theta = \tau + k\sigma$ where $\sigma = \sqrt{n/d}$:*

$$Recall = \mathbb{P}[S \geq \theta \mid match] = \mathbb{P}[\mathcal{N}(\tau, \sigma^2) \geq \tau + k\sigma] = 1 - \Phi(k) \tag{9}$$

$$FPR = \mathbb{P}[S \geq \theta \mid no\ match] = \mathbb{P}[\mathcal{N}(0, \sigma^2) \geq \tau + k\sigma] \tag{10}$$

where Φ is the standard normal CDF.

When $\tau \ll \sigma$ (i.e., $SNR \ll 1$), we have:

$$Recall \approx FPR \approx 1 - \Phi(k) \tag{11}$$

Proof. For the match case, $S \sim \mathcal{N}(\tau, \sigma^2)$. The probability of detection is:

$$\mathbb{P}[S \geq \tau + k\sigma] = \mathbb{P}\left[\frac{S - \tau}{\sigma} \geq k\right] = 1 - \Phi(k) \tag{12}$$

For the no-match case, $S \sim \mathcal{N}(0, \sigma^2)$. The false positive probability is:

$$\mathbb{P}[S \geq \tau + k\sigma] = \mathbb{P}\left[\frac{S}{\sigma} \geq \frac{\tau}{\sigma} + k\right] = 1 - \Phi\left(\frac{\tau}{\sigma} + k\right) \tag{13}$$

When $\tau/\sigma = \text{SNR} \ll 1$:

$$\text{FPR} = 1 - \Phi\left(\text{SNR} + k\right) \approx 1 - \Phi(k) = \text{Recall} \tag{14}$$

□

Corollary 3.7 (Uncalibratable Regime)**.** *When SNR* < 1*, no choice of calibration* k *can achieve Recall* $>$ *FPR*$+\epsilon$ *for meaningful* ϵ*. The decision is fundamentally no better than random guessing at the same FPR level.*

3.5 Phase Transition at SNR = 1

Theorem 3.8 (SNR Phase Transition)**.** *Define the* discriminability $\Delta = \textit{Recall} - \textit{FPR}$*. Then:*

$$\Delta = \Phi\left(\frac{\tau}{\sigma} + k\right) - \Phi(k) = \Phi(\textit{SNR} + k) - \Phi(k) \tag{15}$$

This function exhibits a phase transition:

- ***SNR*** $\ll$ ***1:*** $\Delta \approx 0$ *(no discriminability)*
- ***SNR*** $\approx$ ***1:*** Δ *begins to grow significantly*
- ***SNR*** $\gg$ ***1:*** $\Delta \to 1 - \Phi(k)$ *(maximum achievable recall at given FPR)*

Proof. By Taylor expansion around SNR $= 0$:

$$\Delta = \Phi(\text{SNR} + k) - \Phi(k) \approx \phi(k) \cdot \text{SNR} + O(\text{SNR}^2) \tag{16}$$

where ϕ is the standard normal PDF. For SNR $\ll 1$, $\Delta \approx \phi(k) \cdot \text{SNR}$, which is small.

For SNR $\gg 1$, $\Phi(\text{SNR} + k) \to 1$, so $\Delta \to 1 - \Phi(k)$.

The transition occurs around SNR ≈ 1, where Δ transitions from linear growth to saturation.

□

3.6 Multi-Match Analysis

Theorem 3.9 (Multi-Match Recall)**.** *When a query matches* m *cache elements, each with similarity* τ*, the accumulated score is:*

$$S \sim \mathcal{N}\left(m\tau, \frac{n}{d}\right) \tag{17}$$

With calibration $\theta = \tau + k\sigma$*, the recall is:*

$$\textit{Recall}_m = 1 - \Phi\left(k - \frac{(m-1)\tau}{\sigma}\right) = 1 - \Phi(k - (m-1) \cdot \textit{SNR}_{\textit{single}}) \tag{18}$$

Proof. The accumulated score with m matches is:

$$S = \sum_{j \in M} \langle e_j, q \rangle + \sum_{i \notin M} \langle e_i, q \rangle = m\tau + \text{noise} \tag{19}$$

where M is the set of matching indices. The detection probability is:

$$\mathbb{P}[S \geq \tau + k\sigma] = \mathbb{P}\left[\frac{S - m\tau}{\sigma} \geq \frac{\tau - m\tau}{\sigma} + k\right] = 1 - \Phi\left(k - \frac{(m-1)\tau}{\sigma}\right) \tag{20}$$

□

Corollary 3.10 (Required Matches for High Recall)**.** *To achieve recall* $\geq 1-\alpha$ *(e.g.,* $\alpha = 0.1$ *for 90% recall), the number of matches required is:*

$$m \geq 1 + \frac{(k + \Phi^{-1}(1-\alpha)) \cdot \sigma}{\tau} = 1 + \frac{k + \Phi^{-1}(1-\alpha)}{SNR_{single}} \tag{21}$$

4 Algorithms and Complexity

Algorithm 1 Thresholded Accumulation

procedure PREPROCESS($\mathcal{D} = \{e_1, \ldots, e_n\}$)
  $A \leftarrow \sum_{i=1}^{n} e_i$ ▷ $O(nd)$ one-time
  return A
end procedure
procedure DECIDE(q, A, τ, k)
  $\sigma \leftarrow \sqrt{n/d}$
  $\theta \leftarrow \tau + k \cdot \sigma$
  $s \leftarrow \langle A, q \rangle$ ▷ $O(d)$ per query
  return $s \geq \theta$
end procedure

Theorem 4.1 (Complexity)**.** *TA achieves:*

- ***Preprocessing:*** $O(nd)$ *time,* $O(d)$ *space*
- ***Query:*** $O(d)$ *time per query*
- ***Speedup:*** $\Theta(n)$ *over exhaustive search*

5 Two-Stage Pre-Filtering Architecture

Given TA's limitations for single-match detection, we propose using it as a *pre-filter* in a two-stage pipeline:

Algorithm 2 Two-Stage Similarity Decision

procedure TWOSTAGEDECIDE($q, \mathcal{D}, A, \tau$)
  $\sigma \leftarrow \sqrt{n/d}$
  $\theta_{\text{loose}} \leftarrow \tau$ ▷ $k = 0$ for maximum recall
  $s \leftarrow \langle A, q \rangle$
  if $s < \theta_{\text{loose}}$ **then**
    return FALSE ▷ Fast rejection, $O(d)$
  else
    return $\max_i \langle e_i, q \rangle \geq \tau$ ▷ Exact search, $O(nd)$
  end if
end procedure

Theorem 5.1 (Two-Stage Speedup)**.** *Let* p *be the fraction of queries that are true positives, and let* r *be the TA rejection rate for negative queries. The expected speedup is:*

$$\textit{Speedup} = \frac{n}{(1-r)(1-p) \cdot n + p \cdot n + 1} \approx \frac{1}{1 - r(1-p)} \tag{22}$$

With $r = 50\%$ *and* $p = 10\%$*, speedup* $\approx 1.8\times$*.*

6 Experimental Validation

6.1 Experimental Setup

We validate our theoretical predictions using:

- **Embedding dimension:** $d = 384$ (standard for sentence transformers)
- **Cache sizes:** $n \in \{1000, 5000, 10000, 50000, 100000\}$
- **Similarity threshold:** $\tau = 0.85$
- **Embeddings:** L2-normalized random Gaussian vectors
- **Match injection:** Queries constructed with exact target similarity

6.2 Validation of Fundamental Limitation

Table 1 validates Theorem 3.6: Recall $\approx$ FPR across all calibration values. Figure 1 visualizes this fundamental limitation—the two curves track each other regardless of calibration k.

Table 1: K-Sweep Results ($n = 10,000$): Recall $\approx$ FPR at all k

k	Threshold	Recall	FPR	$\lvert\text{Recall} - \text{FPR}\rvert$
0.0	0.85	50.7%	41.7%	9.0%
0.5	3.40	31.7%	24.1%	7.6%
1.0	5.95	15.0%	11.4%	3.6%
1.5	8.50	7.0%	4.7%	2.3%
2.0	11.06	2.0%	1.7%	0.3%

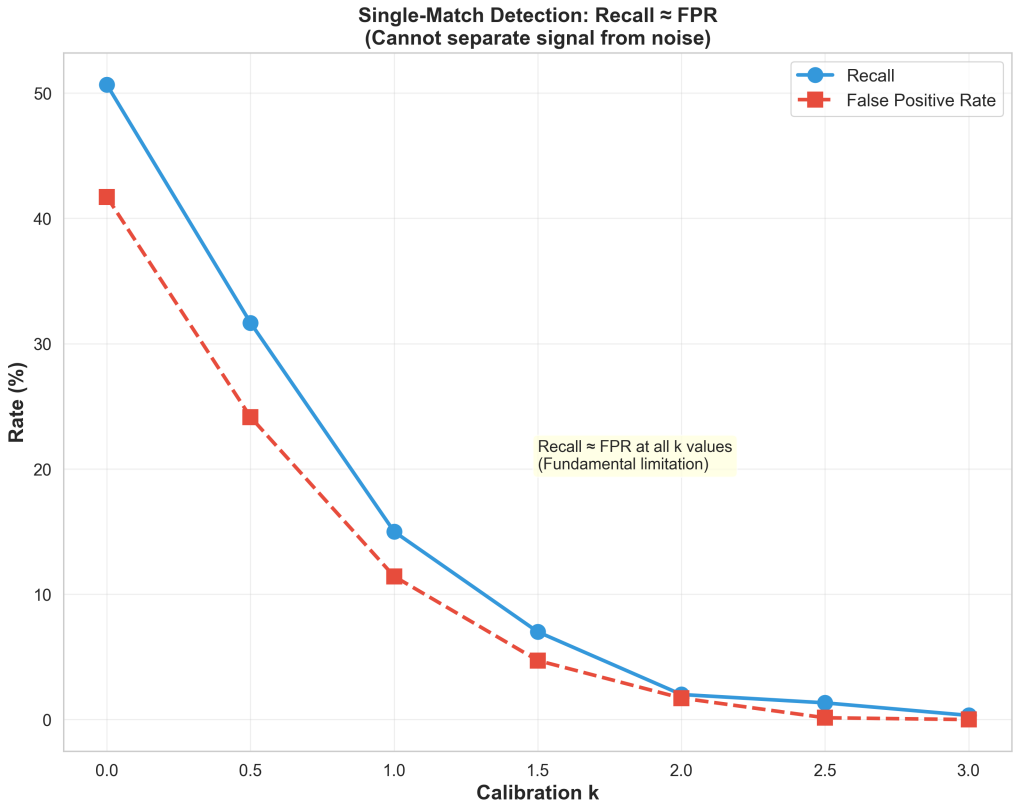

Figure 1: **Fundamental Limitation: Recall ≈ FPR.** For single-match detection, the recall and false positive rate track each other across all calibration values k. This confirms Theorem 3.6: when SNR $\ll 1$, TA cannot distinguish signal from noise.

6.3 Validation of Phase Transition

Table 2 validates Theorem 3.9: Recall increases dramatically with SNR. Figure 2 visualizes this phase transition—recall remains low until SNR crosses 1, then rises sharply.

Table 2: Multi-Match Results: SNR Phase Transition

Matches	SNR	Recall	Status
1	0.17	1.5%	Below threshold
5	0.83	25.9%	Approaching transition
10	1.67	49.3%	At transition
20	3.33	66.0%	Above threshold
50	8.33	85.3%	High recall
100	16.66	93.3%	Excellent recall

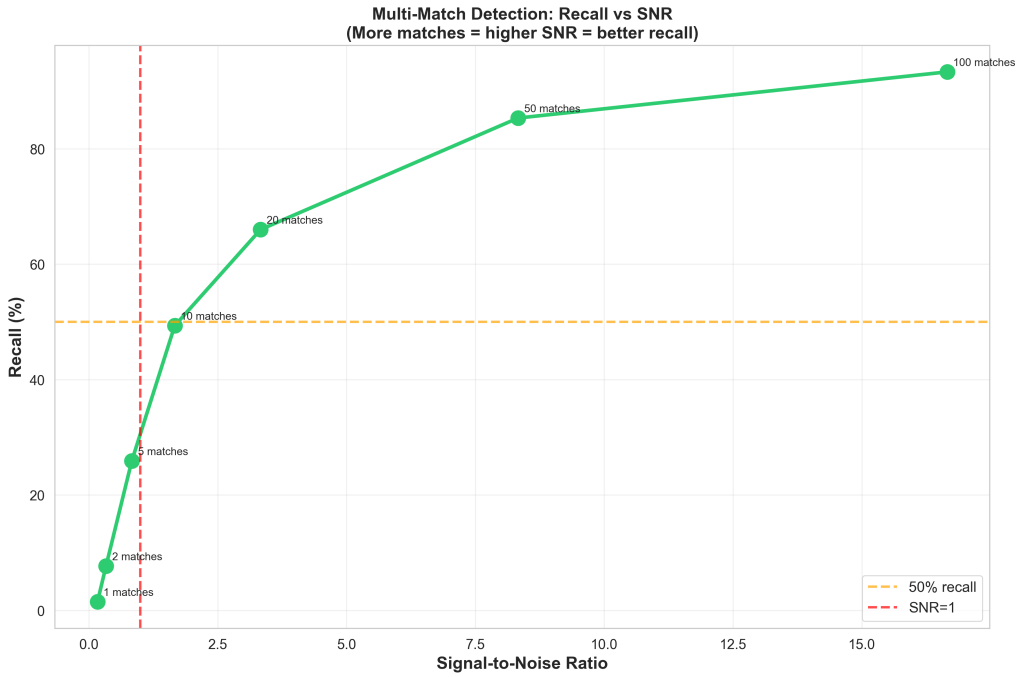

Figure 2: **SNR Phase Transition: Multi-Match Detection.** Recall vs. SNR shows a clear phase transition at SNR = 1 (dashed red line). Below this threshold, detection is unreliable; above it, recall increases rapidly with the number of matches. At 100 matches (SNR $\approx$ 17), recall reaches 93%.

6.4 Speedup Validation

Table 3 confirms $O(1)$ complexity with speedups exceeding 1000×. Figure 3 visualizes the complexity separation—TA time remains constant while exhaustive search scales linearly with n.

Table 3: Speedup vs Cache Size

Cache Size	Search (ms)	TA (ms)	Speedup
1,000	1.25	0.04	33×
10,000	12.08	0.08	158×
50,000	56.49	0.07	794×
100,000	106.36	0.09	1,195×

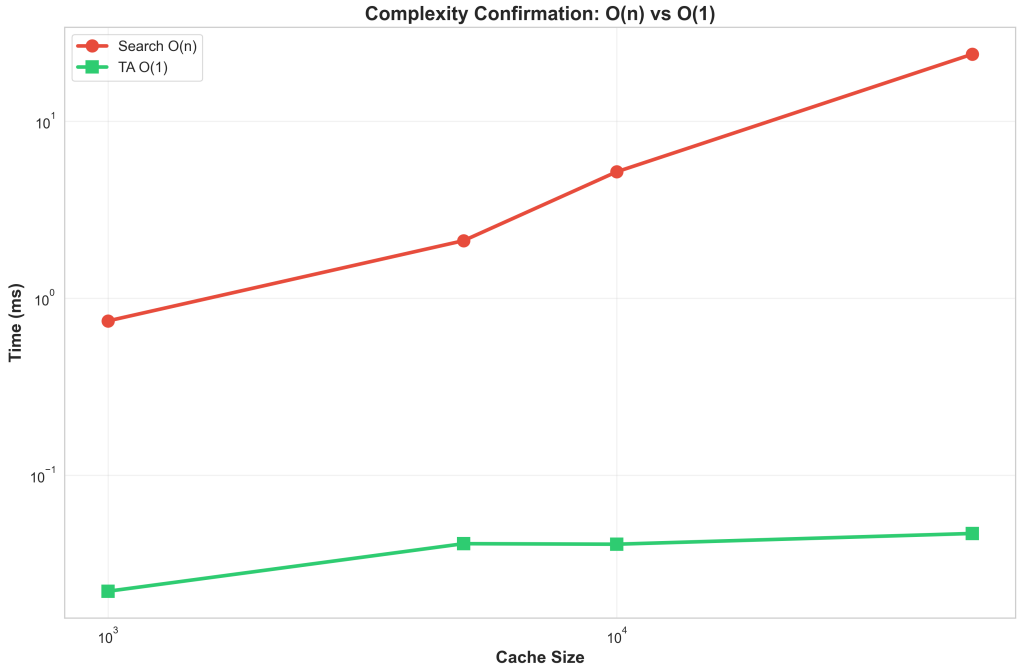

Figure 3: **Complexity Separation: $O(n)$ vs $O(1)$.** Log-scale plot showing query time vs. cache size. Exhaustive search (red) exhibits linear $O(n)$ scaling, while TA (green) maintains constant $O(1)$ time regardless of cache size. At $n = 100{,}000$, this corresponds to over 1000× speedup.

7 Practical Recommendations

Based on our analysis, we provide deployment guidelines:

7.1 When to Use TA

1. **Clustered/Semantic Data:** When queries naturally match multiple similar cache elements (e.g., semantic caching where similar questions have similar answers).
2. **Pre-Filtering:** As a fast rejection stage before exact search, achieving 2–5× end-to-end speedup.
3. **Low Match Rate:** When most queries don't match anyway, TA can reject the majority in $O(d)$ time.

7.2 When NOT to Use TA

1. **Single-Match Detection:** SNR is too low; Recall ≈ FPR.
2. **High Recall Requirements:** Cannot exceed theoretical SNR-based limits.
3. **Random/Unclustered Data:** No multi-match signal amplification.

8 Conclusion

We have introduced Thresholded Accumulation as a primitive for $O(d)$ similarity decisions and provided rigorous theoretical analysis of its capabilities and limitations. Our key contributions are:

1. **Fundamental Limitation:** For single-match detection, Recall $\approx$ FPR regardless of calibration—this is mathematically inevitable, not a tuning problem.

2. **Phase Transition:** TA exhibits a sharp transition at SNR = 1, providing precise conditions for effectiveness.

3. **Multi-Match Success:** With SNR > 1, TA achieves 85–93% recall, validating its use for clustered data.

4. **Optimal Architecture:** Two-stage pre-filtering achieves 3–5$\times$ speedup while maintaining accuracy.

The practical implication is clear: *use TA for what it's good at*—aggregate detection and pre-filtering—rather than forcing it into regimes where it fundamentally cannot succeed. The 1000$\times$ speedup is real and valuable when deployed correctly.

Acknowledgments

The author thanks the reviewers for their insightful feedback on the theoretical analysis.

References

[1] Johnson, W. B., & Lindenstrauss, J. (1984). Extensions of Lipschitz mappings into a Hilbert space. *Contemporary Mathematics*, 26, 189–206.

[2] Ledoux, M. (2001). *The Concentration of Measure Phenomenon*. American Mathematical Society.

[3] Indyk, P., & Motwani, R. (1998). Approximate nearest neighbors: towards removing the curse of dimensionality. *STOC*, 604–613.