Research Note: Nested Similarity-Distance-Magnitude Estimators

Allen Schmaltz
Reexpress AI
allen@re.express

August 2026

Abstract

In this research note, we provide an algorithm and reference implementation for constructing nested SIMILARITY-DISTANCE-MAGNITUDE (SDM) estimators, excluding each higher probability region when running Alg. 1 (Schmaltz, 2026) on descending values of α. The resulting quantities can be useful, for example, to rank lower probability predictions falling outside the HIGH-RELIABILITY region (estimated to have class- and prediction-conditional accuracy at least some value near 1) for further adjudication.

1 Introduction

The work of Schmaltz (2026)[1] introduces an approach for uncertainty quantification over neural networks. The proposed SIMILARITY-DISTANCE-MAGNITUDE (SDM) estimators use the geometry of the feature-representation space of the model against the training (or support) set to estimate a region of the distribution for which the class- and prediction-conditional accuracy is at least α, a value near 1 (e.g., $\alpha = 0.95$). This is referred to as the HIGH-RELIABILITY (HR) region and is intended for selective classification settings for which reliable, robust, and interpretable estimates in the high-probability region are paramount. SDM estimators seek informative (sharp) estimates over in-distribution data, while preferring conservative estimates rather than mis-calibration over covariate-shifted and out-of-distribution data. These advantageous properties come with the tradeoff that an SDM estimator calibrated at a given α provides no information in itself of the relative ranking of the probability of the instances outside the HR region.

For datapoints falling outside the HR region, one approach is to simply triage by $\text{SDM}(\boldsymbol{z}')_{\hat{y}}$ for routing to more resource-intensive processes, such as human review. A disadvantage of such relative ranking of the remaining instances is that it does not provide a constraint on the class-conditional accuracy analogous to the selection criteria of Eq. 9 (from the original work).

An alternative choice for ranking the instances outside the HR region is to simply run the algorithm that partitions the distribution (Alg. 1 in Schmaltz (2026)) for descending α values, successively excluding the datapoints of each higher region that is found (i.e., with a finite q'_{min}). Instances falling in regions partitioned by higher α values are preferable, other things being equal. This follows in a straightforward way from the original work. This brief note serves as background context for the code implementation of this approach, which we provide in the publicly available repo associated with the original paper. Absent additional domain-specific information, for typical language model applications of SDM estimators (e.g., for binary classification of instruction-following), we recommend this approach for ranking the datapoints outside the most conservative HR region.

2 Nested Similarity-Distance-Magnitude Estimators

We seek a relative ranking of the probability of the points outside the HR region, which is the region of the distribution estimated to have class- and prediction-conditional accuracy at least α.

Given a resolution $r \in (0, \frac{1}{2})$, we run Alg. 1 in Schmaltz (2026) for descending values of α spaced by r (starting at $1 - r$), successively removing the datapoints in the calibration set, $\mathcal{D}_{\text{ca}}$, for each region recorded with a finite q'_{min}. Each region is defined by its α, q'_{min}, and output thresholds, $[\psi_1, \ldots, \psi_C]$. The user specifies r, but not all descending α values may have a corresponding region. The procedure appears in Alg. 2.

[1] This brief note is intended to be read as an additional appendix to Schmaltz (2026) and is not a self-contained document. We refer the reader to the original work for additional details, including the notation conventions.

Algorithm 2 Descending α Search Algorithm to Estimate Nested Regions

Input: cached $(q', \text{SDM}(\boldsymbol{z}'))\ \forall\ \boldsymbol{x} \in \mathcal{D}_{\text{ca}}$, resolution $r \in (0, \frac{1}{2})$

procedure ESTIMATE-NESTED-REGIONS(cached $(q', \text{SDM}(\boldsymbol{z}'))\ \forall\ \boldsymbol{x} \in \mathcal{D}_{\text{ca}}$, $r \in (0, \frac{1}{2})$)
  $\mathcal{R} \leftarrow [\,]$ ▷ Ordered list of the $(\alpha, q'_{\min}, [\psi_1, \ldots, \psi_C])$ triples of the recorded regions
  $l \leftarrow 1$
  ${\mathcal{D}_{\text{ca}}}^{(l)} \leftarrow \mathcal{D}_{\text{ca}}$
  while $1 - r \cdot l > \frac{1}{2}$ **do** ▷ $\alpha_l \in (\frac{1}{2}, 1]$, as required by Alg. 1 (Schmaltz, 2026)
    $\alpha_l \leftarrow 1 - r \cdot l$
    ${q'_{\min}}^{(l)}, [\psi_1^{(l)}, \ldots, \psi_C^{(l)}] \leftarrow$ ESTIMATE-HIGH-RELIABILITY-REGION(cached $(q', \text{SDM}(\boldsymbol{z}'))\ \forall\ \boldsymbol{x} \in {\mathcal{D}_{\text{ca}}}^{(l)}, \alpha_l$) ▷ Alg. 1 (Schmaltz, 2026)
    if ${q'_{\min}}^{(l)} < \infty$ **then**
      Append $(\alpha_l, {q'_{\min}}^{(l)}, [\psi_1^{(l)}, \ldots, \psi_C^{(l)}])$ to $\mathcal{R}$
      $\mathcal{T}^{(l)} \leftarrow \{\boldsymbol{x} \in {\mathcal{D}_{\text{ca}}}^{(l)}$ s.t. $q' \geq {q'_{\min}}^{(l)} \wedge \{c : \text{SDM}(\boldsymbol{z}')_c \geq \psi_c^{(l)}\} = \{\hat{y}\}\}$ ▷ Region membership is determined via the selection criteria of Eq. 9 (Schmaltz, 2026), with $\hat{y} = \arg\max_c \boldsymbol{z}'_c$.
      ${\mathcal{D}_{\text{ca}}}^{(l+1)} \leftarrow {\mathcal{D}_{\text{ca}}}^{(l)} \setminus \mathcal{T}^{(l)}$ ▷ Remove the covered instances before the next lower α.
    else
      ${\mathcal{D}_{\text{ca}}}^{(l+1)} \leftarrow {\mathcal{D}_{\text{ca}}}^{(l)}$ ▷ No instances are excluded at an α_l without a finite ${q'_{\min}}^{(l)}$.
    $l \leftarrow l + 1$
  return $\mathcal{R}$

Output: $\mathcal{R}$, which contains the $(\alpha, q'_{\min}, [\psi_1, \ldots, \psi_C])$ triples of the recorded regions (descending in α)

2.1 Test-time

For new, unseen test instances, the application and behavior of the selection criteria of Eq. 9 (Schmaltz, 2026) for the most conservative HR region (nearest 1) *is the same as in the original work*. However, now for points that fall outside that HR region, we perform a search for membership in the highest probability region among the remaining lower probability regions using Eq. 9, but with the $q'_{\min}$ and $[\psi_1, \ldots, \psi_C]$ values for each remaining region.

2.2 Accounting for the Effective Sample Size

The application of the DKW inequality on the distance quantiles is the same as in Appendix A.5 of the original work. We maintain the convention of pinning the α of Eq. 12 to that of the corresponding region. As such, each point assigned to a region is also assigned to a corresponding (and possibly lower probability) *lower* region using Eq. 19, if such a region exists.

3 Discussion

SDM estimators make use of the geometry of the feature-representation space of neural models to robustly estimate the high-probability region of the output distribution. Growing evidence suggests that HR regions with α values near 1 exhibit improved robustness to covariate shifts in natural language data relative to alternatives that marginalize over these metric-learning signals of the epistemic uncertainty, as high SIMILARITY (q) values and DISTANCE quantiles (d) near 1 serve as relatively stable partitions of the distribution. In their own way, datapoints for which $q = 0$ and/or $d = 0$ also admit an "easy" resolution in the decision-making pipeline; they correspond to out-of-distribution points. The challenge, in practice, is how to adjudicate the remaining points outside those extremes.

Addressing this pushes the limits of what we might reasonably hope to reliably estimate without making additional assumptions, or applying domain-specific knowledge, since in practice, as q and/or d decrease toward 0, the resulting partitions of the calibration set provide less robust mappings to new, unseen test instances drawn from covariate shifted distributions, which are not uncommon with high-dimensional data, such as language.

4 Conclusion

In this brief note, we have proposed a straightforward way to rank the relative probability of points outside the HR region introduced in Schmaltz (2026). The HR region retains its interpretation as conservatively having an estimated class- and prediction-conditional accuracy at least the given α value, and the nested regions provide a relative ranking of the probability of the remaining datapoints for routing to additional processes in the decision-making pipeline.

References

Allen Schmaltz. 2026. Similarity-distance-magnitude activations. In *Findings of the Association for Computational Linguistics: ACL 2026*, pages 22037–22057, San Diego, California, United States. Association for Computational Linguistics.