SPARC: Score Prompting and Adaptive Fusion for Zero-Shot Multi-Label Recognition in Vision-Language Models

Supplementary Material

6. Supplementary Overview

We organize our Supplementary Material as follows:

- Sec. 7 provides pseudocode for our compound prompt generation method.
- Sec. 8 extends the noise model analysis from Sec. 3 of the main paper to all datasets and CLIP backbones.
- Sec. 9 shows the per-class performance of our method on all classes from all three datasets.
- Sec. 10 shows expaned results of Rank Fusion ablation with and without the final "merge" step.
- Sec. 11 shows the results of different ablations on the compound prompts in our method, including randomized prompts, cooccurrence filtration, and compound prompt templates.
- Sec. 12 shows and discusses histograms of 1st-max and 2nd-max scores in order to provide some intuition of why the latter provides better separation than the former.
- Sec. 13 offers a theoretical justification for use of a "weakened max" and use of an adaptive fusion strategy. This includes a proof of the theorem that was informally stated at the end of Sec. 3 in the main paper.

7. Compound Prompt Generation Pseudocode

We provide pseudocode for our compound prompt generation method below. Note that it only requires *coarse* knowledge of the cooccurrence probabilities, specifically knowledge of which pairs and triplets have low probability of cooccurring.

Algorithm 2 Compound Prompt Generation

Require: Classnames $c_1, \ldots, c_N$, cooccurrence info $\mathbb{P}$, thresholds τ_2 and τ_3, optional LLM Φ
Ensure: Generated prompt set P

Initialize $P \leftarrow \emptyset$
for $i \in [N-1]$ **do**
  for $j \in [i+1, N]$ **do**
    if $\mathbb{P}(j \mid i) > \tau_2$ **then**
      $P \leftarrow P \cup \{$"c_i and c_j"$\}$
      if $\max_{k \in [N]-\{i,j\}} \mathbb{P}(k \mid i, j) > \tau_3$ **then**
        $k^* \leftarrow \arg\max_{k \in [N]-\{i,j\}} \mathbb{P}(k \mid i, j)$
        $P \leftarrow P \cup \{$"c_i, c_j, and c_{k^*}"$\}$
Remove redundant triplets (i, j, k) vs (i, k, j) by comparing $\mathbb{P}(k|i, j)$ and $\mathbb{P}(j|i, k)$
if optional LLM Φ is provided **then**
  $P \leftarrow P \cup \Phi(P)$
return P

8. Expanded Noise Model Results

We use this section to show the results of running the noise model analysis from Tab. 1 on CLIP similarity scores from all three datasets computed with all nine CLIP backbones. Each columnn of each of Tab. 8, Tab. 9, Tab. 10 represents a separate fit of the noise models. As in the main paper, we report fraction of variance unexplained (FVU), as well as the fitted δ strength of the static AND-bonus and the lower and upper quartiles of the $\delta_{i,j}$ strengths of the variable AND-bonus. We see similar trends to those discussed in the main paper; the OR-only noise model explains significantly more variance than the AND-only model, and the OR-with-AND-bonus models explain most of the variance gap between the "constant" upper-bound and "look-up table" lower-bound. Hence, we find that CLIP scores tend to behave like an OR-gate with an AND-gate "bonus" for many different backbones on multiple datasets.

Noise Model	COCO								
	ViT-L/14 336px	ViT-L/14	ViT-B/16	ViT-B/32	RN50 x64	RN50 x16	RN50 x4	RN101	RN50
constant	0.802	0.812	0.823	0.775	0.809	0.798	0.757	0.771	0.791
only AND	0.535	0.544	0.557	0.517	0.553	0.539	0.501	0.506	0.500
only OR	0.288	0.301	0.292	0.295	0.284	0.281	0.257	0.260	0.280
additive	0.267	0.279	0.276	0.278	0.269	0.264	0.244	0.252	0.263
OR + static AND-bonus	0.263	0.275	0.271	0.273	0.264	0.260	0.239	0.245	0.257
OR + variable AND-bonus	0.248	0.257	0.258	0.259	0.245	0.247	0.228	0.233	0.245
look-up table	0.235	0.243	0.244	0.245	0.230	0.233	0.217	0.221	0.231
static bonus strength	0.560	0.558	0.522	0.502	0.560	0.549	0.506	0.445	0.484
variable bonus strength (lower quartile)	0.422	0.408	0.386	0.366	0.433	0.405	0.389	0.348	0.381
variable bonus strength (upper quartile)	0.531	0.527	0.503	0.489	0.582	0.666	0.656	0.631	0.583

Table 8. Comparison of fidelity of noise models for scoring pairwise compound prompts on COCO for all CLIP backbones. Notice that the OR-only model is a significantly better fit than AND-only, and that the OR+AND-bonus models capture nearly all of the fidelity of the look-up table. Please refer to Tab. 1 for more details on the noise models.

Noise Model	VOC								
	ViT-L/14 336px	ViT-L/14	ViT-B/16	ViT-B/32	RN50 x64	RN50 x16	RN50 x4	RN101	RN50
constant	0.640	0.648	0.661	0.616	0.603	0.631	0.619	0.628	0.653
only AND	0.275	0.276	0.275	0.282	0.276	0.299	0.245	0.230	0.242
only OR	0.144	0.150	0.143	0.136	0.137	0.127	0.123	0.122	0.126
additive	0.128	0.133	0.127	0.123	0.119	0.119	0.111	0.112	0.114
OR + static AND-bonus	0.125	0.130	0.125	0.120	0.118	0.116	0.110	0.111	0.111
OR + variable AND-bonus	0.120	0.123	0.112	0.113	0.109	0.111	0.105	0.104	0.107
look-up table	0.114	0.116	0.109	0.109	0.104	0.105	0.100	0.099	0.101
static bonus strength	0.427	0.412	0.459	0.408	0.557	0.367	0.356	0.441	0.493
variable bonus strength (lower quartile)	0.306	0.274	0.296	0.254	0.429	0.257	0.197	0.330	0.361
variable bonus strength (upper quartile)	0.582	0.555	0.873	0.419	0.680	0.461	0.321	0.833	0.792

Table 9. Comparison of fidelity of noise models for scoring pairwise compound prompts on VOC for all CLIP backbones. Notice that the OR-only model is a significantly better fit than AND-only, and that the OR+AND-bonus models capture nearly all of the fidelity of the look-up table. Please refer to Tab. 1 for more details on the noise models.

9. Per-class performance of SPARC vs vanilla ZSCLIP on all datasets

We show in Fig. 5 that SPARC *consistently* improves performance across all the classes in all three datasets. In fact, we see that there is only one class out of all the classes in all the datasets (181 classes total) where our method does notably worse than ZSCLIP. That class is the "earthquake" class of the NUSWIDE dataset. For all other classes, our method is either the same or (in most cases) notably better than ZSCLIP.

Noise Model	NUSWIDE								
	ViT-L/14 336px	ViT-L/14	ViT-B/16	ViT-B/32	RN50 x64	RN50 x16	RN50 x4	RN101	RN50
constant	0.536	0.531	0.562	0.581	0.565	0.557	0.534	0.534	0.589
only AND	0.330	0.331	0.346	0.358	0.351	0.352	0.344	0.350	0.371
only OR	0.228	0.230	0.234	0.248	0.251	0.239	0.222	0.223	0.256
additive	0.196	0.199	0.200	0.211	0.213	0.207	0.195	0.201	0.226
OR + static AND-bonus	0.193	0.195	0.196	0.207	0.210	0.204	0.191	0.197	0.222
OR + variable AND-bonus	0.175	0.177	0.183	0.190	0.187	0.185	0.180	0.186	0.208
look-up table	0.162	0.164	0.170	0.177	0.173	0.171	0.167	0.173	0.194
static bonus strength	0.604	0.603	0.579	0.584	0.637	0.597	0.581	0.556	0.591
variable bonus strength (lower quartile)	0.488	0.485	0.471	0.457	0.482	0.456	0.435	0.394	0.418
variable bonus strength (upper quartile)	0.647	0.619	0.626	0.620	0.644	0.624	0.663	0.628	0.609

Table 10. Comparison of fidelity of noise models for scoring pairwise compound prompts on NUSWIDE for all CLIP backbones. Notice that the OR-only model is a significantly better fit than AND-only, and that the OR+AND-bonus models capture nearly all of the fidelity of the look-up table. Please refer to Tab. 1 for more details on the noise models.

Use compound	Normalize	Compound prompt type	Rank Fusion Strategy	COCO	VOC	NUSWIDE	Avg
	✓	-	-	65.9	87.7	45.1	66.2
✓	✓	randomized	ours	65.9	87.5	45.1	66.2
✓	✓	randomized	mean	65.9	87.4	45.1	66.1
✓	✓	ours	ours	**68.3**	**89.2**	**47.2**	**68.3**

Table 11. Randomized compound prompt ablation confirms the semantic value of compound prompts. Our ablation replaces compound prompts with prompts that use random characters instead of cooccurrent classes. These prompts offer no benefit over normalized singletons, suggesting that the gain caused by cooccurrent classes is due to semantics, and not just the statistical properties of an ensemble.

10. Expanded results from Rank Fusion Ablation

We showed in the main paper that our Rank Fusion strategy outperforms various handcrafted alternative fusion strategies. Fig. 6 shows these results with and without the final "merge" step where the output of maxVariance is added to the singleton score. We see that this step is critical for performance, pointing to the complementarity of maxVariance and singleton scores.

11. Compound prompt ablation results

We describe some additional ablations on the compound prompts used by our method. We start by comparing the performance of our compound prompts with "randomized" prompts in which the cooccurrent classes are replaced by random characters. We do this in light of the findings of WaffleCLIP [5], which found that randomized prompt ensembles could perform as well as descriptive ones due to the inherent statistical benefits of ensembling. We find that this is not the case for our problem - randomized compound prompts offer no benefit over normalized singletons. We show our results in Tab. 11.

We also consider alternative templates for the formulaic pair prompts. In addition to the "A and B" template used in the main paper, we also try "A or B", "A with B", "A next to B", "A and not B" (alongside "A and B"), and the combination of all templates. For simplicity, we remove the triplet and descriptive compound prompts during this analysis. We report the results in Tab. 12. We find that our original template "A and B" performs the best, although other conjunctive templates do get quite close, while "A or B" does considerably worse. This latter finding suggests that perhaps CLIP does interpret "and" and "or" differently, even if it treats "A and B" primarily as an OR-gate.

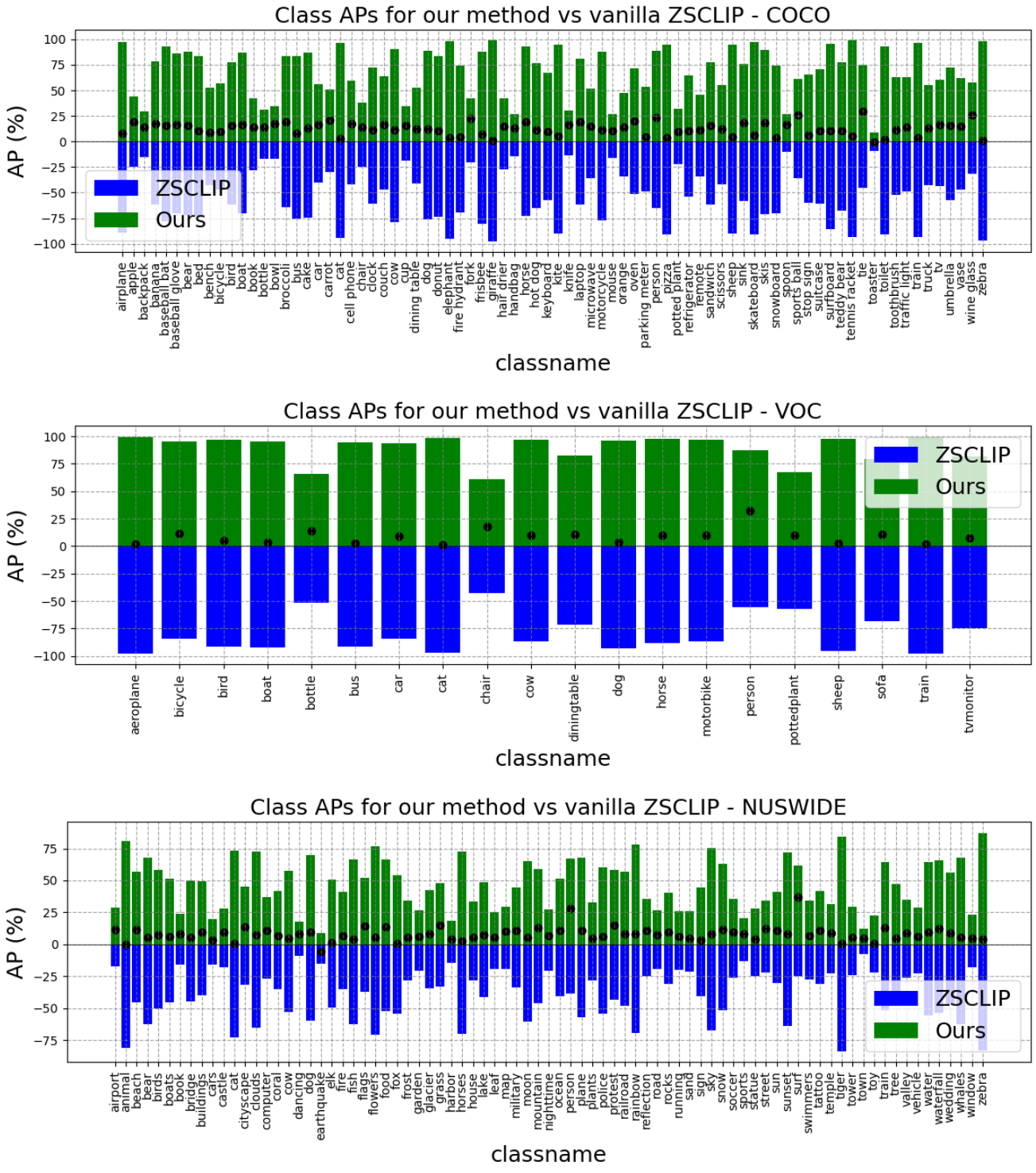

Figure 5. Per-class APs (averaged over all CLIP backbones) for our method vs vanilla ZSCLIP on all three datasets. Our method consistently improves over ZSCLIP for almost every class in all datasets.

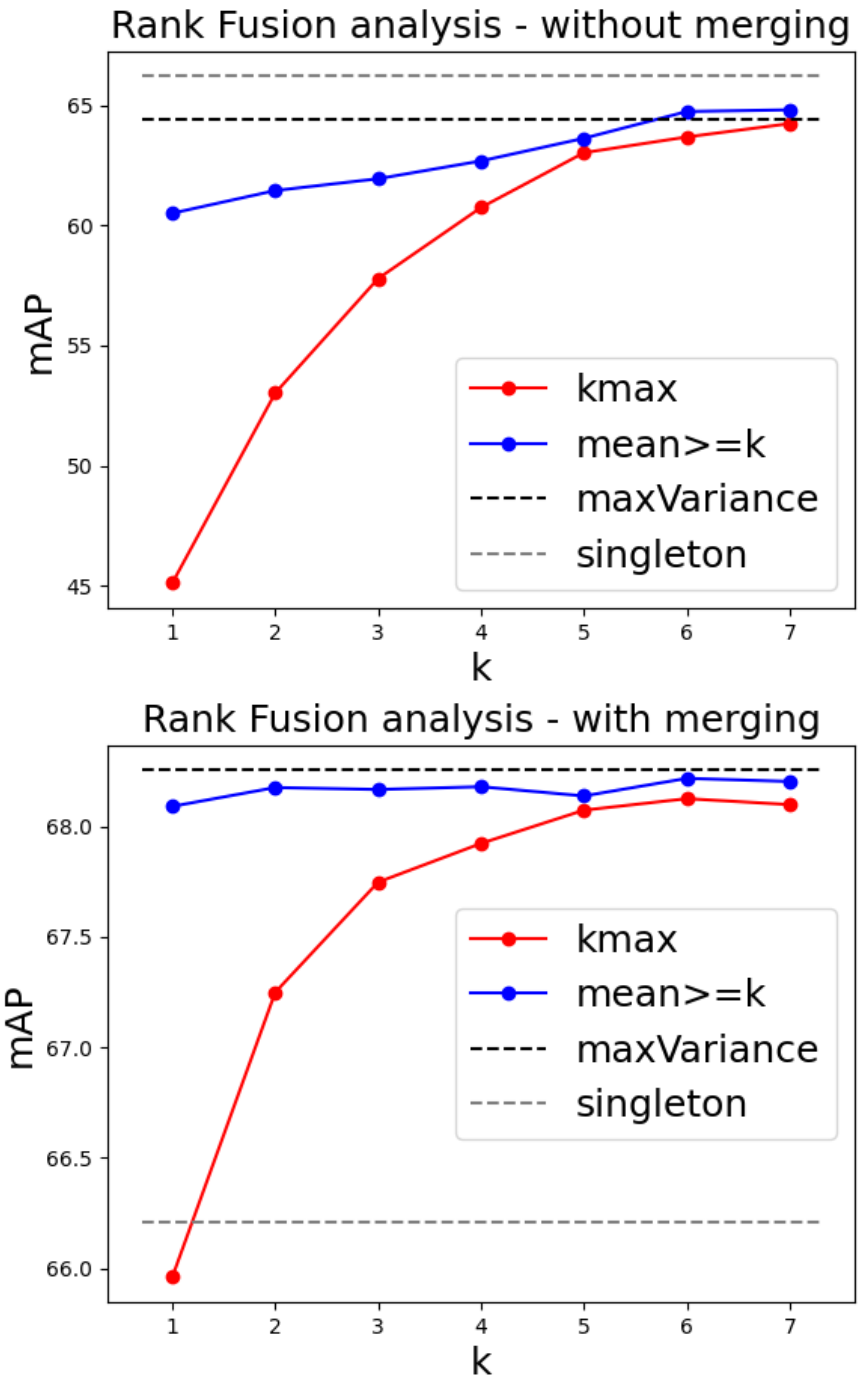

Figure 6. Average mAP for different Rank Fusion strategies, without (top) and with (bottom) the "merge" step demonstrates superiority of adaptive fusion over fixed strategies and the importance of the "merge" step.

12. A qualitative look at the weakened max - histograms

Fig. 7 shows distributions of singleton, 1st-max, and 2nd-max scores for "cat" in COCO, as well as the uniform average of singleton and 2nd-max. We see in the second row that 1st-max lifts a considerable number of negative examples, creating overlap between negatives and positives. Third row shows that 2nd-max causes less overlap, lifting fewer negatives without adversely impacting positives. The last row shows that fusing 2nd-max with singleton leads to good separation.

Normalize	Pair prompt template	Triplets + Descriptive	COCO	VOC	NUSWIDE	Avg
✓	-		65.9	87.7	45.1	66.2
✓	“A and B”		68.1	89.0	47.0	68.0
✓	“A or B”		67.0	88.4	46.2	67.2
✓	“A with B”		67.8	89.0	47.0	67.9
✓	“A next to B”		67.9	88.8	46.7	67.8
✓	“A and not B”		67.9	88.7	46.5	67.7
✓	all templates		67.9	88.7	46.7	67.8
✓	“A and B”	✓	**68.3**	**89.2**	**47.2**	**68.3**

Table 12. Ablations on templates used for formulaic pairwise prompts. We find that our original template “A and B” performs best.

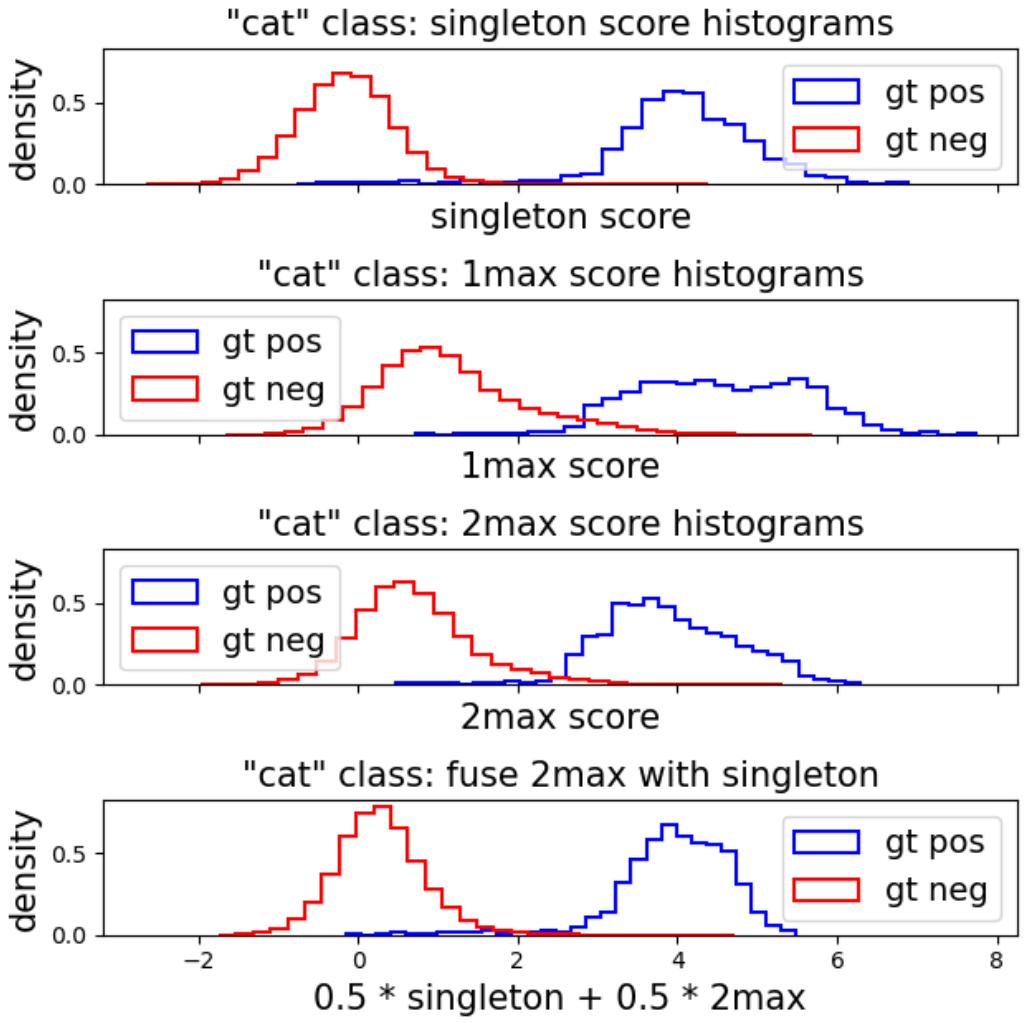

Figure 7. Histograms for singleton, 1st max, and 2nd max scores for “cat” in the COCO dataset. We see that 1st max creates overlap by lifting the scores of some ground-truth negatives. 2nd max does not create these issues and performs well when fused with singleton scores.

13. Theoretical Explanation for Weakened Max and Adaptive Fusion

13.1. Theory Overview

We introduced a noise model in Sec. 3 to explain the behavior of CLIP scores. We will now use this model to come up with a theoretical justification for the use of a "weakened max" instead of an outright maximum of compound scores.

We start with a hypothetical scenario where the goal is to predict the presence or absence of target class 0 given a set of m compound prompts pairing class 0 with each of cooccurring classes $1, ..., m$. We introduce this scenario and our assumptions about it in Sec. 13.2. We find that, given a large enough m, second-max will always have better discriminative performance than first-max, matching our empirical observation from Sec. 4.4.1. We formally state this finding as Theorem 1, which we prove in Sec. 13.4.

We make a further claim in Theorem 2, which states that there are settings for which a sufficiently *small* m will cause the first-max to outperform the second-max, and that the boundary between "sufficiently large" and "sufficiently small" depends on data statistics that are unknowable in any practical setting, even one where exact cooccurrence statistics are available. We prove this theorem in Sec. 13.5. We suspect that we would find similar behavior for other pairs of statistics, such as second-max vs third-max, third vs fourth, fourth vs median, median vs min, etc. In general, **fixed fusion rules are suboptimal** for combining the compound prompt scores.

Together, these theorems justify not only the use of a **"weakened max"**, but also the use of an **adaptive fusion strategy** such as Rank Fusion, which can use the direction of highest variance to figure out which order statistics are most useful for the setting at hand.

13.2. Preliminaries

Suppose we have target class 0 and cooccurring classes $1, .., m$. These have ground-truth presences $y_0, y_1, ..., y_m \in \{0, 1\}$. We make some assumptions about their distribution.

Assumption 1. *The ground-truth distribution has the following properties:*

$$Pr(y_i = 1 \mid y_0 = 1) = \rho \qquad \forall i \in [m] \tag{7}$$

$$Pr(y_i = 1 \mid y_0 = 0) = q \qquad \forall i \in [m] \tag{8}$$

$$1 > \rho > q > 0 \tag{9}$$

$$y_j \perp y_i \mid y_0 \quad \forall i \neq j \in [m] \tag{10}$$

We also introduce variables $\tilde{y}_0, \tilde{y}_1, ..., \tilde{y}_m \in \{0, 1\}$ which are "noisy" versions of the ground-truth. Think of these as indicating whether each object is visible to the VLM. E.g. we might have $y_i = 1$ and $\tilde{y}_i = 0$ if object i was occluded, or we might have $y_i = 0$ and $\tilde{y}_i = 1$ if an spurious object in the image resembled i. For each $i \in [m]$ we have:

$$\tilde{y}_i = \begin{cases} 1 - y_i & \text{with probability } \nu, \\ y_i & \text{with probability } 1 - \nu. \end{cases}$$

We make some assumptions about the distribution of these variables.

Assumption 2. *The distribution of $\tilde{y}_0, \tilde{y}_1, ..., \tilde{y}_m$ has the following properties:*

$$\tilde{y}_i \textit{ depends only on } y_i \tag{11}$$

$$\nu < \frac{1}{2} \tag{12}$$

As mentioned on the main paper, we assume that the score for the prompt "$\{0\}$ and $\{i\}$" is distributed as follows:

$$s_{0,i} = \max(\tilde{y}_0, \tilde{y}_i) + \delta \min(\tilde{y}_0, \tilde{y}_i) + \varepsilon \tag{13}$$

where δ is the strength of the "AND-bonus" described in the main paper, and $\varepsilon \sim \mathcal{W}(\sigma)$ is symmetric, zero-centered, additive noise whose scale is controlled by σ.

From the set $\{s_{0,1}, ..., s_{0,m}\}$ we compute order statistics r_1, r_2, which are the first and second highest elements, respectively.

Now, suppose we independently draw a ground-truth positive sample with $y_0^+ = 1$ and a ground-truth negative sample with $y_0^- = 0$ and compute order statistics r_1^+, r_2^+ and r_1^-, r_2^-. We define "win" events W_1 and W_2 as the events where $r_1^+ > r_1^-$ and $r_2^+ > r_2^-$, respectively.

We will now make an assumption about $\varepsilon \sim \mathcal{W}(\sigma)$ in order to simplify our analysis. In order to state our assumption, we will need a bit more notation.

$$\bar{s}_{0,i} = \max(\tilde{y}_0, \tilde{y}_i) + \delta \min(\tilde{y}_0, \tilde{y}_i) \tag{14}$$

$$\bar{r}_1, \bar{r}_2 \text{ are the first and second highest elements of } \{\bar{s}_{0,1}, ..., \bar{s}_{0,m}\} \tag{15}$$

We are now ready to state our assumption.

Assumption 3. *Assume that σ is small enough for the following to approximately hold for each $k \in \{1, 2\}$*

$$Pr(W_k) \approx \begin{cases} 1 & \text{if } \bar{r}_k^+ > \bar{r}_k^-, \\ 0 & \text{if } \bar{r}_k^+ < \bar{r}_k^-, \\ \frac{1}{2} & \text{if } \bar{r}_k^+ = \bar{r}_k^-, \end{cases}$$

We have now stated all of our assumptions.

Before making our formal theorem statements, we define some shorthand that we will use throughout the proof. First, we note that if we hold $\tilde{y}_0$ fixed, then $\bar{r}_1$ and $\bar{r}_2$ can each take on one of two values. For example, if $\tilde{y}_0 = 0$ then the possible values are $\{0, 1\}$, and if $\tilde{y}_0 = 1$ then the possible values are $\{1, 1 + \delta\}$. As such, we define pairs of complementary events (H_1, L_1) and (H_2, L_2) to denote that $\bar{r}_1$ or $\bar{r}_2$ took the higher or lower of its possible values.

We define some additional shorthand:

$$\rho' := (1 - \nu)\rho \ + \ \nu(1 - \rho) \tag{16}$$

$$q' := (1 - \nu)q \ + \ \nu(1 - q) \tag{17}$$

$$a := 1 - \rho' \tag{18}$$

$$\gamma := \frac{1 - q}{1 - \rho'} \tag{19}$$

$$A := m(1 - a)a^{m-1} \tag{20}$$

$$G := m(1 - \gamma a)(\gamma a)^{m-1} \tag{21}$$

We are now ready to formally state our theorems.

13.3. Formal Theorem Statements

Theorem 1. *Given the assumptions above, plus the additional assumption that $Pr(\tilde{y}_0^+ \neq y_0^+ \bigvee \tilde{y}_0^- \neq y_0^-) > 0$, we can guarantee that $Pr(W_2) > Pr(W_1)$ for sufficiently large m.*

Theorem 2. *There are values of ρ, q, ν and distributions of $(\tilde{y}_0^+, \tilde{y}_0^-)$ which satisfy all the requirements of Theorem 1, for which $Pr(W_2) < Pr(W_1)$ for sufficiently small m. In fact, the value of m at which the inequality reverses depends on label-flip probability ν.*

13.4. Proof of Theorem 1

We start by proving that $\rho' > q'$, i.e. $\Pr(\tilde{y}_i = 1 \mid y_0 = 1) > \Pr(\tilde{y}_i = 1 \mid y_0 = 0)$.

Lemma 1. $1 > \rho' > q' > 0$ *given the above assumptions on ρ, q, and ν.*

Proof. We can use some algebra to prove this from Assumptions 1 and 2.

$$\rho' = \rho + \nu - 2\nu\rho \tag{22}$$
$$q' = q + \nu - 2\nu q \tag{23}$$
$$\rho' - q' = 2(\frac{1}{2} - \nu)(\rho - q) \tag{24}$$
$$> 0 \tag{25}$$

It is trivial to show that $\rho', q' \in (0, 1)$ because they are both mixtures of quantities in that range. □

Our next lemma will derive some probability differences that will be important for our proof.

Lemma 2. *Consider the following probability differences:*

$$d^{HH} := Pr(H_2^+, H_2^-) - Pr(H_1^+, H_1^-) \tag{26}$$
$$d^{HL} := Pr(H_2^+, L_2^-) - Pr(H_1^+, L_1^-) \tag{27}$$
$$d^{LH} := Pr(L_2^+, H_2^-) - Pr(L_1^+, H_1^-) \tag{28}$$
$$d^{LL} := Pr(L_2^+, L_2^-) - Pr(L_1^+, L_1^-) \tag{29}$$

We claim that, regardless of the values or distribution of $\tilde{y}_0^+$ and $\tilde{y}_0^-$, the following is true:

$$d^{HH} = AG - (1 - a^m)G - (1 - (\gamma a)^m)A \tag{30}$$
$$d^{HL} = (1 - a^m)G - AG - (\gamma a)^m A \tag{31}$$
$$d^{LH} = (1 - (\gamma a)^m)A - AG - a^m G \tag{32}$$
$$d^{LL} = AG + a^m G + (\gamma a)^m A \tag{33}$$

Proof. We start by noting that $d^{HH}, d^{HL}, d^{LH}, d^{LL}$ do not depend on $\tilde{y}_0^+, \tilde{y}_0^-$. Although $\tilde{y}_0^+$ and $\tilde{y}_0^-$ affect the specific values that $\bar{r}_1^+, \bar{r}_2^+, \bar{r}_1^-, \bar{r}_2^-$ can take, they do not affect the probabilities of events $H_1^+, L_1^+, H_2^+, L_2^+, H_1^-, L_1^-, H_2^-, L_2^-$. This is because these events only depend on $\tilde{y}_1^+, ..., \tilde{y}_m^+$ and $\tilde{y}_1^-, ..., \tilde{y}_m^-$, which in turn depend on ground truths $y_1^+, ..., y_m^+$ and $y_1^-, ..., y_m^-$, which all depend on y_0^+ and y_0^-, which are fixed, so there is no dependency on $\tilde{y}_0^+$ or $\tilde{y}_0^-$.

We also note that we can factor out the joint probabilities in $d^{HH}, d^{HL}, d^{LH}, d^{LL}$ because the positive and negative samples were drawn independently, hence:

$$d^{HH} = \text{Pr}(H_2^+)\text{Pr}(H_2^-) - \text{Pr}(H_1^+)\text{Pr}(H_1^-) \tag{34}$$
$$d^{HL} = \text{Pr}(H_2^+)\text{Pr}(L_2^-) - \text{Pr}(H_1^+)\text{Pr}(L_1^-) \tag{35}$$
$$d^{LH} = \text{Pr}(L_2^+)\text{Pr}(H_2^-) - \text{Pr}(L_1^+)\text{Pr}(H_1^-) \tag{36}$$
$$d^{LL} = \text{Pr}(L_2^+)\text{Pr}(L_2^-) - \text{Pr}(L_1^+)\text{Pr}(L_1^-) \tag{37}$$

We can work out the probability for event H_1^+, which occurs iff at least one of $\tilde{y}_1^+, ..., \tilde{y}_m^+$ is 1:

$$\text{Pr}(H_1^+) = 1 - (1 - \rho')^m \tag{38}$$
$$= 1 - a^m \tag{39}$$

By similar reasoning, we can say:

$$\text{Pr}(H_1^-) = 1 - (\gamma a)^m \tag{40}$$

Next, we work out the probability for the event H_2^+, which which occurs iff at least two of $\tilde{y}_1^+, ..., \tilde{y}_m^+$ are 1:

$$\text{Pr}(H_2^+) = 1 - (1 - \rho')^m - m\rho'(1 - \rho')^{m-1} \tag{41}$$
$$= 1 - a^m - m(1 - a)a^{m-1} \tag{42}$$
$$= 1 - a^m - A \tag{43}$$

Similar reasoning can be used for H_2^-:

$$\Pr(H_2^-) = 1 - (\gamma a)^m - m(1-\gamma a)(\gamma a)^{m-1} \tag{44}$$
$$= 1 - (\gamma a)^m - G \tag{45}$$

Putting these all together, we get:

$$d^{HH} = \Pr(H_2^+)\Pr(H_2^-) - \Pr(H_1^+)\Pr(H_1^-) \tag{46}$$
$$= (1 - a^m - A)(1 - (\gamma a)^m - G) - (1 - a^m)(1 - (\gamma a)^m) \tag{47}$$
$$= (1 - a^m)(1 - (\gamma a)^m) - (1 - a^m)G - (1 - (\gamma a)^m)A + AG - (1 - a^m)(1 - (\gamma a)^m) \tag{48}$$
$$= AG - (1 - a^m)G - (1 - (\gamma a)^m)A \tag{49}$$
$$d^{HL} = \Pr(H_2^+)\Pr(L_2^-) - \Pr(H_1^+)\Pr(L_1^-) \tag{50}$$
$$= (1 - a^m - A)((\gamma a)^m + G) - (1 - a^m)(\gamma a)^m \tag{51}$$
$$= (1 - a^m)(\gamma a)^m + (1 - a^m)G - (\gamma a)^m A - AG - (1 - a^m)(\gamma a)^m \tag{52}$$
$$= (1 - a^m)G - AG - (\gamma a)^m A \tag{53}$$
$$d^{LH} = \Pr(L_2^+)\Pr(H_2^-) - \Pr(L_1^+)\Pr(H_1^-) \tag{54}$$
$$= (a^m + A)(1 - (\gamma a)^m - G) - (a^m)(1 - (\gamma a)^m) \tag{55}$$
$$= (a^m)(1 - (\gamma a)^m) + (1 - (\gamma a)^m)A - a^m G - AG - (a^m)(1 - (\gamma a)^m) \tag{56}$$
$$= (1 - (\gamma a)^m)A - AG - a^m G \tag{57}$$
$$d^{LL} = \Pr(L_2^+)\Pr(L_2^-) - \Pr(L_1^+)\Pr(L_1^-) \tag{58}$$
$$= (a^m + A)((\gamma a)^m + G) - (a^m)(\gamma a)^m \tag{59}$$
$$= (a^m)(\gamma a)^m + a^m G + (\gamma a)^m A + AG - (a^m)(\gamma a)^m \tag{60}$$
$$= AG + a^m G + (\gamma a)^m A \tag{61}$$

□

Now that we have derived these probability differences, we can use them to work out $\Pr(W_2) - \Pr(W_1)$ for the four possible settings of $(\tilde{y}_0^+, \tilde{y}_0^-)$. We address these case-by-case.

Case 1: $(\tilde{y}_0^+, \tilde{y}_0^-) = (0, 0)$ We can think of this as the **"TN-vs-FN" case**, where object 0 is not visible in either of the ground-truth positive or ground-truth negative images. In this case, we can derive the following expression:

Lemma 3. *If* $(\tilde{y}_0^+, \tilde{y}_0^-) = (0, 0)$ *then* $Pr(W_2) - Pr(W_1) = \frac{1}{2}(G - A)$.

Proof. In this case, we have $\bar{r}_1^+, \bar{r}_2^+, \bar{r}_1^-, \bar{r}_2^- \in \{0, 1\}$, so we can say:

$$\Pr(W_1) = \frac{1}{2}\Pr(H_1^+, H_1^-) + \frac{1}{2}\Pr(L_1^+, L_1^-) + \Pr(H_1^+, L_1^-) \tag{62}$$
$$= \frac{1}{2}\Big(\Pr(H_1^+)\Pr(H_1^-) + \Pr(L_1^+)\Pr(L_1^-) + 2\Pr(H_1^+)\Pr(L_1^-)\Big) \tag{63}$$
$$\Pr(W_2) = \frac{1}{2}\Big(\Pr(H_2^+)\Pr(H_2^-) + \Pr(L_2^+)\Pr(L_2^-) + 2\Pr(H_2^+)\Pr(L_2^-)\Big) \tag{64}$$

Hence, we can express the win-rate difference using $d^{HH}, d^{HL}, d^{LH}, d^{LL}$ as follows:

$$\Pr(W_2) - \Pr(W_1) = \frac{1}{2}\Big(d^{HH} + 2d^{HL} + d^{LL}\Big) \tag{65}$$
$$= \frac{1}{2}\Big(AG - (1-a^m)G - (1-(\gamma a)^m)A + 2((1-a^m)G - AG - (\gamma a)^m A) + AG + a^m G + (\gamma a)^m A\Big) \tag{66}$$
$$= \frac{1}{2}(G - A) \tag{67}$$

We note that this quantity is positive for sufficiently large m because $\frac{G}{A} = \frac{1-\gamma a}{1-a}\gamma^{m-1} = \frac{q'}{\rho'}(\frac{1-q'}{1-\rho'})^{m-1}$, which grows with m because $\rho' > q'$ per Lemma 1. □

Case 2: $(\tilde{y}_0^+, \tilde{y}_0^-) = (1, 1)$ We can think of this as the **"TP-vs-FP" case**, where object 0 is visible (correctly or spuriously) in both the ground-truth positive and ground-truth negative images. This leads to the same win-rate difference as the previous case.

Lemma 4. *If* $(\tilde{y}_0^+, \tilde{y}_0^-) = (1, 1)$ *then* $Pr(W_2) - Pr(W_1) = \frac{1}{2}(G - A)$.

Proof. In this case, we have $\bar{r}_1^+, \bar{r}_2^+, \bar{r}_1^-, \bar{r}_2^- \in \{1, 1+\delta\}$, so we can say:

$$\Pr(W_1) = \frac{1}{2}\Big(\Pr(H_1^+)\Pr(H_1^-) + \Pr(L_1^+)\Pr(L_1^-) + 2\Pr(H_1^+)\Pr(L_1^-)\Big) \tag{68}$$

$$\Pr(W_2) = \frac{1}{2}\Big(\Pr(H_2^+)\Pr(H_2^-) + \Pr(L_2^+)\Pr(L_2^-) + 2\Pr(H_2^+)\Pr(L_2^-)\Big) \tag{69}$$

Hence, the win-rate difference is the same as before, via the same steps as the previous lemma (Lemma 3):

$$\Pr(W_2) - \Pr(W_1) = \frac{1}{2}(G - A) \tag{70}$$

□

Case 3: $(\tilde{y}_0^+, \tilde{y}_0^-) = (0, 1)$ We can think of this as the **"FP-vs-FN" case**, where object 0 is occluded or obscured in the ground-truth positive image and spuriously visible in the ground-truth negative image. This is the most "difficult" case to rectify.

Lemma 5. *If* $(\tilde{y}_0^+, \tilde{y}_0^-) = (0, 1)$ *then* $Pr(W_2) - Pr(W_1) = \frac{1}{2}G\Big(1 - a^{m-1}(1 - \rho' + m\rho' + \rho'\frac{1-q'}{q'})\Big)$.

Proof. In this case, we have $\bar{r}_1^+, \bar{r}_2^+ \in \{0, 1\}$ and $\bar{r}_1^-, \bar{r}_2^- \in \{1, 1+\delta\}$. The best we can hope for is a tie, where the positive example takes on its higher value *and* the negative example takes on its lower value. Hence:

$$\Pr(W_1) = \frac{1}{2}\Pr(H_1^+, L_1^-) = \frac{1}{2}\Pr(H_1^+)\Pr(L_1^-) \tag{71}$$

$$\Pr(W_2) = \frac{1}{2}\Pr(H_2^+, L_2^-) = \frac{1}{2}\Pr(H_2^+)\Pr(L_2^-) \tag{72}$$

Hence, we can express the win-rate difference as:

$$\Pr(W_2) - \Pr(W_1) = \frac{1}{2}d^{HL} \tag{73}$$

$$= \frac{1}{2}\Big((1 - a^m)G - AG - (\gamma a)^m A\Big) \tag{74}$$

$$= \frac{1}{2}G\Big((1 - a^m) - A - (\gamma a)^m \frac{A}{G}\Big) \tag{75}$$

$$= \frac{1}{2}G\Big((1 - a^m) - A - a^m\gamma\frac{1-a}{1-\gamma a}\Big) \tag{76}$$

$$= \frac{1}{2}G\Big((1 - a^m) - m(1-a)a^{m-1} - a^m\gamma\frac{1-a}{1-\gamma a}\Big) \tag{77}$$

$$= \frac{1}{2}G\Big(1 - a^{m-1}(a + m(1-a) + a\gamma\frac{1-a}{1-\gamma a})\Big) \tag{78}$$

$$= \frac{1}{2}G\Big(1 - a^{m-1}(1 - \rho' + m\rho' + \rho'\frac{1-q'}{q'})\Big) \tag{79}$$

We once again note that this quantity is positive for sufficiently large m because $a^{m-1}(1-\rho'+m\rho'+\rho'\frac{1-q'}{q'}) = o(ma^m)$ since $a = (1 - \rho') < 1$ (per Lemma 1). □

Case 4: $(\tilde{y}_0^+, \tilde{y}_0^-) = (1, 0)$ We can think of this as the **"TP-vs-TN" case**, where there is no occlusion or spurious cue for object 0 in either image. This is the "easiest" case to deal with, and it turns out to be the one case where a first-max is actually better than a second-max.

Lemma 6. *If* $(\tilde{y}_0^+, \tilde{y}_0^-) = (1, 0)$ *then* $Pr(W_2) - Pr(W_1) = \frac{1}{2}A\Big(G + a(1-q')^{m-1}(\frac{q'}{\rho'} + \frac{1-q'}{1-\rho'}) - 1\Big)$.

Proof. In this case, we have $\bar{r}_1^+, \bar{r}_2^+ \in \{1, 1+\delta\}$ and $\bar{r}_1^-, \bar{r}_2^- \in \{0, 1\}$. The worst thing that can happen is a "tie", in the event that the positive example gets its lower value *and* the negative example gets its higher one, otherwise we get an outright "win". Hence:

$$\begin{aligned}
\Pr(W_1) &= \Pr(H_1^+, H_1^-) + \Pr(H_1^+, L_1^-) + \frac{1}{2}\Pr(L_1^+, H_1^-) + \Pr(L_1^+, L_1^-) && (80)\\
&= 1 - \frac{1}{2}\Pr(L_1^+, H_1^-) && (81)\\
&= 1 - \frac{1}{2}\Pr(L_1^+)\Pr(H_1^-) && (82)\\
\Pr(W_2) &= 1 - \frac{1}{2}\Pr(L_2^+)\Pr(H_2^-) && (83)
\end{aligned}$$

Hence, we can express the win-rate difference as:

$$\begin{aligned}
\Pr(W_2) - \Pr(W_1) &= -\frac{1}{2}d^{LH} && (84)\\
&= -\frac{1}{2}\Big((1-(\gamma a)^m)A - AG - a^m G\Big) && (85)\\
&= \frac{1}{2}\Big(AG + a^m G - (1-(\gamma a)^m)A\Big) && (86)\\
&= \frac{1}{2}A\Big(G + a^m\frac{G}{A} - (1-(\gamma a)^m)\Big) && (87)\\
&= \frac{1}{2}A\Big(G + a^m\frac{1-\gamma a}{1-a}\gamma^{m-1} + (\gamma a)^m - 1\Big) && (88)\\
&= \frac{1}{2}A\Big(G + a(\gamma a)^{m-1}(\frac{1-\gamma a}{1-a} + \gamma) - 1\Big) && (89)\\
&= \frac{1}{2}A\Big(G + a(1-q')^{m-1}(\frac{q'}{\rho'} + \frac{1-q'}{1-\rho'}) - 1\Big) && (90)
\end{aligned}$$

We note that in this particular case, $\Pr(W_2) - \Pr(W_1)$ actually becomes *negative* as m grows large, because both G and $a(1-q')^{m-1}(\frac{q'}{\rho'} + \frac{1-q'}{1-\rho'})$ shrink exponentially with m, and so $\frac{1}{2}A$ is eventually multiplied by a negative number. However, we will soon see that this negative win-rate difference is outweighed by the positive ones from Lemmas 3, 4, and 5, leading to an overall positive difference. □

We have now worked out the win-rate differences for all four possible values of $(\tilde{y}_0^+, \tilde{y}_0^-)$. We are finally ready to prove our main theorem, which we restate below.

Theorem 1 (restated). *If* $Pr(\tilde{y}_0^+ \neq y_0^+ \bigvee \tilde{y}_0^- \neq y_0^-) > 0$*, then* $Pr(W_2) > Pr(W_1)$ *for sufficiently large* m*.*

Proof. We start by defining a distribution over $(\tilde{y}_0^+, \tilde{y}_0^-)$:

$$\begin{aligned}
\pi^{(0,0)} &:= \Pr(\tilde{y}_0^+ = 0,\ \tilde{y}_0^- = 0) && (91)\\
\pi^{(1,1)} &:= \Pr(\tilde{y}_0^+ = 1,\ \tilde{y}_0^- = 1) && (92)\\
\pi^{(0,1)} &:= \Pr(\tilde{y}_0^+ = 0,\ \tilde{y}_0^- = 1) && (93)\\
\pi^{(1,0)} &:= \Pr(\tilde{y}_0^+ = 1,\ \tilde{y}_0^- = 0) && (94)
\end{aligned}$$

We can combine the results of Lemmas 3, 4, 5, and 6, to get the overall win-rate difference:

$$\begin{aligned}
\Pr(W_2) - \Pr(W_1) \quad &= \quad \frac{1}{2}(\pi^{(0,0)} + \pi^{(1,1)})(G - A) && (95)\\
&+\frac{1}{2}\pi^{(0,1)}G\Big(1 - a^{m-1}(1 - \rho' + m\rho' + \rho'\frac{1-q'}{q'})\Big) && (96)\\
&+\frac{1}{2}\pi^{(1,0)}A\Big(G + a(1-q')^{m-1}(\frac{q'}{\rho'} + \frac{1-q'}{1-\rho'}) - 1\Big) && (97)
\end{aligned}$$

We can lower-bound the last term of this sum with $-A$ to get:

$$\begin{aligned}
\Pr(W_2) - \Pr(W_1) \quad &\geq \quad \frac{1}{2}(\pi^{(0,0)} + \pi^{(1,1)})(G - A) && (98)\\
&+\frac{1}{2}\pi^{(0,1)}G\Big(1 - a^{m-1}(1 - \rho' + m\rho' + \rho'\frac{1-q'}{q'})\Big) && (99)\\
&-\frac{1}{2}\pi^{(1,0)}A && (100)
\end{aligned}$$

If we pick $m > 1 + \frac{\log(2)+\log(1-\rho'+m\rho'+\rho'\frac{1-q'}{q'})}{-\log(1-\rho')}$ then we can lower-bound the second term to get:

$$\begin{aligned}
\Pr(W_2) - \Pr(W_1) \quad &\geq \quad \frac{1}{2}(\pi^{(0,0)} + \pi^{(1,1)})(G - A) && (101)\\
&+\frac{1}{4}\pi^{(0,1)}G && (102)\\
&-\frac{1}{2}\pi^{(1,0)}A && (103)
\end{aligned}$$

which simplifies to:

$$\begin{aligned}
\Pr(W_2) - \Pr(W_1) &\geq \frac{1}{2}\Big((\pi^{(0,0)} + \pi^{(1,1)} + \frac{1}{2}\pi^{(0,1)})G - (\pi^{(0,0)} + \pi^{(1,1)} + \pi^{(1,0)})A\Big) && (104)\\
&= \frac{1}{2}\Big((\pi^{(0,0)} + \pi^{(1,1)} + \frac{1}{2}\pi^{(0,1)})G - (1 - \pi^{(0,1)})A\Big) && (105)
\end{aligned}$$

We note that $\pi^{(0,0)} + \pi^{(1,1)} + \frac{1}{2}\pi^{(0,1)} > 0$ due to our assumption that $\Pr(\tilde{y}_0^+ \neq y_0^+ \bigvee \tilde{y}_0^- \neq y_0^-) > 0$. Hence, we can do some rearrangement to get:

$$\begin{aligned}
\Pr(W_2) - \Pr(W_1) &\geq \frac{1}{2}(\pi^{(0,0)} + \pi^{(1,1)} + \frac{1}{2}\pi^{(0,1)})G\Big(1 - (\frac{1 - \pi^{(0,1)}}{\pi^{(0,0)} + \pi^{(1,1)} + \frac{1}{2}\pi^{(0,1)}})(\frac{A}{G})\Big) && (106)\\
&= \frac{1}{2}(\pi^{(0,0)} + \pi^{(1,1)} + \frac{1}{2}\pi^{(0,1)})G\Big(1 - (\frac{1 - \pi^{(0,1)}}{\pi^{(0,0)} + \pi^{(1,1)} + \frac{1}{2}\pi^{(0,1)}})(\frac{1-a}{1-\gamma a})\gamma^{-(m-1)})\Big) && (107)\\
&= \frac{1}{2}(\pi^{(0,0)} + \pi^{(1,1)} + \frac{1}{2}\pi^{(0,1)})G\Big(1 - (\frac{1 - \pi^{(0,1)}}{\pi^{(0,0)} + \pi^{(1,1)} + \frac{1}{2}\pi^{(0,1)}})(\frac{\rho'}{q'})(\frac{1-\rho'}{1-q'})^{m-1})\Big) && (108)
\end{aligned}$$

We can make this expression positive by picking $m > 1 + \frac{\log(\rho') - \log(q') + \log(1-\pi^{(0,1)}) - \log(\pi^{(0,0)} + \pi^{(1,1)} + \frac{1}{2}\pi^{(0,1)})}{\log(1-q') - \log(1-\rho')}$.

Thus, the second-max will be advantageous over the first-max as long as m satisfies the following lower bounds:

$$m > 1 + \frac{\log(2) + \log(1 - \rho' + m\rho' + \rho'\frac{1-q'}{q'})}{-\log(1-\rho')} \quad (109)$$

$$m > 1 + \frac{\log(\rho') - \log(q') + \log(1 - \pi^{(0,1)}) - \log(\pi^{(0,0)} + \pi^{(1,1)} + \frac{1}{2}\pi^{(0,1)})}{\log(1-q') - \log(1-\rho')} \quad (110)$$

In fact, in the special case where $\pi^{(1,0)} = 0$, these bounds are "tight" in the sense that an m that violates both bounds will lead to the first-max being advantageous over the second-max. Theorem 2 gives further insight into this dependency on m.

□

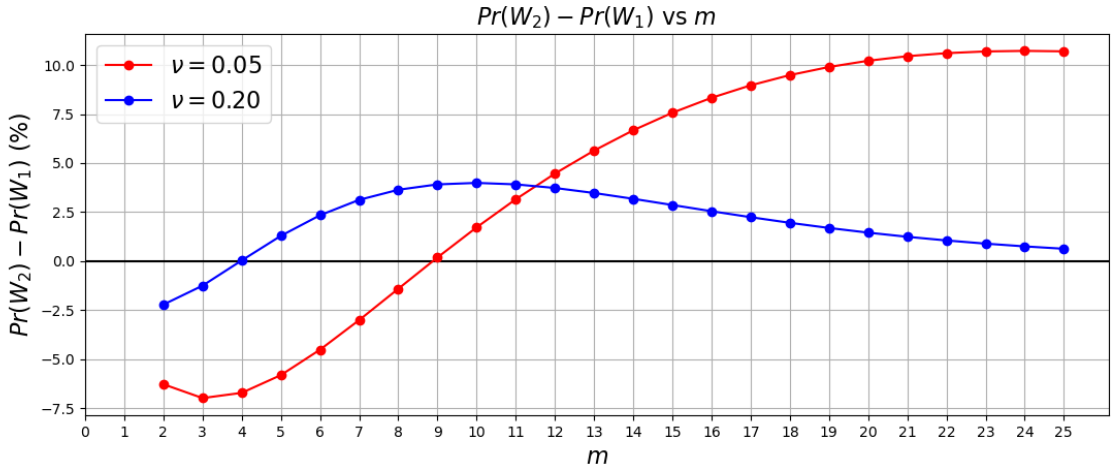

Figure 8. $\Pr(W_2) - \Pr(W_1)$ via Eq. (95) for different values of m, proving by example that first-max can be advantageous for sufficiently small m, and that the point of advantage depends on label-flip probability ν.

13.5. Proof of Theorem 2

Proof. We prove Theorem 2 by example. We give two example settings that fulfill the requirements of Theorem 1 and for which $\Pr(W_2) < \Pr(W_1)$ for sufficiently small m. Furthermore, these two example settings differ only by ν and have different values of m at which the inequality reverses, and so we know that this reversal point depends on ν. Some further examples show that the reversal point also depends on other setting variables such as $\rho, q, \pi^{(0,0)}, \pi^{(1,1)}, \pi^{(0,1)}, \pi^{(1,0)}$, but we limit our analysis to ν for the sake of brevity.

Our two example settings share $\rho = 0.15$, $q = 0.01$, $\pi^{(1,0)} = 0.55^2$, $\pi^{(0,1)} = (1 - 0.55)^2$, and $\pi^{(0,0)} = \pi^{(1,1)} = 0.55 \cdot (1 - 0.55)$. They differ only in their values of ν which are 0.05 and 0.2. We evaluate $\Pr(W_2) - \Pr(W_1)$ via Eq. (95) for multiple values of m under these settings and plot the resulting probability differences in Fig. 8. We see that the claims in Theorem 2 follow from these examples. □