Scalable and Efficient Comparison-based Search without Features Supplementary Material

This supplementary material contains:

1. Proofs of Theorem 1 and Theorem 2.
2. Additional experiments on comparing different embedding techniques.
3. Additional plot of the 2-PCA of the learned embedding in the movie actors face search experiment.

A. Proof of Theorem 1

Proof. Without loss of generality, assume that $\sigma_\varepsilon = 1$. Let y be a binary random variable such that $p(y = 1|\boldsymbol{w}, b, \boldsymbol{x}) = \Phi(\boldsymbol{x}^T\boldsymbol{w} + b)$. Then,

$$\begin{aligned}
&\arg\max_{(\boldsymbol{w},b)\in\mathcal{H}} I[\boldsymbol{x}; y \mid (\boldsymbol{w}, b)] \\
&\quad = \arg\max_{(\boldsymbol{w},b)\in\mathcal{H}} \{1 - \mathbb{E}_{\hat{\boldsymbol{x}}}[H(y \mid \boldsymbol{w}, b, \boldsymbol{x})]\} \qquad (1)\\
&\quad = \arg\min_{(\boldsymbol{w},b)\in\mathcal{H}} \int_{\mathbb{R}^d} H\left[\Phi(\boldsymbol{x}^T\boldsymbol{w} + b)\right]\mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma})d\boldsymbol{x} \\
&\quad = \arg\min_{(\boldsymbol{w},b)\in\mathcal{H}} \int_{\mathbb{R}} H\left[\Phi(t)\right]\mathcal{N}(t; 0, \boldsymbol{w}^T\boldsymbol{\Sigma}\boldsymbol{w})dt \qquad (2)\\
&\quad = \arg\max_{(\boldsymbol{w},b)\in\mathcal{H}} \boldsymbol{w}^T\boldsymbol{\Sigma}\boldsymbol{w}. \qquad (3)
\end{aligned}$$

In (1), we use (2) and the fact that, as the hyperplane is passing through $\boldsymbol{\mu}$,

$$H\left[\int_{\mathbb{R}^d} p(y = 1|\boldsymbol{w}, b, \boldsymbol{x})\mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma})d\boldsymbol{x}\right] = H(1/2) = 1.$$

In (2), we use the fact that $\boldsymbol{x}^T\boldsymbol{w} + b \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{w}^T\boldsymbol{\Sigma}\boldsymbol{w})$, by properties of the Gaussian distribution. Finally, in (3), we start by noting that, for all c_1, c_2 such that $c_1/c_2 > 1$, $H\left[\Phi(c_1 t)\right] \leq H\left[\Phi(c_2 t)\right]$ for all t with equality iff $t = 0$. Hence, if $\tilde{\sigma}^2 > \sigma^2$, then

$$\begin{aligned}
&\int_{\mathbb{R}} H\left[\Phi(t)\right]\mathcal{N}(t; 0, \tilde{\sigma}^2)dt = \int_{\mathbb{R}} H\left[\Phi(\tilde{\sigma}t)\right]\mathcal{N}(t; 0, 1)dt \\
&< \int_{\mathbb{R}} H\left[\Phi(\sigma t)\right]\mathcal{N}(t; 0, 1)dt = \int_{\mathbb{R}} H\left[\Phi(t)\right]\mathcal{N}(t; 0, \sigma^2)dt.
\end{aligned}$$

Therefore, maximizing $\boldsymbol{w}^T\boldsymbol{\Sigma}\boldsymbol{w}$ minimizes the expected entropy of y. □

B. Proof of Theorem 2

For simplicity, we will assume that $d = 1$; Section B.1 explains how to generalize the result to any $d > 1$. Denote by x_t the location of the target object, and let $\mathcal{N}(\hat{x}; \mu_m, \sigma_m)$ be the belief about the target's location after m observations. Without loss of generality, let $\sigma_\varepsilon^2 = 1$. In this case, the updates have the following form.

$$\begin{aligned}
\sigma_{m+1}^2 &= \sigma_m^2 + \beta_m\sigma_m^4, \qquad (4)\\
\mu_{m+1} &= \mu_m + \alpha_m\sigma_m^2 \cdot z_m,
\end{aligned}$$

where $z_m \in \{\pm 1\}$ with $\mathbf{P}(z_m = 1) = \Phi(x_t - \mu_m)$, and

$$\begin{aligned}
\alpha_m &= c/\sqrt{\sigma_m^2 + 1},\\
\beta_m &= -c^2/(\sigma_m^2 + 1),\\
c &= \sqrt{2/\pi}.
\end{aligned}$$

We start with a lemma that essentially states that σ_m^2 decreases as $1/m$.

Lemma 1. *For any initial $\sigma_0^2 > 0$ and for all $m \geq 0$, the posterior variance σ_m^2 can be bounded as*

$$\frac{\min\{0.1, \sigma_0^2\}}{m + 1} \leq \sigma_m^2 \leq \frac{\max\{10, \sigma_0^2\}}{m + 1}.$$

Proof. From (4), we know that

$$\sigma_{m+1}^2 = \left(1 - c^2\frac{\sigma_m^2}{\sigma_m^2 + 1}\right)\sigma_m^2$$

First, we need to show that

$$f(x) = \left(1 - c^2\frac{x}{x + 1}\right)x$$

is increasing on $\mathbf{R}_{>0}$. This is easily verified by checking that

$$f'(x) = \left(1 - c^2\frac{x}{x + 1}\right) + x\left(1 - c^2\frac{1}{(x + 1)^2}\right) \geq 0,$$

for all $x \in \mathbf{R}_{>0}$ Next, we consider the upper bound. Let $b = \max\{10, \sigma_0^2\}$. We will show that $\sigma_m^2 \leq b/(m + 1)$

by induction. The basis step is immediate: by definition, $\sigma_0^2 \leq b$. The induction step is as follows, for $m \geq 1$.

$$\begin{aligned}
\sigma_m^2 &= \left(1 - c^2 \frac{\sigma_{m-1}^2}{\sigma_{m-1}^2 + 1}\right) \sigma_{m-1}^2 \\
&\leq \left(1 - c^2 \frac{b}{b+m}\right) \frac{b}{m} \leq \frac{b}{m+1} \\
&\iff 1 - c^2 \frac{b}{b+m} - \frac{m}{m+1} \leq 0 \\
&\iff b + m - c^2(bm + b) \leq 0 \\
&\iff m(1 - c^2 b) + b(1 - c^2) \leq 1 - b(\underbrace{2c^2 - 1}_{\approx 0.27}) \leq 0,
\end{aligned}$$

where the first inequality holds because $f(x)$ is increasing. The lower bound can be proved in a similar way. □

For completeness, we restate Theorem 2 for $d = 1$.

Theorem 2 (Case $d = 1$)**.** *If the answers follow equation 1, then for any initial μ_0 and $\sigma_0^2 > 0$ and as $m \to \infty$,*

$$\begin{aligned}
\sigma_m^2 &\to 0, \\
\mu_m &\to x_t
\end{aligned}$$

almost surely.

Proof. The first part of the theorem ($\sigma_m^2 \to 0$) is a trivial consequence of Lemma 1. The second part follows from the fact that our update procedure can be cast as the Robbins-Monro algorithm (Robbins & Monro, 1951) applied to $g(\mu) \doteq 2\Phi(x_t - \mu) - 1$, which has a unique root in $\mu = x_t$. Indeed, z_m is a stochastic estimate of the function g at μ_m, i.e., $\mathbf{E}(z_m) = 2\Phi(\mu_m - x_t) - 1$. The remaining conditions to check are as follows.

- the learning rate $\gamma_m = \alpha_m \sigma_m^2$ satisfies

$$\begin{aligned}
\sum_{m=0}^{\infty} \gamma_m &\geq \frac{c \cdot \min\{\sigma_0^2, 0.1\}}{\sqrt{\sigma_0^2 + 1}} \sum_{m}^{\infty} \frac{1}{m} = \infty, \\
\sum_{m=0}^{\infty} \gamma_m^2 &\leq (c \cdot \max\{\sigma_0^2, 10.0\})^2 \sum_{m}^{\infty} \frac{1}{m^2} < \infty
\end{aligned}$$

- $|z_m| \leq 1$ for all m

Almost sure convergence then follows directly from the results derived in (Robbins & Monro, 1951) and (Blum et al., 1954). □

B.1. Extending the proof to $d > 1$

In order to understand how to extend the argument of the proof given above to $d > 1$, the following observation is key: Every query made during a search gives information on the position of the target along a *single* dimension, i.e., the one perpendicular to the bisecting hyperplane. This can be seen, e.g., from the update rule

$$\boldsymbol{\Sigma}_{m+1} = \left(\boldsymbol{\Sigma}_m + \tau \boldsymbol{w}_{i_m, j_m} \otimes \boldsymbol{w}_{i_m, j_m}^T\right)^{-1},$$

which reveals that the precision matrix (i.e., the inverse of the covariance matrix) is affected only in the subspace spanned by $\boldsymbol{w}_{i_m, j_m}$.

Therefore, if we start with $\boldsymbol{\Sigma}_0 = \sigma_0^2 \boldsymbol{I}$, we can (without loss of generality) assume that the search procedure sequentially iterates over the dimensions. At each iteration, the variance shrinks along that dimension only, leaving the other dimensions untouched. Conceptually, we can think of the case $d > 1$ as interleaving d independent one-dimensional search procedures. Each of these one-dimensional searches converges to the corresponding coordinate of the target vector $\boldsymbol{x}_t$, and the variance along the corresponding dimension shrinks to 0.

In general, one should consider the fact that the optimal hyperplane is not always unique, and the chosen hyperplane might not align with the current basis of the space. This case can be taken care of by re-parametrizing the space by using a rotation matrix. However, these technical details do not bring any new insight as to *why* the result holds.

C. Performance of Embedding Methods

We evaluated the quality of the object embedding learned by our embedding technique GAUSSEMB on two real world datasets with crowdsourced triplet comparisons: *Music artists* (Ellis et al., 2002) and *Food* (Wilber et al., 2014).

We compared our model to the state-of-the-art baselines, CKL and t-STE. We measured accuracy—the percentage of satisfied triplets in the learned embeddings on a holdout set using 10-fold cross validation. Since the "true" dimensionality of the feature space is not known a priori, we also vary the dimensionality D of the estimated embedding between 2 and 30. The results are presented in Fig. 1.

Overall, on both datasets, GAUSSEMB showed similar performance to t-STE, correctly modeling between 83% and 85% of triplets, and outperformed CKL. We can conclude that the noise model considered in this paper reflects the real user behaviour on the comparison-like tasks well.

D. Movie Actors Face Search Experiment

We illustrate the results of performing 2-PCA on the learned embedding from the movie actors face search experiment in Fig. 2. It appears that the two principal components align well with gender and race. Note that this embedding was obtained solely from the triplet comparisons collected prior to the experiment (slightly more than 40000 triplets in total).

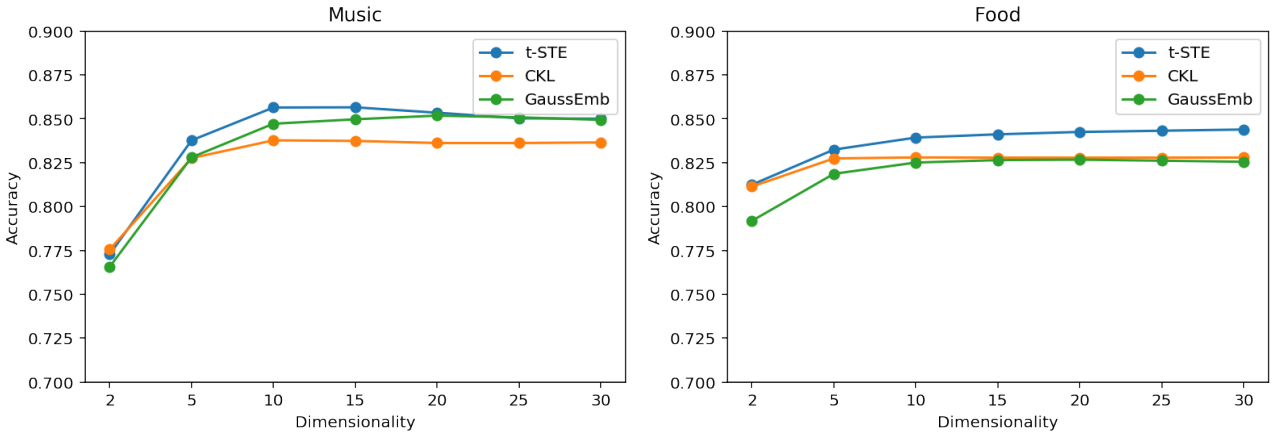

Figure 1. Evaluating embedding methods on two real world datasets: *Music artists*, $n = 400$ music artists with 9090 triplet comparisons after removing repeating and inconsistent triplets, and *Food*, $n = 100$ images of food with 190376 collected unique triplets.

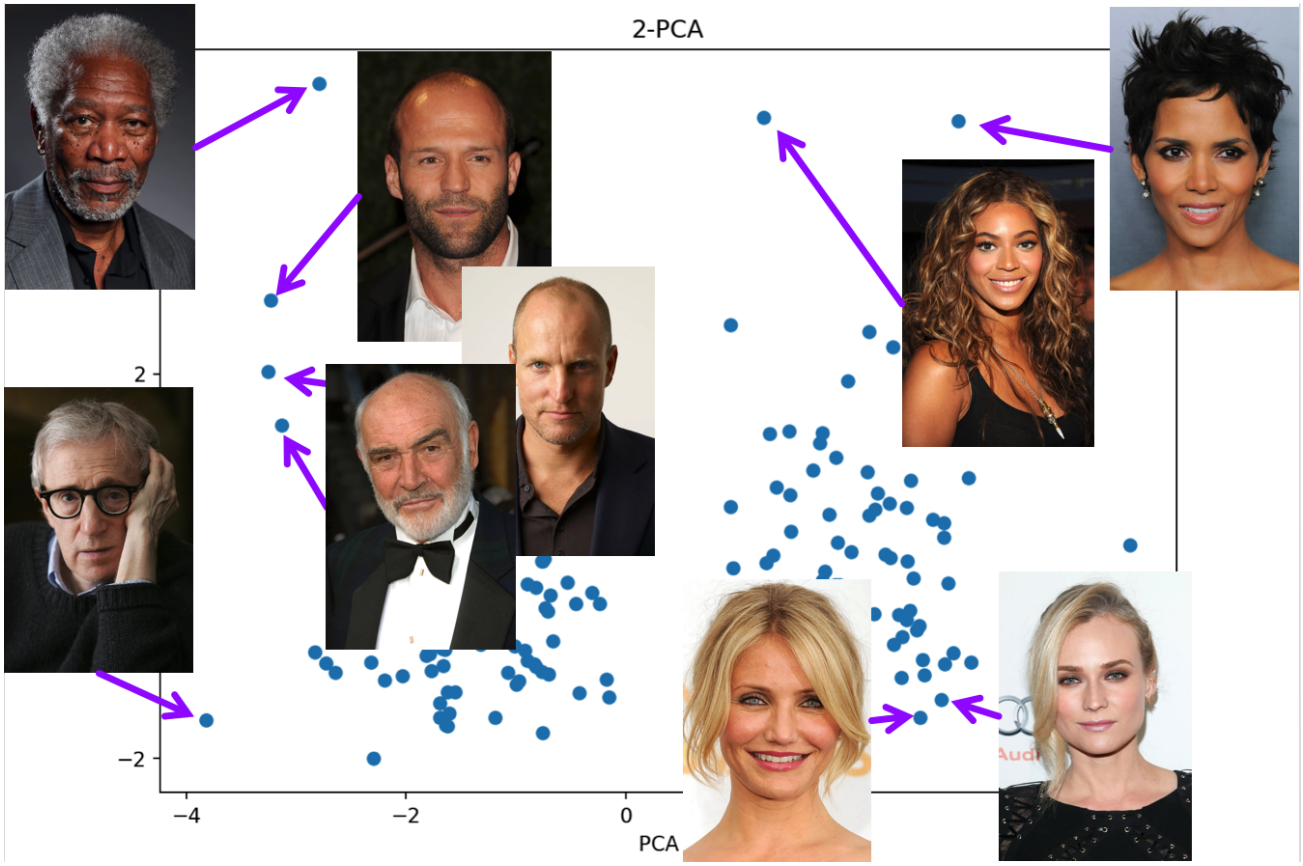

Figure 2. 2-PCA on learned embedding.

References

Blum, J. R. et al. Approximation methods which converge with probability one. *The Annals of Mathematical Statistics*, 25(2):382–386, 1954.

Ellis, D. P., Whitman, B., Berenzweig, A., and Lawrence, S. The quest for ground truth in musical artist similarity. In *ISMIR*. Paris, France, 2002.

Robbins, H. and Monro, S. A stochastic approximation method. *The annals of mathematical statistics*, 22(3): 400–407, 1951.

Wilber, M. J., Kwak, I. S., and Belongie, S. J. Cost-effective hits for relative similarity comparisons. In *Second AAAI conference on human computation and crowdsourcing*, 2014.