A WASSERSTEIN BARYCENTER INDEX FOR EFFICIENT RANK FUSION

Anonymous authors
Paper under double-blind review

ABSTRACT

Information retrieval pipelines routinely aggregate diverse base rankers, e.g., lexical, sparse, and dense, to exploit their complementary strengths. Classical fusion methods score query-document pairs pointwise and, therefore, typically ignore the relationships among retrieved documents in the embedding space. As a result, they treat near-duplicate clusters and genuinely complementary subsets alike. In this work, we recast rank fusion as a Restricted Support Wasserstein Barycenter (RSWB) problem that jointly optimizes the retrieved set and its ranking order. We then fix the support set and cast RSWB as a set-function objective. We further show that it is monotone and restricted submodular, yielding approximation guarantees for greedy maximization. However, the greedy algorithm requires an exhaustive search over the entire corpus during the marginal-gain computation step. Therefore, we approximate the marginal gain using an approximate nearest neighbor (ANN) index amenable to dot-product approximation. This reduces each greedy step to an approximate ANN search, enabling sublinear retrieval via off-the-shelf indices such as DiskANN and Faiss. Our experiments show that the proposed method achieves favorable trade-offs among ranking accuracy, corpus geometry, and latency.

1 INTRODUCTION

In information retrieval (IR), different rankers such as lexical (Robertson et al., 2009), sparse (Formal et al., 2021b;a), dense (Khattab et al., 2020; Karpukhin et al., 2020; Santhanam et al., 2022b;a), and even models within the same family exhibit complementary strengths and weaknesses: queries well handled by one are often missed by another. To exploit this complementarity, IR pipelines often aggregate the outputs of these rankers (called as base rankers) into a unified ranking (Fox et al., 1993; Lee, 1997; Cormack et al., 2009; Bruch et al., 2023; Chen et al., 2021). They operate on ranker outputs, either relevance scores (Fox et al., 1993, CombSUM), (Lee, 1997, CombMNZ) or ranks (Cormack et al., 2009, RRF), and remain agnostic to the underlying models. Rank fusion has several applications, *e.g.*, web search (Cormack et al., 2009; Liu et al., 2025; Mourao et al., 2013), question answering (Rackauckas, 2024), and hybrid retrieval systems (Gao et al., 2021), where the latter typically blend sparse and dense architectures and often rely on rank fusion to merge their outputs (Prajapati, 2026).

Limitations of prior work and the challenge At the outset, the above rank fusion methods (Fox et al., 1993; Lee, 1997; Bruch et al., 2023; Chen et al., 2021) are predominantly arithmetic aggregators: they combine the relevance scores of query–corpus pairs in a pointwise manner, ignoring how the retrieved documents relate to one another in embedding space. As a result, a cluster of near-duplicate documents and a set of genuinely complementary documents are treated with equal likelihood and the pipeline has no mechanism to account for semantic overlap. Addressing this limitation calls for a shift from pointwise scoring to collective retrieval, where relevance is defined over the set of documents via a non-additive objective, rather than through an individual score associated with each corpus item in isolation.

1.1 OUR CONTRIBUTIONS

To address this challenge, we propose a novel Wasserstein barycenter-based rank fusion method (WBFUSION). We first cast rank fusion as a restricted-support Wasserstein barycenter (RSWB) problem, and then develop an approximate nearest neighbor (ANN) based set retrieval framework to efficiently optimize the RSWB objective. Our main contributions are as follows:

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Rank fusion as restricted support Wasserstein barycenter problem Given a query, we characterize each ranking as a probability distribution over the corpus via softmax normalization of the relevance scores. Then we cast the rank fusion problem as computing a restricted-support Wasserstein barycenter (Agueh et al., 2011; Cuturi et al., 2014) (RSWB) over the base rankers, which jointly optimizes for the fused distribution and its support under a cardinality constraint. Here, the underlying transportation cost is built from the pairwise Euclidean distances between corpus embedding vectors. This explicitly captures inter-document relationships in the embedding space and yields a smooth fused ranking.

Reducing the rank fusion problem as a set function optimization In the RSWB problem described above, the support of the barycenter specifies the subset of corpus items to be retrieved. Hence, fixing the RSWB objective to a candidate subset of the corpus induces a set function, which captures the relevance between the query and the corpus subset. We show that this set function is monotone and restricted γ-weakly submodular (Das et al., 2011; Elenberg et al., 2017), which yields approximation guarantees for greedy selection.

ANN for efficient greedy maximization Efficient retrieval at scale relies on approximate nearest neighbor (ANN) search, which in turn requires a structure amenable to indexing. However, the marginal gain, which serves as the relevance score here, does not naturally admit such a form. To overcome this, we derive a surrogate under a softmax model for the barycenter weights and show that it admits a uniform finite-dimensional inner product approximation. This reduces each greedy step to a maximum inner product search (MIPS), enabling us to index corpus items once using off-the-shelf ANN systems (e.g., LSH (Shrivastava et al., 2014; 2015), DiskANN (Jayaram Subramanya et al., 2019), Faiss (Douze et al., 2025)) and query the index at each iteration in sublinear time without computing any Sinkhorn iterations (Cuturi, 2013). Our experiments on several real datasets show that the performance of our method is better than the baselines.

2 PRELIMINARIES

Notations Given an integer N, $[N] = \{1, \cdots, N\}$. We represent a query and a corpus using single-vector embeddings $\boldsymbol{q} \in \mathbb{R}^d$ and $\boldsymbol{x} \in \mathbb{R}^d$, respectively. We denote the set of N corpus items as $\{\boldsymbol{x}_s \,|\, s \in D\}$ with $D = [N]$ (we often refer to a corpus item by its id s rather than $\boldsymbol{x}_s$). For every $S \subseteq D$, let Δ_S define the set of all probability distributions with support in $S \subseteq D$, i.e., $\Delta_S \coloneqq \{\boldsymbol{a} \in \mathbb{R}^N_{\geq 0} : \boldsymbol{a}[s'] = 0 \text{ for all } s' \notin S, \sum_{s \in S} \boldsymbol{a}[s] = 1\}$. Let $\boldsymbol{\mu}_1, \cdots, \boldsymbol{\mu}_B$ denote B different probability distributions in Δ_D. For any two probability distributions $\boldsymbol{\mu}, \boldsymbol{\nu} \in \Delta_D$, let $\Pi(\boldsymbol{\mu}, \boldsymbol{\nu})$ denote the set of all joint distributions with right and left marginals as $\boldsymbol{\mu}$ and $\boldsymbol{\nu}$, respectively. Formally, we define $\Pi(\boldsymbol{\mu}, \boldsymbol{\nu}) \coloneqq \{\boldsymbol{\pi} \in \mathbb{R}^{N \times N}_{\geq 0} : \boldsymbol{\pi}^\top \mathbf{1} = \boldsymbol{\mu}, \boldsymbol{\pi}\mathbf{1} = \boldsymbol{\nu}\}$.

Definition 1 (Entropic Optimal Transport)**.** Given two probability distributions $\boldsymbol{\mu}, \boldsymbol{\nu} \in \Delta_D$, a ground cost matrix $\boldsymbol{C} \in \mathbb{R}^{N \times N}_{\geq 0}$, and $\lambda \geq 0$, the entropic optimal transport (OT) cost (Cuturi, 2013; Peyré et al., 2019) is defined as

$$\mathcal{W}(\boldsymbol{\nu}, \boldsymbol{\mu}) \coloneqq \min_{\boldsymbol{\pi} \in \Pi(\boldsymbol{\mu}, \boldsymbol{\nu})} \sum_{1 \leq i,j \leq N} \boldsymbol{C}[i,j]\boldsymbol{\pi}[i,j] + \lambda \sum_{1 \leq i,j \leq N} \boldsymbol{\pi}[i,j]\,(\ln \boldsymbol{\pi}[i,j] - 1)\,. \tag{1}$$

Definition 2 (Doubly regularized entropic Wasserstein barycenter)**.** Given $\lambda, \tau \geq 0$ and the probability distributions $\boldsymbol{\mu}_1, \cdots, \boldsymbol{\mu}_B \in \Delta_D$, the doubly regularized entropic Wasserstein barycenter objective $W : \Delta_D \to \mathbb{R}$ is defined as (Cuturi et al., 2014; Benamou et al., 2015; Chizat, 2025)

$$W(\boldsymbol{\nu}, \{\boldsymbol{\mu}_r \,|\, r \in [B]\}) \coloneqq \textstyle\sum_{r=1}^B \beta_r \mathcal{W}(\boldsymbol{\nu}, \boldsymbol{\mu}_r) - \tau\, \mathrm{Ent}(\boldsymbol{\nu}), \qquad \text{for all } \nu \in \Delta_D, \tag{2}$$

where $\mathrm{Ent}(\boldsymbol{\nu}) \coloneqq -\sum_{i=1}^N \boldsymbol{\nu}[i] \ln \boldsymbol{\nu}[i]$, for all $\boldsymbol{\nu} \in \Delta_D$, and β_r is the weight assigned for the r^{th} base ranker with $\sum_{r=1}^B \beta_r = 1, \beta_r > 0$, for all $r \in [B]$.

Definition 3 (Monotonicity and restricted submodularity)**.** Let $F : 2^D \to \mathbb{R}$ be a set function defined on the ground set D. For any $S \subseteq D$ and $e \in D \setminus S$, let $F(e \mid S) \coloneqq F(S \cup \{e\}) - F(S)$. Let $A \subseteq B \subseteq D$ be arbitrary. The function F is **monotone** if $F(A) \leq F(B)$. It is **submodular** if given arbitrary $e \in D \setminus B$ it satisfies the diminishing returns property, i.e., $F(e \mid A) \geq F(e \mid B)$. Finally, F is **restricted γ-weakly submodular** (Das et al., 2011) with respect to an integer $k \leq |D|$ and $\gamma \in (0, 1]$ if it is monotone and satisfies $\sum_{e \in S} F(e \mid L) \geq \gamma\big(F(L \cup S) - F(L)\big)$ for all disjoint $L, S \subseteq D$ such that $|L| + |S| \leq k$.

3 PROPOSED APPROACH

Our goal Fix a query $\boldsymbol{q}$. Given a corpus $\{\boldsymbol{x}_s \mid s \in D\}$ and B base rankers $\mathcal{R}_1, \ldots, \mathcal{R}_B$, our goal is twofold: (i) compute a fused ranker $\mathcal{R}_{\text{fuse}}$ that aggregates the base rankers' scores while explicitly accounting for inter-item relationships in the embedding space, and (ii) retrieve a cardinality-constrained subset $S \subseteq D$ that optimizes the resulting fusion objective.

Overview We propose a novel Wasserstein barycenter-based rank fusion method (WBFUSION) to solve the above problem. It consists of five steps:

(1) Given a query, we represent each base ranker as a distribution over the corpus D via softmax normalization of its relevance scores, and formulate rank fusion as a restricted-support doubly regularized entropic Wasserstein barycenter (RSWB) problem over a support Δ_S with $S \subseteq D$, $|S| \leq k$. This induces a set function $F(\boldsymbol{q}, S)$ that measures the collective relevance of S to $\boldsymbol{q}$.

(2) We show that $F(\boldsymbol{q}, \bullet)$ is monotone and restricted γ-weakly submodular, enabling a greedy algorithm with approximation guarantees.

(3) At each iteration, the greedy algorithm selects the corpus item with the largest marginal gain per query $\boldsymbol{q}$. We cast it as a Top-1 retrieval problem and derive a finite-dimensional dot-product approximation of the marginal gain to enable sublinear-time evaluation.

(4) We train the dot-product model and build an ANN index over the corpus.

(5) At each greedy iteration, we query the index using an embedding of $\boldsymbol{q}$ and the current support S to retrieve an item with an approximately high marginal gain. This replaces the exhaustive corpus scan with a sublinear-time lookup. We repeat this $k' \ll N$ times and follow it up with exact-greedy filtering to k items. These items are then reranked to give the final top-k list.

3.1 NOVEL WASSERSTEIN BARYCENTER PROBLEM FOR RANK FUSION

Setup We are given B base rankers $\mathcal{R}_1, \ldots, \mathcal{R}_B$ and a semantic dissimilarity function $\rho : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}_{\geq 0}$. Each base ranker $\mathcal{R}_r$ assigns $\text{score}_r(\boldsymbol{q}, \boldsymbol{x}) \in \mathbb{R}$ to every query-corpus pair $(\boldsymbol{q}, \boldsymbol{x})$. Here, $\rho(\boldsymbol{x}_s, \boldsymbol{x}_t)$ measures the dissimilarity between two corpus items in the embedding space, *e.g.*, $\rho(\boldsymbol{x}_1, \boldsymbol{x}_2) = \|\boldsymbol{x}_1 - \boldsymbol{x}_2\|$ or $\rho(\boldsymbol{x}_1, \boldsymbol{x}_2) = 1 - \cos(\boldsymbol{x}_1, \boldsymbol{x}_2)$ for distance-normalized embeddings. The cost matrix $\boldsymbol{C} \in \mathbb{R}^{N \times N}$ is defined as $\boldsymbol{C}[s, t] = \rho(\boldsymbol{x}_s, \boldsymbol{x}_t)$.

From score to distribution Both the OT cost (1) and Wasserstein barycenter objective (2) operate on probability distributions. We convert ranker scores to probability distributions via softmax on z-score-normalized scores (Details are in Appendix G.4), denoted as $\text{score}_r(\boldsymbol{q}, \boldsymbol{x}_s)$

$$\boldsymbol{\mu}_r(\boldsymbol{q}, D)[s] \coloneqq \frac{\exp\left(\text{score}_r(\boldsymbol{q}, \boldsymbol{x}_s)\right)}{\sum_{t \in D} \exp\left(\text{score}_r(\boldsymbol{q}, \boldsymbol{x}_t)\right)}, \qquad \text{for all } s \in D. \tag{3}$$

Thus, for every fixed query $\boldsymbol{q}$, the B different base rankers induce B different probability distributions $\boldsymbol{\mu}_1(\boldsymbol{q}, D), \ldots, \boldsymbol{\mu}_B(\boldsymbol{q}, D) \in \Delta_D$.

Posing rank fusion as a restricted-support Wasserstein barycenter problem Fix a query $\boldsymbol{q}$ and base rankers $\mathcal{R}_1, \ldots, \mathcal{R}_B$. We fuse the base rankers through a doubly regularized entropic Wasserstein barycenter on the distributions $\boldsymbol{\mu}_1(q, D), \cdots, \boldsymbol{\mu}_B(q, D)$. Given a budget k on the number of retrieved items, we cast rank fusion as retrieving (or selecting) a subset of corpus items $S \subseteq D$ with $|S| = k$ and a distribution $\boldsymbol{\nu} \in \Delta_S$, that minimizes the doubly regularized entropic Wasserstein barycenter objective (2) over all supports of size k:

$$\min_{\boldsymbol{\nu}, S \subseteq D} W(\boldsymbol{\nu}, \{\boldsymbol{\mu}_r(\boldsymbol{q}, D) \mid r \in [B]\}) \quad \text{subject to: } \boldsymbol{\nu} \in \Delta_S,\ |S| \leq k. \tag{4}$$

Here, the support of $\boldsymbol{\nu}$ is restricted to Δ_S, the simplex defined over S, the subset of corpus items to be retrieved. We refer to the above optimization as the restricted-support doubly regularized entropic Wasserstein barycenter (RSWB) problem. For a query $\boldsymbol{q}$ and a fixed subset $S \subseteq D$, let $\boldsymbol{\nu}^*(\boldsymbol{q}, S)$ denote the barycenter, which is the optimal solution of (4) on allowed support S. The solution of the joint optimization is then obtained by solving $\min_{S \subseteq D,\, |S| \leq k} W\big(\boldsymbol{\nu}^*(\boldsymbol{q}, S), \{\boldsymbol{\mu}_r \mid r \in [B]\}\big)$, where the argmin is S^*.

Fused ranking The *fused ranking* is then characterized by $\boldsymbol{\nu}^*(\boldsymbol{q}, S^*)$ for a given query $\boldsymbol{q}$: the barycenter optimal distribution for the RSWB problem (4) supported on the optimal subset S^* whose probability mass on each item $s \in S^*$ defines its relevance score, *i.e.*, $\text{rel}(\boldsymbol{q}, \boldsymbol{x}_s) = \boldsymbol{\nu}^*(\boldsymbol{q}, S^*)[s]$.

Pointwise rank fusion vs. RSWB-based rank fusion Existing rank fusion methods (Fox et al., 1993; Lee, 1997) aggregate scores pointwise amongst base rankers for a given query, corpus pair and are agnostic to inter-item relationships, causing two caveats:

(1) If two similar items receive divergent scores across rankers, the fused ranking inherits the same discontinuity in the embedding space.
(2) High scores on semantically similar items are treated as independent evidence, leaving distinct regions uncovered.

Our RSWB formulation resolves both issues by forming a consensus under the cost $\rho(\boldsymbol{x}_s, \boldsymbol{x}_t)$, so that mass on $\boldsymbol{x}_s$ is interchangeable with mass on $\boldsymbol{x}_t$ at a price proportional to their semantic distance. This yields two advantages:

(1) The fused weight $\boldsymbol{\nu}^*(\boldsymbol{q}, S)[s]$ aggregates each ranker's mass on $\boldsymbol{x}_s$ together with mass across its entire semantic neighborhood, encoding the relational geometry directly into the fused score and guaranteeing smoothness in the embedding space (Dognin et al., 2019, Proposition 1).
(2) The retrieved set contains at least one representative from each highly ranked cluster of similar items, rather than near-duplicates from the same cluster.

Set function based relevance score This naturally leads us to the set function associated with the RSWB objective and the corresponding support selection problem, equivalent to the optimization (4):

$$\max_{S \subseteq D : |S| \leq k} F(\boldsymbol{q}, S) \text{ where, } F(\boldsymbol{q}, S) \coloneqq -W(\boldsymbol{\nu}^*(\boldsymbol{q}, S), \{\boldsymbol{\mu}_r(\boldsymbol{q}, D) \,|\, r \in [B]\}) \tag{5}$$

This viewpoint makes the subset-selection problem intrinsic to our proposed rank fusion approach. For a query $\boldsymbol{q}$, the relevance of a retrieved set S is measured by the value of $F(\boldsymbol{q}, S)$.

3.2 Set theoretic characterization of $F(\boldsymbol{q}, S)$

Having recast the rank fusion objective in (4) as the set-based relevance score F in (5), we now turn to maximizing $F(\boldsymbol{q}, S)$ over S for a given query $\boldsymbol{q}$. To guide the design of an efficient optimization algorithm, we first characterize F in terms of monotonicity and submodularity. We first begin with a toy setting where $\boldsymbol{x}_s \in \mathbb{R}$, *i.e.*, $d = 1$ and RSWB objective $W(\boldsymbol{\nu}, \{\boldsymbol{\mu}_r(\boldsymbol{q}, D) \,|\, r \in [B]\})$ does not contain any entropy term, *i.e.*, we have $\tau = 0$ and $\lambda = 0$, in (2). $F(q, \bullet)$ turns out to be monotone and submodular in this case.

Lemma 4 (Monotonicity and submodularity for special case of scalar embeddings)**.** Assume (1) the corpus items are embedded in real number space, *i.e.*, $\boldsymbol{x}_\bullet \in \mathbb{R}$; (2) $\rho(\boldsymbol{x}_s, \boldsymbol{x}_t) = |\boldsymbol{x}_s - \boldsymbol{x}_t|$; and, (3) $\tau = 0$ in (2) and $\lambda = 0$ in (1). Then, $F(\boldsymbol{q}, S)$ is monotone and submodular in $S \subseteq D$.

Beyond scalar embeddings, however, exact submodularity is no longer guaranteed, so we instead work with the weaker notion of restricted γ-weak submodularity (Definition 3). To estimate submodularity ratio for restricted-size subsets in the general case, we analyze the strong convexity and smoothness of the function $W(\boldsymbol{\nu}, \{\boldsymbol{\mu}_r(\boldsymbol{q}, D) \,|\, r \in [B]\})$ induced by the the RSWB objective, and follow an approach similar to Elenberg et al. (2017) to bound the submodularity ratio. We formally state the result in the following theorem.

Theorem 5 (Monotonicity and restricted γ-weakly submodularity)**.** Let k be an integer, $\rho(\boldsymbol{x}_s, \boldsymbol{x}_t)$ be any valid function used to build the cost matrix $\boldsymbol{C}[s, t] = \rho(\boldsymbol{x}_s, \boldsymbol{x}_t)$ and the cost-oscillation δ be defined as $\delta \coloneqq \max_{s,t,s',t' \in D} |\boldsymbol{C}[s, t] - \boldsymbol{C}[s', t']|$ with $\gamma_k \coloneqq \exp\left(-2\delta/(\tau + \lambda)\right)$. Then $F(\boldsymbol{q}, S)$ is monotone non-decreasing in S and is restriced submodular on all sets S with $|S| \leq k$ with submodularity ratio γ_k.

Note that the γ_k gives a lower bound on the submodularity ratio of the set function $F(\boldsymbol{q}, \cdot)$ induced by the RSWB objective. However, in most real datasets (especially in Information Retrieval), the empirical submodularity ratio is much better than the theoretical lower bound, and the set function is submodular in practice, as shown in Appendix G. Also, the cost oscillation δ is the maximum variation in pairwise distances among the corpus embeddings. Assuming the embeddings to be Gaussian, we show that the cost oscillation is small with high probability in Appendix D.

Greedy maximization Since $F(\boldsymbol{q}, S)$ is monotone and restricted submodular (Theorem 5), we can apply a greedy procedure for the subset selection, which admits a constant-factor approximation guarantee (Das et al., 2011; Elenberg et al., 2017; Gurumoorthy et al., 2019; Manupriya et al., 2024). The greedy procedure is an iterative algorithm, where at each iteration t, we add that element s_t into S_{t-1} which maximizes the marginal gain $F(s \,|\, S_{t-1}, \boldsymbol{q}) = F(\boldsymbol{q}, S_{t-1} \cup s) - F(\boldsymbol{q}, S_{t-1})$ (We define $F(s \,|\, \emptyset, \boldsymbol{q}) = F(\boldsymbol{q}, \{s\})$. Evaluation of the marginal gain $F(s \,|\, S_{t-1}, \boldsymbol{q})$ requires computing $F(\boldsymbol{q}, \bullet)$ on sets S_{t-1} and $S_{t-1} \cup \{s\}$. This, in turn, needs the optimal barycenter $\boldsymbol{\nu}^*(\boldsymbol{q}, S_{t-1})$ on a fixed support $S_{t-1} \subseteq D$. We do this using projected gradient descent (PGD) (Details in Appendix D).

Bottlenecks with greedy algorithm Each greedy step (line 3) must evaluate $F(s \mid S_{t-1}, \boldsymbol{q})$ for every candidate $s \in D \setminus S_{t-1}$, i.e., $O(N)$ calls to Algorithm 5. One such call on a support of size t performs B Sinkhorn solves per projected-gradient iteration, and one such Sinkhorn solve costs $O(tN)$, hence costing $O(BtN)$. The overall complexity is $O(BN^2k^2)$, which quickly becomes intractable at large corpus sizes.

Algorithm 1 Greedy selection: Greedy(D, k)

Require: Query $\boldsymbol{q}$, Information needed to compute $F(\boldsymbol{q}, \bullet)$: (1) rankers $\{\boldsymbol{\mu}_r \mid r \in [B]\}$, (2) λ, (3) τ, budget $k \leq N$

1: Initialize $S_0 \leftarrow \emptyset$
2: **for** $t = 1, \ldots, k$ **do**
3: $\quad s_t \leftarrow \operatorname{argmax}_{s \in D \setminus S_{t-1}} F(s \mid S_{t-1}, \boldsymbol{q})$
4: $\quad S_t \leftarrow S_{t-1} \cup \{s_t\}$
5: **return** S_k

3.3 Marginal Gain Approximation Using Dot Product

Marginal gain maximization as top-1 retrieval problem We seek to avoid the aforementioned bottleneck by working around the step of exhaustive search in line 3 of Algorithm 1 during marginal gain maximization. To this end, we reinterpret the marginal gain maximization task— $\operatorname{argmax}_{s \in D \setminus S_{t-1}} F(s \mid \boldsymbol{q}, S_{t-1})$— as a retrieval problem, where the goal is "Given a query and the subset S_{t-1}, retrieve the top-1 item $s = s_t$ from the corpus D under the relevance score $F(s \mid S, \boldsymbol{q})$." We then seek to design an index over D tailored to this "relevance score" $F(s \mid S_{t-1}, \boldsymbol{q})$, enabling the selection of s_t in time sublinear in $|D|$ without using the costly Sinkhorn solver.

Challenges **(1)** Standard indexing methods (like LSH (Charikar, 2002; Shrivastava et al., 2014; 2015), DiskANN (Jayaram Subramanya et al., 2019), Faiss (Douze et al., 2025) etc.) are built upon Euclidean relevance measures— such as dot product, cosine similarity, and ℓ_1 or ℓ_2 distances. In contrast, $F(s \mid S, \boldsymbol{q})$ combines the item s and the query $\boldsymbol{q}$ in a complex manner, via the minimization of the RSWB objectives over different supports. As a result, they cannot be readily expressed as the above Euclidean relevance measures, which makes them incompatible with existing indexing methods (Douze et al., 2025; Jayaram Subramanya et al., 2019). **(2)** At each iteration, the set S changes, and so does the relevance score $F(s \mid S, \boldsymbol{q})$. This makes indexing challenging, since along with query $\boldsymbol{q}$, the set S is not known in advance and the design has to be compatible with any possible choice of S.
(3) In pairwise measures such as cosine similarity or dot products, the relevance $\text{sim}(\boldsymbol{q}, \boldsymbol{x})$ depends only on the query corpus pair and is unaffected by the rest of the corpus. In contrast, $F(s \mid \boldsymbol{q}, S)$ captures signals from the entire corpus: the query $\boldsymbol{q}$ determines the base rankers' distributions $\mu_1(\boldsymbol{q}, D), \ldots, \mu_B(\boldsymbol{q}, D)$, which in turn interact with every support element $\boldsymbol{x}_c$ through the OT plan.

Existing works (Carbonell et al., 1998; Nair et al., 2026) also approach information retrieval as subset selection. However, Carbonell et al. (1998) do not address the problem of efficient indexing for marginal-gain maximization. Nair et al. (2026) maximize a facility location objective using multi-vector retrieval with indexing, but their settings differ fundamentally from ours: the marginal gain of the facility location function admits an index-friendly decomposition, whereas the RSWB objective involves joint optimization over the support and the transport plan, precluding such a decomposition.

Outline of our approach We tackle the above challenges by decoupling $F(s \mid S, \boldsymbol{q})$ into "(q, S) dependent" and "$\boldsymbol{x}_s$ dependent" components, so that the variations of the subset S can be absorbed by the query, while the corpus remains invariant to the choice of $\boldsymbol{q}, S$, which is key for indexing.

We achieve this goal in three steps: **(1)** We use a KL-based surrogate to approximate the marginal gain $F(s \mid S, \boldsymbol{q}) = F(\boldsymbol{q}, S \cup \{s\}) - F(\boldsymbol{q}, S)$ by analyzing the RSWB objective under small perturbations (Details are in Appendix D). **(2)** We come up with features $\boldsymbol{\zeta}(\boldsymbol{q}, S)$ and $\boldsymbol{\sigma}(\boldsymbol{x}_s)$ whose dot-product well approximates the marginal-gain surrogate, computed in step **(1)**. **(3)** We learn these features using highly expressive neural networks to directly approximate $F(s \mid S, \boldsymbol{q})$ as a learnable dot product with precomputable "(q, S) dependent" and "$\boldsymbol{x}_s$ dependent" components.

Approximating $F(s \mid S, \boldsymbol{q}) \approx \boldsymbol{\zeta}_\theta(\boldsymbol{q}, S)^\top \boldsymbol{\sigma}_\theta(\boldsymbol{x}_s)$ The KL based surrogate for $F(s \mid S, \boldsymbol{q})$ does not readily admit an index-amenable decomposition into $(\boldsymbol{q}, S)$-dependent and $\boldsymbol{x}_s$-dependent terms. To overcome this, we build a neural model that approximates the barycenter $\nu_\theta(\boldsymbol{q}, S) \approx \boldsymbol{\nu}^*(\boldsymbol{q}, S)$ for a given support S. We build it upon a softmax neural network NN_θ using three neural networks u_θ, f_θ, and g_θ on the query $\boldsymbol{q}$, the corpus vectors $\{\boldsymbol{x}_c : c \in S\}$ and $\boldsymbol{x}_s$, the candidate respectively. Then, we perform a simple inner product and perform softmax normalization on $\text{NN}_\theta(\boldsymbol{q}, S, \boldsymbol{x}_s)$ as follows:

$$\nu_\theta(\boldsymbol{q}, S)[i] \propto e^{\text{NN}_\theta(\boldsymbol{q}, S, \boldsymbol{x}_i)},\ \text{NN}_\theta(\boldsymbol{q}, S, \boldsymbol{x}_i) := \left\langle u_\theta(\boldsymbol{q}) - \textstyle\sum_{c \in S} f_\theta(\boldsymbol{x}_c),\ g_\theta(\boldsymbol{x}_i) \right\rangle,\ i \in S \quad (6)$$

Algorithm 2 WBFusion (Indexing)

Require: Queries $\mathcal{Q}$, corpus D, ranker distributions $\{\mu_r(q, D)\}$, Regularizers λ, τ, max set size k.
1: Gather ground truth S_k for $\boldsymbol{q} \in \mathcal{Q}$ (Algorithm 1).
2: Learn $\boldsymbol{\zeta}_\theta, \boldsymbol{\sigma}_\theta$ on $\{(\boldsymbol{q}, S, s, F(s \mid S, \boldsymbol{q}))\}$.
3: Build ANN index $\mathcal{I}$ using $\boldsymbol{\sigma}_\theta$
return Index $\mathcal{I}$, encoders $\boldsymbol{\zeta}_\theta, \boldsymbol{\sigma}_\theta$

Algorithm 3 Reranker training

1: Train reranker score_ϕ by minimizing a pairwise ranking loss on ground truth data $(\boldsymbol{q}, c_+, c_-)$.

Algorithm 4 WBFusion (Efficient retrieval)

Require: Query $\boldsymbol{q}$, index $\mathcal{I}$, final support budget k, pruned set budget k', reranker score_ϕ.
1: Initialize $S_0 \leftarrow \emptyset$
2: `# select pruned set using ANN`
3: **for** $t = 1, \cdots, k'$ **do**
4: $\quad s_t \leftarrow \text{Probe}(\mathcal{I}, \boldsymbol{\zeta}_\theta(\boldsymbol{q}, S_{t-1}))$
5: $\quad S_t \leftarrow S_{t-1} \cup \{s_t\}$
6: $S_k \leftarrow$Greedy$(S_{k'}, k)$ (Algorithm 1)
7: Rerank S_k using score_ϕ
return S_k

Using this softmax neural parametrization $\nu_\theta(\boldsymbol{q}, S)$ to approximate $\boldsymbol{\nu}^*(\boldsymbol{q}, S)$ and using the KL based surrogate for the marginal gain $F(s \mid S, \boldsymbol{q})$, we show that the marginal gain admits a uniform finite-dimensional dot-product approximation $F(s \mid S, \boldsymbol{q}) \approx \boldsymbol{\zeta}_\theta(\boldsymbol{q}, S)^\top \boldsymbol{\sigma}_\theta(\boldsymbol{x}_s)$. We also derive an end-to-end approximation guarantee as a corollary to this.

Theorem 6 (Dot-product approximation of marginal gains)**.** Let $2 \leq k \leq N$. Suppose the three embeddings in (6) take values in $\mathbb{R}^d$ and have uniformly bounded norms. Suppose also that, $\nu_\theta(\boldsymbol{q}, \cdot)$ well approximates the true barycenter, ie $\max_{\boldsymbol{q}, i \in S} |\nu_\theta(\boldsymbol{q}, S) - \boldsymbol{\nu}^*(\boldsymbol{q}, S)| \leq \epsilon$. Suppose both the cost oscillation δ and the entropic regularizer λ be small, specifically $\lambda = O(\epsilon)$ and $\delta = O(\epsilon)$. Then, for every $\eta > 0$, there are finite-dimensional representations $\boldsymbol{\zeta}_\theta(\boldsymbol{q}, S)$ and $\boldsymbol{\sigma}_\theta(\boldsymbol{x}_s)$ such that:

$$|F(s \mid S, \boldsymbol{q}) - \boldsymbol{\zeta}_\theta(\boldsymbol{q}, S)^\top \boldsymbol{\sigma}_\theta(\boldsymbol{x}_c)| = O(\epsilon) + O(\eta), \qquad \text{for } S \neq \emptyset. \tag{7}$$

3.4 Building ANN Index

Training $\boldsymbol{\zeta}_\theta(\boldsymbol{q}, S), \boldsymbol{\sigma}_\theta(\boldsymbol{x}_c)$ We gather training data containing marginal gains for F, *i.e.*, $\{(\boldsymbol{q}, S, s, F(s \mid S, \boldsymbol{q}))\}$ computed once offline via Algorithm 5. Following Theorem 6, we design and train the parametric models $\boldsymbol{\zeta}_\theta(\boldsymbol{q}, S), \boldsymbol{\sigma}_\theta(\boldsymbol{x}_c)$ using a ranking loss with margin-based triplet regularization (Appendix F).

Indexing In the dot-product based approximation (7), the corpus embedding $\boldsymbol{\sigma}_\theta(\boldsymbol{x}_c)$ depends only on the candidate item whereas $\boldsymbol{q}$ and S influence only the state embedding $\boldsymbol{\zeta}_\theta(\boldsymbol{q}, S)$. Since $\boldsymbol{\sigma}_\theta(\boldsymbol{x}_c)$ remains invariant across all choices of $\boldsymbol{q}$ and S, it can be precomputed once and indexed offline, making the above dot product representation amenable to efficient ANN search.

End-to-end approximation error: We now give a bound on the end-to-end approximation error incurred in composing the KL approximation, neural softmax approximation of the barycenter and the resulting approximate index search using the finite-dimensional dot-product approximation of the marginal gain.

3.5 Efficient retrieval

Efficient greedy search Next, we replace the exhaustive search in line 3 of the greedy routine (Algorithm 1) with an ANN-based search. Instead of scanning the entire corpus and performing an $O(|S|N^2)$ search over the marginal gains using a Sinkhorn-based solver, we identify the candidate item that maximizes the marginal gain using a $o(N)$ index look-up and we select that particular item. We do this lookup k' times ($k' \ll N$), each time updating the state embedding to include the latest retrieved item, obtaining a pruned set $S_{k'}$ for the barycenter support without any Sinkhorn calls and avoiding quadratic scaling in corpus size that is associated with the exact greedy setup. Finally, to obtain the set of $k < k'$ items we send to the reranker, we use exact greedy as outlined in Algorithm 1 on $S_{k'}$, which is now much cheaper. The reranker returns the final ranking of the top-k elements found by exact greedy, completing our search.

Theorem 7 (End-to-end approximation for WBFusion)**.** Fix $2 \leq k \leq N$ and a query $\boldsymbol{q}$. Under the assumptions of Theorem 6, suppose the representations used by the index attain its uniform approximation bound. Let $F_k^* := \max_{A \subseteq D, |A| = k} F(\boldsymbol{q}, A)$, and let γ be the submodularity-ratio lower bound from Theorem 5. Starting from a singleton S_1, let $S_{t+1} = S_t \cup \{s_t\}$ for $t = 1, \ldots, k-1$. Suppose each ANN query returns an unselected item with maximum inner product. Then

$$F_k^* - F(\boldsymbol{q}, S_k) \leq O\left(1 - \frac{\gamma}{k}\right)^{k-1} + O(\epsilon + \eta). \tag{8}$$

Reranking the selected support The support set returned by Algorithm 1 or Algorithm 4 optimizes the RSWB objective in Eq. (4) without using ground-truth relevance labels. Consequently, even a support set that is near-optimal for rank fusion may yield a suboptimal ranking. To improve ranking

Method	MS-Marco		NF-Corpus		Hotpot-QA		Fever		SciFact	
	MAP	NDCG	MAP	NDCG	MAP	NDCG	MAP	NDCG	MAP	NDCG
RRF	0.136	0.543	0.119	0.328	0.522	0.624	0.581	0.653	0.623	0.664
CombSUM	0.130	0.545	0.126	0.344	0.551	0.648	0.620	0.685	0.628	0.673
CombMNZ	0.130	0.545	0.125	0.342	0.551	0.648	0.620	0.685	0.628	0.673
MoE	0.153	0.605	0.138	<u>0.376</u>	<u>0.642</u>	<u>0.726</u>	0.759	0.804	<u>0.678</u>	<u>0.720</u>
Ensemble	<u>0.157</u>	<u>0.638</u>	<u>0.141</u>	0.369	0.594	0.685	<u>0.806</u>	<u>0.843</u>	0.658	0.706
NeuralRRF	0.152	0.613	0.130	0.374	0.549	0.649	0.653	0.718	0.646	0.694
Hybrid	0.137	0.566	0.130	0.374	0.494	0.589	0.549	0.616	0.646	0.686
WBFusion	**0.163**	**0.673**	**0.145**	**0.383**	**0.647**	**0.728**	**0.814**	**0.846**	**0.712**	**0.748**

Table 1: Performance comparison, measured in retrieval MAP and NDCG, between WBFusion and other rank-fusion baselines: *viz.*, RRF (Cormack et al., 2009), CombSUM (Fox et al., 1993), CombMNZ (Lee, 1997), MoE (Chen et al., 2021), Ensemble (Borges et al., 2021), NeuralRRF, and Hybrid (Kuzi et al., 2020). First three (last four) are training free (trainable) baselines. Numbers in **bold** (<u>underline</u>) indicate the best (second best) performers. We observe that our method (WBFusion) performs the best.

accuracy, we train a reranker that assigns each selected item $c \in S$ a query- and support-dependent score $\text{score}_\phi(\boldsymbol{q}, S, c)$. We train the reranker on ground-truth relevance labels for each query using a regularized ranking loss.

4 Experiments

Datasets We choose five datasets from BEIR (Thakur et al., 2021) for experiments, *viz.*, **(1)** MS–Marco, **(2)** NF-Corpus, **(3)** Hotpot-QA, **(4)** Fever and **(5)** SciFact.

Base rankers We use six base rankers, *viz.*, **(1)** BM25 (Robertson et al., 2009; Kamphuis et al., 2020), **(2)** TF-IDF (Pedregosa et al., 2018), **(3)** Dense (Reimers et al., 2019), which uses dot product similarity of frozen SBERT embeddings, **(4)** MLP, which trains a multilayer perceptron over SBERT (Reimers et al., 2019) embeddings, **(5)** Cross-encoder (Nogueira et al., 2020), which computes the relevance scores from frozen BERT cross-encoder and **(6)** Query likelihood (QL) (Ponte et al., 1998; Zhai et al., 2001).

Baselines (methods compared) We compare WBFusion against seven baselines. They include training-free methods: **(1)** RRF (Cormack et al., 2009), **(2)** CombSUM (Fox et al., 1993), and **(3)** CombMNZ (Lee, 1997) and trainable methods: **(1)** a query-dependent Mixture of Experts (MoE) (Chen et al., 2021), **(2)** a query-agnostic ensemble model (Ensemble)(Borges et al., 2021). **(3)** an MLP over RRF features (NeuralRRF), and **(4)** Hybrid, a hybrid fusion model (Kuzi et al., 2020).

Evaluation We split the query set Q into training, validation, and test sets according to the public splits (Thakur et al., 2021). After training, we use WBFusion and each baseline to produce a ranked list of corpus items for every test query. We compute average precision and normalized discounted cumulative gain (NDCG) for each list, then average these scores across test queries to report mean average precision (MAP) and mean NDCG. In all cases, we used $k = 10$.

4.1 Results

Ranking accuracy of WBFusion. Table 1 compares WBFusion against baselines on five BEIR datasets and reports mean average precision (MAP) and NDCG. We make the following observations. **(1)** WBFusion achieves best ranking accuracy in all cases. **(2)** training-free baselines, *viz.*, RRF, CombSUM, CombMNZ, consistently perform worse than trainable baselines, indicating the importance of supervision for rank fusion. **(3)** The query-agnostic Ensemble and query-dependent MoE perform well among baselines. Since our base rankers include a cross-encoder which is a strong ranking model, they achieve good accuracy even with pointwise aggregation. WBFusion surpasses them, indicating that RSWB objective-based rank fusion that respects corpus geometry is effective across a wide pool of base rankers.

Ablation on reranker We use a reranker in line 7 in Algorithm 4. To study its utility, we compare a variant of WBFusion, called as WBFusion (no rerank) that does not use a reranker and directly returns S_k after greedy.

Method	MS-Marco	NF-Corpus	Hotpot-QA	Fever	SciFact
RRF	0.136	0.119	0.522	0.581	0.623
CombSUM	0.130	0.126	0.551	0.620	0.628
CombMNZ	0.130	0.125	0.551	0.620	0.628
WBFusion (no rerank)	0.143	0.135	0.592	0.722	0.627
WBFusion	**0.163**	**0.145**	**0.647**	**0.814**	**0.712**

Table 2: MAP for training free baselines, WBFusion and WBFusion (no rerank).

In Table 2, we compare WBFusion (no rerank) against other training-free baselines and WBFusion, in terms of MAP. We observe that **(1)** WBFusion (no rerank) outperforms the training-free baselines on four out of five datasets, showing that using information about corpus geometry helps rank documents better than simple pointwise aggregation. In SciFact, CombSUM and CombMNZ outperform WBFusion (no rerank) by a very small margin (0.001). Supervised reranking helps improve performance in this case. **(2)** WBFusion performs better than WBFusion (no rerank), illustrating the benefit of reranking the greedy support.

Impact of entropic regularization To evaluate the impact of the two entropic regularizers λ and τ, we sweep (λ, τ) in log scale, find the restricted support barycenter at each point and record NDCG of the barycenter-induced ranking (i.e., WBFusion (no rerank)). Figure 3 summarizes the results for NFCorpus. We make the following observations. **(1)** The optimum lies *strictly* within the grid, so neither regularizer is redundant. **(2)** τ is less sensitive hyperparameter than λ, since varying τ with a fixed λ only marginally impacts ranking. **(3)** Performance is robust to small changes in λ and τ, indicating that the regularizers are easy to tune.

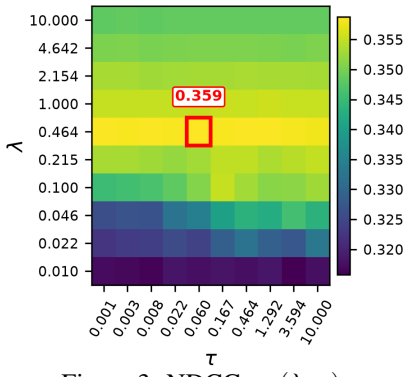


Figure 3: NDCG vs (λ, τ).

Tradeoff between accuracy and fusion quality The RSWB objective $F(\boldsymbol{q}, S)$ (5) aims to enhance rank fusion quality rather than directly optimize performance. Here, we analyze the trade-off between F and retrieval performance in terms of the area under the curve (AUC) between MAP and NDCG by sweeping over hyperparameters. Tables 4 and 5 summarize the results. We make the following observations. **(1)** WBFusion obtains the best AUC among all methods under the constraint that F should reach at least 50% of the optimum (specifically, $\frac{F^\star - F}{|F^\star|} \leq 0.5$, where $F^\star$ is the optimum), indicating that under a predetermined rank-fusion quality floor, WBFusion provides the best tradeoff (Table 4). **(2)** Most other baselines are either significantly worse or don't even reach the required value of F (Table 4). For example, all baselines fail to achieve the minimum F value for Hotpot-QA. **(3)** WBFusion obtains the best $\max F$ under the constraint that optimal retrieval performance must reach at least half of the maximum on four of five datasets (Table 5).

Efficiency vs performance We next analyze the trade-off between retrieval efficiency and (1) rank fusion quality and (2) ranking accuracy (NDCG) of WBFusion (Algorithm 4) and compare with Exact Greedy (Algorithm 1), Lazy Greedy (Minoux, 1978), Lazier than Lazy Greedy (Mirzasoleiman et al., 2015) and Stochastic Greedy (Mirzasoleiman et al., 2015). Table 6 reports the efficiency of all the methods in terms of the speedup with respect to exact greedy algorithm to reach 80% of optimal F (or 80% optimal NDCG). We denote these speedups by $\mathcal{E}_{0.8F}$ and $\mathcal{E}_{0.8\text{NDCG}}$, respectively. We make the following observations. **(1)** WBFusion is significantly faster than other greedy selection baselines in all cases. **(2)** Some greedy variants like stochastic greedy are unable to reach 80% of optimum F or NDCG values. Thus, WBFusion achieves high efficiency while maintaining both ranking accuracy and a competitive rank fusion quality.

	MS-Marco		NF-Corpus		Hotpot-QA		Fever		SciFact	
Method	AUC_{MAP}	AUC_{NDCG}	AUC_{MAP}	AUC_{NDCG}	AUC_{MAP}	AUC_{NDCG}	AUC_{MAP}	AUC_{NDCG}	AUC_{MAP}	AUC_{NDCG}
MoE	**0.008**	0.030	0.183	0.499	—	—	0.048	0.053	—	—
Ensemble	**0.008**	0.032	0.120	0.361	—	—	0.060	0.063	0.086	0.093
NeuralRRF	**0.008**	0.031	0.166	0.499	—	—	0.051	0.056	0.084	0.091
Hybrid	0.007	0.029	0.138	0.437	—	—	0.042	0.047	0.082	0.088
WBFusion	**0.008**	**0.033**	**0.194**	**0.515**	**0.059**	**0.067**	**0.061**	**0.065**	**0.091**	**0.096**

Table 4: Area under the curve (AUC) between ranking metric (MAP and NDCG) and F under the constraint that F should reach at least 50% of the optimum ($F^\star$), i.e., $\frac{F^\star - F}{|F^\star|} \leq 0.5$. We generate the curve by sweeping over the corresponding hyperparameters. Higher is better. Numbers in **bold** (underline) indicate the best (second best) performer. (—) indicates that F never reached 50% of $F^\star$.

Method	MS-Marco		NF-Corpus		Hotpot-QA		Fever		SciFact	
	F_{MAP}	F_{NDCG}	F_{MAP}	F_{NDCG}	F_{MAP}	F_{NDCG}	F_{MAP}	F_{NDCG}	F_{MAP}	F_{NDCG}
MoE	0.062	0.062	2.365	2.365	<u>-0.301</u>	<u>-0.301</u>	<u>0.114</u>	<u>0.114</u>	0.112	0.112
Ensemble	0.062	0.062	2.345	2.345	-0.313	-0.313	0.099	0.099	0.137	0.137
NeuralRRF	<u>0.073</u>	<u>0.073</u>	2.224	2.224	-0.316	-0.316	0.112	0.112	0.142	0.142
Hybrid	0.059	0.059	<u>2.428</u>	<u>2.428</u>	-0.314	-0.314	0.096	0.096	**0.161**	**0.161**
WBFusion	**0.088**	**0.088**	**2.710**	**2.710**	**-0.212**	**-0.212**	**0.134**	**0.134**	<u>0.153</u>	<u>0.153</u>

Table 5: $\max F$ under the constraint that retrieval performance (MAP and NDCG) must reach at least half of the optimum, denoted as F_{MAP} and F_{NDCG} respectively. We generate the curve by sweeping over the corresponding hyperparameters. Higher is better. Numbers in **bold** (<u>underline</u>) indicate the best (second best) performer.

Method	MS-Marco		NF-Corpus		Hotpot-QA		Fever		SciFact	
	$\mathcal{E}_{0.8F}$	$\mathcal{E}_{0.8\text{NDCG}}$	$\mathcal{E}_{0.8F}$	$\mathcal{E}_{0.8\text{NDCG}}$	$\mathcal{E}_{0.8F}$	$\mathcal{E}_{0.8\text{NDCG}}$	$\mathcal{E}_{0.8F}$	$\mathcal{E}_{0.8\text{NDCG}}$	$\mathcal{E}_{0.8F}$	$\mathcal{E}_{0.8\text{NDCG}}$
Exact Greedy	1.00×	1.00×	1.00×	1.00×	1.00×	1.00×	1.00×	1.00×	1.00×	1.00×
Lazy Greedy	<u>2.10×</u>	2.10×	6.24×	<u>6.24×</u>	<u>4.05×</u>	<u>4.05×</u>	3.66×	<u>3.66×</u>	<u>4.18×</u>	<u>4.18×</u>
Lazier than Lazy	2.04×	2.07×	7.18×	5.93×	3.72×	3.72×	<u>3.68×</u>	3.43×	3.94×	3.52×
Stochastic Greedy	—	<u>2.14×</u>	<u>9.12×</u>	—	—	—	—	—	—	—
WBFusion	**2.83×**	**3.79×**	**29.39×**	**16.10×**	**7.22×**	**16.98×**	**5.06×**	**22.50×**	**14.76×**	**22.45×**

Table 6: Efficiency ($\mathcal{E}_{0.8F}$ and $\mathcal{E}_{0.8\text{NDCG}}$)— measured as the speedup over exact greedy (Algorithm 1) in reaching 80% of the F and NDCG values attained by exact greedy— of WBFusion, lazy greedy, lazier than lazy greedy, and stochastic greedy. Higher is better. **Bold** and <u>underlined</u> values indicate the best and second-best results, respectively. WBFusion achieves the highest efficiency.

Scaling with corpus size Finally we analyze how efficiency scales with corpus size N for Lazy, Lazier than Lazy Greedy, Stochastic Greedy and WBFusion (Algorithm 4), compared with exact greedy (Algorithm 1). Table 7 summarizes the results ($\mathcal{E}_{0.85F}$ and $\mathcal{E}_{0.50\text{NDCG}}$) for Fever, with additional results in Appendix G. We observe that **(1)** WBFusion shows the fastest scaling as corpus size grows, and **(2)** among the baselines, Lazy greedy is the only variant that scales meaningfully, but WBFusion shows much better efficiency for large N.

Dataset	N	Lazy		Lazier		Stochastic		WBFusion	
		$\mathcal{E}_{0.85F}$	$\mathcal{E}_{0.5\text{NDCG}}$	$\mathcal{E}_{0.85F}$	$\mathcal{E}_{0.5\text{NDCG}}$	$\mathcal{E}_{0.85F}$	$\mathcal{E}_{0.5\text{NDCG}}$	$\mathcal{E}_{0.85F}$	$\mathcal{E}_{0.5\text{NDCG}}$
Fever	1000	2.99×	2.99×	2.81×	2.81×	—	—	**3.92×**	**13.58×**
	2000	13.28×	13.28×	7.22×	3.78×	—	—	**17.31×**	**56.31×**
	4000	**71.31×**	71.31×	6.12×	3.25×	—	—	<u>60.87×</u>	**258.64×**
	8000	133.08×	133.08×	18.79×	6.38×	—	—	**179.47×**	**480.96×**
	16000	197.18×	197.18×	19.66×	3.22×	—	—	**338.19×**	**664.74×**

Table 7: Efficiency ($\mathcal{E}_{0.85F}$ and $\mathcal{E}_{0.5\text{NDCG}}$)— measured as the speedup over exact greedy (Algorithm 1) in reaching 85% of the F and 50% of the NDCG values attained by exact greedy— of WBFusion, lazy greedy, lazier than lazy greedy, and stochastic greedy, as the corpus size N increases, on Fever. Stochastic greedy did not achieve the required F or NDCG values (—). Higher is better. **Bold** and <u>underlined</u> values indicate the best and second-best results, respectively. WBFusion achieves the best scaling as corpus size N increases.

5 Conclusion

We presented WBFusion, which recasts rank fusion as a restricted-support Wasserstein barycenter problem and shows that the induced set function is monotone and restricted γ-weakly submodular. Using a KL-based surrogate, we establish a uniform finite-dimensional dot-product approximation of the marginal gain under the stated approximation assumptions, which supports ANN-based greedy retrieval with corpus representations that can be indexed offline. Taken together, our results suggest that rank fusion is best treated as a *set-level, geometry-aware* problem, and WBFusion translate this view into consistent gains in ranking accuracy and substantial speedups across five BEIR datasets.

AI USE STATEMENT

In this work, we used generative AI tools for assistance in formulating and proving mathematical claims and implementing methods, with generative AI tools providing significant assistance in coding. We have not used generative AI tools for generating synthetic datasets, development of theoretical models or conceptual frameworks, proposing or refining hypothesis, designing or providing feedback on research methodology or experimentation, supporting qualitative or thematic data analysis or interpretation of results. Use of generative AI for assistance with translation or cleaning and reformatting dataset(s) is not applicable to this work. Additionally, we used generative AI tools for assistance in editing the paper to improve readability, grammatical correctness and clarity of notation. We have reviewed all AI-assisted work. LLM generated code was verified, tested and reviewed by two authors and any edits to the paper were proofread by two authors. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

The primary application of our work is in the domain of information retrieval (IR) where we aim to fuse the rankings of multiple base rankers into one combined ranking. All of our datasets are public and are standard datasets in the IR community. Our rank fusion method may exhibit undesired bias inherited from base rankers or the corpus itself, and we suggest keeping this limitation in mind before using WBFUSION in sensitive or high-stakes settings. More generally, as with any IR work, applying retrieval to sensitive or private data may have ethical concerns. However, such concerns are not specific to our work and apply to any retrieval system.

REPRODUCIBILITY STATEMENT

To ensure reproducibility we provide an anonymized copy of our code as part of supplementary material. We provide complete proofs for our claims in Appendix D and Algorithms 1 2, 3, and 4 provide relevant pseudocode. Pertinent details about our implementation and methods is also provided in Appendix E and Appendix F.

REFERENCES

Martial Agueh and Guillaume Carlier. Barycenters in the wasserstein space. *SIAM Journal on Mathematical Analysis*, 2011.

Jason Altschuler, Jonathan Weed, and Philippe Rigollet. Near-linear time approximation algorithms for optimal transport via sinkhorn iteration, 2018. URL https://arxiv.org/abs/1705.09634.

Jason Altschuler, Francis Bach, Alessandro Rudi, and Jonathan Niles-Weed. Massively scalable sinkhorn distances via the nyström method. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/file/f55cadb97eaff2ba1980e001b0bd9842-Paper.pdf.

Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In *Proceedings of the 34th International Conference on Machine Learning*, pp. 214–223, 2017.

Jean-David Benamou, Guillaume Carlier, Marco Cuturi, Luca Nenna, and Gabriel Peyré. Iterative Bregman projections for regularized transportation problems. *SIAM J. Sci. Comput.*, 37(2):A1111–A1138, 2015.

Mathieu Blondel, Olivier Teboul, Quentin Berthet, and Josip Djolonga. Fast differentiable sorting and ranking, 2020. URL https://arxiv.org/abs/2002.08871.

Luís Borges, Bruno Martins, and Jamie Callan. Assessing the benefits of model ensembles in neural re-ranking for passage retrieval. In *Advances in Information Retrieval: 43rd European Conference on IR Research*, 2021.

Sebastian Bruch, Siyu Gai, and Amir Ingber. An analysis of fusion functions for hybrid retrieval. *ACM Trans. Inf. Syst.*, 2023.

Jaime Carbonell and Jade Goldstein. The use of mmr, diversity-based reranking for reordering documents and producing summaries. In *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 335–336, 1998.

Moses S. Charikar. Similarity estimation techniques from rounding algorithms. In *Proceedings of the Thiry-Fourth Annual ACM Symposium on Theory of Computing*, 2002.

Zhizhong Chen and Carsten Eickhoff. Expertrank: A multi-level coarse-grained expert-based listwise ranking loss. *ArXiv*, abs/2107.13752, 2021. URL https://api.semanticscholar.org/CorpusID:236493621.

L. Chizat. Doubly regularized entropic wasserstein barycenter. *Foundations of Computational Mathematics*, 2025.

Sebastian Claici, Edward Chien, and Justin Solomon. Stochastic wasserstein barycenters, 2018. URL https://arxiv.org/abs/1802.05757.

Gordon V. Cormack, Charles L A Clarke, and Stefan Buettcher. Reciprocal rank fusion outperforms condorcet and individual rank learning methods. In *Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2009.

Marco Cuturi. Sinkhorn distances: lightspeed computation of optimal transport. In *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2*, 2013.

Marco Cuturi and Arnaud Doucet. Fast computation of wasserstein barycenters. In *Proceedings of the 31st International Conference on Machine Learning*, 2014.

Marco Cuturi and Gabriel Peyré. A smoothed dual approach for variational wasserstein problems, 2015. URL https://arxiv.org/abs/1503.02533.

Marco Cuturi, Olivier Teboul, and Jean-Philippe Vert. *Differentiable ranks and sorting using optimal transport*. Curran Associates Inc., 2019.

Abhimanyu Das and David Kempe. Submodular meets spectral: greedy algorithms for subset selection, sparse approximation and dictionary selection. In *Proceedings of the 28th International Conference on International Conference on Machine Learning*, 2011.

Pierre Dognin, Igor Melnyk, Youssef Mroueh, Jerret Ross, Cicero Dos Santos, and Tom Sercu. Wasserstein barycenter model ensembling. In *International Conference on Learning Representations*, 2019.

Matthijs Douze, Alexandr Guzhva, Chengqi Deng, Jeff Johnson, Gergely Szilvasy, Pierre-Emmanuel Mazaré, Maria Lomeli, Lucas Hosseini, and Hervé Jégou. The faiss library. *IEEE Transactions on Big Data*, 2025.

Ethan R. Elenberg, Rajiv Khanna, Alexandros G. Dimakis, and Sahand Negahban. Restricted strong convexity implies weak submodularity, 2017. URL https://arxiv.org/abs/1612.00804.

Dan Feldman and Michael Langberg. A unified framework for approximating and clustering data, 2016. URL https://arxiv.org/abs/1106.1379.

FEVER. Fact Extraction and VERification — fever.ai. https://fever.ai/. [Accessed 24-09-2026].

Jean Feydy, Thibault Séjourné, François-Xavier Vialard, Shun-ichi Amari, Alain Trouve, and Gabriel Peyré. Interpolating between optimal transport and mmd using sinkhorn divergences. In Kamalika Chaudhuri and Masashi Sugiyama (eds.), *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, volume 89 of *Proceedings of Machine Learning Research*, pp. 2681–2690. PMLR, 16–18 Apr 2019. URL https://proceedings.mlr.press/v89/feydy19a.html.

M. L. Fisher, G. L. Nemhauser, and L. A. Wolsey. *An analysis of approximations for maximizing submodular set functions—II*, pp. 73–87. Springer Berlin Heidelberg, Berlin, Heidelberg, 1978. ISBN 978-3-642-00790-3. doi: 10.1007/BFb0121195. URL https://doi.org/10.1007/BFb0121195.

Thibault Formal, Carlos Lassance, Benjamin Piwowarski, and Stéphane Clinchant. Splade v2: Sparse lexical and expansion model for information retrieval. *arXiv preprint arXiv:2109.10086*, 2021a.

Thibault Formal, Benjamin Piwowarski, and Stéphane Clinchant. Splade: Sparse lexical and expansion model for first stage ranking. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2021b.

Edward A. Fox and Joseph A. Shaw. Combination of multiple searches. In *Text Retrieval Conference*, 1993.

Luyu Gao, Zhuyun Dai, Tongfei Chen, Zhen Fan, Benjamin Van Durme, and Jamie Callan. Complementing lexical retrieval with semantic residual embedding, 2021. URL https://arxiv.org/abs/2004.13969.

Aude Genevay, Gabriel Peyré, and Marco Cuturi. Learning generative models with sinkhorn divergences, 2017. URL https://arxiv.org/abs/1706.00292.

K. S. Gurumoorthy, A. Dhurandhar, G. Cecchi, and C. Aggarwal. Efficient data representation by selecting prototypes with importance weights. In *IEEE ICDM*, 2019.

HOTPOTQA. HotpotQA Homepage — hotpotqa.github.io. https://hotpotqa.github.io/. [Accessed 24-09-2026].

Piotr Indyk and Rajeev Motwani. Approximate nearest neighbors: towards removing the curse of dimensionality. In *Proceedings of the Thirtieth Annual ACM Symposium on Theory of Computing*, STOC '98, pp. 604–613, New York, NY, USA, 1998. Association for Computing Machinery. ISBN 0897919629. doi: 10.1145/276698.276876. URL https://doi.org/10.1145/276698.276876.

Suhas Jayaram Subramanya, Fnu Devvrit, Harsha Vardhan Simhadri, Ravishankar Krishnawamy, and Rohan Kadekodi. Diskann: Fast accurate billion-point nearest neighbor search on a single node. *Advances in neural information processing Systems*, 32, 2019.

Chris Kamphuis, Arjen P. de Vries, Leonid Boytsov, and Jimmy Lin. Which bm25 do you mean? a large-scale reproducibility study of scoring variants. In *Advances in Information Retrieval: 42nd European Conference on IR Research, ECIR 2020, Lisbon, Portugal, April 14–17, 2020, Proceedings, Part II*, pp. 28–34, Berlin, Heidelberg, 2020. Springer-Verlag. ISBN 978-3-030-45441-8. doi: 10.1007/978-3-030-45442-5_4. URL https://doi.org/10.1007/978-3-030-45442-5_4.

Athresh Karanam, Krishnateja Killamsetty, Harsha Kokel, and Rishabh Iyer. Orient: Submodular mutual information measures for data subset selection under distribution shift. In *Advances in Neural Information Processing Systems*, 2022.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2020.

Rajiv Khanna, Ethan Elenberg, Alexandros G. Dimakis, Sahand Negahban, and Joydeep Ghosh. Scalable greedy feature selection via weak submodularity, 2017. URL https://arxiv.org/abs/1703.02723.

Omar Khattab and Matei Zaharia. Colbert: Efficient and effective passage search via contextualized late interaction over bert. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2020.

Krishnateja Killamsetty, Durga S, Ganesh Ramakrishnan, Abir De, and Rishabh Iyer. Grad-match: Gradient matching based data subset selection for efficient deep model training. In *Proceedings of the 38th International Conference on Machine Learning*, 2021.

Philip A. Knight. The sinkhorn–knopp algorithm: Convergence and applications. *SIAM Journal on Matrix Analysis and Applications*, 30(1):261–275, 2008.

Matt Kusner, Yu Sun, Nicholas Kolkin, and Kilian Weinberger. From word embeddings to document distances. In Francis Bach and David Blei (eds.), *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pp. 957–966, Lille, France, 07–09 Jul 2015. PMLR. URL https://proceedings.mlr.press/v37/kusnerb15.html.

Saar Kuzi, Mingyang Zhang, Cheng Li, Michael Bendersky, and Marc Najork. Leveraging semantic and lexical matching to improve the recall of document retrieval systems: A hybrid approach. *ArXiv*, 2020.

Joon Ho Lee. Analyses of multiple evidence combination. In *Proceedings of the 20th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 1997.

Lingyuan Liu and Mengxiang Zhang. Exp4fuse: A rank fusion framework for enhanced sparse retrieval using large language model-based query expansion. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 163–173, 2025.

Mario Lucic, Matthew Faulkner, Andreas Krause, and Dan Feldman. Training gaussian mixture models at scale via coresets, 2018. URL https://arxiv.org/abs/1703.08110.

Piyushi Manupriya, Pratik Jawanpuria, Karthik S. Gurumoorthy, Saketha Nath Jagarlapudi, and Bamdev Mishra. Submodular framework for structured-sparse optimal transport. In *International Conference on Machine Learning*, 2024.

Michel Minoux. Accelerated greedy algorithms for maximizing submodular set functions. In J. Stoer (ed.), *Optimization Techniques*, pp. 234–243, Berlin, Heidelberg, 1978. Springer Berlin Heidelberg. ISBN 978-3-540-35890-9.

Baharan Mirzasoleiman, Ashwinkumar Badanidiyuru, Amin Karbasi, Jan Vondrák, and Andreas Krause. Lazier than lazy greedy. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, 2015.

Baharan Mirzasoleiman, Jeff Bilmes, and Jure Leskovec. Coresets for data-efficient training of machine learning models. In *Proceedings of the 37th International Conference on Machine Learning*, 2020.

André Mourao, Flávio Martins, and Joao Magalhaes. Novasearch at trec 2013 federated web search track: Experiments with rank fusion. In *TREC*, 2013.

MSMARCO. MS MARCO — microsoft.github.io. https://microsoft.github.io/msmarco/. [Accessed 24-09-2026].

Kartik Nair, Pritish Chakraborty, Atharva Abhijit Tambat, Indradyumna Roy, Soumen Chakrabarti, Anirban Dasgupta, and Abir De. A dense subset index for collective query coverage. In *The Fourteenth International Conference on Learning Representations*, 2026.

George L. Nemhauser, Laurence A. Wolsey, and Marshall L. Fisher. An analysis of approximations for maximizing submodular set functions—i. *Mathematical Programming*, 1978.

Behnam Neyshabur and Nathan Srebro. On symmetric and asymmetric lshs for inner product search. In *International Conference on Machine Learning*, pp. 1926–1934. PMLR, 2015.

NFCORPUS. NFCorpus: A Full-Text Learning to Rank Dataset for Medical Information Retrieval — cl.uni-heidelberg.de. https://www.cl.uni-heidelberg.de/statnlpgroup/nfcorpus/. [Accessed 24-09-2026].

Rodrigo Nogueira and Kyunghyun Cho. Passage re-ranking with bert, 2020. URL https://arxiv.org/abs/1901.04085.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Andreas Müller, Joel Nothman, Gilles Louppe, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. Scikit-learn: Machine learning in python, 2018. URL https://arxiv.org/abs/1201.0490.

Gabriel Peyré and Marco Cuturi. *Computational Optimal Transport: With Applications to Data Science*. Foundations and Trends in Machine Learning. Now Publishers, 2019.

Jay M. Ponte and W. Bruce Croft. A language modeling approach to information retrieval. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 275–281, 1998. doi: 10.1145/290941.291008.

Harishkumar Kishorkumar Prajapati. Hybrid retrieval for covid-19 literature: Comparing rank fusion and projection fusion with diversity reranking. *arXiv preprint arXiv:2604.13728*, 2026.

Zackary Rackauckas. Rag-fusion: a new take on retrieval-augmented generation. *arXiv preprint arXiv:2402.03367*, 2024.

Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks, 2019. URL https://arxiv.org/abs/1908.10084.

Stephen Robertson and Hugo Zaragoza. The probabilistic relevance framework: Bm25 and beyond. *Found. Trends Inf. Retr.*, 3(4):333–389, April 2009. ISSN 1554-0669. doi: 10.1561/1500000019. URL https://doi.org/10.1561/1500000019.

Indradyumna Roy, Rishi Agarwal, Soumen Chakrabarti, Anirban Dasgupta, and Abir De. Locality sensitive hashing in fourier frequency domain for soft set containment search. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=rUf0GV5CuU.

Filippo Santambrogio. *Optimal Transport for Applied Mathematicians: Calculus of Variations, PDEs, and Modeling*. Springer Nature, 2015.

Keshav Santhanam, Omar Khattab, Christopher Potts, and Matei Zaharia. Plaid: an efficient engine for late interaction retrieval. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, pp. 1747–1756, 2022a.

Keshav Santhanam, Omar Khattab, Jon Saad-Falcon, Christopher Potts, and Matei Zaharia. Colbertv2: Effective and efficient retrieval via lightweight late interaction. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 3715–3734, 2022b.

SCIFACT. GitHub - allenai/scifact: Data and models for the SciFact verification task. — github.com. https://github.com/allenai/scifact. [Accessed 24-09-2026].

Anshumali Shrivastava and Ping Li. Asymmetric lsh (alsh) for sublinear time maximum inner product search (mips). In *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 2*, 2014.

Anshumali Shrivastava and Ping Li. Improved asymmetric locality sensitive hashing (alsh) for maximum inner product search (mips). In *Proceedings of the Thirty-First Conference on Uncertainty in Artificial Intelligence*, 2015.

Justin Solomon, Fernando de Goes, Gabriel Peyré, Marco Cuturi, Adrian Butscher, Andy Nguyen, Tao Du, and Leonidas Guibas. Convolutional wasserstein distances: efficient optimal transportation on geometric domains. *ACM Trans. Graph.*, 34(4), July 2015.

Michael Spivak. *Calculus On Manifolds*. Avalon Publishing, 1965.

Nandan Thakur, Nils Reimers, Andreas Rücklé, Abhishek Srivastava, and Iryna Gurevych. Beir: A heterogenous benchmark for zero-shot evaluation of information retrieval models, 2021. URL https://github.com/beir-cellar/beir.

Cédric Villani. *Optimal Transport: Old and New*. Springer-Verlag Berlin Heidelberg, 2009.

Yu Yang, Hao Kang, and Baharan Mirzasoleiman. Towards sustainable learning: Coresets for data-efficient deep learning. In *Proceedings of the 40th International Conference on Machine Learning*, 2023.

Chengxiang Zhai and John Lafferty. A study of smoothing methods for language models applied to ad hoc information retrieval. In *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 334–342, 2001. doi: 10.1145/383952.384019.

A Wasserstein Barycenter Index For Efficient Rank Fusion

(Appendix)

A Broader Impact

WBFusion introduces a rank fusion framework grounded in Wasserstein barycenter theory, which accounts for the shortcomings of pointwise aggregators such as failing to consider the inter-document relationships. Rank aggregation in general is useful, as one model's weaknesses are covered by another model's strengths, complementing each other. Traditional pointwise rank aggregators are already being used widely in web search, question answering, etc. With WBFusion using a Wasserstein approach, we ensure that there is diversity and reduction of semantic redundancy in the retrieved corpus items. This helps in saving time and compute too, for example, a RAG-based LLM repeatedly processing semantically similar data wastes both time and compute.
We train a ranker based on this obtained barycenter as the plain barycenter is unsupervised and therefore does not necessarily align with the true labeled data's judgments. We hence supervise a new ranker with a KL divergence loss from the barycenter enabling the new ranker to have a balance of empirical relevance signals and distributional regularity.

B Limitations and Future Directions

Despite the advantages, the proposed WBFusion formulation has some limitations worth addressing, opening avenues for future work:

Sinkhorn divergences, introduced by Genevay et al. (Genevay et al., 2017) are not being used by the current WBFusion formulation. Sinkhorn divergences are a variant of entropic OT that corrects the regularization bias. Incorporating this could help improve the model by removing the bias inherent in standard Sinkhorn distances.

The used indexing frameworks such as FAISS (Douze et al., 2025) and DiskANN (Jayaram Subramanya et al., 2019) provide the indexing infrastructure we build upon, but are unaware of the iterative discounting of already-covered content in our query updates $\boldsymbol{\zeta}_\theta(\boldsymbol{q}, S_t)$, motivating future work on dynamic data structures that incorporate the coverage state into their search procedure.

The WBFusion formulation does not account for evolving corpora; a fixed, discrete corpus is taken, and extending this to dynamically evolving corpora, where new items are added or removed over time remains an interesting direction for future work.

Addressing these issues along these directions would help extend WBFusion into a more general, scalable and robust framework for IR problems.

C Related Works

Optimal Transport and Wasserstein Barycenters Optimal transport (OT) provides a geometrically principled framework for comparing and combining probability distributions (Villani, 2009; Peyré et al., 2019). In the context of machine learning, OT has been employed for document classification (Kusner et al., 2015), differentiable sorting and ranking (Cuturi et al., 2019; Blondel et al., 2020), model ensembling (Dognin et al., 2019), and generative modeling (Genevay et al., 2017; Arjovsky et al., 2017). The classical discrete OT formulation (due to Kantorovich) can be posed as a linear program. However, it scales poorly. Hence, (Cuturi, 2013) proposed an entropy-penalized variant of OT that is stricly convex and and admits the matrix-scaling Sinkhorn-Knopp iterations (Knight, 2008) as its solver. Altschuler et al. (2018; 2019) proposed highly efficient entropic-regularized OT solvers via suitable approximations. To remove the bias that entropic OT inherits from its regularizer, Genevay et al. (2017) introduced Sinkhorn divergences, which are debiased and (Feydy et al., 2019) metrized weak convergence.

The formal existence and uniqueness of barycenters in the Wasserstein space were initially established by Agueh et al. (2011). By employing the entropic regularization, (Cuturi et al., 2014) proposed efficient Sinkhorn-Knopp iteration based algorithms for computing (regularized) Wasserstein barycenters. Wasserstein barycenter problem is an instance of the general variational Wasserstein problem, for which smoothed dual approaches (Cuturi et al., 2015) and Bregman-Dykstra iterations (Benamou et al., 2015) based solvers were also developed. (Claici et al., 2018) propose a stochastic algorithm for computing Wasserstein barycenter, which extends to continuous input distributions. Wasserstein barycenters have been used in a variety of machine learning applications such as model ensembling (Dognin et al., 2019), shape interpolation (Solomon et al., 2015), clustering (Cuturi et al., 2014), etc. The variant most directly aligned with our setting is the doubly regularized entropic Wasserstein barycenter studied by Chizat (2025), which regularizes both the transport plans and the barycentric marginals.

Submodularity, Weak Submodularity, and Subset Selection The greedy support selection (Algorithm 1) is grounded in the classical analysis of submodular maximization. Submodular functions are set functions which satisfy the diminishing returns property (Definition 3). In their seminal work, (Nemhauser et al., 1978; Fisher et al., 1978) prove the tight $(1 - 1/e)$ approximation factor of the classical greedy algorithm on monotone submodular maximization problems with cardinality constraint. In our setting, however, the RSWB-induced set function $F(q, \cdot)$ is generally *not* submodular (Theorem 5) and we therefore work with restricted weak submodularity. Existing works (Das et al., 2011; Khanna et al., 2017; Elenberg et al., 2017; Gurumoorthy et al., 2019) have successfully extended these classical guarantees to such approximately submodular objectives. Specifically, (Das et al., 2011) introduced the submodularity ratio (γ) to quantify the degree to which a set function deviates from strict submodularity, proving that the standard greedy algorithm achieves a $(1 - e^{-\gamma})$ approximation factor for cardinality-constrained maximization. Building upon this foundational concept, (Elenberg et al., 2017) established elegant connections between discrete weak submodularity and continuous optimization properties, such as restricted strong convexity and restricted smoothness. Their frameworks demonstrate that greedy methods maintain robust theoretical guarantees in applications like sparse feature selection and dictionary learning. Our approach to support selection is further motivated by the broader literature on efficient dataset representation and matching. For instance, Roy et al. (2023) address the related challenge of efficiently matching query sets against corpus sets using locality-sensitive hashing. In the context of dataset summarization, Feldman et al. (2016) establish a unified theoretical framework for coreset construction, demonstrating that a small subset can approximate the full dataset to within an ϵ-factor for various clustering objectives. Building on these principles to ensure scalability, Lucic et al. (2018) validate the practical construction of coresets for training statistical mixture models on massive datasets. Recently, several works have explored the utility of mini-batch coreset selection for efficient training deep learning models including large language models (Mirzasoleiman et al., 2020; Yang et al., 2023; Karanam et al., 2022; Killamsetty et al., 2021).

Approximate Nearest Neighbour Search Fast approximate nearest neighbour (ANN) search algorithms based on Locality-Sensitive Hashing (LSH) were introduced by Indyk et al. (1998) and extended by Charikar (2002) for a broad class of similarity functions. Shrivastava et al. (2014) proposed the first sublinear-time hashing scheme for maximum inner product search (MIPS) using asymmetric LSH, while Neyshabur et al. (2015) developed an alternative asymmetric hashing framework based

on cosine similarity. Modern ANN libraries such as Douze et al. (2025) and Jayaram Subramanya et al. (2019) provide scalable and easy-to-use indexing frameworks. FAISS unifies several indexing, quantization, and hashing mechanisms within a single library, whereas DiskANN employs a graph-based indexing strategy optimized for large-scale retrieval. More recently, Nair et al. (2026) leveraged ANN indexing for facility location optimization, where a multi-vector index is probed to efficiently optimize marginal gains for each query.

D PROOFS OF ALL TECHNICAL RESULTS

We define some notations and definitions used in the proofs. For every nonempty $S \subseteq D$, we define $\mathcal{N}_S := \{\boldsymbol{h} \in \mathbb{R}^N : \boldsymbol{h}[s'] = 0, \text{ for all } s' \notin S, \sum_{s\in S} \boldsymbol{h}[s] = 0\}$. For any $\eta > 0$ and $S \subseteq D$, we define $\mathcal{D}_\eta(S) := \{\boldsymbol{z} \in \Delta_S : \boldsymbol{z}[i] \geq \eta, \forall i \in S\}$ For any $N \times N$ matrix $\boldsymbol{C}$, we denote the maximum oscillation in the matrix as $\delta_{\boldsymbol{C}} := \max_{s,t,s',t'\in[N]} |\boldsymbol{C}[s,t] - \boldsymbol{C}[s',t']|$ and use $\boldsymbol{C}^\dagger$ to denote the Moore-Penrose pseudoinverse of $\boldsymbol{C}$.

Definition 8 (Monotonic and restricted submodularity)**.** Let $F : 2^D \to \mathbb{R}_{\geq 0}$ be a set function defined on the ground set D. For any $S \subset D$ and $e \notin S$, let $F(e \mid S) := F(S \cup \{e\}) - F(S)$ denote the marginal gain in F upon adding e to S. Then,

(1) F is monotone non-decreasing (or called only as 'monotone') if $F(e \mid S) \geq 0$, for all $S \subset D$ and $e \notin S$.

(2) F is submodular if it exhibits diminishing returns property, i.e., $F(e \mid S') \geq F(e \mid S)$ for all $S' \subseteq S \subset D$ and $e \notin S$.

(3) F is restricted γ-weakly submodular (with $\gamma \in (0, 1]$) with respect to an integer $k < N$ (Das et al., 2011; Elenberg et al., 2017; Gurumoorthy et al., 2019) if its submodularity ratio γ_k satisfies $\gamma_k \geq \gamma$, where the submodularity ratio γ_k is defined as $\gamma_k(F) := \inf_{L\cap S=\emptyset, |L|+|S|\leq k} \frac{\sum_{e\in S} F(e|L)}{F(L\cup S)-F(L)}$. For submodular functions, $\gamma = 1$.

D.1 ANALYSIS OF THE RSWB OBJECTIVE AND THE ASSOCIATED SET FUNCTION

Lemma 9 (Monotonicity and submodularity for special case of scalar embeddings)**.** Assume the corpus items are embedded in real number space, *i.e.*, $\boldsymbol{x}_\bullet \in \mathbb{R}$ with dissimilarity measure $\rho(\boldsymbol{x}_s, \boldsymbol{x}_t) = |\boldsymbol{x}_s - \boldsymbol{x}_t|$; and the entropic regularizers $\lambda, \tau = 0$. Then, $F(\boldsymbol{q}, S)$ is monotone and submodular in $S \subseteq D$.

Proof. We assume that the corpus items are embedded on the real line, with positions $\boldsymbol{x}_1 < \boldsymbol{x}_2 < \cdots < \boldsymbol{x}_N$. We start with an exact representation of the restricted-support optimum.

Proposition 10 (Exact pointwise representation of the restricted-support optimum)**.** For every nonempty support set $S \subseteq D$, we define the function:

$$g(t, \boldsymbol{x}) := \textstyle\sum_{r=1}^B \beta_r \left|\boldsymbol{x} - Q_{\boldsymbol{\mu}_r(q,D)}(t)\right|, \ \boldsymbol{x} \in \mathbb{R}. \tag{9}$$

Then we have that, $F(\boldsymbol{q}, S) = \int_0^1 \max_{\boldsymbol{x}_s : s\in S} \left(-g(t, \boldsymbol{x}_s)\right) \, \mathrm{d}t$.

Proof. Fix a nonempty support set $S \subseteq D$. We prove the identity by showing both lower and upper bounds. We first prove the lower bound.

For a probability vector $\xi \in \Delta_D$ supported on the corpus positions $\boldsymbol{x}_1 < \cdots < \boldsymbol{x}_N$, its CDF is defined as $\mathbb{F}_\xi(\boldsymbol{x}_i) := \sum_{\ell=1}^i \xi[\ell]$, $i \in [N]$, with the convention $\mathbb{F}_\xi(\boldsymbol{x}_0) := 0$, and its *quantile function* $Q_\xi : (0, 1] \to \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N\}$, $Q_\xi(t) := \boldsymbol{x}_i$ whenever $\mathbb{F}_\xi(\boldsymbol{x}_{i-1}) < t \leq \mathbb{F}_\xi(\boldsymbol{x}_i)$. Equivalently, $Q_\xi(t) = \inf\{\boldsymbol{x}_s : s \in D : \mathbb{F}_\xi(\boldsymbol{x}_s) \geq t\}$ is the generalised inverse of $\mathbb{F}_\xi$. It is a non-decreasing, left-continuous step function on $(0, 1]$ taking values in the support $\{\boldsymbol{x}_i : \xi[i] > 0\}$, with each level set $\{t : Q_\xi(t) = \boldsymbol{x}_i\}$ being an interval of length $\xi[i]$. From (Santambrogio, 2015, Proposition 2.17), we get that real-valued random variables ξ, ζ attain the following expression for $1D$ OT:

$$\mathcal{W}^1(\xi, \zeta) = \int_0^1 |Q_\xi(t) - Q_\zeta(t)| \, dt \tag{10}$$

Let $\boldsymbol{\nu} \in \Delta_S$. Since $Q_{\boldsymbol{\nu}}(t) \in S$ for every $t \in (0, 1]$, the quantile representation (10) gives

$$\begin{aligned}
\sum_{r=1}^B \beta_r \, \mathcal{W}^1\left(\boldsymbol{\nu}, \boldsymbol{\mu}_r(q, D)\right) &= \sum_{r=1}^B \beta_r \int_0^1 \left|Q_{\boldsymbol{\nu}}(t) - Q_{\boldsymbol{\mu}_r(q,D)}(t)\right| \, \mathrm{d}t \\
&= \int_0^1 \sum_{r=1}^B \beta_r \left|Q_{\boldsymbol{\nu}}(t) - Q_{\boldsymbol{\mu}_r(q,D)}(t)\right| \, \mathrm{d}t \\
&= \int_0^1 g\left(t, Q_{\boldsymbol{\nu}}(t)\right) \, \mathrm{d}t.
\end{aligned}$$

Since $Q_{\boldsymbol{\nu}}(t) \in S$, we have $g\left(t, Q_{\boldsymbol{\nu}}(t)\right) \geq \min_{\boldsymbol{x}_s : s \in S} g(t, \boldsymbol{x}_s)$, $\forall t \in (0,1]$. Therefore we may say:

$$\begin{aligned} \sum_{r=1}^{B} \beta_r \, \mathcal{W}^1\left(\boldsymbol{\nu}, \boldsymbol{\mu}_r(q, D)\right) &\geq \int_0^1 \min_{\boldsymbol{x}_s : s \in S} g(t, \boldsymbol{x}_s) \, \mathrm{d}t. \\ \implies \min_{\boldsymbol{\nu} \in \Delta_S} \sum_{r=1}^{B} \beta_r \, \mathcal{W}^1\left(\boldsymbol{\nu}, \boldsymbol{\mu}_r(q, D)\right) &\geq \int_0^1 \min_{\boldsymbol{x}_s : s \in S} g(t, \boldsymbol{x}_s) \, \mathrm{d}t. \end{aligned} \tag{11}$$

It remains to prove the reverse inequality. For $i < j$, define the non-increasing map $y \mapsto h_{ij}(y) := |\boldsymbol{x}_j - y| - |\boldsymbol{x}_i - y|$. It is easy to see that h_{ij} is nonincreasing in y. Now define the pointwise selector

$$i_S(t) := \min \left\{ \operatorname*{argmin}_{\boldsymbol{x}_s : s \in S} g(t, \boldsymbol{x}_s) \right\}, \; t \in (0,1]. \tag{12}$$

We claim that the map $t \mapsto i_S(t)$ is nondecreasing. Fix $0 < t_1 < t_2 \leq 1$, and suppose for contradiction that $i_S(t_1) > i_S(t_2)$. Set $i_1 := i_S(t_1)$, $i_2 := i_S(t_2)$, so that $i_2 < i_1$. Since i_2 minimizes $g(t_2, \bullet)$ over S, we have:

$$g(t_2, \boldsymbol{x}_{i_1}) - g(t_2, \boldsymbol{x}_{i_2}) \geq 0. \tag{13}$$

On the other hand, using the definition of $g(t, \bullet)$ and the monotonicity of each quantile function $Q_{\boldsymbol{\mu}_r(q,D)}(t)$ in t, we obtain:

$$\begin{aligned} g(t, \boldsymbol{x}_{i_1}) - g(t, \boldsymbol{x}_{i_2}) &= \sum_{r=1}^{B} \beta_r \left(|x_{i_1} - Q_{\boldsymbol{\mu}_r(q,D)}(t)| - |x_{i_2} - Q_{\boldsymbol{\mu}_r(q,D)}(t)| \right) \\ &= \sum_{r=1}^{B} \beta_r \, h_{i_2 i_1}\left(Q_{\boldsymbol{\mu}_r(q,D)}(t) \right). \end{aligned}$$

Since $h_{i_2 i_1}$ is nonincreasing and $t \mapsto Q_{\boldsymbol{\mu}_r(q,D)}(t)$ is nondecreasing, the map $t \mapsto g(t, \boldsymbol{x}_{i_1}) - g(t, \boldsymbol{x}_{i_2})$ is nonincreasing. Therefore (13) implies $g(t_1, \boldsymbol{x}_{i_1}) - g(t_1, \boldsymbol{x}_{i_2}) \geq 0$. But $i_1 = i_S(t_1)$ is the smallest minimizer at time t_1, so we have $g(t_1, \boldsymbol{x}_{i_1}) - g(t_1, \boldsymbol{x}_{i_2}) \leq 0$. Hence equality holds, and x_{i_2} is also a minimizer of $g(t_1, \bullet)$ on S. Since $i_2 < i_1$, this contradicts the definition of $i_S(t_1)$ as the smallest minimizer. Therefore $t \mapsto i_S(t)$ is nondecreasing.

It follows that the map defined by $Q_S^*(t) := \boldsymbol{x}_{i_S(t)}$, $t \in (0,1]$, is a nondecreasing map taking values in S. Hence Q_S^* is the quantile function of some probability vector $\boldsymbol{\nu}_0 \in \Delta_S$. By construction, $g\left(t, Q_S^*(t)\right) = \min_{x_s : s \in S} g(t, \boldsymbol{x}_s)$, $\forall t \in (0,1]$. Therefore we have:

$$\sum_{r=1}^{B} \beta_r \, \mathcal{W}^1\left(\boldsymbol{\nu}_0, \boldsymbol{\mu}_r(\boldsymbol{q}, D)\right) = \int_0^1 g\left(t, Q_S^*(t)\right) \, \mathrm{d}t = \int_0^1 \min_{\boldsymbol{x}_s : s \in S} g(t, \boldsymbol{x}_s) \, \mathrm{d}t.$$

Combining this with (11) proves that the inequality is in fact an equality. the definition of F in the $\lambda = \tau = 0$ regime, we get that:

$$F(\boldsymbol{q}, S) = -\int_0^1 \min_{\boldsymbol{x}_s : s \in S} g(t, \boldsymbol{x}_s) \, \mathrm{d}t = \int_0^1 \max_{\boldsymbol{x}_s : s \in S} \left(-g(t, \boldsymbol{x}_s)\right) \, \mathrm{d}t$$

□

Thus we may say: $F(\boldsymbol{q}, S) = \int_0^1 \Gamma(t, S) \, \mathrm{d}t$ where $\Gamma(t, S) := \max_{\boldsymbol{x}_s : s \in S} \left(-g(t, \boldsymbol{x}_s)\right)$, We first show that, for each fixed $t \in (0,1]$, the set function Γ_t is monotone nondecreasing and submodular.

Monotonicity is immediate, since $\Gamma(t, \bullet)$ is the maximum over a set. To prove submodularity, let $A \subseteq B \subseteq D$ be nonempty, and let $e \notin B$. Then

$$\begin{aligned} \Gamma(t, A \cup \{e\}) - \Gamma(t, A) &= \max\left(\Gamma(t, A), -g(t, e)\right) - \Gamma(t, A) \\ &= \left[-g(t, e) - \Gamma(t, A)\right]_+, \end{aligned}$$

and similarly $\Gamma(t, B \cup \{e\}) - \Gamma(t, B) = \left[-g(t, e) - \Gamma(t, B)\right]_+$ Since $A \subseteq B$ implies $\Gamma_t(A) \leq \Gamma_t(B)$, and the map $u \mapsto (-g(t, e) - u)_+$ is nonincreasing in u, it follows that $\Gamma(t, A \cup \{e\}) - \Gamma(t, A) \geq \Gamma(t, B \cup \{e\}) - \Gamma(t, B)$. Thus $\Gamma(t, \bullet)$ is submodular.

Finally, integration preserves both monotonicity and submodularity, because these are pointwise linear inequalities. Therefore $F(\boldsymbol{q}, \bullet) = \int_0^1 \Gamma(t, \bullet) \, dt$ is monotone nondecreasing and submodular. □

Now, we prove some fundamental properties of the entropy regulated Wasserstein distance (from eq 1) that we shall used to prove curvature bounds on W and co-ordinate bounds on $\boldsymbol{\nu}^*(\boldsymbol{q}, \bullet)$ for a given query q. We shall use these to prove strong convexity, implying uniqueness of the RSWB optimization problem.

Proposition 11. Let $S \subseteq D$ be nonempty, let $\boldsymbol{\nu} \in \Delta_S$ with $\boldsymbol{\nu}[i] > 0, \forall i \in S$, and let $\boldsymbol{\mu} \in \Delta_D$, with $\boldsymbol{\mu}[i] > 0, \forall i \in [N]$. Then the minimization problem (1) admits a unique optimizer $\boldsymbol{\pi}^* \in \Pi(\boldsymbol{\mu}, \boldsymbol{\nu})$, such that $\boldsymbol{\pi}^*[i, j] = 0,\ \forall i \notin S,\ \forall j \in [N]$ and $\boldsymbol{\pi}^*[i, j] > 0,\ \forall (i, j) \in S \times [N]$.

Proof. The feasible set $\Pi(\boldsymbol{\mu}, \boldsymbol{\nu})$ is nonempty, compact, and convex. The map $x \mapsto x(\ln x - 1)$ is continuous on $[0, \infty)$ and strictly convex on $(0, \infty)$. Hence the objective in (1) is continuous and strictly convex on $\Pi(\boldsymbol{\mu}, \boldsymbol{\nu})$, and therefore has a unique minimizer $\boldsymbol{\pi}^*$. Since the coupling $\boldsymbol{\pi}^*$ is a joint distribution, it is non-negative. Moreover, $\sum_j \boldsymbol{\pi}^*[i, j] = \boldsymbol{\nu}[i]$. Since $\boldsymbol{\nu}[i] = 0,\ \forall i \notin S$, every row of $\boldsymbol{\pi}^*$ outside the ones corresponding to corpus items in S is identically zero.

Assume for contradiction that $\boldsymbol{\pi}^*[i_0, j_0] = 0$ for some $i_0 \in S$ and $j_0 \in [N]$. We have, $\boldsymbol{\nu}[i_0] > 0$. As $\sum_{j=1}^N \boldsymbol{\pi}^*[i_0, j] = \boldsymbol{\nu}[i_0]$, there exists an index $j_1 \in [N]$ with $j_1 \neq j_0$ such that $\boldsymbol{\pi}^*[i_0, j_1] > 0$. Likewise, we have $\boldsymbol{\mu}[j_0] > 0$. As $\sum_{i=1}^N \boldsymbol{\pi}^*[i, j_0] = \boldsymbol{\mu}[j_0]$, there exists an index $i_1 \in [N]$ such that $\boldsymbol{\pi}^*[i_1, j_0] > 0$. Since every row i indexed by $i \notin S$ is identically zero, this implies that $i_1 \in S$. Moreover, we have that, $i_1 \neq i_0$. For $i, j \in [N]$, let $\boldsymbol{E}_{ij} \in \mathbb{R}^{N \times N}$ denote the matrix whose (i, j) entry is equal to 1 and whose other entries are equal to 0. We now define the cyclic perturbation

$$\Gamma := \boldsymbol{E}_{i_0 j_0} + \boldsymbol{E}_{i_1 j_1} - \boldsymbol{E}_{i_0 j_1} - \boldsymbol{E}_{i_1 j_0}. \tag{14}$$

Then, $\Gamma \mathbf{1} = 0,\ \Gamma^\top \mathbf{1} = 0$. Hence, for every $\epsilon > 0$, $\boldsymbol{\pi}^\epsilon := \boldsymbol{\pi}^* + \epsilon\Gamma$ has the same row and column sums as $\boldsymbol{\pi}^*$. For all sufficiently small $\epsilon > 0$, the only entries decreased by the perturbation are $\boldsymbol{\pi}^*[i_0, j_1]$ and $\boldsymbol{\pi}^*[i_1, j_0]$, which are strictly positive. Hence $\boldsymbol{\pi}^\epsilon \in \Pi(\boldsymbol{\mu}, \boldsymbol{\nu})$ for all sufficiently small $\epsilon > 0$.

The linear transport cost changes by $O(\epsilon)$. For the entropic term, the new entry $\boldsymbol{\pi}^\epsilon[i_0, j_0] = \epsilon$ contributes $\lambda\epsilon(\ln \epsilon - 1)$, whereas every other perturbed positive entry contributes only $O(\epsilon)$. Therefore

$$\mathcal{W}(\boldsymbol{\nu}, \boldsymbol{\mu}; \boldsymbol{\pi}^\epsilon) \leq \mathcal{W}(\boldsymbol{\nu}, \boldsymbol{\mu}; \boldsymbol{\pi}^*) + \lambda\epsilon \ln \epsilon + O(\epsilon), \tag{15}$$

where $\mathcal{W}(\boldsymbol{\nu}, \boldsymbol{\mu}; \boldsymbol{\pi})$ denotes the value of the objective in (1) at the feasible plan $\boldsymbol{\pi}$. Since $\epsilon \ln \epsilon < 0$ and $|\epsilon \ln \epsilon| = \omega(\epsilon)$ as $\epsilon \downarrow 0$(equivalently $\epsilon = o(|\epsilon \ln \epsilon|)$), the right-hand side is strictly smaller than $\mathcal{W}(\boldsymbol{\nu}, \boldsymbol{\mu}; \boldsymbol{\pi}^*)$ for all sufficiently small $\epsilon > 0$, since $\lambda\epsilon \ln \epsilon + O(\epsilon) = \lambda\epsilon \ln \epsilon(1 + o(1)) < 0$ This contradicts the optimality of $\boldsymbol{\pi}^*$.

Hence no entry in the block $S \times [N]$ can vanish, and the proof is completed □

We now give an expression of the Hessian of the entropy regularised Wasserstein distance from (1), restricted to the tangent space of valid perturbations to a probability vector.

Lemma 12 (Hessian of the entropic Wasserstein distance on a restricted support)**.** Let $S \subseteq D$ be nonempty, let $\boldsymbol{\nu} \in \Delta_S$, with $\boldsymbol{\nu}[i] > 0, \forall i \in S$ and let $\boldsymbol{\mu} \in \Delta_D$, $\boldsymbol{\mu} > 0$. Let $\boldsymbol{\pi}^*$ be the unique optimizer in (1), and let $\boldsymbol{\pi}^*[S, D] \in \mathbb{R}^{|S| \times N}$ denote the submatrix formed by the rows of $\boldsymbol{\pi}^*$ indexed by the index set of S, and $\boldsymbol{\nu}[S]$ denote the sub-vector formed by the entries of $\boldsymbol{\nu}$ indexed by the index set of S. Define $\mathcal{J}_{\boldsymbol{\nu}} := \operatorname{diag}(\boldsymbol{\nu}[S]) - \boldsymbol{\pi}^*[S, D] \operatorname{diag}(\boldsymbol{\mu}^{-1})(\boldsymbol{\pi}^*[S, D])^\top \in \mathbb{R}^{|S| \times |S|}$. Then: (1) $\mathcal{J}_{\boldsymbol{\nu}}$ is symmetric positive semidefinite; (2) $\ker(\mathcal{J}_{\boldsymbol{\nu}}) = \operatorname{span}\{\mathbf{1}_S\}$; (3) for every $z \in \mathcal{N}_S$, $\nabla^2_{\boldsymbol{\nu}} \mathcal{W}(\boldsymbol{\nu}, \boldsymbol{\mu}) z = \lambda \mathcal{J}^\dagger_{\boldsymbol{\nu}} z$.

Proof. By Proposition 11, every row of $\boldsymbol{\pi}^*$ outside the indices of corpus items in S is identically zero, and every entry of $\boldsymbol{\pi}^*$ in the block $S \times [N]$ is strictly positive. Considering the dual form, and using the KKT conditions for first order optimality, we get vectors $\psi^* \in \mathbb{R}^{|S|}$ and $\phi^* \in \mathbb{R}^N$ such that, for every $i \in S$ and $k \in [N]$, $\boldsymbol{C}[i, k] + \lambda \ln \boldsymbol{\pi}^*[i, k] - \psi^*[i] - \phi^*[k] = 0$. Equivalently,

$$\boldsymbol{\pi}^*[i, k] = \exp\left(\frac{\psi^*[i] + \phi^*[k] - \boldsymbol{C}_{ik}}{\lambda}\right), \qquad i \in S,\ k \in [N]. \tag{16}$$

The pair (ψ^*, ϕ^*) is unique up to the transformation $(\psi^*, \phi^*) \mapsto (\psi^* + c\mathbf{1}_S, \phi^* - c\mathbf{1})$ for some constant $c \in \mathbb{R}$. We fix this by imposing the normalization $\mathbf{1}_S^\top \psi^* = 0$.

We now show that $\mathcal{J}_{\boldsymbol{\nu}}$ is a psd matrix, with the vector $\mathbf{1}$ being it's kernel. For each $k \in [N]$, define $\boldsymbol{p}^{(k)} \in \mathbb{R}^{|S|}$ by $\boldsymbol{p}^{(k)}[i] := \boldsymbol{\pi}^*[i, k]/\boldsymbol{\mu}[k]$ for $i \in S$. Since $\boldsymbol{\mu}[k] > 0$ and $\boldsymbol{\pi}^*[i, k] > 0$ for

every $i \in S$, $k \in [N]$, each $\boldsymbol{p}^{(k)}$ has strictly positive entries. Also, $\sum_{i\in S} \boldsymbol{p}^{(k)}[i] = 1$, because $\sum_{i\in S} \boldsymbol{\pi}^*[i,k] = \boldsymbol{\mu}[k]$, and $\sum_{k=1}^N \boldsymbol{\mu}[k]\boldsymbol{p}^{(k)} = \boldsymbol{\nu}[S]$, because $\sum_{k=1}^N \boldsymbol{\pi}^*[i,k] = \boldsymbol{\nu}[i]$ for every $i \in S$.

Fix any $\boldsymbol{z} \in \mathbb{R}^{|S|}$. Then we have:

$$\begin{aligned}
\boldsymbol{z}^\top \mathcal{J}_{\boldsymbol{\nu}} \boldsymbol{z} &= \boldsymbol{z}^\top \operatorname{diag}(\boldsymbol{\nu}[S])\boldsymbol{z} - ((\boldsymbol{\pi}^*[S,D])^\top \boldsymbol{z})^\top \operatorname{diag}(\boldsymbol{\mu}^{-1})((\boldsymbol{\pi}^*[S,D])^\top \boldsymbol{z}) && (17)\\
&= \sum_{i\in S} \boldsymbol{\nu}[i]\boldsymbol{z}[i]^2 - \sum_{k=1}^N \frac{1}{\boldsymbol{\mu}[k]} \left(\sum_{i\in S} \boldsymbol{\pi}^*[i,k]\boldsymbol{z}[i]\right)^2 \\
&= \sum_{i\in S}\sum_{k=1}^N \boldsymbol{z}[i]^2 \boldsymbol{p}^{(k)}[i]\boldsymbol{\mu}[k] - \sum_{k=1}^N \frac{1}{\boldsymbol{\mu}[k]} \left(\boldsymbol{\mu}[k]\sum_{i\in S} \boldsymbol{p}^{(k)}[i]\boldsymbol{z}[i]\right)^2 && (18)\\
&= \sum_{k=1}^N \boldsymbol{\mu}[k] \left(\sum_{i\in S} \boldsymbol{p}^{(k)}[i]\boldsymbol{z}[i]^2 - \left(\sum_{i\in S} \boldsymbol{p}^{(k)}[i]\boldsymbol{z}[i]\right)^2\right). && (19)
\end{aligned}$$

For each fixed k, using Cauchy Schwarz inequality, we get:

$$\begin{aligned}
&\left(\sum_{i\in S} \boldsymbol{p}^{(k)}[i]\right)\left(\sum_{i\in S} \boldsymbol{p}^{(k)}[i]\boldsymbol{z}[i]^2\right) \geq \left(\sum_{i\in S} \boldsymbol{p}^{(k)}[i]\boldsymbol{z}[i]\right)^2 \\
\implies &\sum_{i\in S} \boldsymbol{p}^{(k)}[i]\boldsymbol{z}[i]^2 \geq \left(\sum_{i\in S} \boldsymbol{p}^{(k)}[i]\boldsymbol{z}[i]\right)^2, \text{ using } \sum_{i\in S} \boldsymbol{p}^{(k)}[i] = 1
\end{aligned} \tag{20}$$

Substituting (20) into (17), we obtain

$$\boldsymbol{z}^\top \mathcal{J}_{\boldsymbol{\nu}} \boldsymbol{z} = \sum_{k=1}^N \boldsymbol{\mu}[k] \left(\sum_{i\in S} \boldsymbol{p}^{(k)}[i]\boldsymbol{z}[i]^2 - \left(\sum_{i\in S} \boldsymbol{p}^{(k)}[i]\boldsymbol{z}[i]\right)^2\right) \geq 0 \tag{21}$$

Thus $\boldsymbol{z}^\top \mathcal{J}_{\boldsymbol{\nu}} \boldsymbol{z} \geq 0$ for all $\boldsymbol{z} \in \mathbb{R}^{|S|}$, so $\mathcal{J}_{\boldsymbol{\nu}}$ is positive semidefinite.

If $\boldsymbol{z}^\top \mathcal{J}_{\boldsymbol{\nu}} \boldsymbol{z} = 0$, then equality must hold for Cauchy Schwarz, this implies that $\boldsymbol{z}[i] = \boldsymbol{z}[\ell]$ for all $i, \ell \in S$. Hence z is constant on the index set of S, so $\ker(\mathcal{J}_{\boldsymbol{\nu}}) = \operatorname{span}\{\mathbf{1}_S\}$.

We now show that the hessian of the Wasserstein objective, restricted to $\mathcal{N}_S$ is $\lambda \mathcal{J}_{\boldsymbol{\nu}}^\dagger$. Fix any $\boldsymbol{z} \in \mathcal{N}_S$, and define $\boldsymbol{\nu}(t) := \boldsymbol{\nu} + t\boldsymbol{z}$ for $|t|$ sufficiently small. Since $\boldsymbol{\nu} \in \operatorname{int}(\Delta_S)$, and $\sum_{i\in S} z[i] = 0$, one has $\boldsymbol{\nu}(t) \in \Delta_S$ and $\boldsymbol{\nu}(t)[i] > 0, \forall i \in S$ for all sufficiently small $|t|$. For each t with $|t|$ sufficiently small, define the map $P : \mathbb{R}^{|S|} \times \mathbb{R}^N \times \mathbb{R} \to \mathbb{R}^{|S|+N}$ by

$$P(\psi, \phi, t) := \left((P^{\text{row}}(\psi,\phi,t)[i])_{i\in S}, \left(P^{\text{col}}(\psi,\phi)[k]\right)_{k=1}^{N-1}, P^{\text{norm}}(\psi)\right), \tag{22}$$

where

$$\begin{aligned}
P^{\text{row}}(\psi,\phi,t)[i] &:= \sum_{k=1}^N \exp\left(\frac{\psi[i] + \phi[k] - \boldsymbol{C}[i,k]}{\lambda}\right) - \boldsymbol{\nu}(t)[i], \qquad i \in S, \\
P^{\text{col}}(\psi,\phi)[k] &:= \sum_{i\in S} \exp\left(\frac{\psi[i] + \phi[k] - \boldsymbol{C}[i,k]}{\lambda}\right) - \boldsymbol{\mu}[k], \qquad k = 1, \ldots, N-1, \\
P^{\text{norm}}(\psi) &:= \mathbf{1}_S^\top \psi.
\end{aligned} \tag{23}$$

The map P is C^∞. Moreover, if (ψ^*, ϕ^*) denotes the normalized dual potentials at $t = 0$, then $P(\psi^*, \phi^*, 0) = 0$. The equation for the last column $k = N$ is not included because it is implied by the others. Hence $P(\psi, \phi, t) = 0$ is a square system in the unknowns $(\psi, \phi) \in \mathbb{R}^{|S|} \times \mathbb{R}^N$.

We now verify that the partial derivative $D_{(\psi,\phi)}P(\psi^*, \phi^*, 0)$ is invertible. Since the domain and codomain have the same dimension, it suffices to prove that its kernel is trivial. Let $(\boldsymbol{u}, \mathbf{v}) \in \mathbb{R}^{|S|} \times \mathbb{R}^N$ satisfy $D_{(\psi,\phi)}P(\psi^*, \phi^*, 0) \cdot (\boldsymbol{u}, \mathbf{v}) = 0$. Writing this condition componentwise yields

the linearized equations

$$\begin{aligned}
&\sum_{k=1}^{N} \frac{\boldsymbol{\pi}^*[i,k]}{\lambda} \left(\boldsymbol{u}[i] + \mathbf{v}[k]\right) = 0, \qquad i \in S, \\
&\sum_{i \in S} \frac{\boldsymbol{\pi}^*[i,k]}{\lambda} \left(\boldsymbol{u}[i] + \mathbf{v}[k]\right) = 0, \qquad k = 1, \ldots, N-1, \\
&\text{and } \mathbf{1}_S^\top \boldsymbol{u} = 0.
\end{aligned} \tag{24}$$

Using $\sum_{k=1}^{N} \boldsymbol{\pi}^*[i,k] = \boldsymbol{\nu}[i]$ for $i \in S$ and $\sum_{i \in S} \boldsymbol{\pi}^*[i,k] = \boldsymbol{\mu}[k]$ for $k \in [N]$,the above equations are equivalent to

$$\begin{aligned}
&\frac{1}{\lambda}\boldsymbol{\nu}[i]\boldsymbol{u}[i] + \frac{1}{\lambda}\sum_{k=1}^{N} \boldsymbol{\pi}^*[i,k]\boldsymbol{v}[k] = 0, \; \forall i \in S \\
&\frac{1}{\lambda}\sum_{i \in S} \pi^*[i,k]\boldsymbol{u}[i] + \frac{1}{\lambda}\boldsymbol{\mu}[k]\boldsymbol{v}[k] = 0, \; \forall k \in \{1, \cdots, N-1\}
\end{aligned} \tag{25}$$

In matrix form the above equations can be written as:

$$\begin{aligned}
&\operatorname{diag}(\boldsymbol{\nu}[S])\boldsymbol{u} + \boldsymbol{\pi}^*[S,D]\mathbf{v} = 0, \\
&(\boldsymbol{\pi}^*[S,[N-1]])^\top \boldsymbol{u} + \operatorname{diag}(\boldsymbol{\mu}[[N-1]])\mathbf{v}[[N-1]] = 0
\end{aligned} \tag{26}$$

We left multiply the first part of (26) by $\mathbf{1}_S$ to obtain:

$$\boldsymbol{\nu}[S]^\top \boldsymbol{u} + \boldsymbol{\mu}^\top \boldsymbol{v} = 0, \text{ using } \mathbf{1}_S^\top \pi^*[S,D] = \boldsymbol{\mu}^\top \tag{27}$$

Now, let $\boldsymbol{r} \in \mathbb{R}^N$ be defined as: $\boldsymbol{r} \coloneqq \operatorname{diag}(\boldsymbol{\mu})\boldsymbol{v} + \boldsymbol{\pi}^*[S,D]^\top \boldsymbol{u}$. Then, left multiplying by $\mathbf{1}_N^\top$ gives us:

$$\begin{aligned}
&\mathbf{1}_N^\top \boldsymbol{r} = (\boldsymbol{\pi}^*[S,D]\mathbf{1}_N)^\top \boldsymbol{u} + \boldsymbol{\mu}^\top \boldsymbol{v} \\
&\implies \mathbf{1}_N^\top \boldsymbol{r} = \boldsymbol{\nu}[S]^\top \boldsymbol{u} + \boldsymbol{\mu}^\top \boldsymbol{v} = \mathbf{0}, \text{ from (27)}
\end{aligned} \tag{28}$$

Therefore we have: $(\boldsymbol{\pi}^*[S,D])^\top \boldsymbol{u} + \operatorname{diag}(\boldsymbol{\mu})\mathbf{v} = 0$. Eliminating $\mathbf{v} = -\operatorname{diag}(\boldsymbol{\mu}^{-1})(\boldsymbol{\pi}^*[S,D])^\top \boldsymbol{u}$ from the first part of (26), we get

$$\left(\operatorname{diag}(\boldsymbol{\nu}[S]) - \boldsymbol{\pi}^*[S,D]\operatorname{diag}(\boldsymbol{\mu}^{-1})(\boldsymbol{\pi}^*[S,D])^\top\right)\boldsymbol{u} = \mathcal{J}_{\boldsymbol{\nu}}\boldsymbol{u} = 0. \tag{29}$$

By the previously established characterization $\ker(\mathcal{J}_{\boldsymbol{\nu}}) = \operatorname{span}\{\mathbf{1}_S\}$, and since $\mathbf{1}_S^\top \boldsymbol{u} = 0$, we conclude that $\boldsymbol{u} = 0$. Then the second part of (26) yields $\operatorname{diag}(\boldsymbol{\mu})\mathbf{v} = 0$, and since $\boldsymbol{\mu} > 0$, the diagonal matrix $\operatorname{diag}(\boldsymbol{\mu})$ is invertible, so $\mathbf{v} = 0$. Thus $\ker\left(D_{(\psi,\phi)}P(\psi^*,\phi^*,0)\right) = \{(0,0)\}$. Hence $D_{(\psi,\phi)}P(\psi^*,\phi^*,0)$ is invertible.

We may therefore apply Implicit function theorem(see (Spivak, 1965)) to the equation $P(\psi,\phi,t) = 0$ at the point $(\psi^*,\phi^*,0)$. It follows that there exist $\epsilon > 0$, an open neighborhood U of (ψ^*,ϕ^*) in $\mathbb{R}^{|S|} \times \mathbb{R}^N$, and a unique C^1 map $(-\epsilon,\epsilon) \ni t \mapsto (\psi^*(t),\phi^*(t)) \in U$, such that $(\psi^*(0),\phi^*(0)) = (\psi^*,\phi^*)$ and $P(\psi^*(t),\phi^*(t),t) = 0$ for all $|t| < \epsilon$. In particular, the normalized dual potentials depend differentiably on t near 0. We may therefore define

$$\boldsymbol{a} \coloneqq \dot{\psi}^*(0), \qquad \boldsymbol{b} \coloneqq \dot{\phi}^*(0). \tag{30}$$

Finally, differentiating the normalization $\mathbf{1}_S^\top \psi^*(t) = 0$ at $t = 0$ gives $\mathbf{1}_S^\top \boldsymbol{a} = 0$, and hence $\boldsymbol{a} \in \mathcal{N}_S$.

For each t, we thus define $\boldsymbol{\pi}[i,k](t) \coloneqq \exp((\psi_i^*(t) + \phi_k^*(t) - \boldsymbol{C}[i,k])/\lambda)$ for $i \in S$ and $k \in [N]$. Differentiating at $t = 0$ yields

$$\dot{\boldsymbol{\pi}}[i,k](0) = \frac{\boldsymbol{\pi}^*[i,k]}{\lambda}\left(\boldsymbol{a}_i + \boldsymbol{b}_k\right), \qquad i \in S, \; k \in [N]. \tag{31}$$

Differentiating the row constraints $\sum_{k=1}^{N} \boldsymbol{\pi}[i,k](t) = \boldsymbol{\nu}[i](t)$ gives

$$\operatorname{diag}(\boldsymbol{\nu}[S])\boldsymbol{a} + \boldsymbol{\pi}^*[S,D]\boldsymbol{b} = \lambda \boldsymbol{z} \tag{32}$$

Differentiating the column constraints $\sum_{i \in S} \boldsymbol{\pi}[i,k](t) = \boldsymbol{\mu}[k]$ gives

$$(\boldsymbol{\pi}^*[S,D])^\top \boldsymbol{a} + \operatorname{diag}(\boldsymbol{\mu})\boldsymbol{b} = 0. \tag{33}$$

Eliminating $\boldsymbol{b} = -\operatorname{diag}(\boldsymbol{\mu}^{-1})(\boldsymbol{\pi}^*[S,D])^\top \boldsymbol{a}$ from (33) and substituting into (32), we obtain $\mathcal{J}_{\boldsymbol{\nu}}\boldsymbol{a} = \lambda \boldsymbol{z}$. Since $\mathcal{J}_{\boldsymbol{\nu}}$ is positive definite on $\mathcal{N}_S$, so the unique solution is $\boldsymbol{a} = \lambda \mathcal{J}_{\boldsymbol{\nu}}^\dagger z$.

It remains to identify the Hessian. The dual formulation of (1), restricted to the active rows, is

$$\mathcal{W}(\boldsymbol{\nu}, \boldsymbol{\mu}) = \max_{\psi \in \mathbb{R}^{|S|},\, \phi \in \mathbb{R}^N} \left(\sum_{i \in S} \psi[i]\boldsymbol{\nu}[i] + \sum_{k=1}^{N} \phi[k]\boldsymbol{\mu}[k] - \lambda \sum_{i \in S} \sum_{k=1}^{N} \exp\left(\frac{\psi[i] + \phi[k] - \boldsymbol{C}[i,k]}{\lambda} \right) \right). \tag{34}$$

At the normalized optimum (ψ^*, ϕ^*), differentiation with respect to $\boldsymbol{\nu}$ in the direction $\boldsymbol{z} \in \mathcal{N}_S$ gives

$$\frac{\mathrm{d}}{\mathrm{d}t} \mathcal{W}(\boldsymbol{\nu} + t\boldsymbol{z}, \boldsymbol{\mu})\Big|_{t=0} = \sum_{i \in S} \psi^*[i]\boldsymbol{z}[i]. \tag{35}$$

Thus the gradient of $\mathcal{W}(\cdot, \boldsymbol{\mu})$ on $\mathcal{N}_S$ is the normalized potential ψ^*. Differentiating once more in the direction $\boldsymbol{z}$, the Hessian applied to $\boldsymbol{z}$ is exactly $\boldsymbol{a}$. Thus, we conclude that, $\nabla^2_{\boldsymbol{\nu}} \mathcal{W}(\boldsymbol{\nu}, \boldsymbol{\mu})\boldsymbol{z} = \lambda \mathcal{J}^\dagger_{\boldsymbol{\nu}} \boldsymbol{z}$, for any given $\boldsymbol{z} \in \mathcal{N}_S$ This proves the lemma on the Hessian. □

Using this, we compute the Hessian of the entire RSWB objective, using the linearity of the Hessian operator.

Corollary 13 (Hessian of the restricted-support barycenter objective)**.** Fix a nonempty support set $S \subseteq D$, let $\boldsymbol{\nu} \in \Delta_S$ with $\boldsymbol{\nu}[i] > 0, \forall \boldsymbol{x}_i \in S$, and for each $r \in [B]$ let $\boldsymbol{\pi}^*_r$ be the optimal coupling for $\mathcal{W}(\boldsymbol{\nu}, \boldsymbol{\mu}_r)$. Define $\mathcal{J}_r(\boldsymbol{\nu}) \coloneqq \operatorname{diag}(\boldsymbol{\nu}) - \boldsymbol{\pi}^*_r \operatorname{diag}(\boldsymbol{\mu}_r^{-1})(\boldsymbol{\pi}^*_r)^\top$. Then the Hessian of the full RSWB objective from (2) restricted to the tangent space $\mathcal{N}_S$ is given by:

$$\nabla^2_{\boldsymbol{\nu}} W(\boldsymbol{\nu})\big|_{\mathcal{N}_S} = \lambda \sum_{r=1}^{B} \beta_r \mathcal{J}_r(\boldsymbol{\nu})^\dagger\big|_{\mathcal{N}_S} + \tau \operatorname{diag}(\boldsymbol{\nu}^{-1})\big|_{\mathcal{N}_S}. \tag{36}$$

Proof. The contribution of the r-th Wasserstein term follows from Lemma 12. Also, $-\tau \operatorname{Ent}(\boldsymbol{\nu}) = \tau \sum_{i=1}^{N} \boldsymbol{\nu}[i] \ln \boldsymbol{\nu}[i]$. So $\nabla^2_{\boldsymbol{\nu}}(-\tau \operatorname{Ent})(\boldsymbol{\nu}) = \tau \operatorname{diag}(\boldsymbol{\nu}^{-1})$. Summing the Hessians with the weights β_r yields (36). □

Proposition 14 (Coordinatewise lower bound on $\boldsymbol{\nu}^*(S)$)**.** Fix a nonempty support set $S \subseteq D$. Then $\boldsymbol{\nu}^*(S) \in \operatorname{int}(\Delta_S)$, and for every $s \in S$, we have:

$$\boldsymbol{\nu}^*(S)[s] \geq \frac{1}{|S|} \exp\left(-\frac{\delta_{\boldsymbol{C}}}{\tau + \lambda} \right). \tag{37}$$

Proof. Here we consider δ_c equal to $\delta \coloneqq \max_{s,t,s',t' \in D} |\boldsymbol{C}[s,t] - \boldsymbol{C}[s',t']|$. We first prove that every coordinate of $\boldsymbol{\nu}^*(S)$ is strictly positive. Suppose, for contradiction, that $\boldsymbol{\nu}^*(S)[u] = 0$ for some item $\boldsymbol{x}_u, u \in S$. Let $\boldsymbol{\nu}^\epsilon \coloneqq (1-\epsilon)\boldsymbol{\nu}^*(S) + \epsilon \boldsymbol{e}_u$, for a given $\epsilon \in (0,1)$, where $\boldsymbol{e}_u$ is the one-hot vector in $\mathbb{R}^N$ with 1 at position u and 0 everywhere else.

For each $r \in [B]$, let $\boldsymbol{\pi}^*_r$ be the optimal coupling for $\mathcal{W}(\boldsymbol{\nu}^*(S), \boldsymbol{\mu}_r)$, and we define

$$\boldsymbol{\pi}^\epsilon_r \coloneqq (1-\epsilon)\boldsymbol{\pi}^*_r + \epsilon \boldsymbol{e}_u \boldsymbol{\mu}_r^\top. \tag{38}$$

Then $\boldsymbol{\pi}^\epsilon_r \geq 0$, and we note that the perturbed coupling is actually a coupling, since

$$\begin{aligned} \boldsymbol{\pi}^\epsilon_r \mathbf{1} &= (1-\epsilon)\boldsymbol{\pi}^*_r \mathbf{1} + \epsilon \boldsymbol{e}_u \boldsymbol{\mu}_r^\top \mathbf{1} \\ &= (1-\epsilon)\boldsymbol{\nu}^*(S) + \epsilon \boldsymbol{e}_u = \boldsymbol{\nu}^\epsilon, \\ (\boldsymbol{\pi}^\epsilon_r)^\top \mathbf{1} &= (1-\epsilon)(\boldsymbol{\pi}^*_r)^\top \mathbf{1} + \epsilon \boldsymbol{\mu}_r \boldsymbol{e}_u^\top \mathbf{1} \\ &= (1-\epsilon)\boldsymbol{\mu}_r + \epsilon \boldsymbol{\mu}_r = \boldsymbol{\mu}_r. \end{aligned}$$

Thus $\boldsymbol{\pi}^\epsilon_r \in \Pi(\boldsymbol{\mu}_r, \boldsymbol{\nu}^\epsilon)$ for every $\epsilon \in (0,1)$. For each $r \in [B]$, define:

$$P(\boldsymbol{\pi}) \coloneqq \sum_{i,j=1}^{N} \boldsymbol{C}[i,j]\boldsymbol{\pi}[i,j] + \lambda \sum_{i,j=1}^{N} \boldsymbol{\pi}[i,j](\ln \boldsymbol{\pi}[i,j] - 1).$$

Then, by definition, $\mathcal{W}(\boldsymbol{\nu}, \boldsymbol{\mu}_r) = \min_{\boldsymbol{\pi} \in \Pi(\boldsymbol{\mu}_r, \boldsymbol{\nu})} P(\boldsymbol{\pi})$. Since $\boldsymbol{\pi}^\epsilon_r \in \Pi(\boldsymbol{\mu}_r, \boldsymbol{\nu}^\epsilon)$, we have $\mathcal{W}(\boldsymbol{\nu}^\epsilon, \boldsymbol{\mu}_r) \leq P(\boldsymbol{\pi}^\epsilon_r)$. Moreover, by the definition of $\boldsymbol{\pi}^\epsilon_r$ and the convexity of P we get:

$$\begin{aligned} P(\boldsymbol{\pi}^\epsilon_r) &= P\left((1-\epsilon)\boldsymbol{\pi}^*_r + \epsilon \boldsymbol{e}_u \boldsymbol{\mu}_r^\top \right) \\ &\leq (1-\epsilon)P(\boldsymbol{\pi}^*_r) + \epsilon P(\boldsymbol{e}_u \boldsymbol{\mu}_r^\top). \end{aligned}$$

Since $\boldsymbol{\pi}^*_r$ is optimal for $\mathcal{W}(\boldsymbol{\nu}^*(S), \boldsymbol{\mu}_r)$, one has $P(\boldsymbol{\pi}^*_r) = \mathcal{W}(\boldsymbol{\nu}^*(S), \boldsymbol{\mu}_r)$. Hence

$$\mathcal{W}(\boldsymbol{\nu}^\epsilon, \boldsymbol{\mu}_r) \leq (1-\epsilon)\mathcal{W}(\boldsymbol{\nu}^*(S), \boldsymbol{\mu}_r) + \epsilon P(\boldsymbol{e}_u \boldsymbol{\mu}_r^\top)$$

$$= \mathcal{W}(\boldsymbol{\nu}^*(S), \boldsymbol{\mu}_r) + \epsilon \left(P(\boldsymbol{e}_u \boldsymbol{\mu}_r^\top) - \mathcal{W}(\boldsymbol{\nu}^*(S), \boldsymbol{\mu}_r) \right). \tag{39}$$

Summing over $r = 1, \ldots, B$, we obtain that,

$$\sum_{r=1}^{B} \beta_r \mathcal{W}(\boldsymbol{\nu}^\epsilon, \boldsymbol{\mu}_r) \leq \sum_{r=1}^{B} \beta_r \mathcal{W}(\boldsymbol{\nu}^*(S), \boldsymbol{\mu}_r) + K_1 \epsilon, \tag{40}$$

where $K_1 := \sum_{r=1}^{B} \beta_r \left(P(\boldsymbol{e}_u \boldsymbol{\mu}_r^\top) - \mathcal{W}(\boldsymbol{\nu}^*(S), \boldsymbol{\mu}_r) \right)$. Since $\boldsymbol{C}$ is finite and $\boldsymbol{\mu}_r \in \operatorname{int}(\Delta_D)$ for every $r \in [B]$, each quantity $P(e_u \boldsymbol{\mu}_r^\top)$ is finite, and therefore $K_1 \in \mathbb{R}$ is finite.

For the entropy term, since $\boldsymbol{\nu}^*(S)[u] = 0$, we have:

$$\sum_{i=1}^{N} \boldsymbol{\nu}^\epsilon[i] \ln \boldsymbol{\nu}^\epsilon[i] = \epsilon \ln \epsilon + \sum_{i \neq u} (1-\epsilon) \boldsymbol{\nu}^*(S)[i] \ln \left((1-\epsilon) \boldsymbol{\nu}^*(S)[i] \right)$$

$$= \sum_{i=1}^{N} \boldsymbol{\nu}^*(S)[i] \ln \boldsymbol{\nu}^*(S)[i] + \epsilon \ln \epsilon + O(\epsilon).$$

Hence there exists a constant $K_2 < \infty$, independent of ϵ, such that

$$-\tau \operatorname{Ent}(\boldsymbol{\nu}^\epsilon) \leq -\tau \operatorname{Ent}(\boldsymbol{\nu}^*(S)) + \tau \epsilon \ln \epsilon + K_2 \epsilon. \tag{41}$$

Combining both the expressions, we obtain

$$W(\boldsymbol{\nu}^\epsilon) \leq W(\boldsymbol{\nu}^*(S)) + \tau \epsilon \ln \epsilon + (K_1 + K_2)\epsilon. \tag{42}$$

Since $\epsilon \ln \epsilon < 0$ and $|\epsilon \ln \epsilon| = \omega(\epsilon)$ as $\epsilon \downarrow 0$ (equivalently $\epsilon = o(|\epsilon \ln \epsilon|)$), the right-hand side is strictly smaller than $W(\boldsymbol{\nu}^*(S))$ for all sufficiently small $\epsilon > 0$ [since $\tau \epsilon \ln \epsilon + O(\epsilon) = \tau \epsilon \ln \epsilon (1 + o(1)) < 0$], contradicting the optimality of $\boldsymbol{\nu}^*(S)$. Hence, $\boldsymbol{\nu}^*(S)[s] > 0, \ \forall s \in S$.

Now, fix $r \in [B]$. Since $\boldsymbol{\nu}^*(S) \in \operatorname{int}(\Delta_S)$ and $\boldsymbol{\mu}_r \in \operatorname{int}(\Delta_D)$, Prop. 11 implies that the optimal coupling $\boldsymbol{\pi}_r^*$ for $\mathcal{W}(\boldsymbol{\nu}^*(S), \boldsymbol{\mu}_r)$ is strictly positive on the active rows. Therefore there exist vectors (from KKT conditions on the dual, following (Cuturi, 2013)) $\boldsymbol{u}_r \in \mathbb{R}^{|S|}, \boldsymbol{v}_r \in \mathbb{R}^N$ such that

$$(\boldsymbol{\pi}_r^*)[i, j] = \exp\left(\frac{\boldsymbol{u}_r[i] + \boldsymbol{v}_r[j] - \boldsymbol{C}[i, j]}{\lambda} \right), \ \forall i \in S, \ j \in [N] \tag{43}$$

We define $\boldsymbol{Z}_1, \cdots, \boldsymbol{Z}_B \in \mathbb{R}^{|S|}$ where $\boldsymbol{Z}_r[i] := \sum_{j=1}^{N} \exp\left(\frac{\boldsymbol{v}_r[j] - \boldsymbol{C}[i,j]}{\lambda} \right)$. Summing (43) over j and using $\sum_j (\boldsymbol{\pi}_r^*)[i, j] = \boldsymbol{\nu}^*(S)[i]$ yields

$$\boldsymbol{\nu}^*(S)[i] = e^{\boldsymbol{u}_r[i]/\lambda} \boldsymbol{Z}_r[i], \qquad \boldsymbol{u}_r[i] = \lambda \ln \boldsymbol{\nu}^*(S)[i] - \lambda \ln \boldsymbol{Z}_r[i]. \tag{44}$$

Since $\boldsymbol{\nu}^*(S) \in \operatorname{int}(\Delta_S)$, first-order optimality for the minimization of $W(\bullet, \{\boldsymbol{\mu}_r(\boldsymbol{q}, D) \,|\, r \in [B]\})$ over Δ_S implies that there exists a scalar slack variable (Lagrange multiplier associated with the equality constraint $\sum_{i \in S} \boldsymbol{\nu}_i = 1$.) $\xi \in \mathbb{R}$ such that

$$\sum_{r=1}^{B} \beta_r \boldsymbol{u}_r[i] + \tau(\ln \boldsymbol{\nu}^*(S)[i] + 1) + \xi = 0, \qquad i \in S. \tag{45}$$

Substituting (44) into (45), we obtain,

$$\begin{aligned} (\tau + \lambda) \ln \boldsymbol{\nu}^*(S)[i] &= \lambda \sum_{r=1}^{B} \beta_r \ln \boldsymbol{Z}_r[i] - (\tau + \xi). \\ \implies \ln \boldsymbol{\nu}^*(S)[i] &= \frac{\lambda}{\tau + \lambda} \sum_{r=1}^{B} \beta_r \ln \boldsymbol{Z}_r[i] - \frac{\tau + \xi}{\tau + \lambda} \end{aligned} \tag{46}$$

We define $\theta_i := \frac{\lambda}{\tau + \lambda} \sum_{r=1}^{B} \beta_r \ln \boldsymbol{Z}_r[i]$, and $c := -\frac{\tau + \xi}{\tau + \lambda}$. Then we have: $\boldsymbol{\nu}^*(S)[i] = e^c e^{\theta_i}$. Since $\sum_{i \in S} \boldsymbol{\nu}^*(S)[i] = 1$, we have, $e^c = \frac{1}{\sum_{k \in S} e^{\theta_k}}$.

Substituting the value of c, we obtain:

$$\boldsymbol{\nu}^*(S)[i] = \frac{e^{\theta_i}}{\sum_{k \in S} e^{\theta_k}}. \tag{47}$$

We next bound the oscillation of $\{\theta_i\}_{i\in S}$. Fix $r \in [B]$ and $i, k \in S$. For every $j \in [N]$,

$$\exp\left(\frac{\boldsymbol{v}_r[j] - \boldsymbol{C}[i,j]}{\lambda}\right) \le e^{\delta_{\boldsymbol{C}}/\lambda} \exp\left(\frac{\boldsymbol{v}_r[j] - \boldsymbol{C}[k,j]}{\lambda}\right), \tag{48}$$

because $|\boldsymbol{C}[i,j] - \boldsymbol{C}[k,j]| \le \delta_{\boldsymbol{C}}$. Summing over j gives $\boldsymbol{Z}_r[i] \le e^{\delta_{\boldsymbol{C}}/\lambda}\boldsymbol{Z}_r[k]$, and symmetrically $\boldsymbol{Z}_r[k] \le e^{\delta_{\boldsymbol{C}}/\lambda}\boldsymbol{Z}_r[i]$ Therefore we have, $|\lambda \ln \boldsymbol{Z}_r[i] - \lambda \ln \boldsymbol{Z}_r[k]| \le \delta_{\boldsymbol{C}}$. Thus, we may write:

$$\begin{aligned}
|\theta_i - \theta_k| &= \frac{\lambda}{\tau+\lambda}\left|\sum_{r=1}^{B} \beta_r \left(\ln \boldsymbol{Z}_r[i] - \ln \boldsymbol{Z}_r[k]\right)\right| \\
&\le \frac{1}{\tau+\lambda}\sum_{r=1}^{B} \beta_r \left|\lambda \ln \boldsymbol{Z}_r[i] - \lambda \ln \boldsymbol{Z}_r[k]\right|, \text{ by triangle ineq.} \\
&\le \frac{1}{\tau+\lambda}\sum_{r=1}^{B} \beta_r \delta_{\boldsymbol{C}} = \frac{\delta_{\boldsymbol{C}}}{\tau+\lambda}
\end{aligned}$$

The above inequality is true for any i, k. Hence we have: $\max_{i\in S}\theta_i - \min_{i\in S}\theta_i \le \frac{\delta_{\boldsymbol{C}}}{\tau+\lambda}$. Using the above bound in (47) we get,

$$\begin{aligned}
\boldsymbol{\nu}^*(S)[i] = \frac{e^{\theta_i}}{\sum_{k\in S} e^{\theta_k}} &\ge \frac{e^{\min_{\ell\in S}\theta_\ell}}{|S| e^{\max_{\ell\in S}\theta_\ell}} \\
&= \frac{1}{|S|}\exp\left(-\left(\max_{\ell\in S}\theta_\ell - \min_{\ell\in S}\theta_\ell\right)\right) \\
&\ge \frac{1}{|S|}\exp\left(-\frac{\delta_{\boldsymbol{C}}}{\tau+\lambda}\right).
\end{aligned}$$

This proves the result. □

We now proceed to show another lower bound, but this time on the entries of the optimal couplings.

Lemma 15 (Lower bound on the entries of the optimal couplings)**.** Fix a nonempty support set $S \subseteq D$, let $\boldsymbol{\nu} \in \mathrm{int}(\Delta_S)$, and for each $r \in [B]$ let $\boldsymbol{\pi}_r^*$ denote the optimal coupling for $\mathcal{W}(\boldsymbol{\nu}, \boldsymbol{\mu}_r)$. Then, for every $r \in [B]$, every $i \in S$, and every $j \in [N]$,

$$(\boldsymbol{\pi}_r^*)[i,j] \ge \boldsymbol{\nu}[i]\boldsymbol{\mu}_r[j]e^{-2\delta_{\boldsymbol{C}}/\lambda}. \tag{49}$$

Proof. Here we consider δ_c equal to $\delta := \max_{s,t,s',t'\in D}|\boldsymbol{C}[s,t] - \boldsymbol{C}[s',t']|$. Fix $r \in [B]$. Since $\boldsymbol{\nu} \in \mathrm{int}(\Delta_S)$ and $\boldsymbol{\mu}_r \in \mathrm{int}(\Delta_D)$, Proposition 11 implies that every row of $\boldsymbol{\pi}_r^*$ outside S is identically zero, and every entry of $\boldsymbol{\pi}_r^*$ in the block $S \times [N]$ is strictly positive. Therefore, from the Sinkhorn factorization (Cuturi, 2013) there exist positive vectors $\boldsymbol{a}_r \in \mathbb{R}_{>0}^{|S|}$ and $\boldsymbol{b}_r \in \mathbb{R}_{>0}^{N}$ for each $r \in [B]$ such that $\boldsymbol{\pi}_r^*[i,j] = \boldsymbol{a}_r[i]\boldsymbol{b}_r[j]e^{-\boldsymbol{C}[i,j]/\lambda}$ for all $i \in S,\ j \in [N]$.

For $i \in S$ and $j \in [N]$, we define

$$p(j|i)_r := \frac{\boldsymbol{\pi}_r^*[i,j]}{\boldsymbol{\nu}[i]}. \tag{50}$$

Since $\sum_{j=1}^{N}(\boldsymbol{\pi}_r^*)[i,j] = \boldsymbol{\nu}[i]$, the vector $\{p(j|i)_r\}_{1\le j\le N}$ is a probability distribution on $[N]$ for each fixed $i \in S$. Using the expression for $\boldsymbol{\pi}_r^*$, we may write

$$p(j|i)_r = \frac{\boldsymbol{b}_r[j]e^{-\boldsymbol{C}[i,j]/\lambda}}{\sum_{\ell=1}^{N}\boldsymbol{b}_r[\ell]e^{-\boldsymbol{C}[i,\ell]/\lambda}}, \quad i \in S,\ j \in [N].$$

Fix $j \in [N]$ and $i, k \in S$. Then we have:

$$\begin{aligned}
\frac{p(j|i)_r}{p(j|k)_r} &= e^{-(\boldsymbol{C}[i,j]-\boldsymbol{C}[k,j])/\lambda}\frac{\sum_{\ell=1}^{N}\boldsymbol{b}_r[\ell]e^{-\boldsymbol{C}[k,\ell]/\lambda}}{\sum_{\ell=1}^{N}\boldsymbol{b}_r[\ell]e^{-\boldsymbol{C}[i,\ell]/\lambda}} \\
&\le e^{-(\boldsymbol{C}[i,j]-\boldsymbol{C}[k,j])/\lambda}\max_{\ell\in[N]} e^{-(\boldsymbol{C}[k,\ell]-\boldsymbol{C}[i,\ell])/\lambda} \\
&= \exp\left(\frac{\max_{\ell\in[N]}\left(\boldsymbol{C}[i,\ell] + \boldsymbol{C}[k,j] - \boldsymbol{C}[k,\ell] - \boldsymbol{C}[i,j]\right)}{\lambda}\right) \le e^{2\delta_{\boldsymbol{C}}/\lambda}.
\end{aligned}$$

Thus, for each fixed $j \in [N]$, any two numbers in the family $\{p(j|i)_r\}_{i\in S}$ differ by at most the multiplicative factor $e^{2\delta_{\boldsymbol{C}}/\lambda}$. Now fix $j \in [N]$, and let $\boldsymbol{m}_r$ be the vector, whose entries are given by $\boldsymbol{m}_r[j] := \min_{i\in S} p(j|i)_r$. By the previous computation, we have

$$p(j|i)_r \le e^{2\delta_{\boldsymbol{C}}/\lambda}\boldsymbol{m}_r[j] \qquad \forall i \in S. \tag{51}$$

Since $\boldsymbol{\pi}_r^*$ has row marginal $\boldsymbol{\nu}$ and column marginal $\boldsymbol{\mu}_r$, we have

$$\boldsymbol{\mu}_r[j] = \sum_{i=1}^{N}(\boldsymbol{\pi}_r^*)[i,j] = \sum_{i\in S}(\boldsymbol{\pi}_r^*)[i,j] = \sum_{i\in S}\boldsymbol{\nu}[i]p(j|i)_r.$$

Using (51) and $\sum_{i\in S}\boldsymbol{\nu}[i] = 1$, it follows that

$$\begin{aligned} \boldsymbol{\mu}_r[j] &\le e^{2\delta_{\boldsymbol{C}}/\lambda}\boldsymbol{m}_r[j]\sum_{i\in S}\boldsymbol{\nu}[i] = e^{2\delta_{\boldsymbol{C}}/\lambda}\boldsymbol{m}_r[j] \\ \implies \boldsymbol{\mu}_r[j]e^{-2\delta_{\boldsymbol{C}}/\lambda} &\le \boldsymbol{m}_r[j] \end{aligned} \tag{52}$$

By the definition of $\boldsymbol{m}_r[j]$, this yields

$$\begin{aligned} p(j|i)_r &\ge \boldsymbol{\mu}_r[j]e^{-2\delta_{\boldsymbol{C}}/\lambda} \qquad \forall i \in S,\ \forall j \in [N] \\ \implies (\boldsymbol{\pi}_r^*)[i,j] = \boldsymbol{\nu}[i]p(j|i)_r &\ge \boldsymbol{\nu}[i]\boldsymbol{\mu}_r[j]e^{-2\delta_{\boldsymbol{C}}/\lambda}, \qquad i \in S,\ j \in [N]. \end{aligned} \tag{53}$$

□

The tangent space $\mathcal{N}_S$ is the space of all valid mass preserving perturbations that takes a probability on Δ_S to another probability measure on Δ_S. Lemma 16 therefore gives face-wise curvature control for the restricted barycenter problem.

Lemma 16 (Strong convexity and local smoothness)**.** Fix a nonempty support set $S \subseteq D$. Let λ and τ be the coefficients of the regularizers in (1) and (2), respectively; and $\mathcal{N}_S$ be the defined as: $\mathcal{N}_S = \left\{\boldsymbol{h} \in \mathbb{R}^N :\ \boldsymbol{h}[s] = 0,\ \text{for all } s \notin S,\ \mathbf{1}^\top \boldsymbol{h} = 0\right\}$. Given $\eta > 0$, there exists positive constants m and L_η, such that: (1) For every $\boldsymbol{\nu} \in \Delta_S$ with $\boldsymbol{\nu}[s] > 0$, for all $s \in S$ and for every $\boldsymbol{z} \in \mathcal{N}_S$, we have $\boldsymbol{z}^\top \nabla^2_{\boldsymbol{\nu}} W(\boldsymbol{\nu}, \{\boldsymbol{\mu}_r(\boldsymbol{q}, D)\,|\,r \in [B]\})\,\boldsymbol{z} \ge m\|\boldsymbol{z}\|^2$, (2) Let $\mathcal{D}_\eta(S) := \{\boldsymbol{z} \in \Delta_S : \boldsymbol{z}[i] \ge \eta,\ \forall i \in S\}$. For every $\boldsymbol{\nu} \in \mathcal{D}_\eta$ and $\boldsymbol{z} \in \mathcal{N}_S$ we have: $\boldsymbol{z}^\top \nabla^2_{\boldsymbol{\nu}} W(\boldsymbol{\nu}, \{\boldsymbol{\mu}_r(\boldsymbol{q}, D)\,|\,r \in [B]\})\,\boldsymbol{z} \le L\|\boldsymbol{z}\|^2$.

Proof. Here we consider δ_c equal to $\delta := \max_{s,t,s',t'\in D} |\boldsymbol{C}[s,t] - \boldsymbol{C}[s',t']|$. Fix $\boldsymbol{\nu} \in \operatorname{int}(\Delta_S)$. For each $r \in [B]$, let $\boldsymbol{\pi}_r^*$ denote the optimal coupling for $\mathcal{W}(\boldsymbol{\nu}, \boldsymbol{\mu}_r)$, and for each $j \in [N]$ define the vector $\boldsymbol{q}_r(j) \in \mathbb{R}^{|S|}$ by

$$\boldsymbol{q}_r(j) := \frac{\boldsymbol{\pi}_r^*[:,j]}{\boldsymbol{\mu}_r[j]}. \tag{54}$$

Since $\boldsymbol{\mu}_r[j] > 0$ and the rows of $\boldsymbol{\pi}_r^*$ outside S are identically zero, $\boldsymbol{q}_r(j)$ is a probability vector supported on S. We define,

$$\mathcal{J}_r(\boldsymbol{\nu}) := \operatorname{diag}(\boldsymbol{\nu}) - \boldsymbol{\pi}_r^* \operatorname{diag}(\boldsymbol{\mu}_r)^{-1}(\boldsymbol{\pi}_r^*)^\top = \sum_{j=1}^{N}\boldsymbol{\mu}_r[j]\left(\operatorname{diag}(\boldsymbol{q}_r(j)) - \boldsymbol{q}_r(j)\boldsymbol{q}_r(j)^\top\right). \tag{55}$$

We first prove the strong convexity bound. By Lemma 12, $\mathcal{J}_r(\boldsymbol{\nu})$ is positive definite on $\mathcal{N}_S$ and $\ker(\mathcal{J}_r(\boldsymbol{\nu})) = \operatorname{span}\{\mathbf{1}_S\}$. Also,

$$\mathcal{J}_r(\boldsymbol{\nu}) = \operatorname{diag}(\boldsymbol{\nu}) - \boldsymbol{\pi}_r^* \operatorname{diag}(\boldsymbol{\mu}_r)^{-1}(\boldsymbol{\pi}_r^*)^\top \preceq \operatorname{diag}(\boldsymbol{\nu}) \preceq \boldsymbol{I} \tag{56}$$

Therefore every eigenvalue of $\mathcal{J}_r(\boldsymbol{\nu})\big|_{\mathcal{N}_S}$ lies in $(0,1]$, and hence every eigenvalue of $\mathcal{J}_r(\boldsymbol{\nu})^\dagger\big|_{\mathcal{N}_S}$ is at least 1. Thus, $\lambda\mathcal{J}_r(\boldsymbol{\nu})^\dagger\big|_{\mathcal{N}_S} \succeq \lambda\boldsymbol{I}_S$. Summing over $r \in [B]$, we obtain

$$\sum_{r=1}^{B}\beta_r\lambda\mathcal{J}_r(\boldsymbol{\nu})^\dagger\big|_{\mathcal{N}_S} \succeq \lambda\boldsymbol{I}_S. \tag{57}$$

Also, since $\boldsymbol{\nu}_i \le 1$ for all $i \in S$, we get: $\tau \operatorname{diag}(\boldsymbol{\nu})^\dagger\big|_{\mathcal{N}_S} \succeq \tau\boldsymbol{I}_S$. Combining this with (36) and (57) gives the strong convexity bound on the Hessian:

$$\sum_{r=1}^{B}\beta_r\lambda\mathcal{J}_r(\boldsymbol{\nu})^\dagger\big|_{\mathcal{N}_S} + \tau\operatorname{diag}(\boldsymbol{\nu})^\dagger\big|_{\mathcal{N}_S} \succeq (\lambda+\tau)\boldsymbol{I}_S. \tag{58}$$

We next prove the smoothness bound. Assume that $\boldsymbol{\nu}_i \geq \eta$ for all $i \in S$, and fix $\boldsymbol{h} \in \mathcal{N}_S$. By Lemma 15 we have, $\boldsymbol{q}_r(j)[i] = \frac{\boldsymbol{\pi}_r^*[i,j]}{\boldsymbol{\mu}_r[j]} \geq \boldsymbol{\nu}[i]e^{-2\delta_C/\lambda}$ for all $i \in S, j \in [N]$. We set $\kappa := e^{-2\delta_C/\lambda}$.

If $\kappa = 1$, then we would thus have $\boldsymbol{q}_r(j) \geq \boldsymbol{\nu}$ componentwise. Since both $\boldsymbol{q}_r(j)$ and $\boldsymbol{\nu}$ are probability vectors on S, this forces $\boldsymbol{q}_r(j) = \boldsymbol{\nu}$. In this case, $\operatorname{diag}(\boldsymbol{q}_r(j)) - \boldsymbol{q}_r(j)\boldsymbol{q}_r(j)^\top = \operatorname{diag}(\boldsymbol{\nu}) - \boldsymbol{\nu}\boldsymbol{\nu}^\top$, and hence

$$\boldsymbol{h}^\top \left(\operatorname{diag}(\boldsymbol{q}_r(j)) - \boldsymbol{q}_r(j)\boldsymbol{q}_r(j)^\top\right)\boldsymbol{h} = \boldsymbol{h}^\top \left(\operatorname{diag}(\boldsymbol{\nu}) - \boldsymbol{\nu}\boldsymbol{\nu}^\top\right)\boldsymbol{h} = \kappa\, \boldsymbol{h}^\top \left(\operatorname{diag}(\boldsymbol{\nu}) - \boldsymbol{\nu}\boldsymbol{\nu}^\top\right)\boldsymbol{h} \tag{59}$$

Assume now that $\kappa < 1$. Since $\boldsymbol{q}_r(j) \geq \kappa\boldsymbol{\nu}$ componentwise and both vectors sum to 1, we have that $\widetilde{\boldsymbol{q}}_r(j) := \frac{\boldsymbol{q}_r(j) - \kappa\boldsymbol{\nu}}{1-\kappa} \in \Delta_S$. Hence we express $\boldsymbol{q}_r(j)$ as a linear combination of probability vectors as $\boldsymbol{q}_r(j) = \kappa\boldsymbol{\nu} + (1-\kappa)\widetilde{\boldsymbol{q}}_r(j)$. Now, for any probability vector $\boldsymbol{p}$ on S and any $\boldsymbol{h} \in \mathcal{N}_S$, we define

$$Q(\boldsymbol{p}, \boldsymbol{h}) := \boldsymbol{h}^\top \left(\operatorname{diag}(\boldsymbol{p}) - \boldsymbol{p}\boldsymbol{p}^\top\right)\boldsymbol{h}. \tag{60}$$

Now, for convex combination of any two probability vectors $\boldsymbol{p}, \boldsymbol{q}$ on Δ_S, we have:

$$\begin{aligned}
Q(\alpha\boldsymbol{p} + (1-\alpha)\boldsymbol{q}, h) &= \sum_{i\in S} \left(\alpha\boldsymbol{p}[i] + (1-\alpha)\boldsymbol{q}[i]\right)(\boldsymbol{h}[i])^2 - \left(\sum_{i\in S} \left(\alpha\boldsymbol{p}[i] + (1-\alpha)\boldsymbol{q}[i]\right)\boldsymbol{h}[i]\right)^2 \\
&= \alpha \sum_{i\in S} \boldsymbol{p}[i](\boldsymbol{h}[i])^2 + (1-\alpha)\sum_{i\in S} \boldsymbol{q}[i](\boldsymbol{h}[i])^2 - \left(\alpha\sum_{i\in S} \boldsymbol{p}[i]\boldsymbol{h}[i] + (1-\alpha)\sum_{i\in S} \boldsymbol{q}[i]\boldsymbol{h}[i]\right)^2 \\
&= \alpha\left(\sum_{i\in S} \boldsymbol{p}[i](\boldsymbol{h}[i])^2 - \left(\sum_{i\in S} \boldsymbol{p}[i]\boldsymbol{h}[i]\right)^2\right) + (1-\alpha)\left(\sum_{i\in S} \boldsymbol{q}[i](\boldsymbol{h}[i])^2 - \left(\sum_{i\in S} \boldsymbol{q}[i]\boldsymbol{h}[i]\right)^2\right) \\
&\quad + \alpha(1-\alpha)\left(\sum_{i\in S} \boldsymbol{p}[i]\boldsymbol{h}[i] - \sum_{i\in S} \boldsymbol{q}[i]\boldsymbol{h}[i]\right)^2 \\
&= \alpha Q(\boldsymbol{p}, \boldsymbol{h}) + (1-\alpha)Q(\boldsymbol{q}, \boldsymbol{h}) + \alpha(1-\alpha)\left(\sum_{i\in S} \boldsymbol{p}[i]\boldsymbol{h}[i] - \sum_{i\in S} \boldsymbol{q}[i]\boldsymbol{h}[i]\right)^2 \\
&\geq \alpha Q(\boldsymbol{p}, \boldsymbol{h}),
\end{aligned} \tag{61}$$

where the last part uses the fact that $\left(\sum_{i\in S} \boldsymbol{p}[i]\boldsymbol{h}[i] - \sum_{i\in S} \boldsymbol{q}[i]\boldsymbol{h}[i]\right)^2$ is non-negative and any matrix of the form $\operatorname{diag}(\boldsymbol{p}) - \boldsymbol{p}\boldsymbol{p}^\top$ is positive semi-definite whenever $\boldsymbol{p}$ is a probability vector. Using the inequality (61) with $\boldsymbol{p} = \boldsymbol{\nu}$, $\boldsymbol{q} = \widetilde{\boldsymbol{q}}_r(j)$, and $\alpha = \kappa$, we obtain:

$$\boldsymbol{h}^\top \left(\operatorname{diag}(\boldsymbol{q}_r(j)) - \boldsymbol{q}_r(j)\boldsymbol{q}_r(j)^\top\right)\boldsymbol{h} \geq \kappa\, \boldsymbol{h}^\top \left(\operatorname{diag}(\boldsymbol{\nu}) - \boldsymbol{\nu}\boldsymbol{\nu}^\top\right)\boldsymbol{h}. \tag{62}$$

Thus from the inequalities (59) and (62), we get that in all cases,

$$\boldsymbol{h}^\top \mathcal{J}_r(\boldsymbol{\nu})\boldsymbol{h} \geq \kappa\, \boldsymbol{h}^\top \left(\operatorname{diag}(\boldsymbol{\nu}) - \boldsymbol{\nu}\boldsymbol{\nu}^\top\right)\boldsymbol{h}. \tag{63}$$

We now lower bound the right-hand side. Since $\boldsymbol{\nu}_i \geq \eta$ for all $i \in S$, we may write $\boldsymbol{\nu} = \eta\mathbf{1}_S + \boldsymbol{w}$, $\boldsymbol{w}[i] \geq 0$. Because $\boldsymbol{h} \in \mathcal{N}_S$, one has $\mathbf{1}_S^\top \boldsymbol{h} = 0$, and therefore

$$\begin{aligned}
\boldsymbol{h}^\top \left(\operatorname{diag}(\boldsymbol{\nu}) - \boldsymbol{\nu}\boldsymbol{\nu}^\top\right)\boldsymbol{h} &= \sum_{i\in S} \boldsymbol{\nu}[i](\boldsymbol{h}[i])^2 - \left(\sum_{i\in S} \boldsymbol{\nu}[i]\boldsymbol{h}[i]\right)^2 \\
&= \sum_{i\in S} \left(\eta + \boldsymbol{w}[i]\right)(\boldsymbol{h}[i])^2 - \left(\sum_{i\in S} \left(\eta + \boldsymbol{w}[i]\right)\boldsymbol{h}[i]\right)^2 \\
&= \eta\sum_{i\in S} (\boldsymbol{h}[i])^2 + \sum_{i\in S} \boldsymbol{w}[i](\boldsymbol{h}[i])^2 - \left(\eta\sum_{i\in S} \boldsymbol{h}[i] + \sum_{i\in S} \boldsymbol{w}[i]\boldsymbol{h}[i]\right)^2 \\
&= \eta\|\boldsymbol{h}\|^2 + \sum_{i\in S} \boldsymbol{w}[i](\boldsymbol{h}[i])^2 - \left(\sum_{i\in S} \boldsymbol{w}[i]\boldsymbol{h}[i]\right)^2, \quad \text{as } \boldsymbol{h} \in \mathcal{N}_S
\end{aligned}$$

Now, applying Cauchy-Schwarz inequality, we get:

$$\left(\sum_{i\in S} \boldsymbol{w}[i](\boldsymbol{h}[i])^2\right)\left(\sum_{i\in S} \boldsymbol{w}[i]\right) \geq \left(\sum_{i\in S} \boldsymbol{w}[i]\boldsymbol{h}[i]\right)^2 \tag{64}$$

But notice that, component-wise $\boldsymbol{w}[i] \leq \boldsymbol{\nu}[i], \forall i \in S$. Summing gives us $\sum_{i\in S} \boldsymbol{w}[i] \leq \sum_{i\in S} \boldsymbol{\nu}[i] = 1$. Using this in the inequality (64) gives us $\sum_{i\in S} \boldsymbol{w}[i](\boldsymbol{h}[i])^2 \geq \left(\sum_{i\in S} \boldsymbol{w}[i]\boldsymbol{h}[i]\right)^2$. Putting this in the expression for $\boldsymbol{h}^\top \left(\operatorname{diag}(\boldsymbol{\nu}) - \boldsymbol{\nu}\boldsymbol{\nu}^\top\right) \boldsymbol{h}$ gives us:

$$\boldsymbol{h}^\top \left(\operatorname{diag}(\boldsymbol{\nu}) - \boldsymbol{\nu}\boldsymbol{\nu}^\top\right) \boldsymbol{h} \geq \eta \|\boldsymbol{h}\|^2 \tag{65}$$

Combining (63) (65) and using $\kappa = e^{-2\delta_{\boldsymbol{C}}/\lambda}$, we obtain

$$\begin{aligned}
&\mathcal{J}_r(\boldsymbol{\nu})\big|_{\mathcal{N}_S} \succeq \eta e^{-2\delta_{\boldsymbol{C}}/\lambda} \boldsymbol{I}_S \\
\implies &\lambda \mathcal{J}_r(\boldsymbol{\nu})^\dagger\big|_{\mathcal{N}_S} \preceq \frac{\lambda e^{2\delta_{\boldsymbol{C}}/\lambda}}{\eta} \boldsymbol{I}_S \\
\implies &\sum_{r=1}^{B} \beta_r \lambda \mathcal{J}_r(\boldsymbol{\nu})^\dagger\big|_{\mathcal{N}_S} \preceq \frac{\lambda e^{2\delta_{\boldsymbol{C}}/\lambda}}{\eta} \boldsymbol{I}_S, \text{ using } \sum_{r=1}^{B} \beta_r = 1 \\
\implies &\sum_{r=1}^{B} \beta_r \lambda \mathcal{J}_r(\boldsymbol{\nu})^\dagger\big|_{\mathcal{N}_S} + \tau \operatorname{diag}(\boldsymbol{\nu})^{-1}\big|_{\mathcal{N}_S} \preceq \frac{\lambda e^{2\delta_{\boldsymbol{C}}/\lambda} + \tau}{\eta} \boldsymbol{I}_S.
\end{aligned} \tag{66}$$

Combining (36), and (66) yields the smoothness bound. Specifically, we get the following smoothness and strong convexity bounds:

$$\begin{aligned}
&\boldsymbol{z}^\top \nabla^2_{\boldsymbol{\nu}} W(\boldsymbol{\nu}) \boldsymbol{z} \leq \frac{\lambda e^{2\delta_{\boldsymbol{C}}/\lambda} + \tau}{\eta} \|\boldsymbol{z}\|^2, \ \forall \boldsymbol{z} \in \mathcal{N}_S, \ \boldsymbol{\nu} \in \{\tilde{\boldsymbol{\nu}} \in \operatorname{int}(\Delta_S) : \tilde{\boldsymbol{\nu}}[i] \geq \eta, \ \forall i \in S\} \\
&\boldsymbol{z}^\top \nabla^2_{\boldsymbol{\nu}} W(\boldsymbol{\nu}) \boldsymbol{z} \geq (\lambda + \tau)\|z\|^2, \ \forall \boldsymbol{z} \in \mathcal{N}_S, \ \boldsymbol{\nu} \in \Delta_S
\end{aligned} \tag{67}$$

□

The following two lemmas establish the properties used in the proof of Theorem 5. We suppress q in intermediate notation and write $\operatorname{TV}(p, p') = \frac{1}{2}\|p - p'\|_1$.

Lemma 17 (Entropy decomposition and variation bound)**.** There is a finite convex function G on Δ_D such that

$$\Phi(p) = ah(p) + G(p), \qquad |G(p) - G(p')| \leq \omega_C \operatorname{TV}(p, p'). \tag{68}$$

Every nonempty face Δ_S has a unique minimizer, positive on S.

Proof. For a coupling π with marginals p, μ, let

$$I(\pi) = \sum_{i,j} \pi_{ij} \log \frac{\pi_{ij}}{p_i \mu_j}.$$

Terms on zero rows or columns contribute zero. The entropy identity

$$\sum_{i,j} \pi_{ij}(\log \pi_{ij} - 1) = h(p) + h(\mu) - 1 + I(\pi)$$

gives $G = \sum_r w_r G_r$, where

$$G_r(p) = \lambda(h(\mu_r) - 1) + \min_{\pi \in \Pi(p, \mu_r)} \{\langle C, \pi \rangle + \lambda I(\pi)\}.$$

For fixed μ_r, $I(\pi) = \operatorname{KL}(\pi \| (\pi \mathbf{1}) \otimes \mu_r)$ is convex in π, by joint convexity of relative entropy. Mixing feasible couplings proves convexity of G_r, hence of G.

For the variation bound, take an optimal coupling (X, Y) for $G_r(p)$. Couple p and p' by matching their common mass first, so that there is a transition from X to X' with marginal p' and $\mathbb{P}(X \neq X') = \operatorname{TV}(p, p')$. Apply this transition independently of Y conditional on X. The new pair (X', Y) is feasible for $G_r(p')$. Its expected transport cost increases by at most $\omega_C \mathbb{P}(X \neq X')$. The log-sum inequality, applied to the same transition on the joint law and on $p \otimes \mu_r$, gives

$$I(X'; Y) \leq I(X; Y).$$

Thus $G_r(p') \leq G_r(p) + \omega_C \operatorname{TV}(p, p')$. Interchanging the marginals and averaging proves (68), including when $\lambda = 0$.

This bound also gives continuity. Compactness gives existence of a minimizer, and the strict convexity of ah gives uniqueness. If a minimizer p on Δ_S has $p_j = 0$ for $j \in S$, set $p(t) = (1-t)p + te_j$. Then

$$\Phi(p(t)) - \Phi(p) \le at \log t + O(t) < 0$$

for all sufficiently small $t > 0$, a contradiction. Here the residual change is at most $\omega_C t$, and the remaining entropy terms are $O(t)$. This argument uses $a > 0$, and does not require $\tau > 0$. □

Lemma 18 (An affine lower bound at a restricted minimizer)**.** If $\lambda > 0$ and $L \neq \varnothing$, there is $g \in \mathbb{R}^D$ with $\max_i g_i - \min_i g_i \le \omega_C$ such that

$$G(p) \ge G(p_L) + g^\top (p - p_L) \quad (p \in \Delta_D), \qquad (p_L)_i = \frac{e^{-g_i/a}}{\sum_{\ell \in L} e^{-g_\ell/a}} \quad (i \in L). \tag{69}$$

Proof. For each r, delete columns on which μ_r is zero. Writing $\pi_{ij} = p_i Q_{ij}$ and imposing the column constraint with multipliers v_j yields

$$G_r(p) = \lambda(h(\mu_r) - 1) + \max_v \left\{ \mu_r^\top v + \sum_i p_i b_{ri}(v) \right\},$$

$$b_{ri}(v) = -\lambda \log \sum_{j:\mu_{rj}>0} \mu_{rj} \exp\left(\frac{v_j - C_{ij}}{\lambda} \right).$$

Indeed, minimizing a conditional row $Q_{i,:}$ over its probability simplex gives exactly $b_{ri}(v)$ by the entropy conjugacy formula. At p_L, all active row and column marginals are positive. The finite entropic transport problem therefore has finite optimal dual potentials, which give a maximizer v_r in this display. Its only ambiguity is addition of a constant: the negative Hessian in v is a positive weighted sum of categorical covariance matrices with positive entries, with kernel equal to the constant vectors.

Termwise comparison inside the logarithms gives, for all i, k,

$$|b_{ri}(v_r) - b_{rk}(v_r)| \le \max_j |C_{ij} - C_{kj}| \le \omega_C.$$

Holding v_r fixed defines a global affine lower bound with equality at p_L. Taking $g_i = \sum_r w_r b_{ri}(v_r)$ proves the first part of (69) and the oscillation bound. On the relative interior of Δ_L, these slopes give the directional derivatives of G: this follows either by differentiating the unique normalized dual optimum or from the active transport KKT equations. Optimality of p_L thus implies $a(\log(p_L)_i + 1) + g_i$ is constant on L. Normalization proves the second part. □

Now, before proceeding to the proof of our main theorem on restricted γ-weakly submodularity, we come up with a counter-example of strict submodularity. In this appendix, we shall use $F_{\lambda,\tau}(\bullet)$ to denote the set function from (5), where the RSWB objective is defined with the regularizers λ, τ.

Proposition 19 (Counterexample for unregularized case)**.** Let $N = 4$, and place the corpus at the corners of the unit square in $\mathbb{R}^2$ on the points $\boldsymbol{x}_1 = (0,0)$, $\boldsymbol{x}_2 = (1,0)$, $\boldsymbol{x}_3 = (0,1)$, $\boldsymbol{x}_4 = (1,1)$, take the ℓ_1 ground cost given by $\boldsymbol{C}[i,j] := \|\boldsymbol{x}_i - \boldsymbol{x}_j\|_1$, (note: $\boldsymbol{C}[i,i] = 0$, $\boldsymbol{C}[i,j] > 0$ for $i \neq j$, and $\boldsymbol{C}$ is a metric). Take $B = 3$ reference measures on $[N]$ as $\boldsymbol{\mu}_1 := \frac{1}{2}(\delta_1 + \delta_2)$, $\boldsymbol{\mu}_2 := \frac{1}{2}(\delta_1 + \delta_3)$, $\boldsymbol{\mu}_3 := \frac{1}{2}(\delta_1 + \delta_4)$, with uniform weights $\beta_1 = \beta_2 = \beta_3 = \frac{1}{3}$. Then $F_{0,0}(S) := -\min_{\boldsymbol{\nu} \in \Delta_S} \frac{1}{3} \sum_{r=1}^3 \mathcal{W}(\boldsymbol{\nu}, \boldsymbol{\mu}_r)$. is not submodular.

Proof. We define $A := \{2,3\}$, $B := \{1,2,3\}$, $e := 4, D := \{1,2,3,4\}$. We compute the four values $F_{0,0}(A), F_{0,0}(A \cup \{e\}), F_{0,0}(B), F_{0,0}(B \cup \{e\})$. We shall show that $F_{0,0}(e \mid A) < F_{0,0}(e \mid B)$, proving the non-submodularity since $A \subset B$. We now compute the values one by one.

Take $\boldsymbol{\nu}_A := \frac{1}{2}(\delta_2 + \delta_3)$. Then $\mathcal{W}(\boldsymbol{\nu}_A, \boldsymbol{\mu}_1) = \frac{1}{2}$, $\mathcal{W}(\boldsymbol{\nu}_A, \boldsymbol{\mu}_2) = \frac{1}{2}$, $\mathcal{W}(\boldsymbol{\nu}_A, \boldsymbol{\mu}_3) = 1$. Indeed, for $\boldsymbol{\mu}_1$ one keeps the mass at 2 and moves the mass at 3 to 1, which costs $\frac{1}{2}\|\boldsymbol{x}_3 - \boldsymbol{x}_1\|_1 = \frac{1}{2}$; the computation for $\boldsymbol{\mu}_2$ is symmetric; and for $\boldsymbol{\mu}_3$ every unit of mass transported from $\{2,3\}$ to $\{1,4\}$ pays cost 1. Define $G_{0,0} := -F_{0,0}$ We have: $G_{0,0}(A) \le \frac{1}{3}\left(\frac{1}{2} + \frac{1}{2} + 1\right) = \frac{2}{3}$.

Conversely, let $\boldsymbol{\nu} \in \Delta_A$ be arbitrary. Since $\boldsymbol{\mu}_3$ is supported on $\{1,4\}$ while $\boldsymbol{\nu}$ is supported on $\{2,3\}$, every unit of mass must move distance at least 1. Therefore $\mathcal{W}(\boldsymbol{\nu}, \boldsymbol{\mu}_3) \ge 1$. Also, by the triangle inequality for Wasserstein distance, $\mathcal{W}(\boldsymbol{\nu}, \boldsymbol{\mu}_1) + \mathcal{W}(\boldsymbol{\nu}, \boldsymbol{\mu}_2) \ge \mathcal{W}(\boldsymbol{\mu}_1, \boldsymbol{\mu}_2)$. Now $\mathcal{W}(\boldsymbol{\mu}_1, \boldsymbol{\mu}_2) = 1$,

since one must move mass $\frac{1}{2}$ from 2 to 3, at cost $\|\boldsymbol{x}_2 - \boldsymbol{x}_3\|_1 = 2$. Thus $\mathcal{W}(\boldsymbol{\nu}, \boldsymbol{\mu}_1) + \mathcal{W}(\boldsymbol{\nu}, \boldsymbol{\mu}_2) + \mathcal{W}(\boldsymbol{\nu}, \boldsymbol{\mu}_3) \geq 2$, so $G_{0,0}(A) \geq \frac{2}{3}$. Hence, $G_{0,0}(A) = \frac{2}{3}$.

Now we compute $G_{0,0}(B)$. Take $\boldsymbol{\nu}_B := \boldsymbol{\mu}_2 = \frac{1}{2}(\delta_1 + \delta_3)$. Then $\mathcal{W}(\boldsymbol{\nu}_B, \boldsymbol{\mu}_2) = 0$, $\mathcal{W}(\boldsymbol{\nu}_B, \boldsymbol{\mu}_1) = 1$, $\mathcal{W}(\boldsymbol{\nu}_B, \boldsymbol{\mu}_3) = \frac{1}{2}$. Hence $G_{0,0}(B) \leq \frac{1}{3}\left(0 + 1 + \frac{1}{2}\right) = \frac{1}{2}$.

For the reverse inequality, let $\boldsymbol{\nu} \in \Delta_B$. Since $4 \notin B$ but $\boldsymbol{\mu}_3$ places mass $\frac{1}{2}$ at 4, at least $\frac{1}{2}$ mass must be transported to 4 from distance at least 1, so $\mathcal{W}(\boldsymbol{\nu}, \boldsymbol{\mu}_3) \geq \frac{1}{2}$. Also, $\mathcal{W}(\boldsymbol{\nu}, \boldsymbol{\mu}_1) + \mathcal{W}(\boldsymbol{\nu}, \boldsymbol{\mu}_2) \geq \mathcal{W}(\boldsymbol{\mu}_1, \boldsymbol{\mu}_2) = 1$. Therefore $\mathcal{W}(\boldsymbol{\nu}, \boldsymbol{\mu}_1) + \mathcal{W}(\boldsymbol{\nu}, \boldsymbol{\mu}_2) + \mathcal{W}(\boldsymbol{\nu}, \boldsymbol{\mu}_3) \geq \frac{3}{2}$, and hence $G_{0,0}(B) \geq \frac{1}{2}$. Thus $G_{0,0}(B) = \frac{1}{2}$.

We now compute $G_{0,0}(D)$. Take $\boldsymbol{\nu}_D := \boldsymbol{\mu}_3 = \frac{1}{2}(\delta_1 + \delta_4)$. Then $\mathcal{W}(\boldsymbol{\nu}_D, \boldsymbol{\mu}_3) = 0$, $\mathcal{W}(\boldsymbol{\nu}_D, \boldsymbol{\mu}_1) = \frac{1}{2}$, $\mathcal{W}(\boldsymbol{\nu}_D, \boldsymbol{\mu}_2) = \frac{1}{2}$. Hence $G_{0,0}(D) \leq \frac{1}{3}$. For the lower bound, let $\boldsymbol{\nu} \in \Delta_D$ be arbitrary. By the triangle inequality on pairwise expressions:

$$\begin{aligned}
\mathcal{W}(\boldsymbol{\mu}_1, \boldsymbol{\mu}_2) &\leq \mathcal{W}(\boldsymbol{\mu}_1, \boldsymbol{\nu}) + \mathcal{W}(\boldsymbol{\nu}, \boldsymbol{\mu}_2), \\
\mathcal{W}(\boldsymbol{\mu}_1, \boldsymbol{\mu}_3) &\leq \mathcal{W}(\boldsymbol{\mu}_1, \boldsymbol{\nu}) + \mathcal{W}(\boldsymbol{\nu}, \boldsymbol{\mu}_3), \\
\mathcal{W}(\boldsymbol{\mu}_2, \boldsymbol{\mu}_3) &\leq \mathcal{W}(\boldsymbol{\mu}_2, \boldsymbol{\nu}) + \mathcal{W}(\boldsymbol{\nu}, \boldsymbol{\mu}_3).
\end{aligned}$$

Summing all three expressions,

$$\mathcal{W}(\boldsymbol{\mu}_1, \boldsymbol{\mu}_2) + \mathcal{W}(\boldsymbol{\mu}_1, \boldsymbol{\mu}_3) + \mathcal{W}(\boldsymbol{\mu}_2, \boldsymbol{\mu}_3) \leq 2 \sum_{r=1}^{3} \mathcal{W}(\boldsymbol{\nu}, \boldsymbol{\mu}_r).$$

Now we have: $\mathcal{W}(\boldsymbol{\mu}_1, \boldsymbol{\mu}_2) = 1$, $\mathcal{W}(\boldsymbol{\mu}_1, \boldsymbol{\mu}_3) = \frac{1}{2}$, $\mathcal{W}(\boldsymbol{\mu}_2, \boldsymbol{\mu}_3) = \frac{1}{2}$. Hence $2 \leq 2\sum_{r=1}^{3} \mathcal{W}(\boldsymbol{\nu}, \boldsymbol{\mu}_r)$, so $\sum_{r=1}^{3} \mathcal{W}(\boldsymbol{\nu}, \boldsymbol{\mu}_r) \geq 1$, and therefore $G_{0,0}(D) \geq \frac{1}{3}$. Thus we get: $G_{0,0}(D) = \frac{1}{3}$ combining both lower and upper bounds

Finally, we compute the value of $G_{0,0}$ on $A \cup \{e\} = \{2, 3, 4\}$. Assume any $\boldsymbol{\nu} \in \Delta_{A \cup \{e\}}$. Then we can write:

$$\boldsymbol{\nu} = a\delta_2 + b\delta_3 + c\delta_4,\ a, b, c \geq 0,\ a + b + c = 1.$$

We shall show that $\sum_{r=1}^{3} \mathcal{W}(\boldsymbol{\nu}, \boldsymbol{\mu}_r) \geq 2$. We use the Kantorovich–Rubinstein dual formulation for metric Wasserstein distance:

$$\mathcal{W}(\boldsymbol{\nu}, \boldsymbol{\mu}) = \sup \left\{ \sum_{i=1}^{4} \boldsymbol{f}[i](\boldsymbol{\nu}[i] - \boldsymbol{\mu}[i]) : |\boldsymbol{f}[i] - \boldsymbol{f}[j]| \leq \boldsymbol{C}[i, j]\ \forall i, j \right\}.$$

For $\boldsymbol{\mu}_1$, consider the dual variable $\boldsymbol{f}_1 := (0, 1, 1, 2)$. It is feasible because $|\boldsymbol{f}_1[i] - \boldsymbol{f}_1[j]| \leq \boldsymbol{C}[i, j]\ \forall i, j$. Hence we may say:

$$\mathcal{W}(\boldsymbol{\nu}, \boldsymbol{\mu}_1) \geq \sum_{i=1}^{4} \boldsymbol{f}_1[i](\boldsymbol{\nu}[i] - \boldsymbol{\mu}_1[i]) = (a + b + 2c) - \frac{1}{2} = \frac{1}{2} + c.$$

Similarly, the same potential yields $\mathcal{W}(\boldsymbol{\nu}, \boldsymbol{\mu}_2) \geq \sum_{i=1}^{4} \boldsymbol{f}_1[i](\boldsymbol{\nu}[i] - \boldsymbol{\mu}_2[i]) = (a+b+2c) - \frac{1}{2} = \frac{1}{2} + c$.

For $\boldsymbol{\mu}_3$, we instead consider the dual potential given by $\boldsymbol{f}_3 := (0, 1, 1, 0)$. Again this is feasible, and therefore $\mathcal{W}(\boldsymbol{\nu}, \boldsymbol{\mu}_3) \geq \sum_{i=1}^{4} \boldsymbol{f}_3[i](\boldsymbol{\nu}[i] - \boldsymbol{\mu}_3[i]) = a + b = 1 - c.$. Summing these three lower bounds, we obtain

$$\sum_{r=1}^{3} \mathcal{W}(\boldsymbol{\nu}, \boldsymbol{\mu}_r) \geq \left(\frac{1}{2} + c\right) + \left(\frac{1}{2} + c\right) + (1 - c) = 2 + c \geq 2. \implies \quad G_{0,0}(A \cup \{e\}) \geq \frac{2}{3}.$$

On the other hand, the measure $\boldsymbol{\nu}_A = \frac{1}{2}(\delta_2 + \delta_3)$ is feasible in $\Delta_{A \cup \{e\}}$ and achieves the value $\frac{2}{3}$ as shown previously. Hence $G_{0,0}(A \cup \{e\}) = \frac{2}{3}$.

Thus, we get that $G_{0,0}(A \cup \{e\}) - G_{0,0}(A) = 0$, also $G_{0,0}(B \cup \{e\}) = \frac{1}{3}$ and $G_{0,0}(B) = \frac{1}{2}$. This implies:

$$\begin{aligned}
G_{0,0}(e \mid B) &= -\frac{1}{6},\ G_{0,0}(e \mid A) = 0 \\
F_{0,0}(e \mid B) &= \frac{1}{6},\ F_{0,0}(e \mid A) = 0
\end{aligned} \tag{70}$$

Thus, $F_{0,0}$ does not satisfy diminishing returns, which violates submodularity of $F_{0,0}$ $\square$

As an extension, we now show that the strict violation above persists for non zero values if λ, τ as well. For a given $\lambda, \tau \geq 0$ we define $G_{\lambda,\tau}(S) := \min_{\boldsymbol{\nu} \in \Delta_S} W_{\lambda,\tau}(\boldsymbol{\nu}, \{\boldsymbol{\mu}_r(\boldsymbol{q}, D)\})$. Then, we show the continuity as $(\lambda, \tau) \to (0, 0)$

Proposition 20 (Uniform continuity of the restricted-support objective)**.** For every nonempty $S \subseteq D$, one has

$$\big|G_{\lambda,\tau}(S) - G_{0,0}(S)\big| \leq \lambda\big(1 + \log 16\big) + \tau \log 4. \tag{71}$$

Proof. Fix $\boldsymbol{\nu} \in \Delta_S$ and $r \in \{1, 2, 3\}$. For every coupling $\boldsymbol{\pi} \in \Pi(\boldsymbol{\nu}, \boldsymbol{\mu}_r)$, one has $\sum_{i,j} \boldsymbol{\pi}[i, j] = 1$, and $\boldsymbol{\pi}$ has at most 16 entries. Hence $-\log 16 \leq \sum_{i,j=1}^{4} \boldsymbol{\pi}[i, j] \log \boldsymbol{\pi}[i, j] \leq 0$, which implies

$$-(1 + \log 16) \leq \sum_{i,j=1}^{4} \boldsymbol{\pi}[i, j](\log \boldsymbol{\pi}[i, j] - 1) \leq -1.$$

Therefore, for every feasible $\boldsymbol{\pi}$ we must have:

$$\langle C, \boldsymbol{\pi}\rangle - \lambda(1 + \log 16) \leq \langle C, \boldsymbol{\pi}\rangle + \lambda \sum_{i,j=1}^{4} \boldsymbol{\pi}[i, j](\log \boldsymbol{\pi}[i, j] - 1) \leq \langle C, \boldsymbol{\pi}\rangle - \lambda.$$

Minimizing over $\boldsymbol{\pi} \in \Pi(\boldsymbol{\nu}, \boldsymbol{\mu}_r)$ that: $\mathcal{W}_0(\boldsymbol{\nu}, \boldsymbol{\mu}_r) - \lambda(1 + \log 16) \leq \mathcal{W}_\lambda(\boldsymbol{\nu}, \boldsymbol{\mu}_r) \leq \mathcal{W}_0(\boldsymbol{\nu}, \boldsymbol{\mu}_r) - \lambda$. In particular, $\big|\mathcal{W}_\lambda(\boldsymbol{\nu}, \boldsymbol{\mu}_r) - \mathcal{W}_0(\boldsymbol{\nu}, \boldsymbol{\mu}_r)\big| \leq \lambda(1 + \log 16)$. Averaging over $r = 1, 2, 3$, we obtain $\Big|\frac{1}{3}\sum_{r=1}^{3} \mathcal{W}_\lambda(\boldsymbol{\nu}, \boldsymbol{\mu}_r) - \frac{1}{3}\sum_{r=1}^{3} \mathcal{W}_0(\boldsymbol{\nu}, \boldsymbol{\mu}_r)\Big| \leq \lambda(1 + \log 16)$.

Next, since $\boldsymbol{\nu} \in \Delta_S$ and $|S| \leq 4$, one has $0 \leq \mathrm{Ent}(\boldsymbol{\nu}) \leq \log |S| \leq \log 4$. Hence $0 \leq \tau\, \mathrm{Ent}(\boldsymbol{\nu}) \leq \tau \log 4$. Therefore, for every $\boldsymbol{\nu} \in \Delta_S$, $\big|W_{\lambda,\tau}(\boldsymbol{\nu}) - W_{0,0}(\boldsymbol{\nu})\big| \leq \lambda(1 + \log 16) + \tau \log 4$.

Since the two objective functions differ uniformly by at most this quantity, their minima differ by at most the same quantity. This proves (71). □

Corollary 21 (Persistence of non-submodularity for positive regularization)**.** Let $A := \{2, 3\}$, $B := \{1, 2, 3\}$, $e := 4$. If $\lambda(1 + \log 16) + \tau \log 4 < \frac{1}{24}$, then

$$G_{\lambda,\tau}(A) - G_{\lambda,\tau}(A \cup \{e\}) < G_{\lambda,\tau}(B) - G_{\lambda,\tau}(B \cup \{e\}), \tag{72}$$

and consequently the set function $F_{\lambda,\tau}$ is not submodular.

Proof. By (70), $G_{0,0}(e \mid A) = 0$, $G_{0,0}(e \mid B) = -\frac{1}{6}$. Hence $G_{0,0}(e \mid B) - G_{0,0}(e \mid A) = -\frac{1}{6}$.

By Proposition (20), each of the four quantities $G_{\lambda,\tau}(A)$, $G_{\lambda,\tau}(A \cup \{e\})$, $G_{\lambda,\tau}(B)$, $G_{\lambda,\tau}(B \cup \{e\})$ differs from its value at $(0, 0)$ by at most $\varepsilon(\lambda, \tau) := \lambda(1 + \log 16) + \tau \log 4$. Therefore

$$\big|G_{\lambda,\tau}(e \mid A) - G_{0,0}(e \mid A)\big| \leq 2\varepsilon(\lambda, \tau), \ \big|G_{\lambda,\tau}(e \mid B) - G_{0,0}(e \mid B)\big| \leq 2\varepsilon(\lambda, \tau).$$

Thus we get:

$$\begin{aligned} G_{\lambda,\tau}(e \mid B) - G_{\lambda,\tau}(e \mid A) &\leq G_{0,0}(e \mid B) + 2\epsilon(\lambda, \tau) - (G_{0,0}(e \mid A) - 2\epsilon(\lambda, \tau)) \\ &= 4\epsilon(\lambda, \tau) + G_{0,0}(e \mid B) - G_{0,0}(e \mid A) \\ &= 4\epsilon(\lambda, \tau) - \frac{1}{6} \end{aligned}$$

Under the hypothesis, the right-hand side is strictly negative. Hence $G_{\lambda,\tau}(e \mid B) < G_{\lambda,\tau}(e \mid A)$, which is exactly (72).

Since $F_{\lambda,\tau} = -G_{\lambda,\tau}$, the above gives us $F_{\lambda,\tau}(e \mid A) < F_{\lambda,\tau}(e \mid B)$ Thus $F_{\lambda,\tau}$ is not submodular. □

Theorem 5 (Monotonicity and restricted γ-weakly submodularity)**.** Let k be an integer, $\rho(\boldsymbol{x}_s, \boldsymbol{x}_t)$ be any valid function used to build the cost matrix $\boldsymbol{C}[s, t] = \rho(\boldsymbol{x}_s, \boldsymbol{x}_t)$ and the cost-oscillation δ be defined as $\delta := \max_{s,t,s',t' \in D} |\boldsymbol{C}[s, t] - \boldsymbol{C}[s', t']|$ with $\gamma_k := \exp\left(-2\delta/(\tau + \lambda)\right)$. Then $F(\boldsymbol{q}, S)$ is monotone non-decreasing in S and is restriced submodular on all sets S with $|S| \leq k$ with submodularity ratio γ_k.

Proof. In this proof, we interchangeably use $\delta_{\boldsymbol{C}}$ and δ to denote the same quantity $\delta := \max_{s,t,s',t' \in D} |\boldsymbol{C}[s,t] - \boldsymbol{C}[s',t']|$. Firstly, we show monotonicity. Note that, $F(\boldsymbol{q}, \bullet)$ is defined as: $-F(\boldsymbol{q}, S) = \min_{\boldsymbol{\nu} \in \Delta_S} W(\boldsymbol{\nu}, \{\boldsymbol{\mu}_r(\boldsymbol{q}, D) \,|\, r \in [B]\})$. So the only S dependent quantity in the definition of $F(\boldsymbol{q}, \bullet)$ is the support on which the optimization of $\boldsymbol{\nu}$ is defined. Note that, if $S \subseteq T$ then $\Delta_S \subseteq \Delta_T$. This implies that, for the optimization over $\boldsymbol{\nu}$, we must have $\min_{\boldsymbol{\nu} \in \Delta_T} W(\boldsymbol{\nu}, \{\boldsymbol{\mu}_r(\boldsymbol{q}, D) \,|\, r \in [B]\}) \leq \min_{\boldsymbol{\nu} \in \Delta_S} W(\boldsymbol{\nu}, \{\boldsymbol{\mu}_r(\boldsymbol{q}, D) \,|\, r \in [B]\})$. This gives us $F(\boldsymbol{q}, S) \leq F(\boldsymbol{q}, T)$, $\forall S \subseteq T$ for a given $\boldsymbol{q}$, and monotonicity follows.

Now, we shall show that, $\exists$ a metric $\rho(\bullet, *)$ such that defining the cost matrix using this metric ρ renders $F(\boldsymbol{q}, \bullet)$ to be non-submodular. Note that, as previously shown in Lemma 9, if the embeddings $\boldsymbol{x}_i$ are in $\mathbb{R}$, then for the special case of $\lambda = \tau = 0$, exact submodularity would follow. But here we show a counter-example where this breaks in higher dimensions. Specifically, for $\mathbb{R}^2$, we construct a set $\mathcal{V}$, a metric ρ, and probability distributions $\mu_1, \cdots, \mu_B$, with support in $\mathbb{R}^2$ such that the set function given by $F(S) := -\min_{\boldsymbol{\nu} \in \Delta_S} W(\boldsymbol{\nu}, \{\boldsymbol{\mu}_r(\boldsymbol{q}, D)\})$is non-submodular. This implies that if the counterexample set $\{\boldsymbol{\mu}_1, \cdots, \boldsymbol{\mu}_N\}$ matches with $\{\boldsymbol{\mu}_r(\boldsymbol{q}, D) \,|\, r \in [B]\}$, for a given $\boldsymbol{q}$, then $F(\boldsymbol{q}, \bullet)$ won't be submodular. The counter-example directly follows from Corollary 21.

Finally, we proceed to prove the restricted submodularity of $F(q, \bullet)$ for any given query q.

Write $\ell = |L|$, $s = |A|$, and $R = \omega_C / a$. Monotonicity follows immediately from inclusion of feasible simplices. We prove stronger quantitative gain bounds.

First fix $j \notin L$, put $p = p_L$, and take the feasible distribution $p(t) = (1-t)p + te_j$ for $0 \leq t \leq 1$. The supports of p and e_j are disjoint, so

$$h(p(t)) = (1-t)h(p) + (1-t)\log(1-t) + t\log t.$$

The variation bound and $h(p) \geq -\log \ell$ imply

$$\begin{aligned} \Phi(p) - \Phi(p(t)) &\geq a\{th(p) - (1-t)\log(1-t) - t\log t\} - \omega_C t \\ &\geq a\{-t\log \ell - (1-t)\log(1-t) - t\log t - Rt\}. \end{aligned}$$

The maximum over $t \in [0,1]$ of $tz - (1-t)\log(1-t) - t\log t$ is $\log(1+e^z)$. Taking $z = -\log \ell - R$ and using optimality on the enlarged face gives

$$F(q, L \cup \{j\}) - F(q, L) \geq a \log\left(1 + \frac{e^{-R}}{\ell}\right) > 0. \tag{73}$$

This also proves strict monotonicity under proper support inclusions.

For the joint gain, first take $\lambda > 0$ and let g be supplied by Lemma 18. Define

$$Z_L = \sum_{i \in L} e^{-g_i/a}, \qquad Z_A = \sum_{j \in A} e^{-g_j/a}, \qquad M(p) = ah(p) + G(p_L) + g^\top (p - p_L).$$

Then $M \leq \Phi$ everywhere, with equality at p_L, and p_L minimizes M on Δ_L. On any nonempty support T,

$$\min_{p \in \Delta_T} \{ah(p) + g^\top p\} = -a \log \sum_{i \in T} e^{-g_i/a}.$$

Consequently

$$\begin{aligned} F(q, L \cup A) - F(q, L) &= \Phi(p_L) - \min_{p \in \Delta_{L \cup A}} \Phi(p) \\ &\leq M(p_L) - \min_{p \in \Delta_{L \cup A}} M(p) \\ &= a \log(1 + Z_A / Z_L) \leq a \log\left(1 + \frac{se^R}{\ell}\right). \end{aligned} \tag{74}$$

The last inequality uses $\operatorname{osc}(g) \leq \omega_C$.

When $\lambda = 0$ and $\tau > 0$, apply this bound with $\lambda_n \downarrow 0$ and fixed τ. For every coupling on $D \times D$,

$$\left| \lambda_n \sum_{i,j} \pi_{ij}(\log \pi_{ij} - 1) \right| \leq \lambda_n (1 + 2\log N).$$

All restricted objective values therefore converge uniformly to their $\lambda = 0$ values. Passing to the limit proves (74) in that case as well.

Combining (73) and (74), whose denominator is positive, gives the stronger size-dependent estimate

$$\frac{\sum_{j\in A}[F(q,L\cup\{j\})-F(q,L)]}{F(q,L\cup A)-F(q,L)} \geq \frac{s\log(1+e^{-R}/\ell)}{\log(1+se^{R}/\ell)}. \tag{75}$$

Set $x = e^R/\ell$ and $c = e^{-2R}$. Concavity gives $\log(1+cx) \geq c\log(1+x)$, while $(1+x)^s \geq 1+sx$ gives $s\log(1+x) \geq \log(1+sx)$. It follows that the right-hand side of (75) is at least $c = e^{-2R}$. This proves the theorem. Notice that $\underline{\gamma}$ is a guaranteed lower bound, not a claimed exact value of $\gamma_k(F;q)$. □

Proposition 22 (Gaussian model for corpus embeddings)**.** Suppose $x_i = m + \sigma g_i$, where $m \in \mathbb{R}^d$, $\sigma > 0$, and $g_1, \ldots, g_N$ are independent $\mathcal{N}(0, I_d)$ vectors. For $N \geq 2$ and $0 < \delta_0 < 1$, define

$$u_{\delta_0} = \log\frac{N(N-1)}{2\delta_0}, \qquad B_{\delta_0} = \sigma\sqrt{2d + 4\sqrt{du_{\delta_0}} + 4u_{\delta_0}}.$$

With probability at least $1-\delta_0$, the Euclidean cost satisfies

$$\delta \leq B_{\delta_0}, \qquad \gamma_k(F;q) \geq \exp\left(-\frac{2B_{\delta_0}}{\lambda+\tau}\right).$$

Proof. For $i \neq j$, the variable $Q = \|x_i - x_j\|_2^2/(2\sigma^2)$ has a chi-square distribution with d degrees of freedom. For $0 < t < 1/2$, its moment generating function and the power series for the logarithm give

$$\log \mathbb{E}e^{t(Q-d)} = -dt - \frac{d}{2}\log(1-2t) \leq \frac{dt^2}{1-2t}.$$

For $u > 0$, applying the Chernoff bound with $t = \sqrt{u}/(\sqrt{d}+2\sqrt{u})$ yields

$$\mathbb{P}\{Q \geq d + 2\sqrt{du} + 2u\} \leq e^{-u}.$$

Take $u = u_{\delta_0}$ and a union bound over the $N(N-1)/2$ pairs. All pairwise distances are then at most B_{δ_0} with probability at least $1-\delta_0$. Since the diagonal costs are zero, $\Omega_C = \max_{i,j} C_{ij}$, and the theorem gives the ratio bound. □

D.2 RESTRICTED SUPPORT BARYCENTER COMPUTATION AND SUPPORT SELECTION

We first state the algorithm used to compute the restricted barycenter, which is turn is used to compute the induced set function $F(\boldsymbol{q}, S)$

Algorithm 5 Barycenter on fixed support

Require: Support S, smoothness coefficients: η, L_η, tolerance $\varepsilon > 0$
1: **function** COMPUTEBARY($S, \{\boldsymbol{\mu}_r\}$)
2: $\boldsymbol{\nu}^{(0)} \leftarrow \mathbf{1}_S/|S|$; $t \leftarrow 0$; $\mathrm{grad}^{(0)} \leftarrow \infty$.
3: **while** $\|\mathrm{grad}^{(t)}\| \geq \varepsilon$ **do**
4: $\mathrm{grad}^{(t)} \leftarrow \mathrm{Proj}_{\mathcal{N}_S}\big(\nabla W(\boldsymbol{\nu}^{(t)}, \{\boldsymbol{\mu}_r\})\big)$
5: $\boldsymbol{\nu}^{(t+1)} \leftarrow \mathrm{Proj}_{\mathcal{D}_\eta}\left(\boldsymbol{\nu}^{(t)} - \frac{1}{L_\eta}\mathrm{grad}^{(t)}\right)$
6: $t \leftarrow t + 1$
7: **return** $\boldsymbol{\nu}^{(t)}, W(\boldsymbol{\nu}^{(t)}, \{\boldsymbol{\mu}_r \mid r \in [B]\})$

Fix a query q and call its nonempty corresponding support set as $S(\boldsymbol{q}) \subseteq D$. Throughout the following section, we denote by $\mathcal{D}_\eta(S)$ for any set $S \neq \emptyset$ and $\eta > 0$ as $\mathcal{D}_\eta(S) \coloneqq \{\boldsymbol{a} \in \Delta_S : \boldsymbol{a}[i] \geq \eta,\ \forall i \in S\}$ and $\Delta_S^\circ \coloneqq \mathrm{int}(\Delta_S)$. In this appendix we make explicit the inner Sinkhorn subroutine used to compute the projected gradient in Algorithm 5. Since $\lambda > 0$, each transport term $\mathcal{W}(\boldsymbol{\nu}, \boldsymbol{\mu}_r(q, D))$ is an entropically regularized OT problem. For the fixed support $S(\boldsymbol{q})$, define the restricted Gibbs kernel

$$K_{S(\boldsymbol{q}),D}[i, j] \coloneqq \exp\Big(-\frac{\boldsymbol{C}[i,j]}{\lambda}\Big), \qquad i \in S(\boldsymbol{q}),\ j \in [N].$$

For the current iterate $\boldsymbol{\nu}^{(t)}(q, S)$ and ranker induced measure $\boldsymbol{\mu}_r(q, D)$, the entropic OT problem seeks a coupling $\boldsymbol{\pi}_r^{(t)}(q, S)$ whose row sums match the current barycenter weights on the support $S(\boldsymbol{q})$ and whose column sums match the reference distribution $\boldsymbol{\mu}_r(q, D)$. The optimal coupling has the form $\boldsymbol{\pi}_r^{(t)}(q, S)[i, j] = u_r[i]\, K_{S(\boldsymbol{q}),D}[i, j]\, v_r[j]$, $i \in S(\boldsymbol{q})$, $j \in [N]$ where the positive vectors $u_r(q, S) \in \mathbb{R}_+^{|S(\boldsymbol{q})|}$ and $v_r(q, S) \in \mathbb{R}_+^N$ are the Lagrange multipliers preserving row and column sum constraints. They are computed so that the coupling has the required marginals $\sum_{j=1}^N \boldsymbol{\pi}_r^{(t)}(q, S)[i, j] = \boldsymbol{\nu}^{(t)}(q, S)[i]$, $i \in S(\boldsymbol{q})$, and $\sum_{\boldsymbol{x}_i \in S(\boldsymbol{q})} \boldsymbol{\pi}_r^{(t)}(q, S)[i, j] = \boldsymbol{\mu}_r(q, D)[j]$, $j \in [N]$. Hence, $u_r(q, S)$ and $v_r(q, S)$ are row and column multipliers that rescale $K_{S(\boldsymbol{q}),D}$ such that the result has row marginal $\boldsymbol{\nu}^{(t)}(q, S)$ and column marginal $\boldsymbol{\mu}_r(q, D)$. These vectors are computed by the standard alternating Sinkhorn updates. Once $u_r(q, S)$ is obtained, the corresponding first dual potential is $\psi_r^{(t)}(q, S)[i] := \lambda \log u_r(q, S)[i]$, $i \in S(\boldsymbol{q})$. The gradient of W on the face $\Delta_{S(\boldsymbol{q})}$ is then

$$\nabla_{\boldsymbol{\nu}} W(\boldsymbol{\nu}^{(t)}(q, S))[i] \coloneqq \sum_{r=1}^B \beta_r\, \psi_r^{(t)}(q, S)[i] + \tau\left(\log \boldsymbol{\nu}^{(t)}(q, S)[i] + 1\right),\ i \in S(\boldsymbol{q}),$$

and its projection onto the tangent space is given by

$$g^{(t)}(q, S)[i] \coloneqq \nabla_{\boldsymbol{\nu}} W(\boldsymbol{\nu}^{(t)}(q, S))[i] - \frac{1}{|S(\boldsymbol{q})|} \sum_{\boldsymbol{x}_j \in S(\boldsymbol{q})} \nabla_{\boldsymbol{\nu}} W(\boldsymbol{\nu}^{(t)}(q, S))[j],\ i \in S(\boldsymbol{q}),$$

with $g^{(t)}(q, S)[i] = 0$ for $i \notin S(\boldsymbol{q})$. The outer iterate is then updated by a projected gradient step inside the η-interior $D_\eta(S(\boldsymbol{q}))$.

Complexity Analysis of entire greedy selection: Since each projected-gradient iteration requires B Sinkhorn solves on an $s \times N$ kernel, the cost of computing the restricted-support barycenter on a support of size $|S|$ using Algorithm 5 is $O(B|S|NT_{\mathrm{Sk}}T_{\mathrm{gd}})$ where $T_{\mathrm{Sk}}, T_{\mathrm{gd}}$ are iteration counts of the inner Sinkhorn and outer gradient descent loop .Consequently, the overall complexity of top-k retrieval using Algorithm 1 becomes $O(BN\sum_{s=1}^k (N - s)s) = O(BN^2k^2)$ upto constants of iteration $T_{\mathrm{Sk}}, T_{\mathrm{gd}}$.

Next, we state the result on the convergence of projected gradient descent for the convex RSWB objective, as in Algorithm 5.

Theorem 23 (Convergence of projected gradient descent on a fixed support)**.** For a given query $\boldsymbol{q}$ and a fixed subset of the corpus $S \subseteq D$ let $\boldsymbol{\nu}^*(\boldsymbol{q}, S)$ denote the unique minimizer of $W(\bullet, \{\boldsymbol{\mu}_r(\boldsymbol{q}, D)\})$

on Δ_S. Algorithm 5 produces iterates $\{\boldsymbol{\nu}^{(t)}\}_{t=1}^{\infty}$. Then there exists constant $m_c \in (0,1)$, such that:

$$W(\boldsymbol{\nu}^{(t)}, \{\boldsymbol{\mu}_r\}) - W(\boldsymbol{\nu}^*(\boldsymbol{q}, S), \{\boldsymbol{\mu}_r\}) \leq (1-m_c)^t \left(W(\boldsymbol{\nu}^{(0)}, \{\boldsymbol{\mu}_r\}) - W(\boldsymbol{\nu}^*(\boldsymbol{q}, S), \{\boldsymbol{\mu}_r\}) \right). \quad (76)$$

Proof. Here we consider δ_c equal to $\delta := \max_{s,t,s',t' \in D} |\boldsymbol{C}[s,t] - \boldsymbol{C}[s',t']|$. By Proposition 14, applied to the query-dependent support $S(\boldsymbol{q})$, the minimizer $\boldsymbol{\nu}^*(\boldsymbol{q}, S(\boldsymbol{q}))$ satisfies $\boldsymbol{\nu}^*(\boldsymbol{q}, S(\boldsymbol{q}))[j] \geq \frac{1}{|S(\boldsymbol{q})|} \exp\left(-\frac{\delta_{\mathcal{C}}}{\tau+\lambda}\right) =: \eta_{S(\boldsymbol{q})}\, \forall j \in S(\boldsymbol{q})$. Hence $\boldsymbol{\nu}^*_{S(\boldsymbol{q})} \in \mathcal{D}_{\eta_{S(\boldsymbol{q})}}(S(\boldsymbol{q}))$.

From the PGD steps in Algorithm 5, we can summarize the updates as:

$$\begin{aligned} g^{(t)} &= \mathrm{Proj}_{\mathcal{N}_S}\left(\nabla W(\boldsymbol{\nu}^{(t)})\right) \\ \boldsymbol{\nu}^{(t+1)} &= \mathrm{Proj}_{\mathcal{D}_\eta}\left(\boldsymbol{\nu}^{(t)} - \frac{1}{L_\eta} g^{(t)}\right) \end{aligned} \quad (77)$$

Now, we show that, for any $\boldsymbol{\nu}, \widetilde{\boldsymbol{\nu}} \in \mathcal{D}_\eta$ we must have:

$$W(\widetilde{\boldsymbol{\nu}}) \leq W(\boldsymbol{\nu}) + \langle \nabla W(\boldsymbol{\nu}), \widetilde{\boldsymbol{\nu}} - \boldsymbol{\nu} \rangle + \frac{L_\eta}{2} \|\widetilde{\boldsymbol{\nu}} - \boldsymbol{\nu}\|^2, \forall \widetilde{\boldsymbol{\nu}}, \boldsymbol{\nu} \in \mathcal{D}_\eta \quad (78)$$

To show this, we define $\boldsymbol{\nu}(y) := \boldsymbol{\nu} + y(\widetilde{\boldsymbol{\nu}} - \boldsymbol{\nu})$, $\forall y \in (0,1)$ and we define $g : [0,1] \to \mathbb{R}$ as $g(y) := W(\boldsymbol{\nu}(y))$. Then from Fundamental theorem of Calculus we have:

$$\begin{aligned} W(\widetilde{\boldsymbol{\nu}}) - W(\boldsymbol{\nu}) &= \int_0^1 g'(y)\, \mathrm{d}y \\ &= \langle \nabla W(\nu), \widetilde{\boldsymbol{\nu}} - \boldsymbol{\nu} \rangle + \int_0^1 (1-y) g''(y)\, \mathrm{d}y \\ &= \langle \nabla W(\nu), \widetilde{\boldsymbol{\nu}} - \boldsymbol{\nu} \rangle + \int_0^1 (1-y)(\widetilde{\boldsymbol{\nu}} - \boldsymbol{\nu})^\top \nabla^2 W(\boldsymbol{\nu}(y))(\widetilde{\boldsymbol{\nu}} - \boldsymbol{\nu})\, \mathrm{d}y \\ &\leq \langle \nabla W(\nu), \widetilde{\boldsymbol{\nu}} - \boldsymbol{\nu} \rangle + L_\eta \|\widetilde{\boldsymbol{\nu}} - \boldsymbol{\nu}\|^2/2 \end{aligned}$$

Similarly we also get by strong convexity that,

$$W(\widetilde{\boldsymbol{\nu}}) \geq W(\boldsymbol{\nu}) + \langle \nabla W(\boldsymbol{\nu}), \widetilde{\boldsymbol{\nu}} - \boldsymbol{\nu} \rangle + \frac{m_0}{2} \|\widetilde{\boldsymbol{\nu}} - \boldsymbol{\nu}\|^2, \forall \widetilde{\boldsymbol{\nu}}, \boldsymbol{\nu} \in \mathrm{int}\, \Delta(S) \quad (79)$$

Now, note that, since $\widetilde{\boldsymbol{\nu}} - \boldsymbol{\nu} \in \mathcal{N}_S$, we have that $\langle \nabla W(\boldsymbol{\nu}), \widetilde{\boldsymbol{\nu}} - \boldsymbol{\nu} \rangle = \left\langle \nabla W(\boldsymbol{\nu})\big|_{\mathcal{N}_S}, \widetilde{\boldsymbol{\nu}} - \boldsymbol{\nu} \right\rangle$

In particular, for every $\widetilde{\boldsymbol{\nu}} \in \mathcal{D}_\eta(S(\boldsymbol{q}))$, we have:

$$\left\langle \nabla W(\boldsymbol{\nu}^{(t)}), \widetilde{\boldsymbol{\nu}} - \boldsymbol{\nu}^{(t)} \right\rangle = \left\langle g^{(t)}, \widetilde{\boldsymbol{\nu}} - \boldsymbol{\nu}^{(t)} \right\rangle. \quad (80)$$

Now define the quadratic upper model at iteration t by

$$Q_t(\widetilde{\boldsymbol{\nu}}) := W(\boldsymbol{\nu}^{(t)}) + \left\langle g^{(t)}, \widetilde{\boldsymbol{\nu}} - \boldsymbol{\nu}^{(t)} \right\rangle + \frac{L_\eta}{2} \|\widetilde{\boldsymbol{\nu}} - \boldsymbol{\nu}^{(t)}\|^2, \quad \widetilde{\boldsymbol{\nu}} \in \mathcal{D}_\eta(S(\boldsymbol{q})). \quad (81)$$

By (78) and (80), we obtain $W(\widetilde{\boldsymbol{\nu}}) \leq_t (\widetilde{\boldsymbol{\nu}})$, $\forall \widetilde{\boldsymbol{\nu}} \in \mathcal{D}_\eta(S(\boldsymbol{q}))$.

Next, by the projected gradient update in (77), we have

$$\begin{aligned} \boldsymbol{\nu}^{(t+1)} &= \mathrm{Proj}_{\mathcal{D}_\eta(S(\boldsymbol{q}))}\left(\boldsymbol{\nu}^{(t)} - \frac{1}{L_\eta} g^{(t)}\right) \\ &= \operatorname*{argmin}_{\widetilde{\boldsymbol{\nu}} \in \mathcal{D}_\eta(S(\boldsymbol{q}))} \|\widetilde{\boldsymbol{\nu}} - \left(\boldsymbol{\nu}^{(t)} - \frac{1}{L_\eta} g^{(t)}\right)\|^2. \end{aligned}$$

Expanding the square and ignoring terms independent of $\widetilde{\boldsymbol{\nu}}$, this is equivalent to $\boldsymbol{\nu}^{(t+1)} = \mathrm{argmin}_{\widetilde{\boldsymbol{\nu}} \in \mathcal{D}_\eta(S(\boldsymbol{q}))} Q_t(\widetilde{\boldsymbol{\nu}})$.. Hence, $Q_t(\boldsymbol{\nu}^{(t+1)}) \leq Q_t(\widetilde{\boldsymbol{\nu}}), \forall \widetilde{\boldsymbol{\nu}} \in \mathcal{D}_\eta(S(\boldsymbol{q}))$

Using smoothness results, we get that:

$$W(\boldsymbol{\nu}^{(t+1)}) \leq Q_t(\widetilde{\boldsymbol{\nu}}), \qquad \forall \widetilde{\boldsymbol{\nu}} \in \mathcal{D}_\eta(S(\boldsymbol{q})). \quad (82)$$

Now we write $\boldsymbol{\nu}^* := \boldsymbol{\nu}^*(\boldsymbol{q}, S(\boldsymbol{q}))$, $\theta := \frac{m_0}{L_\eta}$. Since $m_0 \leq L_\eta$, we have $\theta \in (0,1]$. Since both $\boldsymbol{\nu}^{(t)}$ and $\boldsymbol{\nu}^*$ belong to $\mathcal{D}_\eta(S(\boldsymbol{q}))$, the convex combination

$$\widetilde{\boldsymbol{\nu}}_\theta := (1-\theta)\boldsymbol{\nu}^{(t)} + \theta \boldsymbol{\nu}^* \quad (83)$$

also lies in $\mathcal{D}_\eta(S(\boldsymbol{q}))$. Therefore, using (82) with $\widetilde{\boldsymbol{\nu}} = \widetilde{\boldsymbol{\nu}}_\theta$, we obtain

$$W(\boldsymbol{\nu}^{(t+1)}) \leq Q_t(\widetilde{\boldsymbol{\nu}}_\theta)$$
$$= W(\boldsymbol{\nu}^{(t)}) + \theta \left\langle g^{(t)}, \boldsymbol{\nu}^* - \boldsymbol{\nu}^{(t)} \right\rangle + \frac{L_\eta \theta^2}{2} \|\boldsymbol{\nu}^* - \boldsymbol{\nu}^{(t)}\|^2. \tag{84}$$

We now apply the strong convexity bound (79) with $(\boldsymbol{\nu}, \boldsymbol{\nu}^*)$ since $\boldsymbol{\nu}^* - \boldsymbol{\nu}^{(t)} \in \mathcal{N}_S$, using (80) gives:

$$W(\boldsymbol{\nu}^*) \geq W(\boldsymbol{\nu}^{(t)}) + \left\langle g^{(t)}, \boldsymbol{\nu}^* - \boldsymbol{\nu}^{(t)} \right\rangle + \frac{m_0}{2} \|\boldsymbol{\nu}^* - \boldsymbol{\nu}^{(t)}\|^2. \tag{85}$$

Rearranging (85), we get

$$\left\langle g^{(t)}, \boldsymbol{\nu}^* - \boldsymbol{\nu}^{(t)} \right\rangle \leq W(\boldsymbol{\nu}^*) - W(\boldsymbol{\nu}^{(t)}) - \frac{m_0}{2} \|\boldsymbol{\nu}^* - \boldsymbol{\nu}^{(t)}\|^2. \tag{86}$$

Substituting (86) into (84), we obtain

$$W(\boldsymbol{\nu}^{(t+1)}) \leq W(\boldsymbol{\nu}^{(t)}) + \theta \left(W(\boldsymbol{\nu}^*) - W(\boldsymbol{\nu}^{(t)}) - \frac{m_0}{2} \|\boldsymbol{\nu}^* - \boldsymbol{\nu}^{(t)}\|^2 \right) + \frac{L_\eta \theta^2}{2} \|\boldsymbol{\nu}^* - \boldsymbol{\nu}^{(t)}\|^2$$
$$= (1-\theta) W(\boldsymbol{\nu}^{(t)}) + \theta W(\boldsymbol{\nu}^*) + \frac{1}{2} \left(L_\eta \theta^2 - m_0 \theta \right) \|\boldsymbol{\nu}^* - \boldsymbol{\nu}^{(t)}\|^2.$$

By the choice $\theta = \frac{m_0}{L_\eta}$, we have $L_\eta \theta^2 - m_0 \theta = 0$. Therefore,

$$W(\boldsymbol{\nu}^{(t+1)}) \leq (1-\theta) W(\boldsymbol{\nu}^{(t)}) + \theta W(\boldsymbol{\nu}^*). \tag{87}$$

Subtracting $W(\boldsymbol{\nu}^*)$ from both sides gives

$$W(\boldsymbol{\nu}^{(t+1)}) - W(\boldsymbol{\nu}^*) \leq (1-\theta) \left(W(\boldsymbol{\nu}^{(t)}) - W(\boldsymbol{\nu}^*) \right)$$
$$= \left(1 - \frac{m_0}{L_\eta} \right) \left(W(\boldsymbol{\nu}^{(t)}) - W(\boldsymbol{\nu}^*) \right).$$

Iterating this inequality yields

$$W(\boldsymbol{\nu}^{(t)}) - W(\boldsymbol{\nu}^*(\boldsymbol{q}, S(\boldsymbol{q}))) \leq \left(1 - \frac{m_0}{L_\eta} \right)^t \left(W(\boldsymbol{\nu}^{(0)}) - W(\boldsymbol{\nu}^*(\boldsymbol{q}, S(\boldsymbol{q}))) \right).$$

This proves (76). □

Greedy support selection guarantees: In the following, we state the Nemhauser-syle approximation guarantee of Algorithm 1 using our results from Theorem 5.

Proposition 24 (Approximation guarantee for greedy support selection)**.** Fix $k \leq N$, and a query $\boldsymbol{q}$. Let γ_k be what is defined in Theorem 5. Define: $S_k^* \in \operatorname{argmax}_{S \subseteq D, |S|=k} F(\boldsymbol{q}, S)$. Let $S_1, \ldots, S_k$ be the sequence generated by Algorithm 1. We also assume that $S_1 \in \operatorname{argmax}_{s \in D} F(\boldsymbol{q}, \{s\})$. Then we have:

$$(F(\boldsymbol{q}, S^*) - F(\boldsymbol{q}, S_k)) \leq (1 - \gamma_{2k}/k)^{k-1} \left(F(\boldsymbol{q}, S^*) - \max_{s \in D} F(\boldsymbol{q}, \{s\}) \right) \tag{88}$$

where $\max_{s \in D} F(\boldsymbol{q}, \{s\})$ is the maximum of $F(\boldsymbol{q}, \bullet)$ over singletons.

Proof. In the following proof, we use $S(\boldsymbol{q})_t$ to denote the set selected by Algorithm 1 at step t. Hence we use $F(S(\boldsymbol{q})_t)$ and $F(\boldsymbol{q}, S(\boldsymbol{q})_t)$ interchangeably. For every $t \in \{1, \ldots, k-1\}$, we define $U_t := S_k^* \setminus S(\boldsymbol{q})_t$. If $U_t = \varnothing$, then $S_k^* \subseteq S(\boldsymbol{q})_t$. Since F is monotone, we have $F(S(\boldsymbol{q})_k) \geq F(S(\boldsymbol{q})_t) \geq F(S_k^*)$, and hence $(F(S_k^*) - F(S(\boldsymbol{q})_k) \leq 0$. In this case, the desired bound is immediate. Therefore, it suffices to consider the case where $U_t \neq \varnothing$ for all $t \in \{1, \ldots, k-1\}$.

Fix such a t. Since $S(\boldsymbol{q})_t \cap U_t = \varnothing$, $|S(\boldsymbol{q})_t| \leq k$, and $|U_t| \leq |S_k^*| \leq k$, we have $|S(\boldsymbol{q})_t| + |U_t| \leq 2k$. Hence, by Theorem 5, applied to the disjoint pair $(S(\boldsymbol{q})_t, U_t)$ with cardinality parameter $2k$, we get

$$\sum_{j \in U_t} (F(S(\boldsymbol{q})_t \cup \{j\}) - F(S(\boldsymbol{q})_t)) \geq \gamma_{2k} \left(F(S(\boldsymbol{q})_t \cup U_t) - F(S(\boldsymbol{q})_t) \right).$$

Now, $S(\boldsymbol{q})_t \cup U_t = S(\boldsymbol{q})_t \cup S_k^*$. Since F is monotone, this implies $F(S(\boldsymbol{q})_t \cup U_t) = F(S(\boldsymbol{q})_t \cup S_k^*) \geq F(S_k^*)$. Substituting this into the previous inequality, we obtain:

$$\sum_{j \in U_t} (F(S(\boldsymbol{q})_t \cup \{j\}) - F(S(\boldsymbol{q})_t)) \geq \gamma_{2k} \left(F(S_k^*) - F(S(\boldsymbol{q})_t) \right).$$

We now use the fact that $|U_t| \leq k$. For the sake of contradiction, we assume every $j \in U_t$ satisfies

$$F(S(\boldsymbol{q})_t \cup \{j\}) - F(S(\boldsymbol{q})_t) < \frac{\gamma_{2k}}{k} \left(F(S_k^*) - F(S(\boldsymbol{q})_t) \right),$$

Then summing over $j \in U_t$ we get:

$$\sum_{j \in U_t} \left(F(S(\boldsymbol{q})_t \cup \{j\}) - F(S(\boldsymbol{q})_t)\right) < |U_t| \cdot \frac{\gamma_{2k}}{k} \left(F(S_k^*) - F(S(\boldsymbol{q})_t)\right) \leq \gamma_{2k} \left(F(S_k^*) - F(S(\boldsymbol{q})_t)\right),$$

This contradicts the previous lower bound. Therefore, there exists $j_t^* \in U_t$ such that

$$F(S(\boldsymbol{q})_t \cup \{j_t^*\}) - F(S(\boldsymbol{q})_t) \geq \frac{\gamma_{2k}}{k} \left(F(S_k^*) - F(S(\boldsymbol{q})_t)\right).$$

Algorithm 1 chooses the element with the largest marginal gain among all elements in $[N] \setminus S(\boldsymbol{q})_t$. Hence its chosen element j_t satisfies

$$F(S(\boldsymbol{q})_t \cup \{j_t\}) - F(S(\boldsymbol{q})_t) \geq F(S(\boldsymbol{q})_t \cup \{j_t^*\}) - F(S(\boldsymbol{q})_t)$$
$$\implies F(S(\boldsymbol{q})_{t+1}) - F(S(\boldsymbol{q})_t) \geq \frac{\gamma_{2k}}{k} \left(F(S_k^*) - F(S(\boldsymbol{q})_t)\right).$$

Rearranging the above gives

$$\begin{aligned} F(S_k^*) - F(S(\boldsymbol{q})_{t+1}) &= F(S_k^*) - F(S(\boldsymbol{q})_t) - \left(F(S(\boldsymbol{q})_{t+1}) - F(S(\boldsymbol{q})_t)\right) \\ &\leq F(S_k^*) - F(S(\boldsymbol{q})_t) - \frac{\gamma_{2k}}{k} \left(F(S_k^*) - F(S(\boldsymbol{q})_t)\right) \\ &= \left(1 - \frac{\gamma_{2k}}{k}\right) \left(F(S_k^*) - F(S(\boldsymbol{q})_t)\right). \end{aligned}$$

We now iterate this recursion. We set $E_t := F(S_k^*) - F(S(\boldsymbol{q})_t)$, $q := 1 - \frac{\gamma_{2k}}{k}$. Then the previous inequality states that $E_{t+1} \leq qE_t \, \forall t \in \{1, \ldots, k-1\}$. This immediately implies $E_t \leq q^{t-1} E_1$, $\forall t \in \{1, \ldots, k\}$. Taking $t = k$, we get

$$F(S_k^*) - F(S(\boldsymbol{q})_k) \leq \left(1 - \frac{\gamma_{2k}}{k}\right)^{k-1} \left(F(S_k^*) - F(S(\boldsymbol{q})_1)\right).$$

It remains to compute $F(S(\boldsymbol{q})_1)$. Since $S(\boldsymbol{q})_1$ is the best singleton chosen by Algorithm 1, we have $\hat{S}_1 = \{\hat{\jmath}\}$, $\hat{\jmath} \in \operatorname{argmax}_{j \in [N]} F(\{j\})$. For any fixed $j \in [N]$, the set $\Delta_{\{j\}}$ contains the entire mass 1. Hence, $\boldsymbol{\nu}^*_{\{j\}} = e_j, F(\{j\}) = -W(e_j)$. Also, $\mathrm{Ent}(e_j) = 0$. Therefore $W(e_j) = \sum_{r=1}^{B} \mathcal{W}(e_j, \boldsymbol{\mu}_r)$. For each $r \in [B]$, the constraints $\boldsymbol{\pi} \in \Pi(\boldsymbol{\mu}_r, e_j)$ force $\boldsymbol{\pi} = e_j \boldsymbol{\mu}_r^\top$. Substituting this unique feasible plan into (1), we obtain

$$\mathcal{W}(e_j, \boldsymbol{\mu}_r) = \sum_{\ell=1}^{N} \boldsymbol{C}[j, \ell] \boldsymbol{\mu}_r[\ell] + \lambda \sum_{\ell=1}^{N} \boldsymbol{\mu}_r[\ell] \left(\ln \boldsymbol{\mu}_r[\ell] - 1\right).$$
$$\implies W(e_j) = \sum_{r=1}^{B} \sum_{\ell=1}^{N} \boldsymbol{C}[j, \ell] \boldsymbol{\mu}_r[\ell] + \lambda \sum_{r=1}^{B} \sum_{\ell=1}^{N} \boldsymbol{\mu}_r[\ell] \left(\ln \boldsymbol{\mu}_r[\ell] - 1\right)$$
$$\implies F(\{j\}) = -\sum_{r=1}^{B} \sum_{\ell=1}^{N} \boldsymbol{C}[j, \ell] \boldsymbol{\mu}_r[\ell] - \lambda \sum_{r=1}^{B} \sum_{\ell=1}^{N} \boldsymbol{\mu}_r[\ell] \left(\ln \boldsymbol{\mu}_r[\ell] - 1\right).$$

This implies $F(S(\boldsymbol{q})_1) = \max_{s \in D} F(\boldsymbol{q}, \{s\})$. Substituting this identity into the preceding bound yields

$$\left(F(S_k^*) - F(S(\boldsymbol{q})_k\right) \leq \left(1 - \frac{\gamma_{2k}}{k}\right)^{k-1} \left(F(S_k^*) - \max_{s \in D} F(\boldsymbol{q}, \{s\})\right),$$

which proves the required approximation bound. □

D.3 Marginal gain as Approximate ANN search

Theorem 6 (Dot-product approximation of marginal gains)**.** Let $2 \le k \le N$. Suppose the three embeddings in (6) take values in $\mathbb{R}^d$ and have uniformly bounded norms. Suppose also that, $\nu_\theta(\boldsymbol{q}, \cdot)$ well approximates the true barycenter, ie $\max_{\boldsymbol{q}, i \in S} |\nu_\theta(\boldsymbol{q}, S) - \boldsymbol{\nu}^*(\boldsymbol{q}, S)| \le \epsilon$. Suppose both the cost oscillation δ and the entropic regularizer λ be small, specifically $\lambda = O(\epsilon)$ and $\delta = O(\epsilon)$. Then, for every $\eta > 0$, there are finite-dimensional representations $\boldsymbol{\zeta}_\theta(\boldsymbol{q}, S)$ and $\boldsymbol{\sigma}_\theta(\boldsymbol{x}_s)$ such that:

$$|F(s \mid S, \boldsymbol{q}) - \boldsymbol{\zeta}_\theta(\boldsymbol{q}, S)^\top \boldsymbol{\sigma}_\theta(\boldsymbol{x}_c)| = O(\epsilon) + O(\eta), \qquad \text{for } S \neq \emptyset. \tag{7}$$

Proof. Fix $\boldsymbol{q}$, put $T = S \cup \{s\}$, and write $p = \boldsymbol{\nu}^*(\boldsymbol{q}, S)$, $v = \boldsymbol{\nu}^*(\boldsymbol{q}, T)$, $\widehat{p} = \nu_\theta(\boldsymbol{q}, S)$, and $\widehat{v} = \nu_\theta(\boldsymbol{q}, T)$. Distributions on S are extended by zero to T. Let $a = \lambda + \tau$ and suppress the fixed ranker distributions in W.

By the entropy-decomposition lemma,

$$W(\nu) = a \sum_i \nu[i] \log \nu[i] + G(\nu), \qquad |G(\nu) - G(\nu')| \le \delta \, \mathrm{TV}(\nu, \nu'),$$

where G is convex and $\mathrm{TV}(\nu, \nu') = \frac{1}{2}\|\nu - \nu'\|_1$. The affine-bound lemma at v supplies a vector g with $\mathrm{osc}(g) \le \delta$ such that

$$G(\nu) \ge G(v) + g^\top(\nu - v), \qquad a(\log v[i] + 1) + g[i] \text{ is constant for } i \in T.$$

Using this last identity, the entropy identity, and $\sum_i (p[i] - v[i]) = 0$, gives the exact formula

$$F(s \mid S, \boldsymbol{q}) = a\mathrm{KL}(p\|v) + G(p) - G(v) - g^\top(p - v). \tag{89}$$

The residual is nonnegative by the affine lower bound. Moreover, for any two probability vectors,

$$|g^\top(p - v)| \le \mathrm{osc}(g) \, \mathrm{TV}(p, v).$$

Consequently,

$$0 \le F(s \mid S, \boldsymbol{q}) - a\mathrm{KL}(p\|v) \le 2\delta \, \mathrm{TV}(p, v) \le 2\delta. \tag{90}$$

This argument is valid when p has a zero coordinate at s; the entropy identity uses $0 \log 0 = 0$, and v is positive on T.

Next, bounded contextual-softmax embeddings give a uniform bound M on every logit for supports of size at most k. For example, if their norms are bounded by R_u, R_ϕ, R_ψ, one may take $M = (R_u + kR_\phi)R_\psi$. Every predicted active probability is therefore at least e^{-2M}/k. For sufficiently small ϵ, every true active probability is at least $c = e^{-2M}/(2k) > 0$ as well.

Along the line segments from (p, v) to $(\widehat{p}, \widehat{v})$, all coordinates in the KL sum lie in $[c, 1]$. Differentiating KL along these segments and integrating yields

$$|\mathrm{KL}(\widehat{p}\|\widehat{v}) - \mathrm{KL}(p\|v)| \le \left(k \log \frac{1}{c} + \frac{1}{c}\right) \epsilon.$$

Indeed, the derivative in the first argument is bounded by $k\epsilon \log(1/c)$, with its constant part canceling because the first argument sums to one. The derivative in the second argument is bounded by ϵ/c. Since a is uniformly bounded, (90) implies

$$|F(s \mid S, \boldsymbol{q}) - a\mathrm{KL}(\widehat{p}\|\widehat{v})| \le 2\delta + C\epsilon.$$

For completeness, the model KL admits separated finite-dimensional features by the same polynomial construction as for the midpoint surrogate. Write

$$h = u_\theta(\boldsymbol{q}) - \sum_{i \in S} f_\theta(\boldsymbol{x}_i), \quad z_i = g_\theta(\boldsymbol{x}_i), \quad Z = \sum_{i \in S} e^{h^\top z_i}, \quad w_i = \widehat{p}[i],$$

and let $x = f_\theta(\boldsymbol{x}_s)$, $y = g_\theta(\boldsymbol{x}_s)$. Direct substitution of the two contextual softmaxes gives

$$\mathrm{KL}(\widehat{p}\|\widehat{v}) = x^\top \sum_{i \in S} w_i z_i + \log \left(\sum_{i \in S} w_i e^{-x^\top z_i} + e^{(h-x)^\top y - \log Z} \right).$$

All exponential arguments lie in bounded intervals and the logarithm's argument lies in a compact interval bounded away from zero, uniformly over all admissible states and candidates. Uniform polynomial approximations to the exponential and logarithm therefore approximate the displayed expression, multiplied by a, within any prescribed $\eta > 0$. After expansion, this is a finite polynomial in the candidate coordinates (x, y), with coefficients depending only on $(\boldsymbol{q}, S)$. Use its coefficients as query features and its candidate monomials as corpus features. Their inner product approximates the scaled model KL within η. The triangle inequality completes the proof. □

Theorem 7 (End-to-end approximation for WBFUSION)**.** Fix $2 \le k \le N$ and a query $\boldsymbol{q}$. Under the assumptions of Theorem 6, suppose the representations used by the index attain its uniform approximation bound. Let $F_k^* := \max_{A \subseteq D, |A|=k} F(\boldsymbol{q}, A)$, and let γ be the submodularity-ratio lower bound from Theorem 5. Starting from a singleton S_1, let $S_{t+1} = S_t \cup \{s_t\}$ for $t = 1, \dots, k-1$. Suppose each ANN query returns an unselected item with maximum inner product. Then

$$F_k^* - F(\boldsymbol{q}, S_k) \le O\left(1 - \frac{\gamma}{k}\right)^{k-1} + O(\epsilon + \eta). \tag{8}$$

Proof. For each visited support S_t, write

$$h_t(s) := \boldsymbol{\zeta}_\theta(\boldsymbol{q}, S_t)^\top \boldsymbol{\sigma}_\theta(\boldsymbol{x}_s).$$

The assumed approximation bound gives

$$|h_t(s) - F(s \mid S_t, \boldsymbol{q})| \le C\epsilon + \eta \qquad \text{for every } s \notin S_t.$$

Let s_t^* maximize the true marginal gain at S_t. Since the index returns an unselected item maximizing h_t,

$$\begin{aligned} F(s_t \mid S_t, \boldsymbol{q}) &\ge h_t(s_t) - (C\epsilon + \eta) \\ &\ge h_t(s_t^*) - (C\epsilon + \eta) \\ &\ge \max_{s \notin S_t} F(s \mid S_t, \boldsymbol{q}) - 2(C\epsilon + \eta). \end{aligned}$$

Choose $O_k \subseteq D$ with $|O_k| = k$ and $F(\boldsymbol{q}, O_k) = F_k^*$, and define

$$E_t := F_k^* - F(\boldsymbol{q}, S_t).$$

For $t \le k$, monotonicity allows S_t to be extended to a set of size k, whose value cannot exceed F_k^*. Thus $E_t \ge 0$.

For $t < k$, let $U_t := O_k \setminus S_t$. Then U_t is nonempty, $|U_t| \le k$, and $|S_t \cup U_t| \le \min(N, 2k-1)$. Applying the submodularity-ratio bound to (S_t, U_t), followed by monotonicity, gives

$$\begin{aligned} \sum_{s \in U_t} F(s \mid S_t, \boldsymbol{q}) &\ge \gamma\big[F(\boldsymbol{q}, S_t \cup O_k) - F(\boldsymbol{q}, S_t)\big] \\ &\ge \gamma E_t. \end{aligned}$$

Since there are at most k terms,

$$\max_{s \notin S_t} F(s \mid S_t, \boldsymbol{q}) \ge \frac{\gamma}{k} E_t.$$

Combining this with the preceding selection bound yields

$$E_{t+1} = E_t - F(s_t \mid S_t, \boldsymbol{q}) \le \left(1 - \frac{\gamma}{k}\right) E_t + 2(C\epsilon + \eta).$$

Iterating over $t = 1, \dots, k-1$, we obtain

$$\begin{aligned} E_k &\le \left(1 - \frac{\gamma}{k}\right)^{k-1} E_1 + 2(C\epsilon + \eta) \sum_{j=0}^{k-2} \left(1 - \frac{\gamma}{k}\right)^j \\ &\le \left(1 - \frac{\gamma}{k}\right)^{k-1} E_1 + 2(k-1)(C\epsilon + \eta). \end{aligned} \tag{91}$$

It remains to show that $E_1 = O(1)$ uniformly in the asymptotic family of objectives. We establish the bound

$$E_1 \le \delta + (\lambda + \tau) \log k.$$

Subtracting the smallest entry of $\boldsymbol{C}$ changes W by a support-independent constant, so it leaves all objective gaps unchanged. We may therefore assume $0 \le \boldsymbol{C}[i,j] \le \delta$.

Define the support-independent quantity

$$c_{\boldsymbol{q}} := -\lambda \sum_{r=1}^{B} \beta_r \big[\mathrm{Ent}(\boldsymbol{\mu}_r(\boldsymbol{q}, D)) + 1\big].$$

For any coupling $\boldsymbol{\pi}$ with marginals $\boldsymbol{\nu}$ and $\boldsymbol{\mu}_r(\boldsymbol{q}, D)$, nonnegativity of its KL divergence from the product of its marginals gives

$$\sum_{i,j} \boldsymbol{\pi}[i,j](\log \boldsymbol{\pi}[i,j] - 1) \ge -\mathrm{Ent}(\boldsymbol{\nu}) - \mathrm{Ent}(\boldsymbol{\mu}_r(\boldsymbol{q}, D)) - 1.$$

Using nonnegative transport costs and $\sum_{r=1}^{B} \beta_r = 1$, we obtain

$$W(\boldsymbol{\nu}, \{\boldsymbol{\mu}_r\}) \geq c_{\boldsymbol{q}} - (\lambda + \tau)\,\mathrm{Ent}(\boldsymbol{\nu}).$$

In particular, every probability distribution supported on at most k items satisfies

$$W(\boldsymbol{\nu}, \{\boldsymbol{\mu}_r\}) \geq c_{\boldsymbol{q}} - (\lambda + \tau)\log k.$$

On the singleton $S_1 = \{j\}$, the barycenter is $\mathbf{e}_j$. The only feasible coupling for ranker r is $\mathbf{e}_j \boldsymbol{\mu}_r(\boldsymbol{q}, D)^\top$, so

$$W(\mathbf{e}_j, \{\boldsymbol{\mu}_r\}) \leq c_{\boldsymbol{q}} + \delta.$$

Consequently, using the definition of F,

$$\begin{aligned} E_1 &= W(\mathbf{e}_j, \{\boldsymbol{\mu}_r\}) - W(\boldsymbol{\nu}^*(\boldsymbol{q}, O_k), \{\boldsymbol{\mu}_r\}) \\ &\leq \delta + (\lambda + \tau)\log k. \end{aligned}$$

The assumptions inherited from Theorem 6 include $\lambda = O(\epsilon)$, $\delta = O(\lambda)$, and $\tau = O(1)$. Hence $E_1 = O(1)$ for fixed k. Likewise, $2(k-1)(C\epsilon + \eta) = O(\epsilon + \eta)$. Substituting these estimates into (91) proves

$$F_k^* - F(\boldsymbol{q}, S_k) \leq O\left(\left(1 - \frac{\gamma}{k}\right)^{k-1}\right) + O(\epsilon + \eta).$$

□

E IMPLEMENTATION DETAILS ABOUT WBFUSION

Base Rankers We use six base rankers. Three of them are lexical rankers (BM25, TFIDF, QL), one is a frozen dense embedder (Sentence-BERT encoder `all-MiniLM-L6-v2` with $d = 384$ dimensions) with dot product similarity, one is frozen Sentence-BERT Cross Encoder and one is trainable heads on frozen embeddings. The trainable one is SBERT-MLP, which stacks the query and corpus SBERT embeddings $\boldsymbol{q}, \boldsymbol{x}$ as $\mathbf{vec}(q, c) = [\boldsymbol{q}, \boldsymbol{x}, \boldsymbol{q} \odot \boldsymbol{x}, |\boldsymbol{q} - \boldsymbol{x}|] \in \mathbb{R}^{4d}$, and applies a MLP of the form $4d \to 256 \to 128 \to 1$ with ReLU activations and dropout 0.1.

WBFUSION Query and corpus embeddings are passed through a linear layer, concatenated into a four-token sequence with learnable CLS and SEP token(s) and positional embeddings, and passed through a two-layer Transformer encoder. The CLS token output is layer-normed, and concatenated with base scorers' scores. We apply a small dropout layer on the base scorer scores (Dropout 0.3) to avoid the risk of the model simply learning a MLP on the base scores. The only difference between base and full methods is that for the full method, we also concatenate the support summary obtained by a DeepSet on the selected greedy-support set. A MLP (Linear $\to$ GeLU $\to$ Dropout(0.1) $\to$ Linear) on top of this gives the final scores.

WBFUSION (no rerank) WBFUSION is the RSWB-based KL regularized reranker. WBFUSION (no rerank) is a version where the ranking is done w.r.t. the Wasserstein Barycentric weights. Instead of using a reranker in line 7 of Algorithm 4 we simply return S_k.

OT cost matrix Given a query $\boldsymbol{q}$ we compute the sinkhorn cost matrix using SentenceTransformer embeddings of the candidate documents $\boldsymbol{x}$. Given documents i, j the cost matrix entry $C[i, j] = \|\boldsymbol{x}_i - \boldsymbol{x}_j\|_2$, $C[i, i] = 0 \; \forall \, i$.

E.1 TRAINING PIPELINE

For evaluations of exact-greedy followed by re-ranking using WBFUSION, we take union of top-100 of each base rankers per query and consider this as the corpus pool. For evaluations involving ANN based WBFUSION , we take RRF-truncated top 1000 corpus items per query. In a scaling study done in Appendix G.9, we have expanded this corpus pool as well.

For computational tractability, we restrict our training set throughout for all models only on a set of at most 2000 queries.

Training Base Models SBERT-MLP is trained via in-batch triplet loss.

Targets Cache Having obtained the six base ranker distributions on the reduced support set we solve the RSWB for a chosen (λ, τ) using greedy selection. We cache the K-item support set S plus the associated barycenter ν. This pair provides the distillation targets for WBFUSION. WBFUSION (no rerank) ranks items by the barycenter weights ν. We sweep over (λ, τ) to obtain dataset-specific choices for hyperparameters which optimize downstream ranking metrics for WBFUSION (no rerank).

WBFUSION training The loss has two terms. The triplet loss operates on the whole pool of candidates. The KL term is on a 10-atom subset, with target Wasserstein Barycentric distribution recomputed on this subset at every step using given (λ, τ). The regularization strength β and the margin of the triplet loss is swept over and we choose the set that give optimal downstream metric (NDCG) on a validation set obtained by splitting the training set of 2000 queries in a $90/10$ train/val split.

Training trainable baselines NeuralRRF, Ensemble, Hybrid and MoE are all trained using triplet loss and a Lipschitz-style smoothness regularizer $\mathbb{E}_{(p,n)}|f(\mu_p) - f(\mu_n)|/(||\boldsymbol{x}_p - \boldsymbol{x}_n||_2 \cdot ||\mu_p - \mu_n||_2 + \epsilon)$ where $\boldsymbol{x}_p, \boldsymbol{x}_n$ are positive and negative corpus item embeddings for a query; and μ denotes the 6-dimensional vector of ranker scores.

Surrogate Two-Tower Indexer Our greedy surrogate is a two-tower MLP that approximates the marginal gain in terms of the inner product, $\boldsymbol{\zeta}_\theta(\boldsymbol{q}, S)^\top \boldsymbol{\sigma}_\theta(\boldsymbol{x}_c)$. The query tower $\boldsymbol{\zeta}_\theta(\boldsymbol{q}, S)$ includes the current support set using a DeepSet aggregator, while the corpus tower $\boldsymbol{\sigma}_\theta(\boldsymbol{x}_c)$ is independent of S. Both towers are 3 layer MLPs with GeLU activations. For each train query, we run exact-greedy on the RRF-1000 pool and log every per-step state. The loss we train the surrogate on consists of two terms: a listwise loss and a triplet loss on the ground-truth relevant documents.

At eval time, the surrogate is used as a pre-filter. The corpus tower $\boldsymbol{\sigma}_\theta$ is used to build an index over the whole pool once and the query tower is then called repeatedly one element at a time till we get $K' = 100$ elements. At each step we just find the argmax for the estimated marginal gain with no Sinkhorn iterations. After the prefilter step is done, lazy-greedy runs on the obtained subset to obtain the top-$K = 10$ items. This set is reranked by WBFUSION reranker.

F Additional details about Experimental Setup

Hardware The experiments were run on NVIDIA A6000 and NVIDIA L40 GPUs with AMD EPYC CPU.

Datasets We choose five BEIR datasets (Thakur et al., 2021) for all of our experiments: MS-Marco, NF-Corpus, Hotpot-QA, Fever and SciFact. These datasets span a range of domains, from web search to multi-hop QA, to fact checking, to biomedicine; a range of query styles, from keywords to natural language as well as corpus size.

For computational tractability, we work with at most 2000 queries on the train split for each of the BEIR datasets. For reporting the test numbers, we use the available splits for MS-Marco, NF-Corpus, Hotpot-QA, Fever and SciFact from (MSMARCO), (NFCORPUS), (HOTPOTQA), (FEVER) and (SCIFACT) respectively.

Ranking metrics We measure ranking quality by two metrics: Mean Average Precision (MAP) and Normalized Discounted Cumulative Gain (NDCG, at 10).

MAP Given a query $q \in Q$ and a set of relevant documents $R_q \subseteq D$, for a ranked list $\{d_k\}$, the Average Precision (AP) is measured as:

$$\text{AP}(q) = \frac{1}{|R_q|} \sum_{k=1}^{K} \text{Prec@k} \cdot \mathbf{1}[d_k \in R_q]. \tag{92}$$

Where Prec@k is the fraction of documents that are relevant till position k in the ranked order. Taking a mean of AP(q) for all queries gives us the MAP:

$$\text{MAP} = \frac{1}{|Q|} \sum_{q \in Q} \text{AP(q)}. \tag{93}$$

NDCG Assume that for each query $q \in Q$ we are given *graded* relevance labels (e.g. $0, 1, 2, 3$ corresponding to degree of relevance) for documents, with rel_k denoting the relevance of document at rank k. Then the Discounted Cumulative Gain (DCG) at K is:

$$\text{DCG@K} = \sum_{k=1}^{K} \frac{2^{\text{rel}_k} - 1}{\log_2(k+1)}. \tag{94}$$

Normalizing this by the DCG of the ideal ranking (called the IDCG) gives us the NDCG:

$$\text{NDCG@K} = \frac{\text{DCG@}K}{\text{IDCG@}K}. \tag{95}$$

We report metrics for $K = 10$.

Loss function We train the parametric models $\boldsymbol{\zeta}_\theta(\boldsymbol{q}, S), \boldsymbol{\sigma}_\theta(\boldsymbol{x}_c)$ using a listwise loss with margin-based triplet regularization.

$$\begin{aligned} \mathcal{L}_\theta = \sum_{\boldsymbol{q}, S} \text{CrossEntropy}(\{(F(c \mid S, \boldsymbol{q}), \boldsymbol{\zeta}_\theta(\boldsymbol{q}, S)^\top \boldsymbol{\sigma}_\theta(\boldsymbol{x}_c)) : c \notin S\}) + \\ \lambda_{\text{reg}} \sum_{(\boldsymbol{q}, c_+, c_-)} [\text{margin} + \boldsymbol{\zeta}_\theta(\boldsymbol{q}, S)^\top \boldsymbol{\sigma}_\theta(\boldsymbol{x}_{c_-}) - \boldsymbol{\zeta}_\theta(\boldsymbol{q}, S)^\top \boldsymbol{\sigma}_\theta(\boldsymbol{x}_{c_+})]_+ \end{aligned} \tag{96}$$

where the triplet regularizer runs over $(\boldsymbol{q}, c_+, c_-)$ from ground truth relevance data.

F.1 Rank Fusion Baselines

Given rankers $r \in \mathcal{R}$, we use the following rank fusion baselines, three of which are training-free and the rest are trainable.

RRF (Cormack et al., 2009) For each document d, $\sum_{r=1}^{k} \frac{1}{\text{rank}_r(d)+60}$ where $\text{rank}_r(d)$ is the rank of document d as given by ranker r. Completely training-free.

CombSUM, CombMNZ (Fox et al., 1993; Lee, 1997) CombSUM performs rank fusion by simply adding the scores of the individual rankers. CombMNZ weighs document scores by the number of rankers that returned the corresponding document.

MoE This is a trainable baseline where a linear layer followed by softmax produces per-query mixing weights to combine base ranker scores.

Ensemble Query independent. We stack the scores of each of the rankers and pass it through a MLP.

Hybrid Consider a dense ranker and a sparse ranker (e.g. dense and BM25). We concatenate their scores and pass the 2D vector through a MLP.

NeuralRRF We modify RRF to be trainable as follows. For a document d and a ranker r we call $1/(60 + \text{rank}_r(d))$ as the rank feature of the document for that ranker. We stack all of these per document across rankers and pass it through a MLP. Effectively we are trying to combine RRF features in a nonlinear manner in an attempt to strengthen the RRF baseline(s).

G Additional Experimental Results

G.1 Comparison of WBFusion with baselines

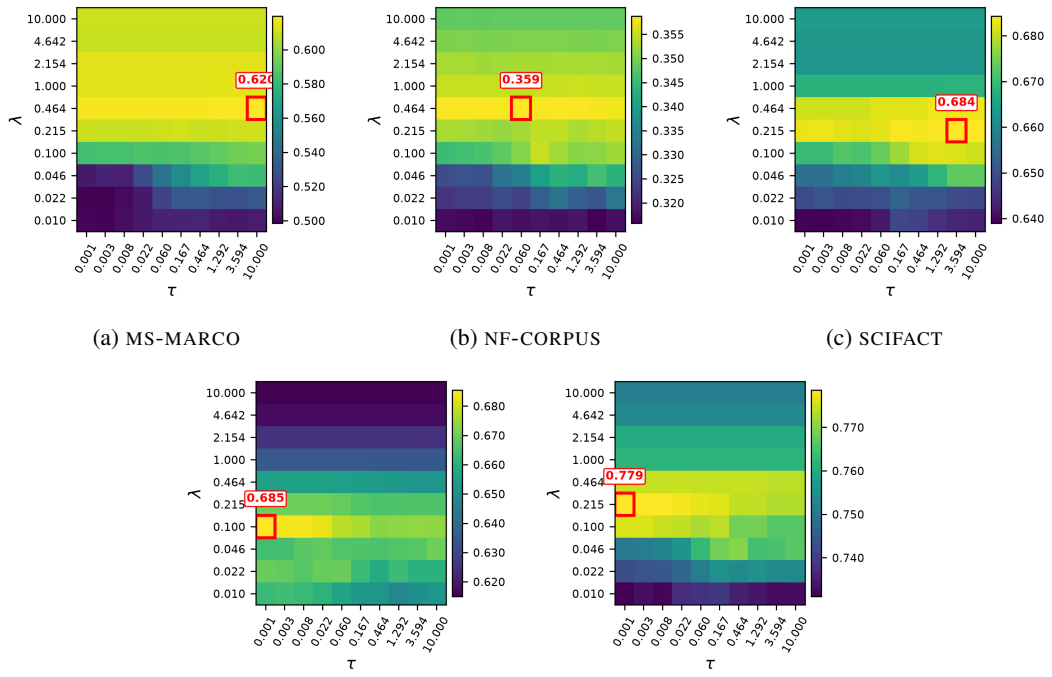
We compare WBFusion against three training-free baselines (RRF, CombSUM and CombMNZ) and four trainable baselines (a query-dependent Mixture of Experts, a query-independent Ensemble model, a neural variant of RRF described in Appendix F, and a Hybrid reranker) for the task of rank fusion. Each rank fusion method used the same six base rankers. WBFusion(no rerank) is a setup that ranks documents by the barycenter weights without reranking. Table 8 shows the comparison. We observe that WBFusion beats all other rank fusion baselines on all five BEIR datasets, with MoE and Ensemble methods sharing second place. WBFusion(no rerank) is the strongest among training-free rank fusion methods.

	MS-Marco		NF-Corpus		Hotpot-QA		Fever		SciFact	
Method	MAP	NDCG	MAP	NDCG	MAP	NDCG	MAP	NDCG	MAP	NDCG
RRF	0.136	0.543	0.119	0.328	0.522	0.624	0.581	0.653	0.623	0.664
CombSUM	0.130	0.545	0.126	0.344	0.551	0.648	0.620	0.685	0.628	0.673
CombMNZ	0.130	0.545	0.125	0.342	0.551	0.648	0.620	0.685	0.628	0.673
WBFusion (no rerank)	0.143	0.585	0.135	0.358	0.592	0.685	0.722	0.774	0.627	0.678
MoE	0.153	0.605	0.138	<u>0.376</u>	0.642	0.726	0.759	0.804	<u>0.678</u>	<u>0.720</u>
Ensemble	0.157	0.638	<u>0.141</u>	0.369	0.594	0.685	<u>0.806</u>	<u>0.843</u>	0.658	0.706
NeuralRRF	0.152	0.613	0.130	0.374	0.549	0.649	0.653	0.718	0.646	0.694
Hybrid	0.137	0.566	0.130	0.374	0.494	0.589	0.549	0.616	0.646	0.686
WBFusion	**0.163**	**0.673**	**0.145**	**0.383**	**0.647**	**0.728**	**0.814**	**0.846**	**0.712**	**0.748**

Table 8: Performance comparison, measured in terms of retreival MAP and NDCG, between WBFusion and other rank fusion baselines. Top: training-free, pointwise rank fusion methods. Bottom: trainable rank fusion methods, *viz*, query-dependent Mixture of Experts (MoE), query-independent ensemble model (Ensemble), a MLP over RRF (NeuralRRF), and a hybrid fusion model (Hybrid). Numbers in **bold** (<u>underline</u>) indicate the best (second best) performers. WBFusion beats all baseline rank fusion methods.

G.2 Dependence on Entropic regularizers

We show heatmaps across a (λ, τ) sweep of NDCG for WBFusion (no rerank) variant for all the datasets (Figure 9). We can see that both λ and τ impact retrieval performance for MS-Marco, NF-Corpus and SciFact. For Hotpot-QA and Fever λ is the dominating hyperparameter. Performance is robust to hyperparameter choice as NDCG remains high around the optimum for four of the five datasets.



(a) MS-MARCO (b) NF-CORPUS (c) SCIFACT

(d) HOTPOTQA (e) FEVER

Figure 9: NDCG of WBFusion (no rerank) over the (λ, τ) grid across five datasets. Red box marks the optimum. NDCG is robust to choice of (λ, τ) for four of five datasets, while for HotpotQA it is sensitive to λ.

G.3 Impact of Regularization strength β

β	MS-Marco	NF-Corpus	Hotpot-QA	Fever	SciFact
0	0.670	0.373	0.722	0.832	0.745
0.001	0.671	0.375	0.726	**0.846**	**0.748**
0.01	**0.673**	0.377	**0.728**	0.830	0.728
0.1	0.672	0.380	0.724	0.835	0.734
1	0.670	**0.383**	0.722	0.831	0.725

Table 10: NDCG over β for WBFusion. The optimal regularization strength is dataset dependent, i.e., it depends on the dataset in question how much impact corpus geometry has on retrieval performance.

The training loss is $\mathcal{L}_\phi = \mathcal{L}_{\text{tri}} + \beta\, D_{\text{KL}}$.

$$\mathcal{L}_\phi := \sum_{(q,c_+,c_-)} [m + \text{score}_\phi(q, S, c_-) - \text{score}_\phi(q, S, c_+)]_+ + \beta \sum_q \text{KL}(\nu^*(q, S) \| \nu_\phi(q, S)).$$

The optimal β drifts across datasets by over an order of magnitude: Fever and SciFact peak at small $\beta \approx 0.001$, while Hotpot-QA and MS-Marco prefer moderate $\beta \approx 0.01$. NF-Corpus prefers large $\beta \approx 1.0$. We observe in Table 10 how ranking accuracy varies with β. For this, at each β, we run a 3×2 grid sweep across learning rates and dropouts, and report the test metric for the best-by-validation checkpoint. The variation in β suggests that the importance of the barycenter, and consequently, of corpus geometry, is dataset-dependent.

G.4 Impact of the Z-score Normalization:

Both the OT cost (1) and the Wasserstein barycenter objective (2) operate on probability distributions, whereas rankers typically output unconstrained scores $\mathrm{score}(\boldsymbol{q}, \boldsymbol{x}) \in \mathbb{R}$. To map these into a probability distribution without changing the induced ranking, for each ranker $\mathcal{R}_r$ with relevance score function $\mathrm{score}_r(\bullet, \bullet)$, we define the induced z-score normalized categorical distribution as:

$$\begin{aligned} \mathrm{z}_r(\boldsymbol{q}, \boldsymbol{x}_s) &\coloneqq \begin{cases} \left(\mathrm{score}_r(\boldsymbol{q}, \boldsymbol{x}_s) - \overline{\mathrm{score}_r(\boldsymbol{q})}\right) / \sigma_r(\boldsymbol{q}), & \sigma_r(\boldsymbol{q}) > 0 \\ 0, & \sigma_r(\boldsymbol{q}) = 0 \end{cases} \\ \boldsymbol{\mu}_r(\boldsymbol{q}, D)[s] &\coloneqq \frac{\exp\left(\mathrm{z}_r(\boldsymbol{q}, \boldsymbol{x}_s)\right)}{\sum_{t \in D} \exp\left(\mathrm{z}_r(\boldsymbol{q}, \boldsymbol{x}_t)\right)}, \qquad \text{for all } s \in D. \end{aligned} \tag{97}$$

In the above, $\overline{\mathrm{score}_r(\boldsymbol{q})}, \sigma_r(\boldsymbol{q})$ are the population mean and standard deviation respectively of raw scores $\{\mathrm{score}_r(\boldsymbol{q}, \boldsymbol{x}_s) : s \in D\}$ of the rth base ranker. Thus, for every fixed query $\boldsymbol{q}$, the B different base rankers induce B different probability distributions $\boldsymbol{\mu}_1(\boldsymbol{q}, D), \ldots, \boldsymbol{\mu}_B(\boldsymbol{q}, D) \in \Delta_D$.

In Table 11 we show how our proposed approach of z-score normalization mitigates the problem that happens with the spiky behavior of some of the base rankers, by reporting entropy of the distribution of base rankers' scores with and without z-score normalization.

	MS-Marco		NF-Corpus		Hotpot-QA		Fever		SciFact	
Ranker	Raw	Z	Raw	Z	Raw	Z	Raw	Z	Raw	Z
BM25	0.30	0.93	0.34	0.64	0.11	0.89	0.21	0.91	0.12	0.69
TF-IDF	1.00	0.89	1.00	0.55	1.00	0.85	1.00	0.87	1.00	0.57
QLM	0.73	0.91	0.69	0.60	0.29	0.81	0.44	0.88	0.35	0.63
Dense	1.00	0.93	1.00	0.91	1.00	0.94	1.00	0.94	1.00	0.90
SBERT-MLP	0.63	0.93	0.94	0.94	0.90	0.93	0.94	0.93	0.99	0.89
Cross-encoder	0.49	0.91	0.32	0.49	0.24	0.80	0.24	0.84	0.22	0.61

Table 11: Entropy of base ranker distributions. (Raw) means entropy before z-score normalization and (Z) means entropy after z-score normalization. Overall, z-score normalization increases entropy, i.e., makes the distributions smoother, especially for sparse retrieval methods like BM25 known to have spiky score distributions.

G.5 Ablation: RSWB anchor vs others

Tables 12 and 13 isolate the exact contribution of RSWB in the WBFusion pieline by comparing it to (i) Pure triplet loss by making the regularization strength $\beta = 0$ (ii) Another rank fusion anchor for the KL regularization, we take RRF. We see that the RRF anchor generally degrades NDCG over no anchor ($\beta = 0$). The barycentric anchor gives the best retrieval quality in terms of NDCG over all datasets, and in terms of MAP for four of five datasets.

KL target	MS-Marco	NF-Corpus	Hotpot-QA	Fever	SciFact
None ($\beta = 0$)	0.6690	0.3564	0.7219	0.8108	0.7446
RRF	0.6670	0.3722	0.7220	0.8012	0.7422
Barycenter	**0.6727**	**0.3830**	**0.7284**	**0.8465**	**0.7482**

Table 12: NDCG of WBFusion with only the KL anchor changed. Barycenter: WBFusion. None ($\beta = 0$): the same configuration with the KL weight set to 0. RRF: the KL target replaced by the RRF distribution of the base rankers over the same sampled support. The RRF anchor improves over None in NF-Corpus and Hotpot-QA but degrades everywhere else. Anchoring on the barycenter, however, imporves performance in all datasets over both None and RRF. The geometric information contained in this anchor helps improve retrieval quality.

KL target	MS-Marco	NF-Corpus	Hotpot-QA	Fever	SciFact
None ($\beta = 0$)	0.1644	0.1361	0.6383	0.7673	0.7083
RRF	**0.1645**	0.1407	0.6386	0.7545	0.7063
Barycenter	0.1633	**0.1447**	**0.6467**	**0.8142**	**0.7122**

Table 13: MAP of WBFusion (SFT) with only the KL anchor changed. Barycenter: WBFusion. None ($\beta = 0$): the same configuration with the KL weight set to 0. RRF: the KL target replaced by the RRF distribution of the base rankers over the same sampled support.

Remark: For the above ablation, we have taken *exact same* configuration of the WBFusion model from table 1 and replace/remove the KL regularizer. It should be noted that the $\beta = 0$ row of table 10 reports the best(by validation) configuration at $\beta = 0$, and the reported test metric is supposed to be $\geq$ the $\beta = 0$ cell for table 12 which runs the overall winner's configuration at $\beta = 0$ for a "true" ablation study.

G.6 Empirical submodularity ratio

Table 14 shows that the empirical submodularity ratio of the induced set function F for all of our datasets exceeds 1. We report the minimum, over 5000 sampled pairs (L, S'), of $\sum_{s \in S'}[F(L \cup \{s\}) - F(L)] \,/\, [F(L \cup S') - F(L)]$ at each (λ, τ). We see that the observed submodularity ratio is greater than 1.

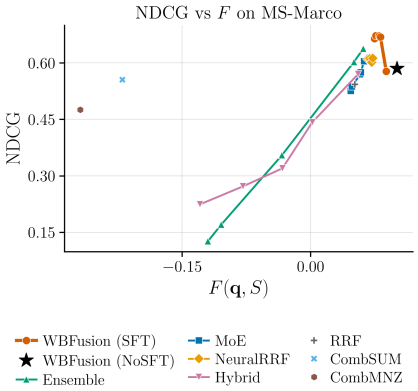
λ	τ	MS-Marco	NF-Corpus	HotpotQA	FEVER	SciFact
0.100	0.001	1.002	1.000	1.000	1.001	1.001
0.100	0.167	1.006	1.006	1.008	1.009	1.008
0.464	0.001	1.013	1.012	1.013	1.015	1.014
0.464	0.167	1.017	1.016	1.017	1.019	1.018
2.154	0.001	1.027	1.027	1.027	1.027	1.027
2.154	0.167	1.027	1.027	1.027	1.028	1.028

Table 14: Empirical submodularity ratio for set function F over all BEIR datasets for different values of entropic regularizers. Minimum over sampled pairs of sets (L, S'). Singleton sets are excluded by construction since their ratio is trivially 1.

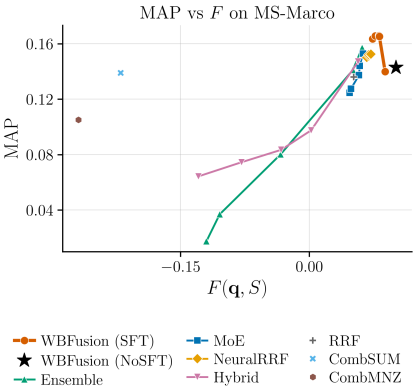
Singletons ($|S'| = 1$) are excluded since in that case the ratio is identically 1.

G.7 TRADEOFF BETWEEN RANK FUSION QUALITY AND RETRIEVAL PERFORMANCE

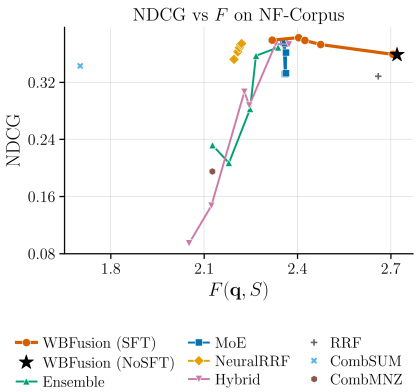
We sweep the regularization hyperparameters for each method, bucket the corresponding values of F into 12 buckets, and select the best-performing methods to plot a scatterplot of F versus the metric. These tradeoff plots are shown in Figures 15 and 16. WBFUSION generally shows the best tradeoffs between retrieval performance (MAP/NDCG) and rank fusion quality F. WBFUSION(no rerank) generally gives better F but trades off some retrieval performance in the process. This shows that the reranking step helps preserve rank fusion quality while improving retrieval performance.



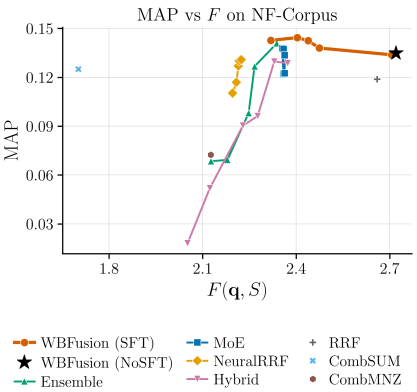
(a) MS-Marco — NDCG



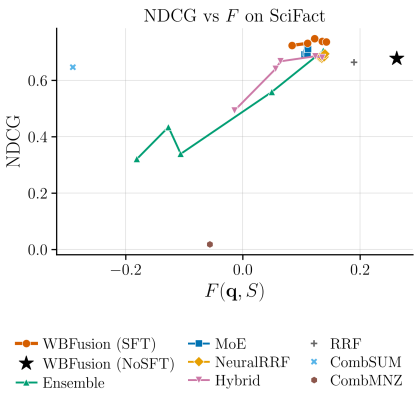
(b) MS-Marco — MAP



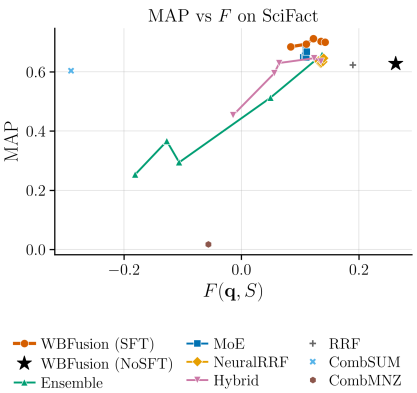
(c) NF-Corpus — NDCG



(d) NF-Corpus — MAP

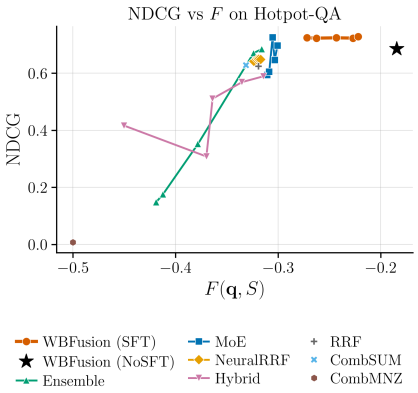


(e) SciFact — NDCG

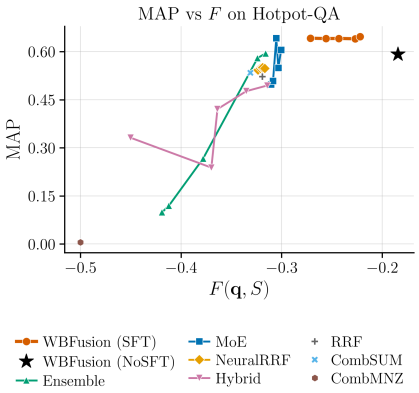


(f) SciFact — MAP

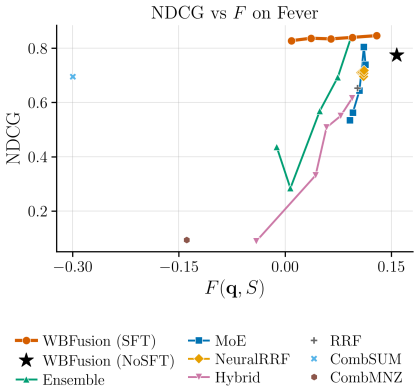
Figure 15: Pareto fronts on MS-Marco, NF-Corpus, and SciFact. **Right column:** MAP vs $F(q, S^*)$. **Left column:** NDCG vs $F(q, S^*)$. WBFUSION gives the best F versus MAP/NDCG tradeoffs.



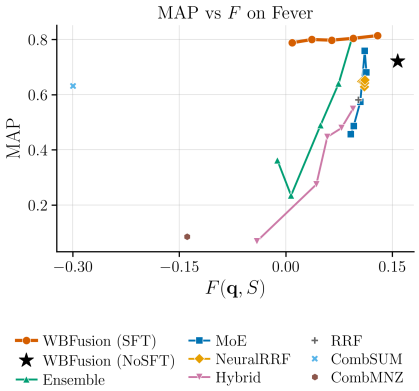
(a) HotPotQA — NDCG



(b) HotPotQA — MAP



(c) Fever — NDCG



(d) Fever — MAP

Figure 16: Pareto fronts on HotPotQA and Fever. **Left column:** NDCG vs $F(q, S^*)$. **Right column:** MAP vs $F(q, S^*)$. WBFUSION gives the best F versus MAP/NDCG tradeoffs.

G.7.1 AVERAGE NDCG AND AVERAGE MAP

To get a better sense of the performance of these models, we perform rigorous thresholding experiments. Given a hyperparameter sweep, in Tables 17, 18,19 we consider thresholding F at $20, 50$, and 80% of its optimal ($F^\star$), i.e., $\frac{F^\star - F}{|F^\star|} \leq 0.8, 0.5,$ and 0.2 respectively, and report the average metric (NDCG or MAP) of the part of the sweep that crosses the threshold. These threshold tables show that as we demand better rank fusion quality, WBFUSION clears the requirements most often while maintaining high retrieval performance.

	MS-Marco		NF-Corpus		Hotpot-QA		Fever		SciFact	
Method	AVG_{MAP}	AVG_{NDCG}	AVG_{MAP}	AVG_{NDCG}	AVG_{MAP}	AVG_{NDCG}	AVG_{MAP}	AVG_{NDCG}	AVG_{MAP}	AVG_{NDCG}
MoE	0.134	0.552	0.126	0.342	0.522	0.618	0.515	0.586	<u>0.654</u>	<u>0.698</u>
Ensemble	<u>0.151</u>	<u>0.625</u>	0.104	0.290	<u>0.587</u>	<u>0.677</u>	<u>0.694</u>	<u>0.745</u>	0.624	0.670
NeuralRRF	<u>0.151</u>	0.608	<u>0.127</u>	<u>0.369</u>	0.544	0.645	0.643	0.708	0.635	0.684
Hybrid	0.134	0.540	0.116	0.333	0.484	0.579	0.495	0.563	0.623	0.666
WBFUSION	**0.165**	**0.669**	**0.142**	**0.376**	**0.641**	**0.724**	**0.777**	**0.819**	**0.695**	**0.734**

Table 17: Average value of ranking metric (MAP and NDCG) under the constraint that F should reach at least 20% of the optimum ($F^\star$), i.e., $\frac{F^\star - F}{|F^\star|} \leq 0.8$. Higher is better. Numbers in **bold** (<u>underline</u>) indicate the best (second best) performer. (—) indicates that F never reached 20% of the optimum. WBFUSION gives the best average MAP and NDCG across datasets.

	MS-Marco		NF-Corpus		Hotpot-QA		Fever		SciFact	
Method	AVG_{MAP}	AVG_{NDCG}	AVG_{MAP}	AVG_{NDCG}	AVG_{MAP}	AVG_{NDCG}	AVG_{MAP}	AVG_{NDCG}	AVG_{MAP}	AVG_{NDCG}
MoE	0.146	0.587	0.126	0.342	—	—	0.515	0.586	—	—
Ensemble	<u>0.151</u>	<u>0.625</u>	0.104	0.290	—	—	**0.786**	**0.826**	<u>0.653</u>	<u>0.698</u>
NeuralRRF	<u>0.151</u>	0.608	<u>0.127</u>	<u>0.369</u>	—	—	0.643	0.708	0.635	0.684
Hybrid	0.141	0.566	0.116	0.333	—	—	0.526	0.596	0.627	0.673
WBFUSION	**0.165**	**0.669**	**0.142**	**0.376**	**0.641**	**0.724**	<u>0.772</u>	<u>0.815</u>	**0.691**	**0.730**

Table 18: Average value of ranking metric (MAP and NDCG) under the constraint that F should reach at least 50% of the optimum ($F^\star$), i.e., $\frac{F^\star - F}{|F^\star|} \leq 0.5$. Higher is better. Numbers in **bold** (<u>underline</u>) indicate the best (second best) performer. (—) indicates that F never reached 50% of the optimum.WBFUSION gives the best average MAP and NDCG across datasets.

	MS-Marco		NF-Corpus		Hotpot-QA		Fever		SciFact	
Method	AVG_{MAP}	AVG_{NDCG}	AVG_{MAP}	AVG_{NDCG}	AVG_{MAP}	AVG_{NDCG}	AVG_{MAP}	AVG_{NDCG}	AVG_{MAP}	AVG_{NDCG}
MoE	—	—	0.126	0.342	—	—	—	—	—	—
Ensemble	—	—	0.107	0.296	—	—	—	—	—	—
NeuralRRF	—	—	<u>0.127</u>	<u>0.369</u>	—	—	—	—	—	—
Hybrid	—	—	0.116	0.333	—	—	—	—	—	—
WBFUSION	**0.165**	**0.668**	**0.142**	**0.376**	**0.640**	**0.723**	**0.804**	**0.839**	—	—

Table 19: Average value of ranking metric (MAP and NDCG) under the constraint that F should reach at least 80% of the optimum ($F^\star$), i.e., $\frac{F^\star - F}{|F^\star|} \leq 0.2$. Higher is better. Numbers in **bold** (<u>underline</u>) indicate the best (second best) performer. (—) indicates that F never reached 80% of the optimum. WBFUSION gives the best average MAP and NDCG across datasets.

G.7.2 AREA UNDER THE CUMULATIVE MEAN CURVE, AUC CUM_MEAN

Given a hyperparameter sweep, we generate a curve between ranking metric (MAP/NDCG) and F using cumulative means, considering only the part of the curve beyond some fraction α of optimal ($F^\star$), i.e., $\frac{F^\star - F}{|F^\star|} \leq 1 - \alpha$ for some $0 < \alpha < 1$. The area under this restricted curve is denoted AUC cum_mean for the chosen threshold of F. Tables 20, 21, 22 summarize the results for AUC cum_mean under different thresholds of F. WBFUSION remains active for the most stringent threshold across more datasets than baselines and consistently has the highest AUC cum_mean, showing that WBFUSION gives the most robust tradeoffs of all the rank fusion methods compared.

	MS-Marco		NF-Corpus		Hotpot-QA		Fever		SciFact	
Method	AUC_{MAP}	AUC_{NDCG}	AUC_{MAP}	AUC_{NDCG}	AUC_{MAP}	AUC_{NDCG}	AUC_{MAP}	AUC_{NDCG}	AUC_{MAP}	AUC_{NDCG}
MoE	0.010	0.044	<u>0.296</u>	<u>0.806</u>	0.076	0.090	0.060	0.069	<u>0.137</u>	<u>0.146</u>
Ensemble	<u>0.012</u>	<u>0.050</u>	0.164	0.527	<u>0.086</u>	<u>0.100</u>	0.076	0.084	0.121	0.131
NeuralRRF	<u>0.012</u>	0.049	0.260	0.792	0.080	0.095	<u>0.081</u>	<u>0.089</u>	0.133	0.143
Hybrid	0.010	0.041	0.206	0.676	0.071	0.085	0.053	0.061	0.130	0.139
WBFUSION	**0.013**	**0.053**	**0.310**	**0.824**	**0.095**	**0.107**	**0.100**	**0.105**	**0.146**	**0.154**

Table 20: Area under the cumulative mean curve (AUC cum_mean) between ranking metric (MAP and NDCG) and F under the constraint that F should reach at least 20% of the optimum ($F^\star$), i.e., $\frac{F^\star - F}{|F^\star|} \leq 0.8$. We generate the curve by sweeping over the hyperparameters and computing the cumulative mean. Higher is better. Numbers in **bold** (<u>underline</u>) indicate the best (second best) performer. (—) indicates that F never reached 20% of the optimum. WBFUSION has the highest AUC cum_mean, showing that its tradeoff is robust.

	MS-Marco		NF-Corpus		Hotpot-QA		Fever		SciFact	
Method	AUC_{MAP}	AUC_{NDCG}	AUC_{MAP}	AUC_{NDCG}	AUC_{MAP}	AUC_{NDCG}	AUC_{MAP}	AUC_{NDCG}	AUC_{MAP}	AUC_{NDCG}
MoE	<u>0.007</u>	0.029	<u>0.183</u>	<u>0.499</u>	—	—	0.039	0.044	—	—
Ensemble	**0.008**	<u>0.031</u>	0.108	0.338	—	—	**0.062**	**0.065**	<u>0.086</u>	<u>0.092</u>
NeuralRRF	**0.008**	<u>0.031</u>	0.164	0.496	—	—	<u>0.051</u>	<u>0.056</u>	0.083	0.090
Hybrid	<u>0.007</u>	0.029	0.133	0.425	—	—	0.041	0.047	0.082	0.088
WBFUSION	**0.008**	**0.033**	**0.194**	**0.515**	**0.059**	**0.067**	**0.062**	**0.065**	**0.091**	**0.096**

Table 21: Area under the cumulative mean curve (AUC cum_mean) between ranking metric (MAP and NDCG) and F under the constraint that F should reach at least 50% of the optimum ($F^\star$), i.e., $\frac{F^\star - F}{|F^\star|} \leq 0.5$. We generate the curve by sweeping over the hyperparameters and computing the cumulative mean. Higher is better. Numbers in **bold** (<u>underline</u>) indicate the best (second best) performer. (—) indicates that F never reached 50% of the optimum. WBFUSION has the highest AUC cum_mean, showing that its tradeoff is robust.

	MS-Marco		NF-Corpus		Hotpot-QA		Fever		SciFact	
Method	AUC_{MAP}	AUC_{NDCG}	AUC_{MAP}	AUC_{NDCG}	AUC_{MAP}	AUC_{NDCG}	AUC_{MAP}	AUC_{NDCG}	AUC_{MAP}	AUC_{NDCG}
MoE	—	—	<u>0.071</u>	0.192	—	—	—	—	—	—
Ensemble	—	—	0.054	0.150	—	—	—	—	—	—
NeuralRRF	—	—	0.069	<u>0.200</u>	—	—	—	—	—	—
Hybrid	—	—	0.059	0.175	—	—	—	—	—	—
WBFUSION	**0.003**	**0.013**	**0.077**	**0.205**	—	—	**0.025**	**0.026**	—	—

Table 22: Area under the cumulative mean curve (AUC cum_mean) between ranking metric (MAP and NDCG) and F under the constraint that F should reach at least 80% of the optimum ($F^\star$), i.e., $\frac{F^\star - F}{|F^\star|} \leq 0.2$. We generate the curve by sweeping over the hyperparameters and computing the cumulative mean. Higher is better. Numbers in **bold** (<u>underline</u>) indicate the best (second best) performer. (—) indicates that F never reached 80% of the optimum. WBFUSION has the highest AUC cum_mean and passes the threshold for most datasets, showing that its tradeoff is robust.

G.7.3 AREA UNDER THE PIECEWISE-LINEAR CURVE, AUC PWLIN

Given a hyperparameter sweep, we generate a curve between ranking metric (MAP/NDCG) and F, considering only the part of the curve beyond some fraction α of optimal ($F^\star$), i.e., $\frac{F^\star - F}{|F^\star|} \leq 1 - \alpha$ for some $0 < \alpha < 1$. The area under this restricted curve is denoted AUC pwlin for the chosen threshold of F. It is dubbed piecewise linear (pwlin) since we interpolate between observed points by straight lines. Tables 23, 24, 25 summarize the results for AUC pwlin under different thresholds of F. WBFUSION has the highest AUC pwlin, indicating a strong retrieval-rank fusion quality tradeoff, and remains active across the most stringent threshold for more datasets than any baseline.

	MS-Marco		NF-Corpus		Hotpot-QA		Fever		SciFact	
Method	AUC_{MAP}	AUC_{NDCG}	AUC_{MAP}	AUC_{NDCG}	AUC_{MAP}	AUC_{NDCG}	AUC_{MAP}	AUC_{NDCG}	AUC_{MAP}	AUC_{NDCG}
MoE	0.011	0.046	<u>0.296</u>	<u>0.806</u>	<u>0.086</u>	<u>0.100</u>	0.070	0.078	<u>0.136</u>	<u>0.145</u>
Ensemble	<u>0.012</u>	<u>0.050</u>	0.175	0.550	<u>0.086</u>	0.099	<u>0.084</u>	<u>0.091</u>	0.131	0.141
NeuralRRF	<u>0.012</u>	0.049	0.262	0.795	0.080	0.095	0.081	0.089	0.134	0.144
Hybrid	0.011	0.044	0.212	0.687	0.073	0.087	0.060	0.068	0.131	0.141
WBFUSION	**0.013**	**0.053**	**0.310**	**0.824**	**0.095**	**0.107**	**0.099**	**0.104**	**0.145**	**0.154**

Table 23: Area under the piecewise linear curve (AUC pwlin) between ranking metric (MAP and NDCG) and F under the constraint that F should reach at least 20% of the optimum ($F^\star$), i.e., $\frac{F^\star - F}{|F^\star|} \leq 0.8$. We generate the curve by sweeping over the hyperparameters and joining points by straight lines. Higher is better. Numbers in **bold** (<u>underline</u>) indicate the best (second best) performer. (—) indicates that F never reached 20% of the optimum. WBFUSION has the largest AUC values, showing that it has the strongest tradeoff between F and retrieval quality.

	MS-Marco		NF-Corpus		Hotpot-QA		Fever		SciFact	
Method	AUC_{MAP}	AUC_{NDCG}	AUC_{MAP}	AUC_{NDCG}	AUC_{MAP}	AUC_{NDCG}	AUC_{MAP}	AUC_{NDCG}	AUC_{MAP}	AUC_{NDCG}
MoE	**0.008**	0.030	<u>0.183</u>	<u>0.499</u>	—	—	0.048	0.053	—	—
Ensemble	**0.008**	<u>0.032</u>	0.120	0.361	—	—	<u>0.060</u>	<u>0.063</u>	<u>0.086</u>	<u>0.093</u>
NeuralRRF	**0.008**	0.031	0.166	<u>0.499</u>	—	—	0.051	0.056	0.084	0.091
Hybrid	<u>0.007</u>	0.029	0.138	0.437	—	—	0.042	0.047	0.082	0.088
WBFUSION	**0.008**	**0.033**	**0.194**	**0.515**	**0.059**	**0.067**	**0.061**	**0.065**	**0.091**	**0.096**

Table 24: Area under the piecewise linear curve (AUC pwlin) between ranking metric (MAP and NDCG) and F under the constraint that F should reach at least 50% of the optimum ($F^\star$), i.e., $\frac{F^\star - F}{|F^\star|} \leq 0.5$. We generate the curve by sweeping over the hyperparameters and joining points by straight lines. Higher is better. Numbers in **bold** (<u>underline</u>) indicate the best (second best) performer. (—) indicates that F never reached 50% of the optimum. WBFUSION has the largest AUC values, showing that it has the strongest tradeoff between F and retrieval quality.

	MS-Marco		NF-Corpus		Hotpot-QA		Fever		SciFact	
Method	AUC_{MAP}	AUC_{NDCG}	AUC_{MAP}	AUC_{NDCG}	AUC_{MAP}	AUC_{NDCG}	AUC_{MAP}	AUC_{NDCG}	AUC_{MAP}	AUC_{NDCG}
MoE	—	—	<u>0.071</u>	0.193	—	—	—	—	—	—
Ensemble	—	—	0.064	0.173	—	—	—	—	—	—
NeuralRRF	—	—	<u>0.071</u>	<u>0.203</u>	—	—	—	—	—	—
Hybrid	—	—	0.064	0.186	—	—	—	—	—	—
WBFUSION	**0.003**	**0.013**	**0.077**	**0.206**	—	—	**0.025**	**0.026**	—	—

Table 25: Area under the piecewise linear curve (AUC pwlin) between ranking metric (MAP and NDCG) and F under the constraint that F should reach at least 80% of the optimum ($F^\star$), i.e., $\frac{F^\star - F}{|F^\star|} \leq 0.2$. We generate the curve by sweeping over the hyperparameters and joining points by straight lines. Higher is better. Numbers in **bold** (<u>underline</u>) indicate the best (second best) performer. (—) indicates that F never reached 80% of the optimum. WBFUSION has the largest AUC values, showing that it has the strongest tradeoff between F and retreival quality. This is supported by it being the only rank fusion method to survive the strict threshold over three of five datasets

G.8 Efficient ANN retrieval and performance trade-offs

The ranking accuracy and trade-off results shown in the previous sections involve composing a greedy top-k retrieval, followed by a re-ranking on the retrieved set. But as discussed before, the cost of retrieving top-k items is $O(k^2N^2)$ which comes from repeated calls to $O(N)$ many Sinkhorn solvers for step-t to compute the marginal gains, each costing $O(tN)$. In this experiment, we evaluate an efficient subset selection procedure based on Algorithm 4.

Model and training details: Based on the recipe from Theorem 6, we train neural embedding models $\boldsymbol{\zeta}_\theta(\boldsymbol{q}, S)$ and $\boldsymbol{\sigma}(\boldsymbol{x}_s)$, whose dot-product is designed to approximate the marginal gain $F(c \mid S, \boldsymbol{q})$, especially the order of the marginal gain. The model architecture of $\boldsymbol{\zeta}_\theta$ and $\boldsymbol{\sigma}_\theta$ follows the Two-Tower architecture described in Appendix E.1. During training we minimize a linear combination of two losses: **(1)** A listwise loss $\mathcal{L}_{\text{list}}$ that is the tempered cross entropy between the target greedy gains $F(c \mid S, \boldsymbol{q})$ and the softmax of model scores $\boldsymbol{\zeta}_\theta(\boldsymbol{q}, S)^\top \boldsymbol{\sigma}(\boldsymbol{x}_c)$; **(2)** A pairwise triplet loss with margin $\mathcal{L}_{\text{qrel}}$, based on ground truth relevance labels for every $(\boldsymbol{q}, c)$ pair.

Inference: During inference, we first index the corpus items using the already trained embeddings $\boldsymbol{\sigma}_\theta$ to get an index $\mathcal{I}$ using Faiss. Then we run Algorithm 4 to probe $\mathcal{I}$ and select top K from the pool of N corpus items and rerank them.

Experiment: We measure efficiency of WBFusion using relative wall-clock time w.r.t. Exact greedy. To compare the competitiveness vs efficiency tradeoff of our method against Greedy and its variants, we threshold both at $x\%$ of exact greedy's values, $\frac{F^\star - F}{|F^\star|} \leq \frac{100-x}{100}$ and $\text{NDCG} \geq x \cdot \text{NDCG}^\star$, for $x \in \{75, 80, 85, 90, 95\}$, to obtain speedups $\mathcal{E}_{xF}$ and $\mathcal{E}_{x\text{NDCG}}$ respectively.

Observation: We benchmark against popular greedy baselines and we observe in Tables 26, 27, 28, 29, and 30 that WBFusion is most efficient across all thresholds. Lazy greedy is consistently the second best while stochastic greedy almost always fails to reach the required minimum F or NDCG.

	MS-Marco		NF-Corpus		Hotpot-QA		Fever		SciFact	
Method	$\mathcal{E}_{0.75F}$	$\mathcal{E}_{0.75\text{NDCG}}$	$\mathcal{E}_{0.75F}$	$\mathcal{E}_{0.75\text{NDCG}}$	$\mathcal{E}_{0.75F}$	$\mathcal{E}_{0.75\text{NDCG}}$	$\mathcal{E}_{0.75F}$	$\mathcal{E}_{0.75\text{NDCG}}$	$\mathcal{E}_{0.75F}$	$\mathcal{E}_{0.75\text{NDCG}}$
Exact Greedy	1.00×	1.00×	1.00×	1.00×	1.00×	1.00×	1.00×	1.00×	1.00×	1.00×
Lazy Greedy	2.10×	2.10×	6.24×	6.24×	4.05×	4.05×	3.66×	3.66×	4.18×	4.18×
Lazier than Lazy	2.04×	2.07×	7.18×	5.93×	3.72×	3.72×	3.68×	3.43×	3.94×	3.52×
Stochastic Greedy	—	2.15×	9.12×	—	—	—	—	—	—	—
WBFusion	**2.95×**	**4.62×**	**29.39×**	**19.00×**	**9.69×**	**24.28×**	**6.80×**	**22.66×**	**21.22×**	**22.45×**

Table 26: Efficiency ($\mathcal{E}_{0.75F}$ and $\mathcal{E}_{0.75\text{NDCG}}$)— measured as the speedup over exact greedy (Algorithm 1) in reaching 75% of the F and NDCG values attained by exact greedy— of WBFusion, lazy greedy, lazier than lazy greedy, and stochastic greedy. Higher is better. **Bold** and underlined values indicate the best and second-best results, respectively. WBFusion achieves the highest efficiency.

	MS-Marco		NF-Corpus		Hotpot-QA		Fever		SciFact	
Method	$\mathcal{E}_{0.8F}$	$\mathcal{E}_{0.8\text{NDCG}}$	$\mathcal{E}_{0.8F}$	$\mathcal{E}_{0.8\text{NDCG}}$	$\mathcal{E}_{0.8F}$	$\mathcal{E}_{0.8\text{NDCG}}$	$\mathcal{E}_{0.8F}$	$\mathcal{E}_{0.8\text{NDCG}}$	$\mathcal{E}_{0.8F}$	$\mathcal{E}_{0.8\text{NDCG}}$
Exact Greedy	1.00×	1.00×	1.00×	1.00×	1.00×	1.00×	1.00×	1.00×	1.00×	1.00×
Lazy Greedy	2.10×	2.10×	6.24×	6.24×	4.05×	4.05×	3.66×	3.66×	4.18×	4.18×
Lazier than Lazy	2.04×	2.07×	7.18×	5.93×	3.72×	3.72×	3.68×	3.43×	3.94×	3.52×
Stochastic Greedy	—	2.14×	9.12×	—	—	—	—	—	—	—
WBFusion	**2.83×**	**3.79×**	**29.39×**	**16.10×**	**7.22×**	**16.98×**	**5.06×**	**22.50×**	**14.76×**	**22.45×**

Table 27: Efficiency ($\mathcal{E}_{0.8F}$ and $\mathcal{E}_{0.8\text{NDCG}}$)— measured as the speedup over exact greedy (Algorithm 1) in reaching 80% of the F and NDCG values attained by exact greedy— of WBFusion, lazy greedy, lazier than lazy greedy, and stochastic greedy. Higher is better. **Bold** and underlined values indicate the best and second-best results, respectively. WBFusion achieves the highest efficiency.

Method	MS-Marco		NF-Corpus		Hotpot-QA		Fever		SciFact	
	$\mathcal{E}_{0.85F}$	$\mathcal{E}_{0.85\text{NDCG}}$	$\mathcal{E}_{0.85F}$	$\mathcal{E}_{0.85\text{NDCG}}$	$\mathcal{E}_{0.85F}$	$\mathcal{E}_{0.85\text{NDCG}}$	$\mathcal{E}_{0.85F}$	$\mathcal{E}_{0.85\text{NDCG}}$	$\mathcal{E}_{0.85F}$	$\mathcal{E}_{0.85\text{NDCG}}$
Exact Greedy	1.00×	1.00×	1.00×	1.00×	1.00×	1.00×	1.00×	1.00×	1.00×	1.00×
Lazy Greedy	**2.10**×	2.10×	6.24×	6.24×	4.05×	4.05×	3.66×	3.66×	4.18×	4.18×
Lazier than Lazy	2.03×	2.07×	7.18×	4.47×	3.72×	3.72×	3.60×	3.43×	3.52×	3.52×
Stochastic Greedy	—	—	9.12×	—	—	—	—	—	—	—
WBFUSION	2.03×	**3.79**×	**29.39**×	**11.79**×	**7.10**×	**16.98**×	**4.95**×	**15.83**×	**9.14**×	**22.45**×

Table 28: Efficiency ($\mathcal{E}_{0.85F}$ and $\mathcal{E}_{0.85\text{NDCG}}$)— measured as the speedup over exact greedy (Algorithm 1) in reaching 85% of the F and NDCG values attained by exact greedy— of WBFUSION, lazy greedy, lazier than lazy greedy, and stochastic greedy. Higher is better. **Bold** and underlined values indicate the best and second-best results, respectively. WBFUSION achieves the highest efficiency.

Method	MS-Marco		NF-Corpus		Hotpot-QA		Fever		SciFact	
	$\mathcal{E}_{0.9F}$	$\mathcal{E}_{0.9\text{NDCG}}$	$\mathcal{E}_{0.9F}$	$\mathcal{E}_{0.9\text{NDCG}}$	$\mathcal{E}_{0.9F}$	$\mathcal{E}_{0.9\text{NDCG}}$	$\mathcal{E}_{0.9F}$	$\mathcal{E}_{0.9\text{NDCG}}$	$\mathcal{E}_{0.9F}$	$\mathcal{E}_{0.9\text{NDCG}}$
Exact Greedy	1.00×	1.00×	1.00×	1.00×	1.00×	1.00×	1.00×	1.00×	1.00×	1.00×
Lazy Greedy	—	2.10×	6.24×	6.24×	—	—	—	3.66×	—	4.18×
Lazier than Lazy	2.03×	2.04×	7.18×	4.47×	3.72×	3.72×	3.43×	3.43×	3.52×	3.52×
Stochastic Greedy	—	—	9.12×	—	—	—	—	—	—	—
WBFUSION	**2.03**×	**2.95**×	**29.39**×	**10.01**×	**5.25**×	**13.81**×	**3.54**×	**15.72**×	**7.35**×	**21.22**×

Table 29: Efficiency ($\mathcal{E}_{0.9F}$ and $\mathcal{E}_{0.9\text{NDCG}}$)— measured as the speedup over exact greedy (Algorithm 1) in reaching 90% of the F and NDCG values attained by exact greedy— of WBFUSION, lazy greedy, lazier than lazy greedy, and stochastic greedy. Higher is better. **Bold** and underlined values indicate the best and second-best results, respectively. WBFUSION achieves the highest efficiency.

Method	MS-Marco		NF-Corpus		Hotpot-QA		Fever		SciFact	
	$\mathcal{E}_{0.95F}$	$\mathcal{E}_{0.95\text{NDCG}}$	$\mathcal{E}_{0.95F}$	$\mathcal{E}_{0.95\text{NDCG}}$	$\mathcal{E}_{0.95F}$	$\mathcal{E}_{0.95\text{NDCG}}$	$\mathcal{E}_{0.95F}$	$\mathcal{E}_{0.95\text{NDCG}}$	$\mathcal{E}_{0.95F}$	$\mathcal{E}_{0.95\text{NDCG}}$
Exact Greedy	1.00×	1.00×	1.00×	1.00×	1.00×	1.00×	1.00×	1.00×	1.00×	1.00×
Lazy Greedy	—	—	6.24×	6.24×	—	—	—	3.66×	—	4.18×
Lazier than Lazy	0.97×	2.03×	7.18×	4.47×	2.62×	3.29×	3.15×	3.15×	3.09×	3.09×
Stochastic Greedy	—	—	8.87×	—	—	—	—	—	—	—
WBFUSION	**2.03**×	**2.03**×	**29.39**×	**7.82**×	**3.63**×	**7.22**×	**3.47**×	**12.68**×	**4.62**×	**14.76**×

Table 30: Efficiency ($\mathcal{E}_{0.95F}$ and $\mathcal{E}_{0.95\text{NDCG}}$)— measured as the speedup over exact greedy (Algorithm 1) in reaching 95% of the F and NDCG values attained by exact greedy— of WBFUSION, lazy greedy, lazier than lazy greedy, and stochastic greedy. Higher is better. **Bold** and underlined values indicate the best and second-best results, respectively. WBFUSION achieves the highest efficiency.

G.9 Scaling of Efficient ANN retrieval with corpus size

In this section, we provide additional measurements for how different greedy variants, *viz.*, Lazy greedy, Lazier than Lazy greedy, Stochastic greedy, and WBFusion (Algorithm 4) scale with the corpus size N. A configuration is counted only if it reaches $F \geq 85\%$ of exact greedy's F or NDCG $\geq 50\%$ of exact greedy's NDCG. To perform this scaling study, we sampled 100 queries from the test set (or all if the test set was smaller) and performed RRF truncation to get to N documents. Table 31 summarizes the results for MS-Marco, NF-Corpus, Fever and SciFact in terms of speedup relative to exact greedy (Algorithm 1). We observe that **(1)** across all four datasets WBFusion almost always has the greatest efficiency gains, **(2)** the gains for WBFusion also increase faster than any baseline as N increases. **(3)** Stochastic Greedy is not able to achieve the minimum required F and NDCG values to be considered, while Lazier than Lazy greedy scales extremely slowly. **(4)** Lazy-greedy scales well, and its speedup is comparable to ours in the case of MS-Marco and Fever for small N, but WBFusion quickly outpaces it as corpus size increases.

		Lazy		Lazier		Stochastic		WBFusion	
Dataset	N	$\mathcal{E}_{0.85F}$	$\mathcal{E}_{0.5\text{NDCG}}$	$\mathcal{E}_{0.85F}$	$\mathcal{E}_{0.5\text{NDCG}}$	$\mathcal{E}_{0.85F}$	$\mathcal{E}_{0.5\text{NDCG}}$	$\mathcal{E}_{0.85F}$	$\mathcal{E}_{0.5\text{NDCG}}$
MS-Marco	1000	**2.14**×	2.14×	2.04×	2.08×	—	2.08×	2.04×	**10.30**×
	2000	**10.02**×	10.02×	6.44×	9.39×	—	9.47×	9.94×	**40.31**×
	4000	35.95×	35.95×	7.52×	21.49×	—	34.29×	**38.48**×	**132.66**×
	8000	92.35×	92.35×	6.45×	22.60×	—	—	**109.80**×	**279.51**×
	16000	156.64×	156.64×	6.35×	20.04×	—	—	**243.66**×	**424.22**×
NF-Corpus	1000	7.43×	7.43×	7.27×	7.27×	9.80×	8.24×	**29.54**×	**29.54**×
	2000	23.68×	23.68×	15.89×	15.89×	31.19×	—	**88.50**×	**87.30**×
	3000	41.10×	41.10×	18.73×	18.73×	51.38×	—	**150.98**×	**146.83**×
Fever	1000	2.99×	2.99×	2.81×	2.81×	—	—	**3.92**×	**13.58**×
	2000	13.28×	13.28×	7.22×	3.78×	—	—	**17.31**×	**56.31**×
	4000	**71.31**×	71.31×	6.12×	3.25×	—	—	60.87×	**258.64**×
	8000	133.08×	133.08×	18.79×	6.38×	—	—	**179.47**×	**480.96**×
	16000	197.18×	197.18×	19.66×	3.22×	—	—	**338.19**×	**664.74**×
SciFact	1000	4.31×	4.31×	3.42×	3.42×	—	—	**9.48**×	**21.21**×
	2000	13.23×	13.23×	5.00×	10.29×	—	—	**25.04**×	**59.94**×
	4000	41.04×	41.04×	5.87×	5.87×	—	—	**50.96**×	**173.83**×

Table 31: Efficiency ($\mathcal{E}_{0.85F}$ and $\mathcal{E}_{0.5\text{NDCG}}$)— measured as the speedup over exact greedy (Algorithm 1) in reaching 85% of the F and 50% of the NDCG values attained by exact greedy— of WBFusion, lazy greedy, lazier than lazy greedy, and stochastic greedy, as the corpus size N increases. Higher is better, (—) indicates the method did not achieve the required F or NDCG values. **Bold** and underlined values indicate the best and second-best results, respectively. WBFusion achieves the best scaling across datasets as corpus size N increases.