LOW-BUDGET ACTIVE LEARNING THROUGH ENTROPIC OPTIMAL TRANSPORT

Anonymous authors
Paper under double-blind review

ABSTRACT

We consider low-budget active learning, which consists of selecting a limited number of points, the coreset, such that a model can be trained to high accuracy on the selection only. This problem is particularly relevant in contexts where labeling requires costly expert intervention, as in medical applications. We leverage features extracted from a pretrained self-supervised model to represent the data, and perform coreset selection directly in this feature space. In this paper, we use entropic optimal transport, specifically the Sinkhorn divergence, as the coreset selection criterion, which first allows us to get dimension-free sample complexity results, and second admits computationally efficient gradient evaluations. This opens the way to using gradient-based algorithms to rapidly compute solution candidates, further improved by a swap-based local search, with guarantees on the solution quality. Experiments on image benchmarks and medical datasets show that our method outperforms state-of-the-art heuristics in low-budget settings.

1 INTRODUCTION

Active learning is a machine learning paradigm designed to reduce the amount of labeled data required to train a model by selecting the most informative samples for annotation. Instead of labeling a large dataset uniformly, the algorithm selects a subset of unlabeled data points whose labels are expected to provide the greatest improvement to the model (Settles, 2009). In some key applications, labeling is expensive, and by strategically querying labels for a small number of examples, active learning can achieve performance comparable to that of training on a much larger labeled dataset (Wang et al., 2024). In the low-budget setting, the problem relies critically on defining a small set of informative samples. Self-supervised learning models have provided increasingly powerful representations in recent years, in which train to map data points to a feature space where Euclidean distances can be used to evaluate their proximity (Gui et al., 2024). This improvement helped low-budget active learning reach higher levels of accuracies.

Active learning strategies broadly fall into two categories, *uncertainty-based* methods (Settles, 2009) which query points where the current model is least confident but require an already-trained model and typically a high labeling budget, and *representation-based* methods which select a *coreset*, a subset of the data that best summarizes the whole dataset, independently of the current learning task. Classical approaches include k-medoids (Kaufman & Rousseeuw, 1990) and k-centers (Sener & Savarese, 2018), and more recently, TypiClust (Hacohen et al., 2022), ProbCover (Yehuda et al., 2022), and MaxHerding (Bae et al., 2024). The challenge of representation-based active learning thus lies in defining a criterion that captures the most representative data points for the learning task.

Optimal Transport (OT) provides a natural framework for defining a criterion that measures the discrepancy between the coreset and the whole dataset for the downstream learning task. Given two empirical distributions supported on the dataset, OT defines the minimum cost required to transport mass from one distribution to the other. This cost, known as the Wasserstein distance, can be expressed as a linear program over transport plans (i.e., probability distributions in the product space subject to marginal constraints), and was exploited in particular in Mahmood et al. (2022) in the context of active learning.

One main reason for the success of OT in learning is the use of the entropic regularization of OT, pioneered by Cuturi (2013) and presented across a range of applications in Peyré et al. (2019). Entropic OT first offers an efficient and scalable method, through the so-called Sinkhorn algorithm,

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.

to approximate Wasserstein distances. Furthermore, entropic OT has been investigated in its own right, notably offering better statistical properties than the Wasserstein distance (Genevay et al., 2019; Rigollet & Stromme, 2025). Related work on the use of entropic OT for coreset selection include Kokot & Luedtke (2025); Yin et al. (2025). First, Kokot & Luedtke (2025) consider a similar coreset selection problem for a compression task, which differs from the low-budget active learning setting. Second, Yin et al. (2025) propose to use the Sinkhorn loss for coreset selection across both low- and high-budget regimes, without addressing the combinatorial hardness of the discrete selection problem in the low-budget setting. Our work is then complementary to theirs by investigating the symmetric Sinkhorn divergence and developing algorithmic tools tailored for low budgets.

Contributions. We provide an entropic OT framework for coreset selection in low-budget active learning, with the following central contributions:

- **Sinkhorn divergence criterion.** We propose to measure the quality of the selection by the Sinkhorn divergence. We illustrate the fact that the bias of the usual Sinkhorn *cost* of entropic OT may lead to degenerate selections, and that the Sinkhorn *divergence* (Genevay et al. (2018); Feydy et al. (2019)), by correcting this bias, recovers the natural selections. We then establish a dimension-free generalization bound for general models trained on the resulting coreset.
- **Algorithm and guarantees.** We propose an algorithm (**FW-Swap**) to approximately solve the selection problem, by leveraging the computationally cheap gradients of the Sinkhorn divergence. More specifically, **FW-Swap** warm-starts a first-order swap-based local search, with candidates from the active set produced by a Frank-Wolfe algorithm. We establish a guarantee on a worst-case suboptimality bound for this local algorithm.
- **Experiments.** We evaluate the empirical performance of our approach on vision and EEG benchmarks, against previously proposed methods and show competitive or superior accuracy across all low-budget settings.

2 Active learning from the Sinkhorn divergence

We detail in this section our approach using the Sinkhorn divergence for coreset selection. We start by introducing, in Section 2.1, the notions of OT used in our developments and, in Section 2.2, our active set framework. We explain our methodological approach by discussing, in Section 2.3, a revealing example and by presenting, in Section 2.4, our generalization results.

2.1 Preliminaries on optimal transport

We quickly recall some essential notions of OT; we refer to Peyré et al. (2019) for a comprehensive introduction to computational optimal transport and to Chewi et al. (2025) for its use in statistics.

The optimal transport distance (or Wasserstein distance) between two measures μ and ν quantifies the minimum transport cost to displace the mass from μ to ν given a ground cost c. Mathematically,

$$W(\mu, \nu) = \min_{\boldsymbol{\pi} \in \Pi(\mu,\nu)} \int c(\boldsymbol{y}, \boldsymbol{z})\, d\boldsymbol{\pi}$$

where $\Pi(\mu, \nu)$ is the set of "transport plans" (or joint distributions on the product space) such that the first marginal is μ and the second marginal is ν. The entropic OT cost is then introduced as the regularization of the Wasserstein distance by the Kullback-Leibler divergence from the product measure $\pi_0 = \mu \otimes \nu$: for a regularization parameter $\varepsilon > 0$,

$$S_\varepsilon(\mu, \nu) = \min_{\boldsymbol{\pi} \in \Pi(\mu,\nu)} \int c(\boldsymbol{y}, \boldsymbol{z})\, d\boldsymbol{\pi} + \varepsilon \mathrm{KL}(\boldsymbol{\pi}, \mu \otimes \nu).$$

Though computationally more tractable thanks to the so-called Sinkhorn algorithm, the entropic OT cost $S_\varepsilon(\mu, \nu)$ is not a distance nor even a divergence, unlike $W(\mu, \nu)$ (assuming that c defines a metric). In particular, it introduces a significant bias, since one notably has $S_\varepsilon(\mu, \mu) > 0$. To correct this bias, the Sinkhorn divergence was introduced in Genevay et al. (2018) (and further studied in

Feydy et al. (2019); Séjourné et al. (2019); Lavenant et al. (2025)):

$$D_\varepsilon(\mu,\nu) = S_\varepsilon(\mu,\nu) - \frac{1}{2}S_\varepsilon(\mu,\mu) - \frac{1}{2}S_\varepsilon(\nu,\nu). \tag{1}$$

We finish these recalls by the case of empirical measures over n points, identified with elements of Δ_n the probability simplex of dimension $n-1$. In the sequel, we use the shorthand $[n] = \{1 \dots n\}$. For $\boldsymbol{a} \in \Delta_n$, $\boldsymbol{b} \in \Delta_m$, and the cost matrix of entries $C_{ij} = c(\boldsymbol{a}_i, \boldsymbol{b}_j)$, computing $W(\boldsymbol{a}, \boldsymbol{b})$ amounts to solving a linear optimization problem, which can be expressed in primal form over the transport plan $\boldsymbol{\pi}$, and in dual form over the dual variables $\mathbf{f} \in \mathbb{R}^n$ and $\mathbf{g} \in \mathbb{R}^m$ (called potentials):

$$\begin{aligned} W(\boldsymbol{a}, \boldsymbol{b}) = \max_{\mathbf{f}\in\mathbb{R}^n, \mathbf{g}\in\mathbb{R}^m} \ & \langle \mathbf{f}, \boldsymbol{a}\rangle + \langle \mathbf{g}, \boldsymbol{b}\rangle \\ \text{s.t. } & \mathbf{f}_i + \mathbf{g}_j \leq C_{ij} \quad \forall i \in [n], j \in [m] \end{aligned}$$

Similarly, for the entropic OT cost, we have

$$S_\varepsilon(\boldsymbol{a}, \boldsymbol{b}) = \max_{\mathbf{f}\in\mathbb{R}^n, \mathbf{g}\in\mathbb{R}^m} \langle \mathbf{f}, \boldsymbol{a}\rangle + \langle \mathbf{g}, \boldsymbol{b}\rangle - \varepsilon \sum_{(i,j)\in[n]\times[m]} \boldsymbol{a}_i \boldsymbol{b}_j \left(\exp\left(\frac{1}{\varepsilon}(\mathbf{f}_i + \mathbf{g}_j - C_{ij})\right) - 1 \right).$$

We will use that S_ε is a differentiable function and that its gradient is given by $\nabla S_\varepsilon(\boldsymbol{a}, \boldsymbol{b}) = (\mathbf{f}^*, \mathbf{g}^*)$, for the optimal solutions $\mathbf{f}^*$, $\mathbf{g}^*$ of the above problem (under the normalization $\sum_i \mathbf{f}_i^* = 0$). Finally, considering also $\mathbf{p}^*$ as the optimal potential for the symmetric problem $S_\varepsilon(\boldsymbol{a}, \boldsymbol{a})$, the gradient of the Sinkhorn divergence with respect to its first marginal is given by

$$\nabla_{\boldsymbol{a}} D_\varepsilon(\boldsymbol{a}, \boldsymbol{b}) = \mathbf{f}^* - \mathbf{p}^*. \tag{2}$$

This gradient expression (obtained as a by-product of the two runs of the Sinkhorn algorithm (Feydy et al., 2019)) will be a key ingredient in the algorithm we propose in Section 3.1.

2.2 Entropic optimal transport framework for active learning

Consider a dataset of n points $\mathcal{D} = \{\boldsymbol{y}_i\}_{i\in[n]}$ and the empirical distribution over this dataset:

$$\hat{\mu}^{\mathcal{D}} = \sum_{i\in[n]} \frac{1}{n}\delta_{\boldsymbol{y}_i} \qquad \text{(where } \delta_{\boldsymbol{y}} \text{ is the Dirac measure at the point } \boldsymbol{y})$$

For $\mathbf{x} \in \{0,1\}^n$, we introduce $\hat{\mu}_{\mathbf{x}}$ as the empirical distribution over points whose incidence vector is $\mathbf{x}$. Given a budget $B \leq n$, we write $\mathcal{V}_{n,B}$ the set of $\{0,1\}$-vectors of dimension n with B nonzero entries, and $\Delta_{n,B}$ for the hypersimplex in dimension n of parameter B:

$$\mathcal{V}_{n,B} = \left\{ \mathbf{x} \in \{0,1\}^n \ : \ \sum_{i\in[n]} \mathbf{x}_i = B \right\} \quad \text{and} \quad \Delta_{n,B} = \left\{ \mathbf{x} \in [0,1]^n \ : \ \sum_{i\in[n]} \mathbf{x}_i = B \right\}.$$

The method we develop in this paper uses the Sinkhorn divergence D_ε for coreset selection. More specifically, we propose to select B points within the dataset $\mathcal{D}$, as the best selection vector $\mathbf{x} \in \mathcal{V}_{n,B}$ such that $\hat{\mu}_{\mathbf{x}}$ minimizes the Sinkhorn divergence from $\hat{\mu}_{\mathbf{x}}$ to the empirical measure $\hat{\mu}^{\mathcal{D}}$. The coreset selection problem thus writes:

$$\min_{\mathbf{x}\in\mathcal{V}_{n,B}} D_\varepsilon(\hat{\mu}_{\mathbf{x}}, \hat{\mu}^{\mathcal{D}}). \tag{3}$$

Note that this formulation relates to existing approaches from the literature for the coreset selection problem. First, replacing D_ε by W with an Euclidean ground cost c, we recover the approach of Mahmood et al. (2022). Moreover, this OT-based formulation also captures two seemingly unrelated previous methods: the problems underlying ProbCover (Yehuda et al., 2022) and k-medoid (Kaufman & Rousseeuw, 1990) can indeed be recast as a similar problem with a suitable ground cost. This is formalized in Proposition 2 in Section A.

The active learning problem can thus be formalized as the combinatorial problem (3), which aims at minimizing, over the discrete set $\mathcal{V}_{n,B}$, the function D_ε, which is differentiable and convex in its first argument. In the next subsections, we explain why D_ε is preferable over S_ε and what generalization properties we can expect. Section 3 will then present our solution method. Before moving on, we highlight in the next remark that the combinatorial structure of the problem is essential, since its convex relaxation is trivial.

Remark 1. *In general, one way to approximate a combinatorial problem is to solve its convex relaxation: in our situation, this would mean replacing $\mathcal{V}_{n,B}$ with its convex hull, the hypersimplex $\Delta_{n,B}$,*

$$\min_{\mathbf{x}\in\Delta_{n,B}} D_\varepsilon(\hat{\mu}_{\mathbf{x}}, \hat{\mu}^{\mathcal{D}}). \tag{4}$$

However, we easily see that this relaxation is uninformative: solving this relaxation yields both a zero optimal value and a constant optimal solution, regardless of the instance. This is formalized, for any divergence, in Proposition 3 in Section A. Still, in Section 3.1, we will use continuous optimization methods, but in an original way, to get relevant information on the combinatorial problem.

2.3 Sinkhorn bias on active learning instances

In this section, we highlight why it is essential to use the Sinkhorn divergence D_ε for our active learning problem, rather than simply the entropic OT cost S_ε. Indeed, the Sinkhorn algorithm is known to introduce a bias (Feydy et al., 2019; Séjourné et al., 2019), corrected by the symmetrization of D_ε. Here we characterize the effect of the Sinkhorn bias on a simple parameterized family of instances (with two clusters and a single central outlier), showing that using S_ε can be problematic for the coreset selection task and active learning in particular.

Let us consider the simple family of instances called $\texttt{CCO}(\Delta, m)$ defined by two clusters of $m \geq 2$ points each, symmetric around the origin and at distance Δ from the origin, and a single point, the outlier, at the origin, see Figure 1 for the illustration. The empirical measure is

$$\hat{\mu}^{\mathcal{D}} = \frac{1}{2m+1}\delta_0 + \frac{m}{2m+1}\delta_{-\Delta} + \frac{m}{2m+1}\delta_{\Delta}.$$

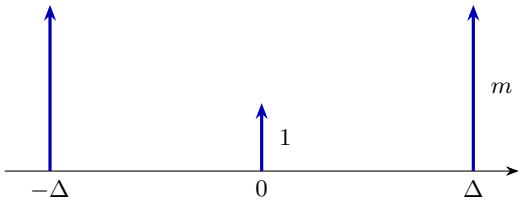

Figure 1: $\hat{\mu}^{\mathcal{D}}$ for the $\texttt{CCO}(\Delta, m)$ instances. For $B \leq 2$, we would like to select one point in $-\Delta$ and another in Δ (the *extreme* configuration).

There are four distinct configurations up to symmetries: the *central* configuration with measure α_C corresponds to the mass concentrated at the origin, the *left* configuration with measure α_L corresponds to the mass evenly split between the origin and the left cluster, the *extreme* configuration with measure α_E corresponds to the mass evenly split between the two clusters, the *one-sided* configuration with all of the mass concentrated on the left cluster. The last one is always trivially dominated and we discard it. With $m \gg 1$, this instance corresponds to a dataset with two clusters and one central outlier separated from both clusters. Proposition 1 shows that using the Sinkhorn cost with a large regularization ε misses the right configuration.

Proposition 1 (Entropic bias on CCO instances)**.** *Take $B \leq 2$ and $\|\boldsymbol{y}_1 - \boldsymbol{y}_2\|_2^2$ as the ground cost. For any value $\Delta > 0, m \geq 2$, the $\texttt{CCO}(\Delta, m)$ instance has the following properties:*

- *For any $\varepsilon > 0$, selecting the coreset with D_ε favors the extreme configuration α_E, which is the desirable outcome, and identical to selecting the coreset with Wasserstein.*
- *However there exists a finite $\bar{\varepsilon} > 0$ such that for any $\varepsilon > \bar{\varepsilon}$, selecting the coreset based on S_ε leads to favoring the central configuration α_C.*

In Section D, we present the quantitative version of this proposition and its proof, precisely comparing the three configurations above. Although intuitive, this example illustrates and quantifies the effect of the Sinkhorn bias in our active learning setting.

2.4 Generalization bound for the Sinkhorn divergence coreset

In this section, we establish a generalization result for our approach, based on the elegant dimension-independent complexity result of Rigollet & Stromme (2025) (in the case of the squared Euclidean norm as the ground cost). We provide here the framework and the main result, and we postpone to Section C the full proof.

We assume that the data is i.i.d samples from a "true" data distribution μ, i.e. $\mathcal{D} = \{\boldsymbol{y}_i\}_{i\in[n]} \sim \mu$. For a parametric family of loss functions $f_{\boldsymbol{\theta}}$ of parameter $\boldsymbol{\theta}$, we define the risk of a parameter $\boldsymbol{\theta}$

under a distribution ν as

$$R_\nu(\boldsymbol{\theta}) = \mathbb{E}_{\boldsymbol{y}\sim\nu}[f_{\boldsymbol{\theta}}(\boldsymbol{y})]$$

We then consider three risks: the true risk $R(\boldsymbol{\theta})$, the empirical risk $\hat{R}_n(\boldsymbol{\theta})$, and the coreset empirical risk $\hat{R}_{S^*}(\boldsymbol{\theta})$ (when $\nu = \hat{\mu}_{S^*}$ with the subset $S^* \subset [n]$ supported on B). As typical in active learning (see e.g., Sener & Savarese (2018)), we decompose the risk error of our method into a generalization term and a coreset loss term:

$$R(\boldsymbol{\theta}) - \mathbb{E}[\hat{R}_{S^*}(\boldsymbol{\theta})] = \mathbb{E}\left[R(\boldsymbol{\theta}) - \hat{R}_n(\boldsymbol{\theta})\right] + \mathbb{E}\left[\hat{R}_n(\boldsymbol{\theta}) - \hat{R}_{S^*}(\boldsymbol{\theta})\right]. \tag{5}$$

The next result shows that we can bound each term, independently of the dimension, when the loss functions are Lipschitz and the ground cost is the squared Euclidean norm.

Theorem 1 (Generalization bound for the Sinkhorn divergence coreset)**.** *Consider the Sinkhorn divergence D_ε with ground cost $c(\boldsymbol{y}, \boldsymbol{z}) = \|\boldsymbol{y} - \boldsymbol{z}\|_2^2$. Suppose that the loss function $f_{\boldsymbol{\theta}}$ is L-Lipschitz for any $\boldsymbol{\theta}$. Then the coreset loss is bounded uniformly in $\boldsymbol{\theta}$ by*

$$\hat{R}_n(\boldsymbol{\theta}) - \hat{R}_{S^*}(\boldsymbol{\theta}) \leq L\sqrt{D_\varepsilon(\hat{\mu}_{S^*}, \hat{\mu}^{\mathcal{D}}) + \mathrm{Var}[\hat{\mu}_{S^*}] + \mathrm{Var}[\hat{\mu}^{\mathcal{D}}]}.$$

Moreover, if the underlying data distribution μ satisfies $\|\boldsymbol{y}\|_2 \leq \frac{1}{2}$ for μ-almost every sample $\boldsymbol{y}$, then, for any estimator $\hat{\boldsymbol{\theta}}$, the generalization with respect to the true risk is bounded by

$$\mathbb{E}\left[R(\hat{\boldsymbol{\theta}}) - \hat{R}_n(\hat{\boldsymbol{\theta}})\right] \leq L\sqrt{\frac{\kappa(\varepsilon)}{n} + \frac{2n-1}{n}\mathrm{Var}[\mu]}$$

where $\kappa(\varepsilon) > 0$ is a constant depending only on the regularization parameter.

The first part of this theorem establishes that the coreset loss is upper-bounded by a term involving the Sinkhorn divergence, that is minimized during learning. The second part, based on recent sample complexity results from Rigollet & Stromme (2025), gives a dimension-free bound. Putting the two bounds together thus gives a bound on the quantity in (5) and thus provides theoretical grounds to use the Sinkhorn divergence as a proxy for the representativeness of the coreset.

The key hypothesis of this result is the Lipschitzness of the loss functions, which could be a restriction for a full learning model. However, low-budget active learning relies on two distinct learning phases: (1) the self-supervised learning model and (2) the final classifier trained on the small number of examples. Our coreset selection problem is solved in the feature space built by the self-supervised learning model and the Lipschitz bound is only needed on the final classifier.

3 Solution methods for the Sinkhorn divergence coreset

This section presents (in Section 3.1) and analyzes (in Section 3.2) the proposed algorithm for solving the Sinkhorn divergence selection problem (3). An analysis of the runtime of our algorithm is provided in Section 3.3. This algorithm is based on an original combination of gradient-based continuous optimization methods with an efficient swap heuristic.

3.1 The Frank-Wolfe Swap Algorithm

The Frank-Wolfe (FW) algorithm (Jaggi, 2013) produces sparse solutions to convex constrained optimization, by constructing each iterate as a convex combination of a limited number of extreme points in the constraint set. In our situation, despite Remark 1, we will apply a variant of FW to the convexified problem (4), to obtain a relevant *active set* $\mathcal{A}$, a set of vertices whose convex combination construct the current iterate.

We performed tests with random initialization candidates that yield worse results and get stuck in local minima, which we provide in the empirical analysis in Section E. Empirically, we find that the active set returned after running FW for a limited number of iterations always contains vertices with low divergence. We then refine these good solutions, by running a local swap algorithm exploiting gradient information to move through vertices.

Our algorithm **FW-Swap** runs in two stages: first, in the *continuous phase*, it executes the FW variant algorithm, blended pairwise conditional gradient (Tsuji et al., 2022), for a fixed number of iterations on the hypersimplex, returning the active set. Then, in the *swap phase*, it runs a local swap algorithm: starting from a vertex, the swap algorithm computes the best single swap minimizing in the direction of the gradient, jumping to an adjacent vertex of the vertex-edge graph of the hypersimplex, and stopping when it has created a cycle. The vertex returned by each swap step is the one having the lowest divergence on the whole cycle.

Algorithm 1 FW-Swap

Require: Cost matrix C, budget B, regularization ε
Define $f(\mathbf{x}) = D_\varepsilon(\hat{\mu}_\mathbf{x}, \hat{\mu}^\mathcal{D})$
(1) Continuous phase
Run FW for $\min_{\mathbf{x}\in\Delta_{n,B}} f(\mathbf{x})$ for T iterations
Obtain fractional solution $\bar{\mathbf{x}}$ and active set $\mathcal{A}$
(2) Swap phase
Initialize candidates $\mathcal{C} \leftarrow \{\}$
for $\boldsymbol{a} \in \mathcal{A}$ **do**
  Initialize $\mathbf{x} \leftarrow \boldsymbol{a}$
  Initialize a solution pool $\mathcal{S} \leftarrow \{\boldsymbol{a}\}$
  while no cycle created **do**
    $\mathbf{x}_\text{new} \leftarrow \arg\min_{\boldsymbol{v}\in\text{swap}(\mathbf{x})} \langle \nabla f(\mathbf{x}), \boldsymbol{v}\rangle$
    if $\mathbf{x}_\text{new} \in \mathcal{S}$ **then**
      break
    end if
    $\mathcal{S} \leftarrow \mathcal{S} \cup \{\mathbf{x}_\text{new}\}$
    $\mathbf{x} \leftarrow \mathbf{x}_\text{new}$
  end while
  $\mathcal{C} \leftarrow \mathcal{C} \cup \{\arg\min_{\mathbf{x}\in\mathcal{S}} f(\mathbf{x})\}$
end for
return $\arg\min_{\mathbf{x}\in\mathcal{C}} f(\mathbf{x})$

We detail **FW-Swap** in Algorithm 1. The description involves the operator swap, which, given $\mathbf{x} \in \mathcal{V}_{n,B}$, produces $\text{swap}(\mathbf{x}) \subset \mathcal{V}_{n,B}$ the subset of vertices that differ from $\mathbf{x}$ by exactly two entries, i.e. by swapping one nonzero entry with a zero entry. The proposed algorithm combines core ideas from two groups of methods: a) local search algorithms for combinatorial problems typically evaluate the objective function on all neighbors of the current vertex, including for instance the analysis of Arya et al. (2001) for k-median and facility location; b) "follow-the-gradient" methods which use the linear minimization oracle of the complete polytope as studied in Baes et al. (2012), akin to a Frank-Wolfe algorithm with a unit step size. Taking the useful aspects of these methods, our algorithm avoids their weaknesses for our context.[1]

Remark 2 (Additional mixture term)**.** *The Sinkhorn divergence is smooth only for* $\mathbf{x}$ *in the relative interior of the hypersimplex; as a consequence, at the boundary, the potentials cannot be obtained as fixed points of the Sinkhorn iterations. In Yin et al. (2025), this issue is handled by a smooth optimization scheme that short-circuits the Sinkhorn algorithm. In our work, we prefer to stabilize the evaluations with a mixture term* λ*, which we rely on both numerically and for the analysis of Section 3.2. The empirical distribution on the selection is then defined as* $\alpha^\lambda_\mathbf{x} = (1-\lambda)\hat{\mu}_\mathbf{x} + \lambda\hat{\mu}^\mathcal{D}$*. This stabilization avoids issues in the computation of the potentials at the boundary of the hypersimplex; For simplicity of exposition, we present* ***FW-Swap*** *in Algorithm 1 without* λ*.*

3.2 Approximation guarantees

In this section, we analyze the swapping procedure of **FW-Swap** that is run on the active set constructed by an initial run of the Frank-Wolfe algorithm. The next theorem establishes a suboptimality guarantee on the quality of the computed solutions. We defer the proof and relevant empirical experiments to Section E.

Theorem 2 (Suboptimality of Swap Descent on Sinkhorn Divergence)**.** *Consider the execution of the swap algorithm with mixture parameter* $\lambda \in (0,1)$ *and denote by* $\bar{\mathbf{x}}$ *its output. Consider* $\mathbf{x}^*$ *a minimizer of the selection problem* (3)*, and* $k^* \leq B$ *the number of swaps required from* $\bar{\mathbf{x}}$ *to* $\mathbf{x}^*$*. There exists a smoothness constant* $L^\text{swap}_{\varepsilon,\lambda}$ *depending on* ε *and* λ*, such that the following suboptimality bound holds:*

$$D_\varepsilon(\hat{\mu}_{\bar{\mathbf{x}}}, \hat{\mu}^\mathcal{D}) - D_\varepsilon(\hat{\mu}_{\mathbf{x}^*}, \hat{\mu}^\mathcal{D}) \leq \frac{k^* L^\text{swap}_{\varepsilon,\lambda}}{2} + 4\lambda\left(1 - \frac{B}{n}\right)\|C\|_\infty . \tag{6}$$

The bound on the quality of the solution thus decomposes into the first term bounding suboptimality of the local search and the second term controlling the bias introduced by the mixture term λ. Considering $\lambda \ll 1$, our bound directly relates the suboptimality of the output $\bar{\mathbf{x}}$ to $k^* \leq B$, the

[1] We also evaluated empirically: zero-th order local search which is prohibitively expensive, and follow-the-gradient which never provided improvements over the candidates from the active set.

number of swaps required from $\bar{\mathbf{x}}$ to the optimal solution $\mathbf{x}^*$. The expression of the constant $L_{\varepsilon,\lambda}^{\text{swap}}$ appearing in this bound is provided with the full proof in Section E.

3.3 Runtime analysis

Per iteration cost. Each swap iteration requires one Sinkhorn OT computation, while the symmetric OT correction is comparatively inexpensive (Feydy et al., 2019). Thus, the Sinkhorn OT computation dominates the per iteration cost and has complexity $O(n^2)$. In contrast, a local search strategy requires $O\left(B(n-B)\right)$ function evaluations per iteration, which was prohibitively expensive in our experiments.

Number of swap iterations. This is observed empirically and explained by a trajectory consisting of a long tail followed by a short cycle. In particular, the descent inequality bounds the tail in terms of the objective range and the swap gap, while the cycle can only persist once the swap gap has reached the scale of the curvature constant. We further make the following assumption.

Assumption 1 (Existence of a core)**.** *Note S the support of $\mathbf{x}$. There exists a set S_0 with $|S_0| = B_0 < B$ such that, along the swap trajectory, every point in S_0 is acquired before any point outside S_0, and once acquired, no point in S_0 is subsequently removed.*

Under Assumption 1, the core is fully acquired in at most B_0 iterations, after which the optimization reduces to a hypersimplex with $n - B_0$ candidate points and a budget $B - B_0$. The preceding tail and cycle bounds then apply to this reduced problem. The detailed analysis is given in Section B.

4 Computational experiments

In this section, we evaluate the empirical performances of our method **FW-Swap**, and we compare it to previously proposed methods.

4.1 Experimental setup

Datasets. We evaluate our active learning methods on the following datasets:

- **STL-10** (Coates et al., 2011): a standard image classification benchmark containing 13 000 labeled images across 10 classes.
- **SVHN** (Netzer et al., 2011): a digit classification dataset consisting of 99 289 real-world images of house numbers. We use the standard training and test sets with 10 classes.
- **GasHisSDB** (Hu et al., 2022): a gastric histopathology image dataset containing 2 classes: normal and abnormal tissue samples. We use 33 284 images corresponding to 160×160 cropped patches.
- **EEGMMIDB** (Zhang & Yao, 2021): an EEG dataset containing recordings from 109 subjects with 3 131 recordings in total. We consider the binary classification task corresponding to left- versus right-hand movement.
- **WBCIC-SHU** (Yang et al., 2025): another EEG dataset containing recordings from 52 subjects with 30 591 recordings in total, with the same binary classes as EEGMMIDB.

Phase 1: Feature learning. We first train a self-supervised learning model on the raw input data (images or signals) without their labels and extract the features[2] then passed to the active learning methods. We rely on either SimCLR (Chen et al., 2020) or DINOv3 (Siméoni et al., 2025) for images. Specifically, we use the ViT-S+/16 distilled model pretrained on web dataset LVD-1689M with DINOv3 for the STL-10 dataset. The frozen pretrained weights allow us to directly access the backbone, and thus, the feature space, without any training or fine tuning. For lower resolution image datasets (32×32), such as GasHisSDB and SVHN, we prefer to fine tune a SimCLR model, as initial observations show that it provides more suitable feature representations. We use the SimCLR-based model implemented in the `SelfEEG` library (Del Pup et al., 2024) for the two EEG datasets.

[2]The choice of feature representation is critical, as the quality of the learned embedding directly impacts downstream performance. In particular, using raw images systematically leads to very low accuracies, consistent with previous studies (Yehuda et al., 2022). We assess the effect of the features extracted from SimCLR or DINOv3 in Section F.

Phase 2: Coreset selection. We compare our algorithm **FW-Swap** to existing approaches, specifically a random selection baseline, the classical heuristics k-medoids and k-centers, and recent low-budget active learning methods: MaxHerding (Bae et al., 2024), TypiClust (Hacohen et al., 2022), and ProbCover (Yehuda et al., 2022). We also experimented with the Wasserstein integer approach of Mahmood et al. (2022), but consistently with their results and those of Yehuda et al. (2022), we stop the method after three hours without obtaining competitive solutions, see appendix Section F. ProbCover's distance hyperparameter δ is set to the mean of the entries of the cost matrix, as tuning this parameter requires grid search on the problem itself without good heuristics. We fix the parameter associated to MaxHerding to one to match the original paper (Bae et al., 2024).

We take the ground cost matrix C defined by the Euclidean distance between feature vectors. For the regularization parameter, we take a fixed value $\varepsilon = 1.5$ for the vision tasks and $\varepsilon = 0.5$ for the EEG tasks. Preliminary experiments with values in $[0.05, 1.5]$ show that, within this range, the parameter does not impact our methods. However, larger or smaller values for ε drastically change the performance[3].

Phase 3: Active learning. We train a linear classifier on the selected subsets of the training data and evaluate the classification accuracy on the associated test sets. The final active learning classifier is implemented using `Flux` v0.16.10 (Innes et al., 2018) and consists of a single linear layer mapping the feature representations to the target classes. Model training is performed using a momentum-based optimizer defined with a learning rate 0.1 and a momentum coefficient 0.9. The classifier is trained for 30 epochs at each active learning round using mini-batches of size 32. Training relies on the logit cross-entropy loss.

Hardware specification. The experiments are run on a cluster of Debian servers with NVIDIA GTX 1080 Ti GPUs and Intel Xeon E5-2620 v4 CPUs with 256Go of RAM, using `Julia` version 1.12.5 (Bezanson et al., 2012).

4.2 Empirical results

We now present the main experimental results (additional results are provided in the appendix Section G). The figures below show the accuracies across the different budgets and methods for STL-10 (Figure 3), for SVHN (Figure 2) and for GasHisSDB Figure 4, and the two tables present the results for both EEG datasets (Tables 1 and 2). The reported values correspond to the mean and standard deviations across 10 seeds. The seeds affect both randomness in the learning phase and selection from randomized methods (k-medoids, k-center, and random selection).

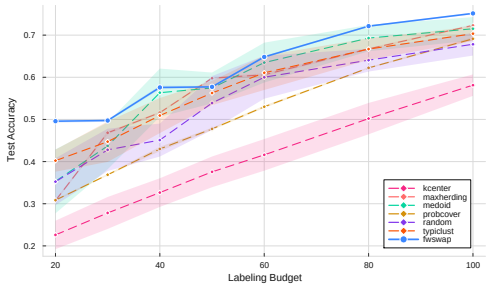

Figure 2: Test accuracy (mean $\pm$ std) on SVHN.

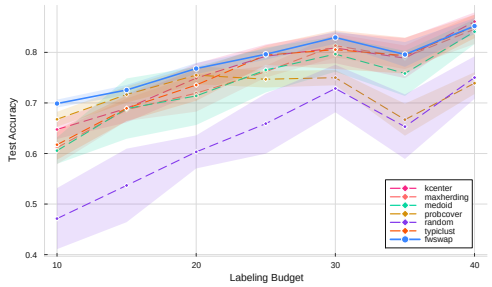

Figure 3: Test accuracy (mean $\pm$ std) on STL-10.

For the SVHN dataset, shown in Figure 2, however, we consistently achieve the best performance across all low budget settings. For the STL-10 dataset, as shown in Figure 3, **FW-Swap** yields the best results on low budgets. For higher budgets, starting $B = 40$, **FW-Swap** is comparable to the best methods. Results for more budgets are provided in Section G.

[3]It is well-known in OT that extreme values of ε change the geometry of the problem. Putting our numerical observations in perspective with Feydy et al. (2019), this suggests that true interpolation between extreme cases (Wasserstein and maximum-mean discrepancy) is the suitable regime for the coreset selection problem. This is also the regime used by Rigollet & Stromme (2025) for their sample complexity result.

For the medical GasHisSDB dataset in Figure 4, our method achieves the best results across small budgets. Note that the original paper for the dataset reports an accuracy of 85.99% when training the full dataset using color-histogram features and a Random Forest classifier (Hu et al., 2022). We achieve similar results with only $B = 10$ images.

On both EEG datasets, **FW-Swap** achieves consistently better accuracy on all budgets as seen in Tables 1 and 2 with a remarkable difference at $B = 2$.

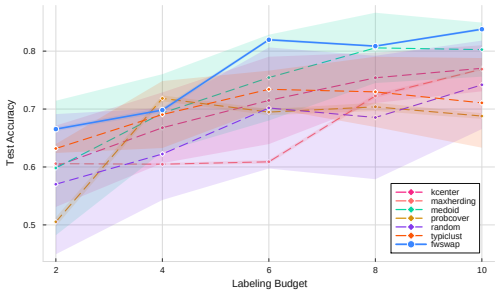

Figure 4: Test accuracy (mean $\pm$ std) on GasHisSDB.

Table 1: Test accuracy (mean $\pm$ std) on EEGMMIDB dataset.

Method	**2**	**4**	**6**	**8**	**10**
FW-Swap	$\mathbf{0.805 \pm 0.002}$	$\mathbf{0.714 \pm 0.003}$	$\mathbf{0.783 \pm 0.003}$	$\mathbf{0.748 \pm 0.004}$	$\mathbf{0.783 \pm 0.002}$
MaxHerding	0.487 ± 0.002	0.687 ± 0.003	0.749 ± 0.002	0.705 ± 0.002	0.714 ± 0.003
TypiClust	0.501 ± 0.002	0.642 ± 0.014	0.674 ± 0.006	0.697 ± 0.027	0.732 ± 0.020
Medoid	0.588 ± 0.106	$\mathbf{0.716 \pm 0.063}$	0.735 ± 0.050	0.744 ± 0.048	0.765 ± 0.038
k-Center	0.500 ± 0.001	0.547 ± 0.058	0.574 ± 0.055	0.592 ± 0.057	0.629 ± 0.057
ProbCover	0.515 ± 0.001	0.564 ± 0.002	0.645 ± 0.002	0.653 ± 0.002	0.659 ± 0.002
Random	0.535 ± 0.078	0.602 ± 0.107	0.660 ± 0.085	0.719 ± 0.063	0.744 ± 0.033

Table 2: Test accuracy (mean $\pm$ std) on the WBCIC-SHU Motor Imagery dataset.

Method	**2**	**4**	**6**	**8**	**10**
FW-Swap	$\mathbf{0.758 \pm 0.079}$	$\mathbf{0.773 \pm 0.041}$	$\mathbf{0.811 \pm 0.029}$	$\mathbf{0.819 \pm 0.010}$	$\mathbf{0.829 \pm 0.033}$
MaxHerding	0.499 ± 0.003	0.460 ± 0.012	0.488 ± 0.023	0.538 ± 0.026	0.596 ± 0.037
TypiClust	0.499 ± 0.008	0.507 ± 0.003	0.511 ± 0.022	0.514 ± 0.012	0.556 ± 0.066
Medoid	0.550 ± 0.114	0.664 ± 0.119	0.757 ± 0.061	0.740 ± 0.086	0.737 ± 0.090
k-Center	0.498 ± 0.037	0.521 ± 0.025	0.524 ± 0.034	0.515 ± 0.029	0.546 ± 0.050
ProbCover	0.503 ± 0.008	0.673 ± 0.053	0.645 ± 0.060	0.656 ± 0.049	0.662 ± 0.060
Random	0.529 ± 0.043	0.676 ± 0.130	0.739 ± 0.105	0.717 ± 0.111	0.772 ± 0.041

The key takeaway of these experiments is that **FW-Swap** consistently yields the best performing results on the low budget. Furthermore, we also note greater stability of our results: across all experiments, the variance of the performance of **FW-Swap** is lower than those of other methods.

5 Limitations

We highlight three limitations of the proposed approach that open perspective for future work. First, as with other representation-based active set methods, our approach relies on the quality of the learned representation features (see discussion in Section F). Second, the underlying cost matrix and corresponding Sinkhorn kernel can be computationally expensive for large datasets, which would require further research on the computational OT side to scale. Finally, while **FW-Swap** is competitive at a small budget, its performance becomes less consistent when the budget increases (see discussion in Section G).

6 Conclusion

We propose a theoretically grounded approach to low-budget active learning, combining a principled criterion, the Sinkhorn divergence, with an algorithm designed for the combinatorial structure of the problem. Our algorithm exploits the convexity and differentiability of the criterion in a combinatorial local search, with quantitative suboptimality bounds and statistical guarantees. Empirically, **FW-Swap** achieves superior accuracy in the low-budget regime, with the advantage most pronounced at very small budgets where representativeness matters most. In particular, the results show that our method can be highly suitable to medical applications as EEG signals and radio imaging.

AI USE STATEMENT

In this work, we used generative AI tools to provide critical ingredients for proving mathematical claims. We have not used generative AI tools to generate synthetic data sets, help develop theoretical models or conceptual frameworks, formulate mathematical claims, assist in the writing of proofs, propose or refine hypotheses, design or provide feedback on research methodology or experiments, implement methods, assist with translation, clean and reformat dataset, support qualitative and thematic data analysis, or interpret results.

Additionally, we used generative AI tools for sourcing information, editing software code, and editing the paper to improve readability. We have reviewed all AI-assisted work. For example, we revised some passages of the manuscript with the use of LLMs to improve the presentation of the core ideas. We also reviewed LLM-assisted code and verified by authors before use, including code intended to improve computational efficiency. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

We have made efforts to ensure that the theoretical and experimental results presented in this work are reproducible. The source code of the experiments will be available for review. All assumptions and formal definitions underlying the theoretical findings are stated in the main paper, with complete proofs provided in the appendix.

REFERENCES

Vijay Arya, Naveen Garg, Rohit Khandekar, Adam Meyerson, Kamesh Munagala, and Vinayaka Pandit. Local search heuristic for k-median and facility location problems. In *Proceedings of the thirty-third annual ACM symposium on Theory of computing*, pp. 21–29, 2001.

Wonho Bae, Junhyug Noh, and Danica J Sutherland. Generalized coverage for more robust low-budget active learning. In *European conference on computer vision*, pp. 318–334. Springer, 2024.

Michel Baes, Alberto Del Pia, Yurii Nesterov, Shmuel Onn, and Robert Weismantel. Minimizing Lipschitz-continuous strongly convex functions over integer points in polytopes. *Mathematical programming*, 134(1):305–322, 2012.

Jeff Bezanson, Stefan Karpinski, Viral B Shah, and Alan Edelman. Julia: A fast dynamic language for technical computing. *arXiv preprint arXiv:1209.5145*, 2012.

Guillaume Carlier, Lénaïc Chizat, and Maxime Laborde. Displacement smoothness of entropic optimal transport. *ESAIM: Control, Optimisation and Calculus of Variations*, 30:25, 2024.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pp. 1597–1607. PmLR, 2020.

Sinho Chewi, Jonathan Niles-Weed, and Philippe Rigollet. *Statistical optimal transport*. Springer, 2025.

Adam Coates, Andrew Ng, and Honglak Lee. An analysis of single-layer networks in unsupervised feature learning. In *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, pp. 215–223. JMLR Workshop and Conference Proceedings, 2011.

Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. *Advances in neural information processing systems*, 26, 2013.

Federico Del Pup, Andrea Zanola, Louis Fabrice Tshimanga, Paolo Emilio Mazzon, and Manfredo Atzori. SelfEEG: A python library for self-supervised learning in electroencephalography. *Journal of Open Source Software*, 9(95):6224, 2024. doi: 10.21105/joss.06224. URL https://doi.org/10.21105/joss.06224.

George Deligiannidis, Valentin De Bortoli, and Arnaud Doucet. Quantitative uniform stability of the iterative proportional fitting procedure. *The Annals of Applied Probability*, 34(1A):501–516, 2024.

Jean Feydy, Thibault Séjourné, François-Xavier Vialard, Shun-ichi Amari, Alain Trouvé, and Gabriel Peyré. Interpolating between optimal transport and MMD using Sinkhorn divergences. In *The 22nd international conference on artificial intelligence and statistics*, pp. 2681–2690. PMLR, 2019.

Aude Genevay, Gabriel Peyré, and Marco Cuturi. Learning generative models with Sinkhorn divergences. In *International Conference on Artificial Intelligence and Statistics*, pp. 1608–1617. PMLR, 2018.

Aude Genevay, Lénaic Chizat, Francis Bach, Marco Cuturi, and Gabriel Peyré. Sample complexity of Sinkhorn divergences. In *The 22nd international conference on artificial intelligence and statistics*, pp. 1574–1583. PMLR, 2019.

Jie Gui, Tuo Chen, Jing Zhang, Qiong Cao, Zhenan Sun, Hao Luo, and Dacheng Tao. A survey on self-supervised learning: Algorithms, applications, and future trends. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(12):9052–9071, 2024.

Guy Hacohen, Avihu Dekel, and Daphna Weinshall. Active learning on a budget: Opposite strategies suit high and low budgets. *arXiv preprint arXiv:2202.02794*, 2022.

Weiming Hu, Chen Li, Xiaoyan Li, Md Mamunur Rahaman, Jiquan Ma, Yong Zhang, Haoyuan Chen, Wanli Liu, Changhao Sun, Yudong Yao, et al. Gashissdb: A new gastric histopathology image dataset for computer aided diagnosis of gastric cancer. *Computers in biology and medicine*, 142:105207, 2022.

Michael Innes, Elliot Saba, Keno Fischer, Dhairya Gandhi, Marco Concetto Rudilosso, Neethu Mariya Joy, Tejan Karmali, Avik Pal, and Viral Shah. Fashionable modelling with Flux, 2018. URL https://arxiv.org/abs/1811.01457.

Martin Jaggi. Revisiting Frank-Wolfe: Projection-free sparse convex optimization. In *International conference on machine learning*, pp. 427–435. PMLR, 2013.

Leonard Kaufman and Peter J Rousseeuw. *Finding groups in data: an introduction to cluster analysis*. John Wiley & Sons, 1990.

Alex Kokot and Alex Luedtke. Coreset selection for the sinkhorn divergence and generic smooth divergences. *arXiv preprint arXiv:2504.20194*, 2025.

Hugo Lavenant, Jonas Luckhardt, Gilles Mordant, Bernhard Schmitzer, and Luca Tamanini. The Riemannian geometry of Sinkhorn divergences. *Ann. Inst. H. Poincaré C Anal. Non Linéaire*, 2025.

Rafid Mahmood, Sanja Fidler, and Marc T Law. Low-budget active learning via Wasserstein distance: An integer programming approach. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=v8OlxjGn23S.

Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Baolin Wu, Andrew Y Ng, et al. Reading digits in natural images with unsupervised feature learning. In *NIPS workshop on deep learning and unsupervised feature learning*, volume 2011, pp. 4. Granada, 2011.

Marcel Nutz and Johannes Wiesel. Stability of Schrödinger potentials and convergence of Sinkhorn's algorithm. *The Annals of Probability*, 51(2):699–722, 2023.

Gabriel Peyré, Marco Cuturi, et al. Computational optimal transport with applications to data sciences. *Foundations and Trends® in Machine Learning*, 2019.

Philippe Rigollet and Austin J Stromme. On the sample complexity of entropic optimal transport. *The Annals of Statistics*, 53(1):61–90, 2025.

Thibault Séjourné, Jean Feydy, François-Xavier Vialard, Alain Trouvé, and Gabriel Peyré. Sinkhorn divergences for unbalanced optimal transport. *arXiv preprint arXiv:1910.12958*, 2019.

Ozan Sener and Silvio Savarese. Active learning for convolutional neural networks: A core-set approach. In *International Conference on Learning Representations*, 2018.

Burr Settles. Active learning literature survey. Technical report, University of Wisconsin-Madison Department of Computer Sciences, 2009.

Oriane Siméoni, Huy V Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, et al. Dinov3. *arXiv preprint arXiv:2508.10104*, 2025.

Kazuma K Tsuji, Ken'ichiro Tanaka, and Sebastian Pokutta. Pairwise conditional gradients without swap steps and sparser kernel herding. In *International Conference on Machine Learning*, pp. 21864–21883. PMLR, 2022.

Haoran Wang, Qiuye Jin, Shiman Li, Siyu Liu, Manning Wang, and Zhijian Song. A comprehensive survey on deep active learning in medical image analysis. *Medical Image Analysis*, 95:103201, 2024.

Banghua Yang, Fenqi Rong, Yunlong Xie, Du Li, Jiayang Zhang, Fu Li, Guangming Shi, and Xiaorong Gao. A multi-day and high-quality eeg dataset for motor imagery brain-computer interface. *Scientific Data*, 12(1):488, 2025.

Ofer Yehuda, Avihu Dekel, Guy Hacohen, and Daphna Weinshall. Active learning through a covering lens. *Advances in Neural Information Processing Systems*, 35:22354–22367, 2022.

Haoyun Yin, Yixuan Qiu, and Xiao Wang. Wasserstein coreset via Sinkhorn loss. *Transactions on Machine Learning Research*, 2025.

Xiang Zhang and Lina Yao. *Deep learning for EEG-based brain–computer interfaces: representations, algorithms and applications*. World Scientific, 2021.

APPENDIX

The appendix complements the main paper with proofs of the results, as well as auxiliary technical developments and additional computational results. Note that the sections are ordered by technical difficulty rather than their order as presented in the main paper. More precisely, we detail:

- in Section A, the **statements** and proofs of the two results mentioned in Section 2.2,
- in Section B, the runtime analysis of the **FW-Swap** algorithm.
- in Section C, the proof and extension of the **generalization bound** with dimension-free guarantees stated in Section 2.4,
- in Section D, the **analysis of Sinkhorn bias** provided of cluster-central-outlier example,
- in Section E, the proof of **theoretical guarantees** of our algorithm,
- in Section F, a comparison based on the **feature space** defined by the SSL model,
- in Section G, **experimental results** complementary to the empirical observations in Section 4.

A AUXILIARY RESULTS

In this section, we prove two auxiliary results on reformulating previous active learning methods in an OT framework, and on the triviality of the convex relaxation of the selection problem, regardless of the divergence.

Proposition 2 (Previous methods within a OT framework)**.** *The optimization problems underlying the ProbCover and k-medoid methods can be cast as minimizing a Wasserstein distance from the dataset to a measure supported on B points, with k-medoids equipped with the Euclidean distance and ProbCover equipped with the ground cost:*

$$c^{\mathrm{PC}}(\boldsymbol{y}_1, \boldsymbol{y}_2) = \begin{cases} 0 & \textit{if } \|\boldsymbol{y}_1 - \boldsymbol{y}_2\|_2 \leq \delta \\ 1 & \textit{otherwise.} \end{cases}$$

Proof. We show the criterion optimized by both methods can, for any $\mathbf{x}$, be expressed as the value of an OT problem in which the selected marginal is free over the points in the support of $\mathbf{x}$.

The k-medoids criterion for a selection $\mathbf{x} \in \mathcal{V}_{n,B}$ is $\sum_{j\in[n]} \min_{i\in\mathrm{supp}(\mathbf{x})} C_{ij}$ which assigns each dataset point to its nearest medoid at cost C_{ij}. This is exactly the minimum Wasserstein distance under a free second marginal

$$\min_{\substack{\boldsymbol{m}\in\Delta_n \\ \boldsymbol{m}_i \leq \mathbf{x}_i}} W_C\Big(\sum_{i\in[n]} \boldsymbol{m}_i \delta_{\boldsymbol{y}_i}, \hat{\mu}^{\mathcal{D}}\Big).$$

Since the optimal transport plan assigns the full mass of each point to its nearest medoid, minimizing over $\mathbf{x} \in \mathcal{V}_{n,B}$ recovers k-medoids exactly.

A similar formulation can be proposed for ProbCover, for which the underlying optimization problem can be expressed as

$$\min_{\mathbf{x}\in\mathcal{V}_{n,B}} \frac{1}{n} \sum_{i\in[n]} \mathbf{1}\Big(\min_{j:\,\mathbf{x}_j=1} \|\boldsymbol{y}_i - \boldsymbol{y}_j\|_2 > \delta\Big).$$

Consider the optimal transport problem

$$W_{c^{\mathrm{PC}}}(\mu, \hat{\mu}^{\mathcal{D}}) = \min_{\boldsymbol{\pi}\in\Pi(\mu,\hat{\mu}^{\mathcal{D}})} \sum_{(i,j)\in[n]\times[n]} \boldsymbol{\pi}_{ij}\, c^{\mathrm{PC}}(\boldsymbol{y}_i, \boldsymbol{y}_j),$$

where $\Pi(\mu, \hat{\mu}^{\mathcal{D}})$ is the set of couplings with marginals μ and $\hat{\mu}^{\mathcal{D}}$. Because the cost c^{PC} is binary, any mass transported between points i and j with $\|\boldsymbol{y}_i - \boldsymbol{y}_j\|_2 \leq \delta$ incurs zero cost, and any other assignment incurs a unit cost.
For each $\boldsymbol{y}_i$, if there exists a selected point $\boldsymbol{y}_j$ such that $\|\boldsymbol{y}_i - \boldsymbol{y}_j\|_2 \leq \delta$, then assigning its full mass $1/n$ to such a point yields zero cost. Furthermore, a selected center can be attributed a mass

k/n, with k the number of points it captures. All points outside the circle around any of the centers needs to be transported to one of the centers arbitrarily, inducing a constant cost of one. Hence, the optimal transport cost satisfies for any $\mathbf{x} \in \mathcal{V}_{n,B}$

$$\min_{\substack{\boldsymbol{m}\in\Delta_n \\ \boldsymbol{m}_i \le \mathbf{x}_i}} W_{c^{\mathrm{PC}}}\Big(\sum_{i\in[n]} \delta_{\boldsymbol{y}_i \boldsymbol{u}_i}, \hat{\mu}^{\mathcal{D}}\Big) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}\Big(\min_{j:\,\mathbf{x}_j=1} \|\boldsymbol{y}_i - \boldsymbol{y}_j\|_2 > \delta\Big),$$

recovering the ProbCover loss. The problems underlying both methods can be expressed as the minimization of a Wasserstein distance from a distribution supported on the coreset to $\hat{\mu}^{\mathcal{D}}$. □

Proposition 3 (Convex relaxation over the hypersimplex yields a trivial bound)**.** *For any divergence D, the convex relaxation of the coreset minimization problem* (3) *is the following trivial optimization problem over the hypersimplex $\Delta_{n,B}$*

$$\min_{\mathbf{x}\in\Delta_{n,B}} D(\hat{\mu}_{\mathbf{x}}, \hat{\mu}^{\mathcal{D}}) = 0,$$

which attains its minimum at $\mathbf{x} = \frac{B}{n}\mathbf{1}$ and is independent of the data.

Proof. By symmetry, the point $\frac{B}{n}\mathbf{1}$ is the center of $\Delta_{n,B}$. The corresponding probability distribution, once rescaled, is exactly $\hat{\mu}^{\mathcal{D}}$. Since D is a divergence, the unique optimum over $\Delta_{n,B}$, which is a convex combination of all vertices of $\Delta_{n,B}$ with equal weights, trivially achieves an objective value of zero. □

B Number of swap iterations analysis

We now focus on the number of iterations of the swap algorithm, with three key observations noted O1, O2, O3.

Note $r(\mathbf{x})$ the gradient of the divergence at $\mathbf{x}$. Swap operated a deterministic map from one vertex to another of the hypersimplex so the swap chain first forms a tail, and then a cycle. Let $\gamma(\mathbf{x})$ be the swap gap at the current vertex:

$$\gamma(\mathbf{x}) = -\min_{\boldsymbol{v}\in\mathrm{swap}(\mathbf{x})} \langle r(\mathbf{x}^c), \boldsymbol{v} - \mathbf{x}^c\rangle > 0.$$

Positivity comes from no vertices being optimal for the continuous problem.

O1. Average gap on the cycle is bounded above. Let T_c be the length of the cycle. We have, with L the Lipschitz constant, along iterations of the cycle:

$$f_\lambda(\mathbf{x}^{(c+1)} \le f_\lambda(\mathbf{x}^c) - \gamma(\mathbf{x}^c) + \frac{L}{2}.$$

Telescoping, $0 < \frac{1}{T_c\left(\sum_{c=1}^{T_c}\gamma(\mathbf{x}^c)\right)} \le \frac{L}{2}$. A cycle can only occur once the swap gap has reduced to the scale of the curvature constant $\frac{L}{2}$, so, where the objective function is already "flat".

O2. The length of the tail T_t is bounded by the range R (max-min function value on the vertices). Telescope the descent inequality on the tail:

$$\sum_{c=1}^{T_t}\left(\gamma(\mathbf{x}^c) - \frac{L}{2}\right) \le f_\lambda(\mathbf{x}^0) - f_\lambda(\mathbf{x}^{T_t}).$$

Defining $\gamma_{\min}$ as the minimum gap on the whole tail, we have whenever $\gamma_{\min} > \frac{L}{2}$, the inequality $T_t \le \frac{R}{\gamma_{\min} - \frac{L}{2}}$. This bounds the tail with the function values of the Sinkhorn divergence.

O3. The core is fully acquired in at most $B_0 < B$ iterations, after which the problem is reduced to a hypersimplex of parameters $n - B_0$, $B - B_0$, on which O1 and O2 apply. This is formalized in Assumption 1.

Assumption 1 (Existence of a core)**.** *Note S the support of* $\mathbf{x}$*. There exists a set S_0 with $|S_0| = B_0 < B$ such that, along the swap trajectory, every point in S_0 is acquired before any point outside S_0, and once acquired, no point in S_0 is subsequently removed.*

Assumption 1 can be verified empirically. S_0 is a set of points that are unambiguously central for a cluster in the data, and $S_0 = \emptyset$ means several candidates are close to the optimal for each cluster. We report in Table 3 the number of iterations of **FW-Swap** over multiple starts on STL-10. We report separately the two parts of the trajectory: the tail (vertices visited once) from the cycle. We observe a tail roughly proportional to the budget on average, while the cycle length remains small and constant.

Table 3: Details of **FW-Swap** trajectory for different budgets.

Budget	End Cycle	Start Cycle	Detected Cycle	Detected Std	Cycle Length
10	11.0	12.0	14.0	1.22474	3.0
20	22.6	23.6	25.6	2.19089	3.0
30	33.2	34.2	36.2	1.30384	3.0
40	44.8	45.8	47.8	2.68328	3.0
50	54.8	55.8	57.8	2.77489	3.0
60	64.2	65.2	67.2	4.91935	3.0

Table 3 shows a long tail and a short cycle, which coincides with what happens under a large core assumption. We ran additional experiments to test **O3** and we could confirm that the length of the tail is linked primarily to the acquisition of the core S_0.

Table 4: Number of acquired points and initial points retained in the final selection.

Budget	Acquired + Initial	Percentage
10	$(1 + 8 = 9)$	90%
20	$(2 + 16 = 18)$	90%
30	$(5 + 24 = 29)$	96.67%
40	$(8 + 31 = 39)$	97.5%
50	$(8 + 38 = 46)$	92%
60	$(9 + 48 = 57)$	95%

Table 4 reports the number of acquired-once points during the swap and the number of original points from the initialization that have not been swapped, which together form a core. On all budgets, the core accounts for more than 90% of the final selection, serving as empirical evidence for Assumption 1.

C GENERALIZATION BOUND

We detail here the different results supporting Theorem 1. We make the following assumptions:

Assumption 2. *The data are i.i.d. samples from a distribution μ such that for μ-almost every $\boldsymbol{y}$, we have $\|\boldsymbol{y}\|_2 \leq \frac{1}{2}$.*

Our analysis is based on the following sample complexity result for the estimation of entropic OT from samples.

Theorem 3 (Strong duality, Theorem 1 of Rigollet & Stromme (2025))**.** *Let α, β be probability measures with bounded support on $\mathbb{R}^d$ and let $\varepsilon > 0$. For bounded measurable $f, g : \mathbb{R}^d \to \mathbb{R}$ set the dual objective*

$$\Phi_{f,g}(\alpha, \beta) \;=\; \langle f, \alpha\rangle + \langle g, \beta\rangle \;-\; \varepsilon \left\langle \alpha \otimes \beta, e^{(f\oplus g-c)/\varepsilon} - 1 \right\rangle, \tag{7}$$

where $(f \oplus g)(\boldsymbol{y}, \boldsymbol{z}) = f(\boldsymbol{y}) + g(\boldsymbol{z})$ (we define $\otimes$ similarly). Then the dual entropic OT problem is

$$S_\varepsilon(\alpha, \beta) \;=\; \sup_{(f,g)\in L^\infty(\alpha)\times L^\infty(\beta)} \Phi_{f,g}(\alpha, \beta), \tag{8}$$

and the supremum is attained at a pair (f, g), unique up to $(f, g) \mapsto (f + a, g - a)$, $a \in \mathbb{R}$. Moreover, primal and dual solutions are linked as follows: for a pair (f, g), let π be the measure with density $e^{(f\oplus g-c)/\varepsilon}$ with respect to $\alpha \otimes \beta$. Then (f, g) is optimal for (8) *if and only if $\pi \in \Pi(\alpha, \beta)$, in which case π is the optimal plan.*

Theorem 4 (Theorem 3, Rigollet & Stromme (2025)). *Let μ and ν be two distributions satisfying Assumption 2. Let $S_\varepsilon(\mu_n, \nu_n)$ be the entropic optimal transport cost between two empirical distributions with n mutually independent samples from μ and ν respectively. Then, there exists $\kappa_0(\varepsilon) > 0$ depending (exponentially) only on the regularization parameter ε and such that*

$$|\mathbb{E}[S_\varepsilon(\hat{\mu}_n, \hat{\nu}_n)] - S_\varepsilon(\mu, \nu)| \leq \frac{\kappa_0(\varepsilon)}{n}. \tag{9}$$

As a direct consequence, we can establish a similar result for the Sinkhorn divergence $D_\varepsilon(\mu, \hat{\mu}_n)$.

Lemma 1 (Sample complexity of the Sinkhorn divergence). *Let μ satisfy Assumption 2 and let $\hat{\mu}_n$ be the empirical measure of n i.i.d. samples from μ. Set $M = \sup_{\boldsymbol{y},\boldsymbol{z}\in\mathrm{supp}(\mu)} c(\boldsymbol{y}, \boldsymbol{z})$, so that $M \leq 1$ under Assumption 2. Then*

$$\mathbb{E}\big[D_\varepsilon(\mu, \hat{\mu}_n)\big] \;\leq\; \frac{1}{n}\Big(\kappa_0(\varepsilon) + \tfrac{\varepsilon}{2}\big(e^{M/\varepsilon} - 1\big)\Big) \;\leq\; \frac{\kappa(\varepsilon)}{n}, \tag{10}$$

with $\kappa(\varepsilon) \;:=\; \kappa_0(\varepsilon) + \frac{\varepsilon}{2}\big(e^{1/\varepsilon} - 1\big)$, and $\kappa_0(\varepsilon)$ the constant of Theorem 4. In particular, κ depends only (exponentially) on $1/\varepsilon$.

Proof. By the definition of the Sinkhorn divergence,

$$\mathbb{E}\big[D_\varepsilon(\mu, \hat{\mu}_n)\big] = \mathbb{E}\big[S_\varepsilon(\mu, \hat{\mu}_n)\big] - \tfrac{1}{2}S_\varepsilon(\mu, \mu) - \tfrac{1}{2}\,\mathbb{E}\big[S_\varepsilon(\hat{\mu}_n, \hat{\mu}_n)\big]. \tag{11}$$

We bound the first term from above and the third from below. Throughout, $0 \leq S_\varepsilon(\alpha, \beta) \leq M$ for α, β supported in $\mathrm{supp}(\mu)$: the lower bound because $c \geq 0$ and $\mathrm{KL} \geq 0$, the upper one by evaluating the primal objective with the product plan $\pi = \alpha \otimes \beta$. Every S_ε that depends on the sample has an empirical measure in at least one argument, $\Phi_{f,g}$ in (8) is then a continuous function of the positions of the atoms of this empirical measure.

We first bound the cross term. Let β be any probability measure supported on $\mathrm{supp}(\mu)$ and let (f, g) be an optimal pair of potentials for $S_\varepsilon(\mu, \beta)$ from Theorem 3. Integrating the last term of (7) in the second argument $\boldsymbol{z}$ first,

$$\Phi_{f,g}(\alpha, \beta) = \langle h, \alpha\rangle + \langle g, \beta\rangle\,,$$
$$h(\boldsymbol{y}) := f(\boldsymbol{y}) - \varepsilon \int \Big(e^{(f(\boldsymbol{y})+g(\boldsymbol{z})-c(\boldsymbol{y},\boldsymbol{z}))/\varepsilon} - 1\Big)\mathrm{d}\beta(\boldsymbol{z}),$$

so that $\alpha \mapsto \Phi_{f,g}(\alpha, \beta)$ is affine. Let $\hat{\mu}'_n$ be the empirical measure of a sample $X'_1, \ldots, X'_n \sim \mu$ and let $\mathbb{E}'$ denote the expectation over it. Since $\mathbb{E}'\,\langle h, \hat{\mu}'_n\rangle = \langle h, \mu\rangle$, we have

$$S_\varepsilon(\mu, \beta) = \Phi_{f,g}(\mu, \beta) = \mathbb{E}'\big[\Phi_{f,g}(\hat{\mu}'_n, \beta)\big] \;\leq\; \mathbb{E}'\big[S_\varepsilon(\hat{\mu}'_n, \beta)\big],$$

the first equality by optimality of (f, g), the second by affineness, and the inequality because f and g are bounded, hence feasible solutions to (8) for the pair $(\hat{\mu}'_n, \beta)$. Running the same argument on a mixture $\alpha = (1-t)\alpha_0 + t\alpha_1$ shows that $\alpha \mapsto S_\varepsilon(\alpha, \beta)$ is convex, and the inequality above corresponds to Jensen's. Applying it with $\beta = \hat{\mu}_n$, which is independent of $\hat{\mu}'_n$, and averaging,

$$\mathbb{E}\big[S_\varepsilon(\mu, \hat{\mu}_n)\big] \;\leq\; \mathbb{E}\big[S_\varepsilon(\hat{\mu}'_n, \hat{\mu}_n)\big] \;\leq\; S_\varepsilon(\mu, \mu) + \frac{\kappa_0(\varepsilon)}{n}, \tag{12}$$

the last inequality being Theorem 4 applied with $\nu = \mu$: the two samples are i.i.d. from μ, of equal size n, and mutually independent as required.

We can now bound the self-transport term. Let (f, g) be an optimal pair for $S_\varepsilon(\mu, \mu)$, again given by Theorem 3; it is deterministic. Write $\Psi := e^{(f\oplus g-c)/\varepsilon} - 1$. By Theorem 3 the measure with density $\Psi + 1$ with respect to $\mu \otimes \mu$ has first marginal μ, that is, its density integrates to one in the second variable:

$$\int e^{(f(\boldsymbol{y})+g(\boldsymbol{z})-c(\boldsymbol{y},\boldsymbol{z}))/\varepsilon}\,\mathrm{d}\mu(\boldsymbol{z}) = 1 \qquad \text{for } \mu\text{-a.e. } \boldsymbol{y}. \tag{13}$$

Integrating (13) against μ gives $\langle \mu \otimes \mu, \Psi\rangle = 0$, hence

$$S_\varepsilon(\mu, \mu) = \Phi_{f,g}(\mu, \mu) = \langle f + g, \mu\rangle\,. \tag{14}$$

Since f and g are bounded, they are admissible in (8) for the pair $(\hat{\mu}_n, \hat{\mu}_n)$, so

$$S_\varepsilon(\hat{\mu}_n, \hat{\mu}_n) \;\geq\; \Phi_{f,g}(\hat{\mu}_n, \hat{\mu}_n) = \langle f + g, \hat{\mu}_n \rangle - \varepsilon \, \langle \hat{\mu}_n \otimes \hat{\mu}_n, \Psi \rangle \, .$$

Since $X_i \sim \mu$ we have $\mathbb{E}\,\langle f + g, \hat{\mu}_n \rangle = \langle f + g, \mu \rangle$. Separating diagonal from off-diagonal pairs,

$$\begin{aligned} \mathbb{E}\,\langle \hat{\mu}_n \otimes \hat{\mu}_n, \Psi \rangle &= \frac{1}{n^2} \sum_{i \neq j} \mathbb{E}\big[\Psi(X_i, X_j)\big] + \frac{1}{n^2} \sum_i \mathbb{E}\big[\Psi(X_i, X_i)\big] \\ &= \Big(1 - \tfrac{1}{n}\Big) \underbrace{\langle \mu \otimes \mu, \Psi \rangle}_{=0} + \frac{1}{n} \Big(\mathbb{E}_\mu \big[e^{(f+g)/\varepsilon} \big] - 1 \Big), \end{aligned}$$

where we used that X_i and X_j are independent for $i \neq j$, and $\Psi(\boldsymbol{y}, \boldsymbol{y}) = e^{(f(\boldsymbol{y}) + g(\boldsymbol{y}))/\varepsilon} - 1$ since $c(\boldsymbol{y}, \boldsymbol{y}) = 0$. Combining with (14),

$$\mathbb{E}\big[S_\varepsilon(\hat{\mu}_n, \hat{\mu}_n)\big] \;\geq\; S_\varepsilon(\mu, \mu) - \frac{\varepsilon}{n} \Big(\mathbb{E}_\mu \big[e^{(f+g)/\varepsilon} \big] - 1 \Big). \tag{15}$$

We can now bound the second quantity of the right-hand side of (15). Set $I_g = \int e^{g/\varepsilon} \mathrm{d}\mu \in (0, \infty)$. By (13) and $c \leq M$ on $\mathrm{supp}(\mu) \times \mathrm{supp}(\mu)$,

$$\begin{aligned} 1 &= \int e^{(f(\boldsymbol{y}) + g(\boldsymbol{z}) - c(\boldsymbol{y}, \boldsymbol{z}))/\varepsilon} \mathrm{d}\mu(\boldsymbol{z}) \;\geq\; e^{-M/\varepsilon} \, e^{f(\boldsymbol{y})/\varepsilon} I_g, \\ e^{f(\boldsymbol{y})/\varepsilon} &\leq \frac{e^{M/\varepsilon}}{I_g} \qquad \text{for } \mu\text{-a.e. } \boldsymbol{y} \end{aligned}$$

Integrating the remaining factor,

$$\mathbb{E}_\mu \big[e^{(f+g)/\varepsilon} \big] = \int e^{f(\boldsymbol{y})/\varepsilon} \, e^{g(\boldsymbol{y})/\varepsilon} \, \mathrm{d}\mu(\boldsymbol{y}) \;\leq\; \frac{e^{M/\varepsilon}}{I_g} \int e^{g/\varepsilon} \mathrm{d}\mu \;=\; e^{M/\varepsilon}, \tag{16}$$

and therefore $\mathbb{E}_\mu[e^{(f+g)/\varepsilon}] - 1 \leq e^{M/\varepsilon} - 1$.

Substituting (12), (15) and (16) into (11), the three terms $S_\varepsilon(\mu, \mu)$ cancel and

$$\mathbb{E}\big[D_\varepsilon(\mu, \hat{\mu}_n)\big] \;\leq\; \frac{\kappa_0(\varepsilon)}{n} + \frac{\varepsilon}{2n} \big(e^{M/\varepsilon} - 1 \big),$$

which is (10); the final bound follows from $M \leq 1$. □

Lemma 2. *The entropic self-transport is bounded by two times the variance: $S_\varepsilon(\mu, \mu) \leq 2\mathrm{Var}[\mu]$. Moreover if μ satisfying Assumption 2, we have*

$$S_\varepsilon(\mu, \mu) \leq 2\mathrm{Var}[\mu] \leq 1.$$

Proof. The entropic self-transport cost is bounded by the cost of any valid coupling; we consider in particular the product $\mu \otimes \mu$

$$S_\varepsilon(\mu, \mu) \leq \langle C, \mu \otimes \mu \rangle + \mathrm{KL}(\mu \otimes \mu, \mu \otimes \mu).$$

The KL divergence is zero when evaluated on the same two measures and the product coupling gives

$$\langle C, \mu \otimes \mu \rangle = \int \int \|\boldsymbol{z} - \boldsymbol{y}\|_2^2 \, d\mu(\boldsymbol{z}) d\mu(\boldsymbol{y}) = 2\mathrm{Var}[\mu].$$

By assumption, we have $\|\boldsymbol{z} - \boldsymbol{y}\|_2^2 \leq 1$ pointwise almost everywhere, yielding the upper bound of one. □

We can now restate and prove our main generalization theorem.

Theorem 1 (Generalization bound for the Sinkhorn divergence coreset)**.** *Consider the Sinkhorn divergence D_ε with ground cost $c(\boldsymbol{y}, \boldsymbol{z}) = \|\boldsymbol{y} - \boldsymbol{z}\|_2^2$. Suppose that the loss function $f_{\boldsymbol{\theta}}$ is L-Lipschitz for any $\boldsymbol{\theta}$. Then the coreset loss is bounded uniformly in $\boldsymbol{\theta}$ by*

$$\hat{R}_n(\boldsymbol{\theta}) - \hat{R}_{S^*}(\boldsymbol{\theta}) \leq L \sqrt{D_\varepsilon(\hat{\mu}_{S^*}, \hat{\mu}^{\mathcal{D}}) + \mathrm{Var}[\hat{\mu}_{S^*}] + \mathrm{Var}[\hat{\mu}^{\mathcal{D}}]}.$$

Moreover, if the underlying data distribution μ satisfies $\|\boldsymbol{y}\|_2 \leq \frac{1}{2}$ for μ-almost every sample $\boldsymbol{y}$, then, for any estimator $\hat{\boldsymbol{\theta}}$, the generalization with respect to the true risk is bounded by

$$\mathbb{E}\left[R(\hat{\boldsymbol{\theta}}) - \hat{R}_n(\hat{\boldsymbol{\theta}})\right] \leq L\sqrt{\frac{\kappa(\varepsilon)}{n} + \frac{2n-1}{n}\mathrm{Var}[\mu]}$$

where $\kappa(\varepsilon) > 0$ is a constant depending only on the regularization parameter.

Proof. For any two measures α, β respecting Assumption 2, let $\boldsymbol{\pi}$ be the optimal entropic coupling minimizing $S_\varepsilon(\alpha, \beta)$, we can derive

$$\begin{aligned}
\int f_{\boldsymbol{\theta}} d\alpha - \int f_{\boldsymbol{\theta}} d\beta &= \int (f_{\boldsymbol{\theta}}(\boldsymbol{z}) - f_{\boldsymbol{\theta}}(\boldsymbol{y}))d\boldsymbol{\pi}(\boldsymbol{z}, \boldsymbol{y}) && \text{(17a)}\\
&\leq L \int \|\boldsymbol{z} - \boldsymbol{y}\|_2 \, d\boldsymbol{\pi}(\boldsymbol{z}, \boldsymbol{y}) && \text{(17b)}\\
&\leq L\sqrt{\int \|\boldsymbol{z} - \boldsymbol{y}\|_2^2 \, d\boldsymbol{\pi}(\boldsymbol{z}, \boldsymbol{y})} && \text{(17c)}\\
&\leq L\sqrt{S_\varepsilon(\alpha, \beta)}, && \text{(17d)}\\
&= L\sqrt{D_\varepsilon(\alpha, \beta) + \frac{1}{2}S_\varepsilon(\alpha, \alpha) + \frac{1}{2}S_\varepsilon(\beta, \beta)} && \text{(17e)}\\
&\leq L\sqrt{D_\varepsilon(\alpha, \beta) + \mathrm{Var}[\alpha] + \mathrm{Var}[\beta]} && \text{(17f)}
\end{aligned}$$

where the first equality (17a) uses that $\boldsymbol{\pi}$ is a coupling, (17b) uses the Lipschitz bound, (17c) applies Cauchy-Schwarz and (17d) uses the nonnegativity of $\mathrm{KL}(\cdot, \alpha \otimes \beta)$, (17e) uses the definition of the divergence and (17f) uses Lemma 2. Applying this inequality with $\hat{\mu}_{S^*}$ and $\hat{\mu}^{\mathcal{D}}$, we get the first bound on the empirical risk and the coreset risk. Applying the inequality a second time with μ and $\hat{\mu}^{\mathcal{D}}$ gives

$$\begin{aligned}
R(\boldsymbol{\theta}) - \hat{R}_n(\boldsymbol{\theta}) &\leq L\sqrt{D_\varepsilon(\mu, \hat{\mu}^{\mathcal{D}}) + \mathrm{Var}[\mu] + \mathrm{Var}[\hat{\mu}^{\mathcal{D}}]}\\
\mathbb{E}[R(\boldsymbol{\theta}) - \hat{R}_n(\boldsymbol{\theta})] &\leq L\sqrt{\frac{\kappa(\varepsilon)}{n} + \frac{2n-1}{n}\mathrm{Var}[\mu]},
\end{aligned}$$

where the second line takes the expectation over possible realizations of $\hat{\mu}^{\mathcal{D}}$, applies Lemma 1 and Jensen's inequality to $\sqrt{\cdot}$, along with the expectation of the empirical variance. □

D Sinkhorn bias for Cluster-Central-Outlier instances

In this section, we analyze the family of instances defined in Section 2.3, and quantify the bias of S_ε on the three configurations presented there.

We start with closed forms for Wasserstein costs of the three choices, thanks to the simplicity of one-dimensional optimal transport.

Proposition 4. *The Wasserstein cost of the three possible selections are:*

$$W(\alpha_C, \hat{\mu}^{\mathcal{D}}) = \frac{2m\Delta^2}{2m+1}, \quad W(\alpha_L, \hat{\mu}^{\mathcal{D}}) = \frac{\Delta^2}{2}, \quad \textit{and} \quad W(\alpha_E, \hat{\mu}^{\mathcal{D}}) = \frac{\Delta^2}{2m+1}$$

with α_E being the configuration with minimum cost.

Proof. To transport α_C to $\hat{\mu}^{\mathcal{D}}$, the same mass must be transported to both the left and right points. The reverse is performed to transport α_E to $\hat{\mu}^{\mathcal{D}}$, with $\frac{1}{2} - \frac{m}{2m+1}$ being the excess mass at each extreme point that has to be transported to the origin. To transport α_L, the left point has an excess of mass that needs to be transported of $\frac{1}{2} - \frac{m}{2m+1} = \frac{1}{2(2m+1)}$. Because the ground cost is the Euclidean squared distance, the optimal plan transports this excess to the origin, and displaces to the right point enough mass from the origin to match its weight in $\hat{\mu}^{\mathcal{D}}$ which is $\frac{m}{2m+1}$. The total transport cost is $\Delta^2\left(\frac{1}{2(2m+1)} + \frac{m}{2m+1}\right) = \frac{\Delta^2}{2}$. □

The ranking of the configurations of Sinkhorn costs is identical for low values of ε, but we show in the following proposition that it inverts for ε above a certain threshold.

Proposition 5. *The limit of the Sinkhorn objective values of the three configurations at $\varepsilon \to \infty$:*

$$\mathcal{S}_\infty(\alpha_C, \hat{\mu}^{\mathcal{D}}) = \frac{2m\Delta^2}{2m+1}, \quad \mathcal{S}_\infty(\alpha_L, \hat{\mu}^{\mathcal{D}}) = \frac{(6m+1)\Delta^2}{2(2m+1)}, \quad \mathcal{S}_\infty(\alpha_E, \hat{\mu}^{\mathcal{D}}) = \frac{(4m+1)\Delta^2}{2m+1}.$$

There exists a finite ε for which α_C results in the smallest loss.

Proof. The Sinkhorn cost of a transport plan is the sum of the inner product with the distance matrix plus the regularization term, ε multiplied by the KL divergence. Driving $\varepsilon \to \infty$ forces the KL divergence to zero, meaning the transport plan becomes the outer product of the two input weight vectors, and the cost becomes $\sum_{i\in[n_\alpha], j\in\{1,2,3\}}(x_i - y_j)^2 \alpha_i \hat{\mu}_j^{\mathcal{D}}$, with x_i the i-th point of the α distribution and y_j the j-th point of the dataset. We can compute the corresponding expression for each of the configurations. Since the Sinkhorn objective is continuous with respect to ε and $\mathcal{S}_{+\infty}(\alpha_C, \hat{\mu}^{\mathcal{D}})$ is strictly smaller than the loss of the two other configurations, there exists a finite $\bar{\varepsilon}$ above which the central configuration is always preferred. □

We highlight here that the inherent bias of the Sinkhorn cost also produces a paradox for the active learning application: it can lead to selecting *fewer* points to train on than the allocated budget, which in general will go against the expected and empirical behavior that one should train on as many points as the budget allows.

Lemma 3. *The entropy losses of the extreme configuration can be expressed as*

$$S_\varepsilon(\alpha_E, \hat{\mu}^{\mathcal{D}}) = W(\alpha_E, \hat{\mu}^{\mathcal{D}}) + \frac{2m}{2m+1} S_\varepsilon(\alpha_E, \alpha_E). \tag{18}$$

Proof. The coupling $\boldsymbol{\pi}$ is a 2×3 matrix with rows corresponding to the mass supported on $\{-\Delta, \Delta\}$ and columns corresponding to the mass supported on $\{-\Delta, 0, \Delta\}$, with marginals $(\frac{1}{2}, \frac{1}{2})$ and $(\frac{m}{2m+1}, \frac{1}{2m+1}, \frac{m}{2m+1})$. Both marginals and the ground costs are symmetric around the origin, the optimal plan is also symmetric and can be expressed as:

$$\boldsymbol{\pi}^* = \begin{bmatrix} a & b & d \\ d & b & a \end{bmatrix}.$$

The marginal constraint of the central column results in $2b = \frac{1}{2m+1} \Rightarrow b = \frac{1}{4m+2}$. This value corresponds to the product of the probabilities of the marginals for the whole column, the corresponding term in the KL divergence is thus zero. The contribution to the transport cost is the constant $2b\Delta^2 = \Delta^2/(2m+1)$.

Using $p = \frac{m}{2m+1}$ for convenience and setting $d = p - a$, joint optimization over the whole transport plan reduces to determining the optimal a. The weights of the marginal $\hat{\mu}^{\mathcal{D}}$ on the remaining 2×2 block are (p, p), which is proportional to α_E. Using $\mathrm{KL}(p\mu, \nu) = \lambda \mathrm{KL}(p\mu, \nu)$ for $p > 0$, we can reformulate the objective as:

$$F(a) = (1-2p)\Delta^2 + 2p\left[4\Delta^2 \cdot \tfrac{d}{2p} + \varepsilon\,\mathrm{KL}(\tilde{\boldsymbol{\pi}}, \alpha_E \otimes \alpha_E)\right] = (1-2p)\Delta^2 + 2p \cdot G\Big(\tfrac{d}{2p}\Big)$$

where $G(s) = 4\Delta^2 s + \varepsilon\,\mathrm{KL}(\tilde{\boldsymbol{\pi}}(s), \alpha_E \otimes \alpha_E)$.

$G(s)$ is exactly the Sinkhorn objective for $S_\varepsilon(\alpha_E, \alpha_E)$ parameterized by its off-diagonal entry s. Combining with the expression of the Wasserstein cost in Proposition 4, we can identify the expression (18) and conclude. □

Finally, we show that the Sinkhorn divergence, in contrast, avoids this paradox by removing the centrality bias of the Sinkhorn objective.

Proposition 6. *The Sinkhorn divergence D_ε leads to the extreme configuration being preferred for any value of ε.*

Proof. We first derive the expression of the symmetric optimal transport cost for any ε. The measure α_C has a single atom, the transport cost and the KL divergence are both zero, $S_\varepsilon(\alpha_C, \alpha_C) = 0 \, \forall \varepsilon$. The measures α_L and α_E are both supported on two atoms with uniform weights $(1/2, 1/2)$, the analysis is similar for both.

Defining the two-point measure as $1/2(\delta_a + \delta_b)$, setting $c = (a-b)^2$ and noting that the transport plan $\boldsymbol{\pi}$ is symmetric, it is fully parameterized by a single scalar t such that $\boldsymbol{\pi}_{aa} = \boldsymbol{\pi}_{bb} = \frac{1}{2} - t$, $\boldsymbol{\pi}_{ab} = \boldsymbol{\pi}_{ba} = t$. The expression $\langle C, \boldsymbol{\pi} \rangle$ reduces to $2tc$. The product measure has weight $1/4$ on all entries, the KL divergence becomes

$$2(\frac{1}{2} - t) \log \left(\frac{\frac{1}{2} - t}{\frac{1}{4}} \right) + 2t \log \left(\frac{t}{\frac{1}{4}} \right)$$

and we can derive the complete expression as a function of t, and the optimal t^* from first-order optimality:

$$\begin{aligned}
&S_\varepsilon(\alpha, \alpha) = \phi(t) = 2tc + \varepsilon \left((1-2t) \log(2-4t) + t2 \log(4t) \right) \\
&t^* = \frac{1}{2(1 + e^{c/\varepsilon})} \\
&\phi(t^*) = \frac{c}{1 + e^{c/\varepsilon}} + \frac{c e^{c/\varepsilon}}{1 + e^{c/\varepsilon}} + \varepsilon \log \frac{2}{1 + e^{c/\varepsilon}} \\
&\phi(t^*) = c + \varepsilon \log \left(\frac{2}{1 + e^{c/\varepsilon}} \right) = \varepsilon \log \left(\frac{2}{1 + e^{-c/\varepsilon}} \right).
\end{aligned}$$

Specializing for $\alpha = \alpha_L$ with $c = \Delta^2$ and α_E with $c = 4\Delta^2$, we obtain:

$$S_\varepsilon(\alpha_L, \alpha_L) = \varepsilon \log \frac{2}{1 + e^{-\Delta^2/\varepsilon}} \tag{19}$$

$$S_\varepsilon(\alpha_E, \alpha_E) = \varepsilon \log \frac{2}{1 + e^{-4\Delta^2/\varepsilon}}. \tag{20}$$

We can now compare the divergences pairwise between the different configurations, starting with the extreme and central ones.

$$\begin{aligned}
D_\varepsilon(\alpha_E, \hat{\mu}^{\mathcal{D}}) - D_\varepsilon(\alpha_C, \hat{\mu}^{\mathcal{D}}) &= S_\varepsilon(\alpha_E, \hat{\mu}^{\mathcal{D}}) - \frac{1}{2} S_\varepsilon(\alpha_E, \alpha_E) - W(\alpha_C, \hat{\mu}^{\mathcal{D}}) && \text{(21a)} \\
&= W(\alpha_E, \hat{\mu}^{\mathcal{D}}) + (2p - \frac{1}{2}) S_\varepsilon(\alpha_E, \alpha_E) - W(\alpha_C, \hat{\mu}^{\mathcal{D}}) && \text{(21b)} \\
&= \frac{\Delta^2}{2m+1} + (2p - \frac{1}{2}) S_\varepsilon(\alpha_E, \alpha_E) - \frac{2m\Delta^2}{2m+1} && \text{(21c)} \\
&= \frac{2m-1}{2(2m+1)} \left(S_\varepsilon(\alpha_E, \alpha_E) - 2\Delta^2 \right) && \text{(21d)} \\
&= \frac{2m-1}{2(2m+1)} \left(\varepsilon \log \frac{2}{1 + e^{-4\Delta^2/\varepsilon}} - 2\Delta^2 \right) < 0 && \text{(21e)}
\end{aligned}$$

where (21b) uses Lemma 3, (21c) uses the Wasserstein cost expressions from Proposition 4, (21d) rearranges the terms and (21e) uses the expression (20) for the symmetric Sinkhorn. Negativity of the expression comes from $\varepsilon \log \frac{2}{1+e^{-4\Delta^2/\varepsilon}} - 2\Delta^2$ being valid for any finite ε.

We now establish the inequality for the extreme and left configurations. The pair of dual potentials (f, g) with $f = (\Delta^2, 0)$, $g = (-\Delta^2, 0, \Delta^2)$ is optimal for the Wasserstein dual problem $W(\alpha_L, \hat{\mu}^{\mathcal{D}})$, since it results in an objective $\frac{\Delta^2}{2}$ matching the primal and the potentials are feasible.

Weak duality for the Sinkhorn problem then yields

$$\begin{aligned}
S_\varepsilon(\alpha_L, \hat{\mu}^{\mathcal{D}}) \geq & \langle f, \alpha_L \rangle + \langle g, \hat{\mu} \rangle - \varepsilon \sum_{i,j} (\alpha_L)_i \hat{\mu}_j^{\mathcal{D}} \left(e^{(f_i + g_j - C_{ij})/\varepsilon} - 1 \right) \\
\geq & \frac{\Delta^2}{2} + \varepsilon p (1 - e^{-\frac{2\Delta^2}{\varepsilon}}).
\end{aligned}$$

Using the expression (19) for $S_\varepsilon(\alpha_L, \alpha_L)$ and the fact that $\log\left(\frac{2}{1+e^{-\Delta^2/\varepsilon}}\right) \leq 1 - e^{-\Delta^2/\varepsilon}$, we have

$$\frac{1}{2}S_\varepsilon(\alpha_L, \alpha_L) \leq \frac{\varepsilon}{2}(1 - e^{-\Delta^2/\varepsilon}).$$

Factoring $1 - e^{-2\Delta^2/\varepsilon} = (1 - e^{-\Delta^2/\varepsilon})(1 + e^{-\Delta^2/\varepsilon})$, we have

$$\varepsilon p(1 - e^{-2\Delta^2/\varepsilon}) - \tfrac{1}{2}S_\varepsilon(\alpha_L, \alpha_L) \geq \varepsilon(1 - e^{-\Delta^2/\varepsilon})\Big[p(1 + e^{-\Delta^2/\varepsilon}) - \tfrac{1}{2}\Big] \geq \varepsilon(1 - e^{-\Delta^2/\varepsilon}) \cdot \tfrac{2m-1}{2(2m+1)},$$

where the last step uses $p(1 + e^{-x}) - \frac{1}{2} \geq p - \frac{1}{2} = \frac{2m-1}{2(2m+1)} > 0$. Combining the different inequalities, we conclude

$$D_\varepsilon(\alpha_L, \hat{\mu}^{\mathcal{D}}) - D_\varepsilon(\alpha_E, \hat{\mu}^{\mathcal{D}}) \geq \frac{2m-1}{2(2m+1)}\left(\Delta^2 + \varepsilon(1 - e^{-\Delta^2/\varepsilon}) - \varepsilon \log \frac{2}{1 + e^{-4\Delta^2/\varepsilon}}\right).$$

Setting $h(t) = t + 1 - e^{-t} - \log\left(\frac{2}{1+e^{-4t}}\right)$, the expression is proportional to $\varepsilon h(\Delta^2/\varepsilon)$. We can establish $h(0) = 0$, $h'(0) = 0$, $h'(t) > 0\, \forall x > 0$, so the expression is positive for $\Delta > 0$. For any value of ε, the divergence of the extreme configuration is lower than that of the left or central configurations, concluding the proof. □

E Analysis of the vertex descent algorithms

In this section, we analyze the performance of Algorithm 1, including the distribution mixture and the analysis of the bias it introduces. We first derive a directional Lipschitz bound on D_ε., which we then use to state the main result on the algorithm performance.

Lemma 4 (Swap-Directional Lipschitz constant)**.** *Let $C \in \mathbb{R}^{n\times n}$ be the symmetric cost matrix computed from the pairwise distances of points in the dataset with $\|C\|_\infty = M$. Let $\lambda \in (0, 1)$ be a mixture parameter and define the mixed distribution as*

$$\alpha^\lambda_{\mathbf{x}} = (1 - \lambda)\hat{\mu}_{\mathbf{x}} + \lambda\hat{\mu}^{\mathcal{D}} \quad \textit{with weight vector} \quad \boldsymbol{a}^\lambda = \frac{1-\lambda}{B}\mathbf{x} + \frac{\lambda}{n}\mathbf{1}.$$

Fix any $\mathbf{x} \in \mathcal{V}_{n,B}$, $\mathbf{x}' \in \mathrm{swap}(\mathbf{x})$ and set $\boldsymbol{d} = \mathbf{x}' - \mathbf{x}$. Define the univariate function

$$\phi_{\boldsymbol{d},\lambda}(t) = D_\varepsilon(\alpha^\lambda(\mathbf{x} + t\boldsymbol{d}), \hat{\mu}^{\mathcal{D}}).$$

There is a constant $L^{\mathrm{swap}}_{\varepsilon,\lambda}$ depending on ε and the mixture parameter λ such that $\phi''_{\boldsymbol{d},\lambda}(t) \leq L^{\mathrm{swap}}_{\varepsilon,\lambda}$.

Proof. We first show that the second derivative is a Hessian bilinear form and then that the Hessian is bounded by the Lipschitz constant of the Sinkhorn potentials.

Let $\boldsymbol{a}^\lambda(\mathbf{x} + t\boldsymbol{d})$ be the vector of weights along the path from $\mathbf{x}$ to $\mathbf{x}'$, which is an affine function of t with $\partial\boldsymbol{a}^\lambda/\partial t = (1 - \lambda)\boldsymbol{d}/B$, and define $h : \Delta_n \to \mathbb{R}_+$, $h(\boldsymbol{a}) = D_\varepsilon(\sum_{i\in[n]} \boldsymbol{a}_i\delta_{\boldsymbol{y}_i}, \hat{\mu}^{\mathcal{D}})$.

By the chain rule,

$$\phi''_{\boldsymbol{d},\lambda}(t) = \frac{(1-\lambda)^2}{B^2}\boldsymbol{d}^\top\nabla^2 h(\boldsymbol{a}^\lambda(\mathbf{x} + t\boldsymbol{d}))\boldsymbol{d}.$$

We recall from Equation (2) that the gradient of h w.r.t. the weights is given by $\nabla h(\boldsymbol{a}) = \mathbf{f}^*_{\boldsymbol{a}} - \mathbf{p}^*_{\boldsymbol{a}}$.

Define the Sinkhorn marginal Lipschitz constant

$$L^{\mathrm{f,p}}_\varepsilon(\lambda_0) := \sup_{\substack{\boldsymbol{a},\boldsymbol{a}'\in\Delta_n \\ \min_k \boldsymbol{a}_k, \boldsymbol{a}'_k \geq \lambda_0}} \frac{\|\mathbf{f}^\star_{\boldsymbol{a}} - \mathbf{f}^\star_{\boldsymbol{a}'}\|_\infty + \|\mathbf{p}^\star_{\boldsymbol{a}} - \mathbf{p}^\star_{\boldsymbol{a}'}\|_\infty}{\|\boldsymbol{a} - \boldsymbol{a}'\|_1},$$

which is well-defined and finite, from the implicit function theorem on the system of equations defining the fixed point of the Sinkhorn algorithm see, e.g., in Nutz & Wiesel (2023); Deligiannidis et al. (2024); Carlier et al. (2024) for results on the stability of Sinkhorn potentials w.r.t. their

marginals up to normalization. This quantity depends on $\exp(M/\varepsilon)$ which is not problematic in our setup considering large regularization and a small diameter.

The swap direction has only two nonzero entries, thus

$$\boldsymbol{d}^\top \nabla^2 h(\boldsymbol{a}) d = H_{ii} + H_{jj} - 2H_{ij}$$
$$|\boldsymbol{d}^\top \nabla^2 h(\boldsymbol{a}) \boldsymbol{d}| \leq 4 \max_{(i,j)\in[n]\times[n]} |H_{ij}|$$

with H the Hessian of h at $\boldsymbol{a}$. Each entry is bounded in absolute value by the Lipschitz constant of ∇h, and using $\lambda_0 = \lambda/n$, the minimum weight from mixture, we have

$$\boldsymbol{d}^\top \nabla^2 h(\boldsymbol{a}) d \leq 4L_\varepsilon^{\mathrm{f,p}}(\lambda/n)$$
$$\phi''_{\boldsymbol{d},\lambda}(t) \leq L_{\varepsilon,\lambda}^{\mathrm{swap}} = \frac{4(1-\lambda)^2 L_\varepsilon^{\mathrm{f,p}}(\lambda/n)}{B^2}.$$

□

Lemma 5 (Potential bound on vertices)**.** *Let $C \in \mathbb{R}^{n\times n}$ be the symmetric cost matrix with $\|C\|_\infty \leq M$, then the following bound on the optimal potentials holds at any $\mathbf{x}$ in the relative interior of $\Delta_{n,B}$, under the normalization $\sum_i \mathbf{f}_i = 0$:*

$$\|\mathbf{f} - \mathbf{p}\|_\infty \leq 2M. \tag{22}$$

Proof. Using the expression of the optimal plan $\boldsymbol{\pi}$ of the EOT problem from the potentials:

$$\boldsymbol{\pi}_{ij} = e^{(\mathbf{f}_i + \mathbf{g}_j - C_{ij})/\varepsilon} \frac{\mathbf{x}_i}{nB} \; \forall i \in [n], j \in [n],$$

the fact that $\mathbf{x}_i > 0 \, \forall i \in [n]$, and the marginal constraint $\sum_{j\in[n]} \boldsymbol{\pi}_{ij} = \frac{\mathbf{x}_i}{B}$ results in

$$\sum_{j\in[n]} e^{(\mathbf{f}_i + \mathbf{g}_j - C_{ij})/\varepsilon} \frac{\mathbf{x}_i}{nB} = \frac{\mathbf{x}_i}{B} \quad \forall i \in [n]$$
$$e^{\mathbf{f}_i/\varepsilon} \sum_{j\in[n]} e^{(\mathbf{g}_j - C_{ij})/\varepsilon} = n \quad \forall i \in [n]$$
$$\mathbf{f}_i - \mathbf{f}_{i'} = -\varepsilon \log \frac{\sum_j e^{(\mathbf{g}_j - C_{ij})/\varepsilon}}{\sum_j e^{(\mathbf{g}_j - C_{i'j})/\varepsilon}}$$
$$\max_{i,i'} |\mathbf{f}_{i'} - \mathbf{f}_i| \leq M$$

where the last inequality uses cancellation of the $e^{\mathbf{g}_j/\varepsilon}$ terms and $C_{ij} \leq M$. Under the normalization $\sum_i \mathbf{f}_i = 0$ and applying the same derivation to the self-transport problem for $\mathbf{p}$, we obtain the desired inequality (22) □

Theorem 2 (Suboptimality of Swap Descent on Sinkhorn Divergence)**.** *Consider the execution of the swap algorithm with mixture parameter $\lambda \in (0,1)$ and denote by $\bar{\mathbf{x}}$ its output. Consider $\mathbf{x}^*$ a minimizer of the selection problem* (3)*, and $k^* \leq B$ the number of swaps required from $\bar{\mathbf{x}}$ to $\mathbf{x}^*$. There exists a smoothness constant $L_{\varepsilon,\lambda}^{\mathrm{swap}}$ depending on ε and λ, such that the following suboptimality bound holds:*

$$D_\varepsilon(\hat{\mu}_{\bar{\mathbf{x}}}, \hat{\mu}^{\mathcal{D}}) - D_\varepsilon(\hat{\mu}_{\mathbf{x}^*}, \hat{\mu}^{\mathcal{D}}) \leq \frac{k^* L_{\varepsilon,\lambda}^{\mathrm{swap}}}{2} + 4\lambda \left(1 - \frac{B}{n}\right) \|C\|_\infty . \tag{6}$$

Proof. The swap descent algorithm terminates when the new iterate was already visited in a previous iteration, creating a cycle in the vertex graph of the hypersimplex. On this cycle, let $\mathbf{x}^{(c)}$ be the minimizer of the function and $\mathbf{x}^{(c+1)}$ be its immediate successor in the cycle, and set $\boldsymbol{d}_c = \mathbf{x}^{(c+1)} - \mathbf{x}^{(c)}$. We will denote $M = \|C\|_\infty$ throughout the proof.

Define $f_\lambda : \Delta_{n,B} \to \mathbb{R}_+$, $f_\lambda(\mathbf{x}) = D_\varepsilon(\alpha_\lambda(\mathbf{x}), \hat{\mu}^{\mathcal{D}})$, which is convex, and note that $f_\lambda(\mathbf{x} + t\boldsymbol{d}) = \phi_{\boldsymbol{d},\lambda}(t)$ if $\boldsymbol{d}$ is a swap direction, with $\phi_{\boldsymbol{d},\lambda}$ defined as in Lemma 4. The Lipschitz smoothness along $\boldsymbol{d}_c$ yields

$$f_\lambda(\mathbf{x}^{(c+1)}) \leq f_\lambda(\mathbf{x}^{(c)}) + \left\langle \nabla f_\lambda(\mathbf{x}^{(c)}), \boldsymbol{d}_c \right\rangle + \frac{L_{\varepsilon,\lambda}^{\mathrm{swap}}}{2}.$$

Using the cycle-optimality of $\mathbf{x}^{(c)}$, and re-arranging,

$$0 \leq f_\lambda(\mathbf{x}^{(c+1)}) - f_\lambda(\mathbf{x}^{(c)}) \leq \left\langle \nabla f_\lambda(\mathbf{x}^{(c)}), \boldsymbol{d}_c \right\rangle + \frac{L_{\varepsilon,\lambda}^{\text{swap}}}{2}$$
$$\left\langle \nabla f_\lambda(\mathbf{x}^{(c)}), \boldsymbol{d}_c \right\rangle \geq -\frac{L_{\varepsilon,\lambda}^{\text{swap}}}{2}.$$

The last inequality also holds for any swap direction by optimality of $\boldsymbol{d}_c$. By convexity, we have

$$f_\lambda(\mathbf{x}^{(c)}) - f_\lambda(\mathbf{x}^*) \leq -\left\langle \nabla f_\lambda(\mathbf{x}^{(c)}), \mathbf{x}^* - \mathbf{x}^{(c)} \right\rangle$$

Decompose $\mathbf{x}^*$ into $k^* \leq B$ swaps from $\mathbf{x}^{(c)}$:

$$\mathbf{x}^* - \mathbf{x}^{(c)} = \sum_{l=1}^{k^*} \boldsymbol{d}_l$$

with each $\boldsymbol{d}_l$ corresponding to a single swap from the support of $\mathbf{x}^{(c)}$ to the support of $\mathbf{x}^*$. We have

$$f_\lambda(\mathbf{x}^{(c)}) - f_\lambda(\mathbf{x}^*) \leq \sum_{l=1}^{k^*} -\left\langle \nabla f_\lambda(\mathbf{x}^{(c)}), \boldsymbol{d}_l \right\rangle \leq k^* \frac{L_{\varepsilon,\lambda}^{\text{swap}}}{2}. \tag{23}$$

Conservatively, we can bound $k^* \leq B$ for an instance-independent guarantee.

Finally, we bound the bias introduced by the mixture parameter λ. We start by noting that $D_\varepsilon(\alpha_{\mathbf{x}}^\lambda, \hat{\mu}^{\mathcal{D}})$ is a convex decreasing function of lambda on $[0, 1]$. Using this fact with the expression of the gradient of the Sinkhorn divergence with respect to the weights, and the continuity of $D_\varepsilon(\cdot, \hat{\mu}^{\mathcal{D}})$,

$$D_\varepsilon(\alpha_{\mathbf{x}}^\lambda, \hat{\mu}^{\mathcal{D}}) - D_\varepsilon(\alpha_{\mathbf{x}}^0, \hat{\mu}^{\mathcal{D}}) = \lim_{\tau \to 0} \lambda \int_\tau^1 \langle \mathbf{f}_t - \mathbf{p}_t, \hat{\mu}^{\mathcal{D}} - \alpha_{\mathbf{x}}^0 \rangle \, dt$$
$$|D_\varepsilon(\alpha_{\mathbf{x}}^\lambda, \hat{\mu}^{\mathcal{D}}) - D_\varepsilon(\alpha_{\mathbf{x}}^0, \hat{\mu}^{\mathcal{D}})| \leq \lim_{\tau \to 0} \lambda \left\| \hat{\mu}^{\mathcal{D}} - \alpha_{\mathbf{x}}^\tau \right\|_{\text{TV}} \left\| \mathbf{f}_\tau - \mathbf{p}_\tau \right\|_\infty$$
$$|D_\varepsilon(\alpha_{\mathbf{x}}^\lambda, \hat{\mu}^{\mathcal{D}}) - D_\varepsilon(\alpha_{\mathbf{x}}^0, \hat{\mu}^{\mathcal{D}})| \leq 2\lambda M \left\| \hat{\mu}^{\mathcal{D}} - \alpha_{\mathbf{x}}^0 \right\|_{\text{TV}}$$

with $\mathbf{f}_t$, $\mathbf{p}_t$, $\mathbf{f}$, $\mathbf{p}$ the potentials of the Sinkhorn divergence evaluated at $\alpha_{\mathbf{x}}^t$, $\alpha_{\mathbf{x}}^0$ respectively. The first step uses the fact that the divergence is convex and monotonically decreasing in t, such that its derivative w.r.t. and attains its greatest absolute value at $t = \tau$, together with Hölder's inequality. Since $\alpha_{\mathbf{x}}^0$ is a measure supported on B points, $\left\| \hat{\mu}^{\mathcal{D}} - \alpha_{\mathbf{x}}^0 \right\|_{\text{TV}} = 2\left(1 - \frac{B}{n}\right)$. We can therefore bound the mixing bias by

$$|D_\varepsilon(\alpha_{\mathbf{x}}^\lambda, \hat{\mu}^{\mathcal{D}}) - D_\varepsilon(\alpha_{\mathbf{x}}^0, \hat{\mu}^{\mathcal{D}})| \leq 4\lambda M \left(1 - \frac{B}{n}\right). \tag{24}$$

We can combine inequality (23) with the control of the mixture bias (24) applied at $\bar{\mathbf{x}}$, together with the fact that the bias is non-positive at $\mathbf{x}^*$ and can be dropped from the inequality, providing the desired result (6). □

Empirically, we perform multiple runs on a synthetic dataset of 500 points and 2 classes (in order to be able to compute the true minimizer by complete enumeration of the hypersimplex). We present the results in Table 5, where the gap is that of **FW-Swap**, Rand-Swap is the swap algorithm initialized with a random vertex, and the random subset is the average divergence of vertices.

Table 5: Comparison of the objective values obtained by different subset selections over three runs.

Run	True Minimum	FW-Swap	Gap	Rand-Swap	Random Subset
1	0.02316	0.02549	0.00233	0.04523	0.06378
2	0.02522	0.02794	0.00272	0.04341	0.14044
3	0.02436	0.02676	0.00239	0.03632	0.12600

We see a small gap between the true minimum and the one obtained from **FW-Swap**, proving the effectiveness of the algorithm. Furthermore, Table 5 shows the effect of the FW initialization compared to the random one as it outputs closer values to the true minimum.

F COMPARISON OF THE SELF-SUPERVISED LEARNING REPRESENTATION

As mentioned, the choice of the feature space heavily impacts the results. In fact, when comparing a traditional SSL model, SimCLR, to DINOv3, we notice a difference across all methods in performance. In particular, for **FW-Swap**, who relies on the geometry of this feature space. We provide below in Tables 6 and 7 the results obtained for STL-10 and CIFAR-10 when training the active learning model on features extracted from SimCLR.

Table 6: Test accuracy (mean $\pm$ std) across active learning methods and budgets on STL-10 with feature space defined by SimCLR.

Method	10	20	30	40	50	60
FW-Swap	$\mathbf{0.521 \pm 0.018}$	$\mathbf{0.569 \pm 0.008}$	0.611 ± 0.008	0.613 ± 0.020	0.652 ± 0.012	0.625 ± 0.006
MaxHerding	0.462 ± 0.042	0.547 ± 0.027	0.605 ± 0.032	$\mathbf{0.642 \pm 0.032}$	$\mathbf{0.674 \pm 0.021}$	0.679 ± 0.021
TypiClust	0.483 ± 0.035	0.553 ± 0.047	0.614 ± 0.020	0.637 ± 0.029	0.664 ± 0.025	$\mathbf{0.691 \pm 0.016}$
Medoid	0.440 ± 0.062	0.566 ± 0.057	$\mathbf{0.635 \pm 0.021}$	0.638 ± 0.031	0.669 ± 0.026	0.683 ± 0.013
k-Center	0.444 ± 0.037	0.568 ± 0.028	0.612 ± 0.037	0.631 ± 0.027	0.673 ± 0.020	0.688 ± 0.015
ProbCover	0.432 ± 0.010	0.483 ± 0.006	0.574 ± 0.003	0.598 ± 0.012	0.653 ± 0.008	0.679 ± 0.005
Wasserstein	0.402 ± 0.047	–	–	–	–	–
Random	0.383 ± 0.051	0.514 ± 0.045	0.579 ± 0.025	0.618 ± 0.040	0.659 ± 0.024	0.670 ± 0.028

The Wasserstein method is dashed after $B = 20$ in Table 6 because each run had to be terminated after 3 hours, as reported in the original paper.

Table 7: Test accuracy (mean $\pm$ std) across active learning methods and budgets on CIFAR-10 with feature space defined by SimCLR.

Method	10	20	30	40	50	60
FW-Swap	$\mathbf{0.537 \pm 0.016}$	0.606 ± 0.018	0.639 ± 0.010	0.683 ± 0.020	$\mathbf{0.744 \pm 0.009}$	0.749 ± 0.008
MaxHerding	0.468 ± 0.027	0.601 ± 0.023	0.689 ± 0.008	0.704 ± 0.014	0.734 ± 0.011	$\mathbf{0.763 \pm 0.007}$
TypiClust	0.477 ± 0.040	0.553 ± 0.035	0.639 ± 0.034	0.668 ± 0.028	0.693 ± 0.041	0.732 ± 0.012
Medoid	0.518 ± 0.044	$\mathbf{0.630 \pm 0.043}$	$\mathbf{0.714 \pm 0.033}$	$\mathbf{0.723 \pm 0.020}$	0.740 ± 0.025	0.757 ± 0.022
k-Center	0.444 ± 0.059	0.610 ± 0.062	0.670 ± 0.037	0.715 ± 0.037	0.723 ± 0.025	0.741 ± 0.017
ProbCover	0.454 ± 0.014	0.561 ± 0.012	0.633 ± 0.011	0.670 ± 0.026	0.721 ± 0.010	0.746 ± 0.005
Random	0.384 ± 0.073	0.531 ± 0.044	0.604 ± 0.013	0.631 ± 0.042	0.684 ± 0.033	0.732 ± 0.018

G DETAILED EXPERIMENTAL RESULTS

This section presents the tables corresponding to the experiments of Section 4. Tables 8 to 10 show the detailed results of Figures 2 to 4 respectively.

Table 8: Test accuracy (mean $\pm$ std) across active learning methods and budgets on STL-10.

Method	10	20	30	40	50	60
FW-Swap	$\mathbf{0.699 \pm 0.008}$	$\mathbf{0.768 \pm 0.011}$	$\mathbf{0.829 \pm 0.012}$	0.852 ± 0.016	0.879 ± 0.008	0.906 ± 0.004
MaxHerding	0.612 ± 0.033	0.719 ± 0.037	0.813 ± 0.023	0.846 ± 0.029	0.890 ± 0.018	0.899 ± 0.018
TypiClust	0.617 ± 0.030	0.735 ± 0.031	0.805 ± 0.038	0.851 ± 0.023	0.890 ± 0.014	$\mathbf{0.911 \pm 0.014}$
Medoid	0.606 ± 0.026	0.714 ± 0.058	0.797 ± 0.035	0.841 ± 0.025	0.869 ± 0.042	0.902 ± 0.022
k-Center	0.648 ± 0.019	0.749 ± 0.029	0.809 ± 0.030	$\mathbf{0.861 \pm 0.018}$	$\mathbf{0.899 \pm 0.011}$	0.910 ± 0.006
ProbCover	0.668 ± 0.015	0.755 ± 0.018	0.750 ± 0.015	0.739 ± 0.022	0.743 ± 0.023	0.724 ± 0.013
Random	0.471 ± 0.060	0.603 ± 0.033	0.729 ± 0.047	0.750 ± 0.042	0.797 ± 0.043	0.854 ± 0.026

Table 9: Test accuracy (mean $\pm$ std) across active learning methods and budgets on SVHN.

Method	10	20	30	40	50	60
FW-Swap	$\mathbf{0.307 \pm 0.002}$	$\mathbf{0.496 \pm 0.004}$	$\mathbf{0.497 \pm 0.003}$	$\mathbf{0.576 \pm 0.004}$	0.577 ± 0.002	$\mathbf{0.648 \pm 0.003}$
MaxHerding	0.110 ± 0.002	0.307 ± 0.003	0.468 ± 0.003	0.516 ± 0.002	$\mathbf{0.598 \pm 0.002}$	0.606 ± 0.004
TypiClust	0.283 ± 0.010	0.402 ± 0.026	0.447 ± 0.048	0.509 ± 0.042	0.563 ± 0.029	0.599 ± 0.053
Medoid	0.245 ± 0.038	0.353 ± 0.076	0.437 ± 0.055	0.563 ± 0.057	0.575 ± 0.037	0.635 ± 0.047
k-Center	0.156 ± 0.043	0.226 ± 0.034	0.278 ± 0.038	0.326 ± 0.034	0.376 ± 0.036	0.416 ± 0.038
ProbCover	0.256 ± 0.004	0.309 ± 0.002	0.369 ± 0.002	0.430 ± 0.005	0.477 ± 0.003	0.530 ± 0.003
Random	0.261 ± 0.065	0.352 ± 0.056	0.428 ± 0.050	0.451 ± 0.039	0.539 ± 0.064	0.600 ± 0.051

Method	80	100
FW-Swap	$\mathbf{0.721 \pm 0.002}$	$\mathbf{0.752 \pm 0.002}$
MaxHerding	0.667 ± 0.002	0.723 ± 0.003
TypiClust	0.666 ± 0.025	0.708 ± 0.015
Medoid	0.693 ± 0.031	0.715 ± 0.028
k-Center	0.502 ± 0.037	0.581 ± 0.025
ProbCover	0.622 ± 0.002	0.691 ± 0.003
Random	0.641 ± 0.027	0.678 ± 0.027

Table 10: Test accuracy (mean $\pm$ std) across active learning methods and budgets on GasHisSDB.

Method	2	4	6	8	10
FW-Swap	$\mathbf{0.665 \pm 0.003}$	0.698 ± 0.004	$\mathbf{0.820 \pm 0.002}$	$\mathbf{0.809 \pm 0.002}$	$\mathbf{0.838 \pm 0.002}$
MaxHerding	0.606 ± 0.001	0.605 ± 0.002	0.609 ± 0.003	0.723 ± 0.004	0.769 ± 0.002
TypiClust	0.632 ± 0.008	0.691 ± 0.058	0.734 ± 0.032	0.730 ± 0.061	0.711 ± 0.077
Medoid	0.598 ± 0.116	0.693 ± 0.067	0.754 ± 0.074	0.806 ± 0.061	0.803 ± 0.047
k-Center	0.601 ± 0.070	0.668 ± 0.061	0.715 ± 0.075	0.754 ± 0.044	0.770 ± 0.040
ProbCover	0.505 ± 0.008	$\mathbf{0.718 \pm 0.005}$	0.695 ± 0.005	0.704 ± 0.007	0.688 ± 0.005
Random	0.570 ± 0.121	0.622 ± 0.079	0.702 ± 0.105	0.685 ± 0.107	0.742 ± 0.076

The detailed results reported in the tables above further support our observations that **FW-Swap** consistently performs strongly at low budgets. It achieves the highest test accuracy in most cases and remains competitive with the strongest-performing methods when it does not.