Wasserstein gradient flow for optimal probability measure decomposition

Jiangze Han,* Christopher Thomas Ryan,† Xin T. Tong‡

May 16, 2024

Abstract

We examine the infinite-dimensional optimization problem of finding a decomposition of a probability measure into K probability sub-measures to minimize specific loss functions inspired by applications in clustering and user grouping. We analytically explore the structures of the support of optimal sub-measures and introduce algorithms based on Wasserstein gradient flow, demonstrating their convergence. Numerical results illustrate the implementability of our algorithms and provide further insights.

Keywords: probability measure decomposition, Wasserstein gradient flow, optimal transport, infinite-dimensional optimization

1 Introduction

With the rapid advancement of AI, automated algorithms are increasingly being used to solve routine problems. Particularly intriguing are the applications of AI in social organizations, which have the potential to benefit both private and public sectors. These applications include the organization of markets, allocation of resources, and mechanism design, among others (Agrawal et al. 2023, Chen et al. 2021, Dai and Jordan 2021, Niazadeh et al. 2023, Zhalechian et al. 2022). This paper studies a new problem of how to decompose a population of customers or clients into groups to optimize a generic quantitive criterion.

Consider the following probability measure decomposition problem. Later, we will show how this problem can arise in applications. Individuals in a population are represented by their feature vectors $\mathbf{x} \in \mathbb{R}^d$. Feature vectors are distributed according to a probability distribution π. Let $\mathcal{P}_2(\mathbb{R}^d)$ be the space of probability measures defined on $\mathbb{R}^d$ with finite second moment; that is, $\mathcal{P}_2(\mathbb{R}^d) = \{\mu : \int_{\mathbb{R}^d} \|\mathbf{x}\|^2 d\mu(\mathbf{x}) < \infty\}$. When a measure $\mu \in \mathcal{P}_2(\mathbb{R}^d)$ is absolutely continuous with respect to the Lebesgue measure, we use the same symbol μ to represent the measure's associated probability density function. We define a decomposition of π as follows.

Definition 1 (Probability measure decomposition)**.** Given a probability measure $\pi \in \mathcal{P}_2(\mathbb{R}^d)$, we say the vector $\boldsymbol{\mu} \doteq (\mu_1, \mu_2, \ldots, \mu_K) \in \mathcal{P}_2(\mathbb{R}^d)^{\otimes K}$ of probability measures with weight vector

*University of British Columbia, email: jiangze.han@sauder.ubc.ca
†University of British Columbia, email: chris.ryan@sauder.ubc.ca
‡National University of Singapore, email: xin.t.tong@nus.edu.sg

$\mathbf{p} = (p_1, \ldots, p_K) \in \mathbb{R}^{\otimes K}$ is a *decomposition* of π, if $(\boldsymbol{\mu}, \mathbf{p}) \in \mathcal{P}_\pi$, where

$$\mathcal{P}_\pi \doteq \left\{ (\boldsymbol{\mu}, \mathbf{p}) : \sum_{k\in[K]} p_k = 1, p_k \geq 0, \sum_{k\in[K]} p_k \mu_k = \pi \right\} \tag{1}$$

and $[K] \doteq \{1, 2, \ldots, K\}$. The equality $\sum_{k\in[K]} p_k \mu_k = \pi$ holds in duality with the space $C_c^\infty(\mathbb{R}^d)$ of compactly supported smooth (i.e., has infinitely many derivatives) functions; that is, for all $f \in C_c^\infty(\mathbb{R}^d)$,

$$\sum_{k\in[K]} \int_{\mathbb{R}^d} f(x) d\mu_k(x) = \int_{\mathbb{R}^d} f(x) d\pi(x).$$

Intuitively, we decompose the population (distributed according to π) of feature vectors into K sub-populations (distributed according to $\mu_1, \ldots, \mu_K$). Within each sub-population $k \in [K]$, individual features are distributed according to probability measure μ_k. The population weight of the whole population is normalized to be 1. Each sub-population $k \in [K]$ has weight p_k.

Among all decompositions of the feature distribution π, we seek one that minimizes (i) a weighted sum of distribution loss function $L : \mathcal{P}_2(\mathbb{R}^d) \to \mathbb{R}$ associated with feature distribution μ_k of each sub-population, and (ii) a weight loss function $R : \mathbb{R} \to \mathbb{R}$ associated with the population weight p_k of each sub-population. The purpose of this loss is to penalize a sub-population with a small weight, which can be impractical for different reasons detailed in the examples below. Formally, we consider the following optimal decomposition problem.

Problem 1 (Optimal decomposition problem)**.** Let $L : \mathcal{P}_2(\mathbb{R}^d) \to \mathbb{R}$ be a distribution loss function and $R : \mathbb{R} \to \mathbb{R}$ a weight loss function. Given a target of K sub-populations in a population with distribution π, solve

$$\min_{(\boldsymbol{\mu},\mathbf{p})\in\mathcal{P}_\pi} F(\boldsymbol{\mu}, \mathbf{p}), \quad F(\boldsymbol{\mu}, \mathbf{p}) \doteq \sum_{k\in[K]} (p_k L(\mu_k) + R(p_k)), \tag{2}$$

where the feasible region $\mathcal{P}_\pi$ is defined in (1).

We consider the following family of distribution loss functions L.

Definition 2 (Coupled loss function)**.** We say $L : \mathcal{P}_2(\mathbb{R}^d) \to \mathbb{R}$ is a *coupled loss function* if

$$L(\mu) = \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} \ell(\mathbf{x}, \mathbf{y}) d\mu(\mathbf{x}) d\mu(\mathbf{y})$$

for some continuously differentiable function $\ell : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ satisfying

$$|\ell(\mathbf{z}, \mathbf{x}) - \ell(\mathbf{z}, \mathbf{y})| \leq \|\mathbf{x} - \mathbf{y}\|$$

for all $\mathbf{x}, \mathbf{y}, \mathbf{z} \in \mathbb{R}^d$. We call ℓ the *kernel* of L.

Definition 3 (Weight loss function)**.** For some $\theta, \beta > 0$, define $R : (0, 1) \to \mathbb{R}$ as

$$R(p) \doteq \frac{\theta}{p^\beta}.$$

We present two applications of this general setup.

Example 1 (League design with Elo rating system (Elo and Sloan 1978))**.** In many competition-based online games, players are grouped into different "leagues" based on their skill levels, and only players from the same league can compete with each other. League design aims to create competitive gaming environments where players are not overwhelmed by strong opponents or bored by weaker ones. One way to quantify the skill level and competitiveness of games is the

Elo-type system.[1] For simplicity, we focus on one-on-one competitions, similar to chess. In the Elo-type system, each player is given a skill level $x \in (0, \infty)$ (sometimes called Elo score). The probability of winning for a player with skill level x against a player with skill level y is taken to be $x/(x+y)$.[2] A game is deemed more competitive as each player's win rate gets closer to 50%. A common practice is to minimize the difference of each player's winning probability with 50%. For example, Simonov et al. (2023) show in their study using data from the game streaming platform Twitch that the expected game length and viewership can be increased by making the round-win probabilities of games closer to a balanced distribution of 50%-50%.

Suppose the skill-level distribution of all players has density π, and the goal is to decompose players into K leagues. Suppose in each league $k \in [K]$, players arrive to join a game according to a Poisson process with arrival rate p_k (i.e., the expected waiting time for players in sub-population k is $1/p_k$). Our decomposition aims to maximize the competitiveness of games and minimize the waiting time of each league. This can be achieved by solving (2) with weight loss function $R(p) = 1/p$ and distribution loss function

$$\begin{aligned} L(\mu) &= \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} \left(\frac{x}{x+y} - \frac{1}{2} \right)^2 d\mu(x) d\mu(y) \\ &= \frac{1}{2} \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} \frac{x^2+y^2}{(x+y)^2} d\mu(x) d\mu(y) - \frac{1}{4} \end{aligned} \tag{3}$$

Note that L is a coupled loss function with kernel $\ell(x, y) = \frac{1}{2} \frac{x^2+y^2}{(x+y)^2} - \frac{1}{4}$. We call this distribution loss function L the *Elo loss*.

Example 2 (Generalized clustering)**.** In clustering, the goal is to generate sub-populations according to specific criteria. As in our base setup, suppose a population's feature vectors $\mathbf{x} \in \mathbb{R}^d$ are distributed according to $\pi \in \mathcal{P}_2(\mathbb{R}^d)$. This means that the feature vector X of a randomly sampled individual in the population is a random variable with law π. Suppose a designer aims to decompose this population into sub-populations to maximize a sense of similarity in certain feature dimensions while concurrently maximizing a sense of diversity in other dimensions within each sub-population. Accordingly, we can define a loss function L as follows. Let W be a diagonal matrix with nonzero diagonal entries. Define a distribution loss L by

$$\begin{aligned} L(\mu) &= \int_{\mathbb{R}^d} \langle \mathbf{x} - \mathbb{E}_\mu[X], W(\mathbf{x} - \mathbb{E}_\mu[X]) \rangle d\mu(\mathbf{x}) \\ &= \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} \langle \mathbf{x} - \mathbf{y}, W(\mathbf{x} - \mathbf{y}) \rangle d\mu(\mathbf{x}) d\mu(\mathbf{y}), \end{aligned} \tag{4}$$

where $\langle \mathbf{x}, \mathbf{y} \rangle$ denotes the standard inner product in $\mathbb{R}^d$. Note that L is a coupled loss function with the kernel $\ell(\mathbf{x}, \mathbf{y}) = \langle \mathbf{x} - \mathbf{y}, W(\mathbf{x} - \mathbf{y}) \rangle$. We call this distribution loss function L the *variance loss*.

The matrix W specifies the weights assigned to each dimension. When $W_{i,i} > 0$, this minimizes the dissimilarity of features in dimension i. Conversely, when $W_{i,i} < 0$, this maximizes the diversity of features in dimension i. Notably, if W is the identity matrix, then $L(\mu)$ corresponds to the trace of the covariance matrix. In this special case, we decompose the distribution π into

[1] For descriptions about Elo rating system, please see https://en.wikipedia.org/wiki/Elo_rating_system.

[2] In other variants of the Elo system, people use $\log(x)/\alpha$ to represent skill level for some game specific parameter α, which is equivalent to our setting by a change-of-variable argument.

K sub-distributions μ_k to minimize the variance of each sub-distribution μ_k.

To create sub-populations with sufficiently large sizes, we can also impose penalties for selecting smaller sub-populations. This can be achieved, for example, by setting the weight loss function R to $R(p) = 1/p$.

In certain scenarios, the population size p_k of each sub-population $k \in [K]$ is predefined. For example, to make the decomposition more balanced, one can require $p_k = 1/K$ for all $k \in [K]$. In this scenario, we can define the set of feasible decompositions for a given $\mathbf{p}$:

$$\mathcal{P}_{\pi,\mathbf{p}} \doteq \{\boldsymbol{\mu} \in \mathcal{P}_2(\mathbb{R}^d)^{\otimes K} : \sum_{k\in[K]} p_k \mu_k = \pi\}. \tag{5}$$

In this case, we consider the following simplified problem.

Problem 2 (Optimal decomposition problem with specified weights)**.** Let $L : \mathcal{P}_2(\mathbb{R}^d) \to \mathbb{R}$ be a given distribution loss function. Given the number K of sub-populations, sub-population weights $\mathbf{p} = (p_1, \ldots, p_K)$, and feature distribution π, solve

$$\min_{\boldsymbol{\mu}\in\mathcal{P}_{\pi,\mathbf{p}}} F_{\mathbf{p}}(\boldsymbol{\mu}), \quad F_{\mathbf{p}}(\boldsymbol{\mu}) \doteq \sum_{k\in[K]} p_k L(\mu_k). \tag{6}$$

Problem 2 can be thought of a sub-problem of Problem 1. Recall that we denote the objective functions in Problem 1 and Problem 2 by $F(\boldsymbol{\mu}, \mathbf{p})$ and $F_{\mathbf{p}}(\boldsymbol{\mu})$, respectively. Problem 1 can be decomposed as follows

$$\min_{(\boldsymbol{\mu},\mathbf{p})\in\mathcal{P}_\pi} F(\boldsymbol{\mu}, \mathbf{p}) = \min_{\mathbf{p}\in\mathcal{S}} \min_{\boldsymbol{\mu}\in\mathcal{P}_{\pi,\mathbf{p}}} F_{\mathbf{p}}(\boldsymbol{\mu}),$$

where $\mathcal{S} = \{\mathbf{p} : \sum_{k\in[K]} p_k = 1, p_k \geq 0\}$. The inner problem in the above decomposition is exactly Problem 2.

1.1 Literature review. In recent years, distribution-oriented optimization has been an active direction in operations research. Its applications span over robust optimization, matching of social networks, and stochastic games; see for examples Bertsimas et al. (2019), Hu et al. (2022), Light and Weintraub (2022). Our paper falls in this general research direction. Our discussion relies on the methodologies of optimal transport theory and Wasserstein gradient flow. Optimal transport theory introduces a metric—known as the Wasserstein metric—on the set of probability measures, effectively transforming it into a metric space called Wasserstein space. This metric enables the development of calculus and geometric concepts, such as gradient and geodesics, on the space of probability measures. The standard references for optimal transport are Santambrogio (2015), Villani (2009). With calculus and geometry defined by optimal transport theory, we can establish the concept of gradient flow on the Wasserstein space. Wasserstein gradient flow has emerged as a popular methodology for tackling optimization problems formulated on the space of probability measures. The standard references are Ambrosio et al. (2005), Santambrogio (2017). Various optimization algorithms leveraging Wasserstein gradient flow have been developed to address diverse optimization problems (Chewi et al. 2020, Chizat and Bach 2018, Salim et al. 2020, Zhang et al. 2018). In contrast to prior research, one of our methodological contributions is introducing a Wasserstein gradient flow tailored to a constrained optimization problem. We demonstrate that, in the limit, the proposed Wasserstein gradient flow converges to a feasible solution satisfying the optimality condition.

We also give two applications of our paper in Examples 1 and 2. In Example 1 and Section 6.2,

we show that our model and algorithm can be applied to address the league design problem. It is a common practice for online games to group players into different groups (called "league" in our paper) (Agarwal and Lorch 2009, Francillette et al. 2013, Manweiler et al. 2011). One important consideration for grouping is to form groups where players share similar skill levels while ensuring that sessions have an adequate number of players (Francillette et al. 2013), which is the major setting we consider.

Another closely related problem is clustering, where N data points, $\mathbf{x}_1, \ldots, \mathbf{x}_N$, need to be split into K groups. The most well known version is K-means clustering problem, which seeks the optimal label $l : [N] \to [K]$ that minimizes

$$L_K(l) \doteq \sum_{k\in[N]} \|\mathbf{x}_k - \mathbf{m}_{l(k)}\|^2, \quad \mathbf{m}_i \doteq \frac{\sum_{k\in[K]} 1_{l(k)=i}\mathbf{x}_k}{\sum_{k\in[K]} 1_{l(k)=i}}.$$

Suppose we view each cluster as a decomposition component. In that case, K-means clustering is similar to the problem in Example 2 with $W = I_d$ and $R = 0$, where I_d is a d by d identity matrix. The main difference is that K-means clustering focuses on handling problems where finitely many data samples are present, while our decomposition problem focuses on the case where the data distribution is available. Given its popularity, algorithmic studies of K-means began in the 1960s (MacQueen 1967) and remain active until today. Interested readers can refer to a recent survey Ikotun et al. (2023) for its development. Due to its combinatorial nature, algorithms for K-means clustering often rely on greedy heuristic arguments. To the best of our knowledge, the associated optimality conditions of these algorithms are seldom available, except for continuous relaxation of K-means (Blömer et al. 2020). This is also a difference between our problem and the classical K-means clustering problem, as we will formulate and prove the associated optimality conditions.

Our paper investigates the problem of optimally decomposing a probability measure into a finite number of probability measures based on a specific objective function. In statistics, a related problem is the identifiability of a finite mixture model. A finite mixture model can be seen as a probability density function that is a convex combination of K probability density functions called "component densities". A mixture model is said to be identifiable if this convex combination can be uniquely determined. For literature in this direction, see Kim and Lindsay (2015), McLachlan et al. (2019), Teicher (1963), Yakowitz and Spragins (1968). Our research problem differs in two key aspects. First, our problem focuses on finding the optimal decomposition of a probability measure that minimizes a specific loss functional among all possible decompositions, whereas "identifiability" concerns the uniqueness of the decomposition. Second, the "identifiability" literature mainly focuses on parametric probability densities such as Gaussian and Gamma distributions, whereas our problem deals with nonparametric probability measures.

The main algorithm we proposed is a constraint controlled gradient flow (CCGF) type of algorithm. The notion of a CCGF was first proposed in Liu et al. (2021) to generate samples with moment constraints. The decomposition problems we study here imposes a very different kind of constraint, which has not been studied before in the literature. Moreover, we provide a construction of CCGF (more details in Remark 1) and explain it in the Euclidean space setting.

1.2 Our contributions. In short, this work's main contributions are threefold. First, we formulate the optimal decomposition problems and discuss the geometric properties of the optimal solution. Second, we develop the constrained Wasserstein optimality condition for the optimal decomposition problem. Third, we design gradient flow-based algorithms that can approximately achieve the optimality conditions and numerically test their efficacy on various problems. To the best of our knowledge, these aspects have not been studied in the literature.

The paper is organized as follows. In Section 2, we prove several structural properties of the optimal solutions to Problems 1 and 2. In Section 3, we introduce the necessary technical ingredients to understand our analysis, including gradient flow, calculus in Wasserstein space, and geodesic geometry of Wasserstein space. In Section 4, we show that existing optimality conditions do not readily apply to our setting, and we propose our own optimality conditions. In Section 5, we develop Wasserstein gradient flow procedures to solve Problems 1 and 2 and show their convergence. In Section 6, we implement Wasserstein gradient flow in the settings described by Examples 1 and 2. Section 7 concludes and points to future research directions. Proofs of all technical results can be found in the appendix.

2 Structural results and interpretation

Before transitioning to our gradient flow design, we take some time to reflect on the type of structures of the solutions to Problems 1 and 2 we hope to explore. In particular, we are interested in the geometric properties we might see arise in optimality. We will see these geometric properties clearly in numerical examples in Section 6.

We first note that $\boldsymbol{\mu}$ defined by $\mu_k = \pi$ for all $k \in [K]$ is a trivial feasible solution to Problems 1 and 2. We do not want this "boring" solution in practice. Hence, a natural question is: to what extent, and in what sense, are the optimal sub-population measures μ_k^* different? For simplicity, we focus on the optimal solution $\boldsymbol{\mu}^*$ to Problem 2 in this section, although the results also hold for the optimal solutions to Problem 1.

We answer this question by analyzing a generalized notion of the "support" (i.e., (δ, c)-interior support explained later) of an optimal solution. First, we check if the "supports" of the μ_k intersect or are disjoint. If the "supports" are indeed disjoint, we can already rule out the trivial solution $\mu_k = \pi$ for all $k \in [K]$. Moreover, in this case, when we decompose the probability density π in Problem 2. We are effectively partitioning the underlying space $\mathbb{R}^d$.

We are also interested in the relative positions of these "supports" in $\mathbb{R}^d$. For example, in the one-dimensional case, is it possible that the "support" of μ_1 is $[0,1] \cup [2,3]$ and the "support" of μ_2 is $(1,2)$? Motivated by this example, we say a collection of sets $\{S_k\}_{k\in[K]}$, where $S_k \subseteq \mathbb{R}^d$ for $k \in [K]$, is *convex in pairs* if $\text{conv}(S_i) \cap S_j = \emptyset$ for any pair $i \neq j \in [K]$. In the previous example, the "support" $[0,1] \cup [2,3]$ of μ_1 and the "support" $(1,2)$ of μ_2 are not convex in pairs. This concept speaks to the practical implementability of the results that arise from our approach. For instance, in Example 1, a league design that is not convex in pairs would be unnatural because it implies grouping high-end and low-end players into one league while placing middle-level players in another. This type of design would be hard for game designers and players to justify.

We formally define our notion of (δ, c)-interior support as follows.

Definition 4 ((δ, c)-interior support)**.** For real numbers $\delta, c > 0$, $S(\delta, c) \subseteq \mathbb{R}^d$ is the (δ, c)-

interior support of a probability density $\mu \in \mathcal{P}_{2,ac}(\mathbb{R}^d)$ if for any $\mathbf{x} \in S(\delta, c)$,

$$\mu(\mathbf{y}) > c, \quad \forall \mathbf{y} \in \mathbb{R}^d \text{ s.t. } \|\mathbf{x} - \mathbf{y}\| \leq \delta.$$

Recall that the *support* of a probability density function $\mu : \mathbb{R}^d \to \mathbb{R}_+$ is the set $S \doteq \{\mathbf{x} \in \mathbb{R}^d : \mu(\mathbf{x}) > 0\}$. Note that $S(\delta, c) \subseteq S$ for any $\delta, c > 0$. Our Definition 4 of $S(\delta, c)$ generalizes the support S of a density function in the following two ways. First, in $S(\delta, c)$ we only consider points with a density larger than a threshold c. We use this threshold to determine if the density μ is too small. A point $\mathbf{x}$ is called *c-negligible* if $\mu(\mathbf{x}) \leq c$. Points in the support S are 0-negligible while points in (δ, c)-interior supports are c-negligible.

Second, a point $\mathbf{x} \in \mathbb{R}^d$ is in $S(\delta, c)$ if all points $\mathbf{y}$ in the δ-neighbourhood are not negligible (i.e., $\mu(\mathbf{y}) > c$). In comparison, a point $\mathbf{x} \in \mathbb{R}^d$ is in the support S as long as this point itself is not negligible. Hence, the support can be seen as a (δ, c)-interior support with $\delta = 0$ and $c = 0$.

Let $S_k(\delta, c)$ be the (δ, c)-interior support of the optimal solution μ_k to Problem 2 with the distribution loss functions presented in Examples 1 and 2. In the rest of this section, we check under what conditions the $\{S_k(\delta, c)\}_{k\in[K]}$ are disjoint and convex in pairs. We first show that $\{S_k(\delta, c)\}_{k\in[K]}$ are disjoint under mild conditions.

Proposition 1. Let $\{\mu_k\}_{k\in[K]}$ be the optimal solution to Problem 2. Suppose the distribution loss function L is a coupled loss function (Definition 2) with kernel $\ell : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ and let $\delta > 0$ and $c > 0$ be given. Suppose $\nabla^2_{x,y}\ell(\mathbf{x}_0, \mathbf{x}_0)$ is not positive semi-definite for any $\mathbf{x}_0 \in \mathbb{R}^d$. Then, $S_i(\delta, c) \cap S_j(\delta, c) = \emptyset$ for any $i \neq j \in [K]$.

Corollary 1. Suppose L is either Elo loss (equation (3)) or variance loss (equation (4)). Then, the (δ, c)-interior supports of the optimal densities to Problem 2 are disjoint for any $\delta, c > 0$; that is, $S_i(\delta, c) \cap S_j(\delta, c) = \emptyset$ for any $\delta, c > 0$ and any $i, j \in [K]$ such that $i \neq j$.

Hence, when we decompose π into $\mu_1, \ldots, \mu_K$, we in fact almost partition (note that $\cup_{k=1}^K S_k(\delta, c)$ may not be $\mathbb{R}^d$) the feature space $\mathbb{R}^d$ into K disjoint sets.

Next, we show in Example 3 below that under Elo loss, $\{S_k(\delta, c)\}_{k\in[K]}$ may not be convex in pairs; that is, for any two points $\mathbf{x}, \mathbf{y} \in S_i(\delta, c)$, there are possibly some points in the line segement $\{\lambda\mathbf{x} + (1 - \lambda)\mathbf{y} : \lambda \in (0, 1)\}$ that is not in $S_i(\delta, c)$ but in $S_j(\delta, c)$ for some $j \neq i$. Note that this also implies $S_i(\delta, c)$ is not convex.

Example 3 (Non-convexity under Elo loss)**.** We model the application described in Example 1 as an instance of Problem 2. Let $K = 2$ and $p_1 = p_2 = 1/2$. That is, we group players into two leagues with equal sub-populations. Consider the Elo loss function L (equation (3)). Suppose the skill level distribution π is given by

$$\pi = \frac{1}{4}\delta_{a-1} + \frac{1}{2}\delta_a + \frac{1}{4}\delta_{a+1}, \quad a > 1,$$

where δ_x is the dirac measure at point x. That is, only 3 possible skill levels are given by $a-1, a$, and $a+1$. The optimal solution is $\mu_1^* = \delta_a$ (with support $S_1 \doteq \{a\}$) and $\mu_2^* = \frac{1}{2}\delta_{a-1} + \frac{1}{2}\delta_{a+1}$ (with support $S_2 \doteq \{a-1, a+1\}$). Note that $\frac{1}{2}\mu_1^* + \frac{1}{2}\mu_2^* = \pi$. Hence, $\frac{1}{2}(a-1) + \frac{1}{2}(a+1) \in S_1$, while $a - 1, a + 1 \in S_2$. The proof is presented in the appendix.

Note that the discrete distribution π in the example does not admit a probability density function, and hence, not a (δ, c)-interior support. However, the same outcome can be shown if

we smooth the discrete distribution to get a density function. Our numerical experiments (Example 11) illustrate this using a mixed lognormal distribution with three peaks. In Example 11 below we show that under a certain distribution π, it is indeed better to group high-end players and low-end players into one league while middle-level players in another league.

We show in the following lemma that, under a mild condition, $\{S_k(\delta, c)\}_{k\in[K]}$ are convex in pairs if the distribution loss in Problem 2 is the variance loss (4).

Lemma 1. Suppose $L(\mu)$ is the distribution loss (equation (4)) in Example 2 with a positive semidefinite W. Given any $\delta, c > 0$, the collection of sets $\{S_k(\delta, c)\}_{k\in[K]}$ are convex in pairs.

Note that Lemma 1 does not imply that $S_i(\delta, c)$ is convex. Some points in the line segement $\{\lambda\mathbf{x} + (1-\lambda)\mathbf{y} : \lambda \in (0,1)\}$ for some $\mathbf{x}, \mathbf{y} \in S_i(\delta, c)$ may not be in $S_i(\delta, c)$. This may happen if the support of π is not convex by itself. However, Lemma 1 shows that at least the line segment joining any two points $\mathbf{x}, \mathbf{y} \in S_i(\delta, c)$ is not in any other $S_j(\delta, c)$.

3 Preliminaries on optimal transport and gradient flow

Let $\mathcal{P}_{2,ac}(\mathbb{R}^d)$ be all probability measures that are absolutely continuous with respect to the Lebesgue measure m and have a finite second moment, i.e.

$$\mathcal{P}_{2,ac}(\mathbb{R}^d) = \left\{\mu : \mu \ll m, \quad \int_{\mathbb{R}^d} \|\mathbf{x}\|^2 d\mu(\mathbf{x}) < \infty\right\}.$$

Let $(\mathbb{R}^d, \Sigma)$ be the Boreal measurable space. Given a measurable mapping $T : (\mathbb{R}^d, \Sigma) \to \mathbb{R}^d$ and a measure $\mu : \Sigma \to [0, \infty]$, the pushforward measure $T\sharp\mu : \Sigma \to [0, \infty]$ of μ under T is defined to be the measure given by $T\sharp\mu(B) = \mu(T^{-1}(B))$ for $B \in \Sigma$. The main property we need for the push-forward measure is the change-of-variables formula; that is, $\mathbb{E}_{T\sharp\mu}[f] = \mathbb{E}_\mu[f \circ T]$ for any measurable function $f : \mathbb{R}^d \to \mathbb{R}$ and where $\circ$ denotes function composition.

This paper assumes that $\mathbb{R}^d$ has the usual topology. Let $\mathcal{L}^2(\mathbb{R}^d; \mathbb{R}^d)$ be the L_2 space of functions $f : (\mathbb{R}^d, \mathcal{B}) \to (\mathbb{R}^d, \mathcal{B}, m)$, where $\mathcal{B}$ is the Borel σ-algebra and m is the Lebesgue measure; that is, $\mathcal{L}^2 \doteq \mathcal{L}^2(\mathbb{R}^d, \mathbb{R}^d) \doteq \{f : \mathbb{R}^d \to \mathbb{R}^d : \int_{\mathbb{R}^d} \|f(\mathbf{x})\|^2 dm(\mathbf{x}) < \infty\}$, where $\|\cdot\|$ is the 2-norm in $\mathbb{R}^d$.

Let $\mathcal{L}^2_\mu(\mathbb{R}^d; \mathbb{R}^d)$ be the L_2 space of functions $f : (\mathbb{R}^d, \mathcal{B}) \to (\mathbb{R}^d, \mathcal{B}, \mu)$; that is,

$$\mathcal{L}^2_\mu(\mathbb{R}^d; \mathbb{R}^d) \doteq \{f : \mathbb{R}^d \to \mathbb{R}^d : \int_{\mathbb{R}^d} \|f(\mathbf{x})\|^2 d\mu(\mathbf{x}) < \infty\}.$$

The inner product in $\mathcal{L}^2_\mu \doteq \mathcal{L}^2_\mu(\mathbb{R}^d; \mathbb{R}^d)$ is defined as $\langle f, g\rangle_\mu \doteq \int_{\mathbb{R}^d} \langle f(\mathbf{x}), g(\mathbf{x})\rangle d\mu(\mathbf{x})$ for any $f, g \in \mathcal{L}^2_\mu$. The $\mathcal{L}^2_\mu$ norm is defined as $\|f\|_\mu \doteq \sqrt{\int_{\mathbb{R}^d} \|f(\mathbf{x})\|^2 d\mu(\mathbf{x})}$ for any function $f \in \mathcal{L}^2_\mu(\mathbb{R}^d; \mathbb{R}^d)$.

We also need the following product space: for a vector $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_K)$ of Borel measures,

$$\mathcal{L}^2_{\boldsymbol{\mu}}(\mathbb{R}^d; \mathbb{R}^d) \doteq \prod_{k\in[K]} \mathcal{L}^2_{\mu_k}(\mathbb{R}^d; \mathbb{R}^d) = \{(f_1, \ldots, f_K) : f_k : (\mathbb{R}^d, \mathcal{B}) \to (\mathbb{R}^d, \mathcal{B}, \mu_k) \in \mathcal{L}^2_{\mu_k}, \forall k \in [K]\}.$$

The inner product in $\mathcal{L}^2_{\boldsymbol{\mu}}(\mathbb{R}^d; \mathbb{R}^d)$ is $\langle f, g\rangle_{\boldsymbol{\mu}} \doteq \sum_{k\in[K]} \int_{\mathbb{R}^d} \langle f(\mathbf{x}), g(\mathbf{x})\rangle d\mu_k(\mathbf{x})$ for any $f, g \in \mathcal{L}^2_{\boldsymbol{\mu}}$. The corresponding norm is given by $\|(f_1, \ldots, f_K)\|_{\boldsymbol{\mu}} = \sqrt{\sum_{k\in[K]} \|f_k\|^2_{\mu_k}}$.

3.1 Preliminaries on gradient flow. Standard references on gradient flow include Ambrosio et al. (2005), Santambrogio (2015, 2017). Our treatment will be contained, only setting the necessary development to make sense of our algorithms.

Given a time-changing velocity field $\phi : [0, 1] \times \mathbb{R}^d \to \mathbb{R}^d$, suppose a particle is moving according to this velocity field $\phi(t, \mathbf{x})$; that is, at time t and position $\mathbf{x}$, the velocity of this

particle is given by $\phi(t, \mathbf{x})$. Then, the particle's trajectory is the solution to the ordinary differential equation (ODE): $\dot{\mathbf{x}}(t) = \phi(t, \mathbf{x}(t))$. In particular, if $\phi(t, \mathbf{x})$ is the gradient of some function $F : \mathbb{R}^d \to \mathbb{R}$, then the solution function $\mathbf{x} : [0, 1] \to \mathbb{R}^d$ is called the *gradient flow* of F.

Now, suppose a population of particles, whose positions in $\mathbb{R}^d$ are initially distributed according to probability measure $\mu(0)$, are moving together according to the velocity field $\phi(t, \mathbf{x})$; that is, the trajectory $\mathbf{x}(t)$ of a particle initially positioned at point $\mathbf{x}_0$ is given by the solution to

$$\dot{\mathbf{x}}(t) = \phi(t, \mathbf{x}(t)), \quad \mathbf{x}(0) = \mathbf{x}_0. \tag{7}$$

Under this velocity field, $\phi(t, \mathbf{x})$, particles at different initial positions $\mathbf{x}_0$ have different trajectories. Hence, we can define a mapping $T_\phi^t : \mathbf{x}(0) \mapsto \mathbf{x}(t)$, which maps the initial position $\mathbf{x}(0)$ of a particle to its position $\mathbf{x}(t)$ at time t. Here, $\mathbf{x}(t)$ is the solution to ODE (7).

With this mapping T_ϕ^t, we can also study the change in the distribution of the particles' positions. Since these particles are initially distributed according to $\mu(0)$, and their positions at time t are given by the map T_ϕ^t, their positions at time t are distributed according to the pushforward measure $\mu(t) = T_\phi^t \sharp \mu(0)$. It turns out, as a curve in the space of probability measures, $\{\mu(t) : t \in [0, 1]\}$ can be characterized by the following partial differential equation (PDE) called the *continuity equation*, which is expressed as follows:

$$\frac{d}{dt}\mu(t, \mathbf{x}) = -\nabla \cdot (\phi(t, \mathbf{x})\mu(t, \mathbf{x})), \tag{8}$$

where $\nabla \cdot : \mathbb{R}^d \to \mathbb{R}$ denotes the divergence operator and where we may think of $\mu(t, \mathbf{x})$ as a density function evaluated at the point $\mathbf{x}$. This equation should be interpreted with some caution. Each term in the continuity equation (8) has no meaning in isolation since each term is not even defined if $\mu(t)$ is a measure (and not a density function). The continuity equation only holds in the following weak sense; that is, for any test function $f \in C_c^\infty(\mathbb{R}^d)$,

$$\frac{d}{dt}\int_{\mathbb{R}^d} f(\mathbf{x}) d\mu(t, \mathbf{x}) = \int_{\mathbb{R}^d} \langle \nabla f(\mathbf{x}), \phi(t, \mathbf{x}) \rangle d\mu(t, \mathbf{x}), \tag{9}$$

where $C_c^\infty(\mathbb{R}^d)$ are smooth functions with compact support. We note that this equation (9) is well-defined even if $\mu(t)$ is a measure, where here $\mu(t, \mathbf{x})$ is short-hand for $\mu(t)(\mathbf{x})$. We formalize this argument in the following lemma.

Lemma 2 (Theorem 1.3.17 in (Chewi 2023))**.** Let $\phi(t, \mathbf{x}) : [0, 1] \times \mathbb{R}^d \to \mathbb{R}^d$ be a time-dependent velocity field. Suppose random variables $t \mapsto X_t$ evolve according to $\dot{X}_t = \phi(t, X_t)$. Then, the law μ_t of X_t solves the continuity equation (8) in the weak sense; that is, equation (8) holds for $(\mu_t, \phi(t, \cdot))$ in the sense of equation (9).

3.2 Calculus in Wasserstein space. To develop gradient flow on our space $\mathcal{P}_2(\mathbb{R}^d)$ of probability measures, we must define an appropriate notion of "gradient". The standard definition is inadequate since it is only defined in vector spaces, but $\mathcal{P}_2(\mathbb{R}^d)$ is not a vector space. We define our notion of gradient over the space of probability measures using optimal transport theory, as we explain now.

Although a large part of optimal transport theory can be developed in a more general framework, we focus on the $\mathcal{P}_2(\mathbb{R}^d)$ for simplicity. The *Wasserstein-2 distance* on $\mathcal{P}_2(\mathbb{R}^d)$ is defined as follows.

Definition 5 (Wasserstein-2 distance)**.** For $\mu, \nu \in \mathcal{P}_2(\mathbb{R}^d)$,

$$W_2^2(\mu,\nu) \doteq \min\left\{\int_{\mathbb{R}^d\times\mathbb{R}^d} \|\mathbf{x}-\mathbf{y}\|^2 d\gamma : \gamma \in \Pi(\mu,\nu)\right\}, \tag{10}$$

where $\|\cdot\|$ is the standard Euclidean norm and $\Pi(\mu,\nu)$ is the set of *transport plans* defined as

$$\Pi(\mu,\nu) \doteq \{\gamma \in \mathcal{P}(\mathbb{R}^d \times \mathbb{R}^d) : (\pi_{\mathbf{x}})\sharp\gamma = \mu, (\pi_{\mathbf{y}})\sharp\gamma = \nu\}, \tag{11}$$

where $\pi_{\mathbf{x}}$ and $\pi_{\mathbf{y}}$ are the two coordinate projections of $\mathbb{R}^d\times\mathbb{R}^d$ onto $\mathbb{R}^d$. The minimizer γ^* to (10) is called the *optiaml transport plan*. Moreover, if there exists a measurable map $T : \mathbb{R}^d \to \mathbb{R}^d$ such that $\gamma^* = (\mathrm{id}, T)\sharp\mu$, then this map T is called the *optimal transport map*.

The set $\mathcal{P}_2(\mathbb{R}^d)$ is called *Wasserstein space* when endowed with the Wasserstein-2 metric W_2. It is well-known that $(\mathcal{P}_2(\mathbb{R}^d), W_2)$ is a metric space but not a vector space since the sum of two probability measures need not be a probability measure (see Proposition 5.1 in Santambrogio (2015)).

In Euclidean space, the gradient of a differentiable function $f : \mathbb{R}^d \to \mathbb{R}$ is defined to be the unique vector field $\nabla f : \mathbb{R}^d \to \mathbb{R}^d$ such that, for any $\mathbf{x}_0, \mathbf{v} \in \mathbb{R}^d$,

$$\lim_{t\to 0} \frac{f(x_0 + t\mathbf{v}) - f(\mathbf{x}_0)}{t} = \langle \nabla f(\mathbf{x}_0), \mathbf{v}\rangle. \tag{12}$$

The left-hand side is the directional derivative of function f at the point $\mathbf{x}_0$ along direction $\mathbf{v}$. The right-hand side is the vector product of gradient $\nabla f(\mathbf{x}_0)$ at point $\mathbf{x}_0$ and the direction $\mathbf{v}$. That is, in Euclidean space, the instantaneous change of function value (i.e., left-hand side of equation (12)) along any direction $\mathbf{v}$ can be easily computed by the inner product between gradient and the direction (i.e., right-hand side of equation (12)).

To define the Wasserstein gradient, we first define a similar notion of directional derivative in Wasserstein space. In Euclidean space, the term $\mathbf{x}(t) \doteq \mathbf{x}_0 + t\mathbf{v}$ in the directional derivative (12) defines a trajectory of movement $\mathbf{x}(t) : [0,1] \to \mathbb{R}^d$ that starts at $\mathbf{x}_0$ and moves in direction $\mathbf{v}$. In Wasserstein space, the vector addition $\mathbf{x}_0 + t\mathbf{v}$ is not valid since Wasserstein space is not a vector space. Instead of using vector addition, we use absolutely continuous curves $\mu(t) : [0,1] \to (\mathcal{P}_2(\mathbb{R}^d), W_2)$ to describe the trajectory of movement. This "absolutely continuous curve" is defined using the Wasserstein metric. We can define a notion of "direction" of the curve $\mu(t)$ similarly to how $\mathbf{v}$ defines a direction for $\mathbf{x}(t)$.

The direction of the trajectory $\mu(t)$ can be described as follows. As shown in the following Lemma 3, in Wasserstein space, any absolutely continuous curve can be characterized by the continuity equation (8). Moreover, there exists a vector field that can be understood as the "direction" of the trajectory $\mu(t)$.

Lemma 3 (Theorem 8.3.1 (Ambrosio et al. 2005))**.** Let $\mu(t) : (0,1) \to (\mathcal{P}_2(\mathbb{R}^d), W_2)$ be an absolutely continuous curve. Then there exists a vector field $\phi : (0,1)\times\mathbb{R}^d \to \mathbb{R}^d$ such that the continuity equation (8) holds in the weak sense, and for m-a.e. $t \in (0,1)$,

$$\phi(t,\cdot) \in \overline{\{\nabla\Psi : \Psi \in C_c^\infty(\mathbb{R}^d)\}}^{\mathcal{L}^2_{\mu_t}},$$

where $\bar{A}^{\mathcal{L}^2_{\mu_t}}$ represents the topological closure of set A in $\mathcal{L}^2_{\mu_t} \doteq \{f : \mathbb{R}^d \to \mathbb{R}^d : \int_{\mathbb{R}^d} \|f(\mathbf{x})\|^2 d\mu_t(\mathbf{x}) < \infty\}$.

In Lemma 3, at any $t \in (0,1)$, the vector field $v(t) : \mathbb{R}^d \to \mathbb{R}^d$ is called the "tangent vector" to the curve $\mu(t)$ at time t because it describes the direction of trajectory $\mu(t)$ through the

continuity equation (8). Lemma 3 indicates that starting at any point $\mu_0 \in (\mathcal{P}_2(\mathbb{R}^d), W_2)$, all directions that we can move along in Wasserstein space should be contained in the following set

$$\mathcal{TP}_2(\mu_0) = \overline{\{\nabla\Psi : \Psi \in C_c^\infty(\mathbb{R}^d)\}}^{\mathcal{L}^2_{\mu_0}}.$$

The set $\mathcal{TP}_2(\mu_0)$ is called the tangent space, which is known to be a vector space (Ambrosio et al. 2005, Lemma 8.4.2). With this tangent space, the Wasserstein-2 gradient can be defined through equation (12) as follows.

Definition 6 (Wasserstein-2 gradient)**.** Given a functional $F : \mathcal{P}_2(\mathbb{R}^d) \to \mathbb{R}$, the Wasserstein gradient of F at $\mu_0 \in \mathcal{P}_2(\mathbb{R}^d)$ is defined to be the unique element $\nabla F(\mu_0) \in \mathcal{TP}_2(\mu_0)$ such that, for every curve $\mu(\cdot) : [0, \infty) \to \mathcal{P}_2(\mathbb{R}^d)$ with $\mu(0) = \mu_0$ and tangent vector $v_0 \in \mathcal{TP}_2(\mu_0)$ at time 0, it holds that

$$\frac{d}{dt} F(\mu(t))|_{t=0} = \langle \nabla F(\mu_0), v_0 \rangle_{\mu_0}, \tag{13}$$

where $\langle \nabla F(\mu_0), v_0 \rangle_{\mu_0} \doteq \int_{\mathbb{R}^d} \langle \nabla F(\mu_0), v_0 \rangle d\mu_0$.

Note that equation (13) recovers the nice property described in (12) for Euclidean space. Next, we compute the Wasserstein-2 gradient of the following relevant functions.

Lemma 4. The Wasserstein gradient $\nabla L : \mathbb{R}^d \to \mathbb{R}^d \in \mathcal{L}^2_\mu$ of the coupled loss function L (defined in Definition 2) at $\mu \in \mathcal{P}_2(\mathbb{R}^d)$ is

$$\nabla L(\mu)(\mathbf{x}) = \int_{\mathbb{R}^d} (\nabla_1 \ell(\mathbf{x}, \mathbf{z}) + \nabla_2 \ell(\mathbf{z}, \mathbf{x})) d\mu(\mathbf{z}).$$

In our work, we are optimally deciding a vector $\boldsymbol{\mu} = (\mu_1, \mu_2, \ldots, \mu_K) \in \mathcal{P}_2(\mathbb{R}^d)^{\otimes K}$ of probability measures. It is straightforward to generalize the notion of Wasserstein calculus onto $\mathcal{P}_2(\mathbb{R}^d)^{\otimes K}$. In particular, the tangent space at $\boldsymbol{\mu}$ is

$$\mathcal{TP}_2^{\otimes K}(\boldsymbol{\mu}) = \bigotimes_{k \in [K]} \overline{\{\nabla\Psi : \Psi \in C_c^\infty(\mathbb{R}^d)\}}^{\mathcal{L}^2_{\mu_k}}.$$

Given a functional $F : \mathcal{P}_2(\mathbb{R}^d)^{\otimes K} \to \mathbb{R}$, we say it is differentiable at $\boldsymbol{\mu}_0$ with gradient $\nabla F(\boldsymbol{\mu}_0) = (\nabla_{\mu_1} F(\boldsymbol{\mu}_0), \ldots, \nabla_{\mu_K} F(\boldsymbol{\mu}_0))$ if for every curve $\boldsymbol{\mu}(\cdot) : [0, \infty) \to \mathcal{P}_2^{\otimes K}$ with $\boldsymbol{\mu}(0) = \boldsymbol{\mu}_0$ and tangent vector $\mathbf{v}_0 \in \mathcal{TP}_2^{\otimes K}(\boldsymbol{\mu}_0)$ at time 0, it holds that

$$\frac{d}{dt} F(\boldsymbol{\mu}(t))|_{t=0} = \langle \nabla F(\boldsymbol{\mu}_0), \mathbf{v}_0 \rangle_{\boldsymbol{\mu}_0}, \tag{14}$$

where $\langle \nabla F(\boldsymbol{\mu}_0), \mathbf{v}_0 \rangle_{\boldsymbol{\mu}_0} \doteq \sum_{k \in [K]} \langle \nabla_{\mu_k} F(\boldsymbol{\mu}_0), (\mathbf{v}_0)_k \rangle_{(\boldsymbol{\mu}_0)_k}$.

By the definition of objective function $F_{\mathbf{p}}(\boldsymbol{\mu})$ of Problem 2 and $F(\boldsymbol{\mu}, \mathbf{p})$ of Problem 1, the Wasserstein gradient $\nabla_{\boldsymbol{\mu}} F_{\mathbf{p}}(\boldsymbol{\mu}_0)$ and $\nabla_{\boldsymbol{\mu}} F(\boldsymbol{\mu}_0, \mathbf{p}_0)$ at $\boldsymbol{\mu}_0 \in \mathcal{P}_2^{\otimes K}$ and $\mathbf{p}_0$ are given by

$$\nabla_{\mu_k} F(\boldsymbol{\mu}_0, \mathbf{p}_0) = \nabla_{\mu_k} F_{\mathbf{p}}(\boldsymbol{\mu}_0) = p_k \nabla L(\mu_k).$$

In later development, we will have occasion to use Kullback-Leibler (KL) divergence to measure constraint violations in our problem. Therefore, we refresh the readers' understanding of this concept and derive its Wasserstein gradient.

Recall that the Kullback-Leibler (KL) divergence between two probability densities μ and π is defined as

$$\mathrm{KL}(\mu \| \pi) \doteq \int_{\mathbb{R}^d} \mu(x) \log \frac{\mu(x)}{\pi(x)} dx.$$

KL divergence is a type of statistical distance of how one probability density μ differs from a second, reference probability density π. It is 0 if and only if $\mu = \pi$ almost surely.

Lemma 5. For any $\boldsymbol{\mu} \in \mathcal{P}_{2,ac}^{\otimes K}(\mathbb{R}^d)$ and $\mathbf{p} \in \mathbb{R}_+^d$ with $\sum_{k\in[K]} p_k = 1$, define $\bar{\mu} \doteq \sum_{k\in[K]} p_k \mu_k$. For any $\pi \in \mathcal{P}_{2,ac}(\mathbb{R}^d)$, the Wasserstein gradient $\nabla_{\boldsymbol{\mu}} \mathrm{KL}(\bar{\mu}\|\pi) = (\nabla_{\mu_k} \mathrm{KL}(\bar{\mu}\|\pi))_{k\in[K]}$ is given by

$$\nabla_{\mu_k} \mathrm{KL}(\bar{\mu}\|\pi)(x) = p_k(s_{\bar{\mu}}(x) - s_\pi(x)),$$

for all $x \in \mathbb{R}^d$, where $s_{\bar{\mu}}(x) \doteq \nabla \log \bar{\mu}(x)$ and $s_\pi(x) \doteq \nabla \log \pi(x)$.

3.3 Geodesic geometry of Wasserstein space. Our work also needs a concept analogous to the vector-space notion of a "convex set" but defined in Wasserstein space. Recall that in Euclidean space, a set M is convex if for any $\mathbf{x}, \mathbf{y} \in M$, the whole line segment $(1-t)\mathbf{x} + t\mathbf{y}$ for $t \in [0,1]$ connecting $\mathbf{x}$ and $\mathbf{y}$ lies in the set M. Since Wasserstein space is not a vector space, we cannot define the "line segment" by the vector addition $(1-t)\mathbf{x} + t\mathbf{y}$. Nevertheless, we can define an analogous notion of "line segment" called a "geodesic."

In a nutshell, a geodesic $\mu(t) : [0,1] \to \mathcal{P}_2(\mathbb{R}^d)$ in Wasserstein space is the shortest path connecting $\mu(0) \in \mathcal{P}_2(\mathbb{R}^d)$ and $\mu(1) \in \mathcal{P}_2(\mathbb{R}^d)$. The term "shortest path" refers to the fact that, if we measure the "total length" of this curve $\mu(t)$ with Wasserstein metric, it is the shortest one among all curves connecting $\mu(0)$ and $\mu(1)$. This generalizes the fact that a line segment is the shortest path connecting two points in Euclidean space. In Wasserstein space, geodesics are characterized as follows.

Lemma 6 (Theorem 5.27 (Santambrogio 2015))**.** Suppose $T : \mathbb{R}^d \to \mathbb{R}^d$ is the optimal transport map (to problem (10)) from $\mu \in \mathcal{P}_{2,ac}(\mathbb{R}^d)$ to $\nu \in \mathcal{P}_2(\mathbb{R}^d)$. Then the curve $\mu(t) : [0,1] \to \mathcal{P}_2(\mathbb{R}^d)$ defined by $\mu(t) \doteq ((1-t)\mathrm{id} + tT)\sharp\mu$ is a geodesic connecting μ and ν, where $\mathrm{id} : \mathbb{R}^d \to \mathbb{R}^d$ is the identity map, i.e., $\mathrm{id}(\mathbf{x}) = \mathbf{x}$.

A set $M \subseteq \mathcal{P}_2(\mathbb{R}^d)$ is *geodesically convex* in Wasserstein space if, for any $\mu, \nu \in M$, the geodesic $\{\mu(t) : t \in [0,1]\}$ connecting μ and ν is contained in the set M.

4 Optimality condition

In this section, we derive optimality conditions for Problems 1 and 2. First, we note below that the classical Karush-Kuhn-Tucker (KKT) conditions do not apply in our setting. Accordingly, we propose two new optimality conditions (i.e., Propositions 2 and 3) for Problems 1 and 2. The proposed optimality conditions generalize the geometric form of the KKT condition in Euclidean space.

We begin with Problem 2. Recall that the feasible set of Problem 2 is

$$\mathcal{P}_{\pi,\mathbf{p}} = \left\{ (\mu_1, \ldots, \mu_K) \in \mathcal{P}_2(\mathbb{R}^d)^{\otimes K} : \sum_{k\in[K]} p_k \mu_k = \pi \right\}.$$

Note that $\mathcal{P}_{\pi,\mathbf{p}}$ is not geodesically convex in Wasserstein space $(\mathcal{P}_2(\mathbb{R}^d), W_2)$, even though the constraint looks very linear (but not linear in Wasserstein space!). To see this, consider the following simple example.

Example 4 (Non-convexity of the feasible set $\mathcal{P}_{\pi,\mathbf{p}}$)**.** Let $K = 2$, $\mathbf{p} = (\frac{1}{2}, \frac{1}{2})$, and $\pi = \mathrm{Uniform}[-1,1]$. We have two simple decompositions $\boldsymbol{\mu} = (\mu_1, \mu_2) \in \mathcal{P}_{\pi,\mathbf{p}}$ and $\boldsymbol{\nu} = (\nu_1, \nu_2) \in \mathcal{P}_{\pi,\mathbf{p}}$ given by

$$\mu_1 = \mathrm{Uniform}[-1,0], \quad \mu_2 = \mathrm{Uniform}[0,1], \quad \text{and}$$
$$\nu_1 = \mathrm{Uniform}[0,1], \quad \nu_2 = \mathrm{Uniform}[-1,0].$$

The optimal transport map $T_1 : \mathbb{R}^d \to \mathbb{R}^d$ from μ_1 to ν_1 is given by $T_1(x) = x + 1$, and the optimal transport map $T_2 : \mathbb{R}^d \to \mathbb{R}^d$ from μ_2 to ν_2 is $T_2(x) = x - 1$. Hence, the geodesic $(\mu_1, \mu_2)(\cdot) : [0, 1] \to \mathcal{P}_2^{\otimes 2}$ connecting $\boldsymbol{\mu}$ and $\boldsymbol{\nu}$ are given by

$$\mu_1(t) = \text{Uniform}[-1 + t, t], \quad \mu_2(t) = \text{Uniform}[-t, 1 - t].$$

It is easy to check that $(\mu_1(t), \mu_2(t)) \in \mathcal{P}_{\pi,\mathbf{p}}$ only when $t = 0$ or 1. Hence, the feasible set $\mathcal{P}_{\pi,\mathbf{p}}$ is not geodesically convex in Wasserstein space.

Despite this unusual geometry, we can still consider a KKT-type optimality condition on $\mathcal{P}_{\pi,\mathbf{p}}$. Our optimality condition is motivated by the following observation about the usual KKT condition in Euclidean space.

Suppose we seek to minimize a smooth function $f : \mathbb{R}^d \to \mathbb{R}$ subject to constraint $g(\mathbf{x}) = 0$, where $g : \mathbb{R}^d \to \mathbb{R}$ is also a smooth function. The standard KKT condition is given by

$$\nabla f(\mathbf{x}) + \lambda \nabla g(\mathbf{x}) = 0, \tag{15}$$

for some $\lambda \in \mathbb{R}$. We call equation (15) the algebraic form of the KKT condition. The algebraic form can be extended to optimization problems in infinite-dimensional vector spaces (Luenberger 1997, Theorem 1, Section 9.3). However, this version does not apply to our setting since Wasserstein space is not a vector space. Hence, we need to propose our own KKT condition.

The standard KKT condition in Euclidean space also has a geometric interpretation due to the manifold structure of the feasible region, which we now describe. The constraint $g(\mathbf{x}) = 0$ defines a smooth manifold $M \doteq g^{-1}(0) \cap \{\mathbf{x} \in \mathbb{R}^d : \nabla g(\mathbf{x}) \neq 0\}$. For each point $\mathbf{x}$ on M, the manifold M has a *tangent space* $\mathcal{T}M(\mathbf{x})$ which is a vector space consisting of all tangent vectors at $\mathbf{x}$ to differentiable curves on M passing through $\mathbf{x}$. Moreover, the tangent space can be charaterized by $\mathcal{T}M(\mathbf{x}) = \{\mathbf{v} \in \mathbb{R}^d : \mathbf{v}^\top \nabla g(\mathbf{x}) = 0\}$. For details of above discussion about manifold structure, see (Edwards 1994, Theorem 5.4, 5.5). From a geometric perspective, equation (15) represents the fact that, at optimal solution $\mathbf{x}$, $\nabla f(\mathbf{x})$ is in the straight line $\{\mathbf{v} \in \mathbb{R}^d : \mathbf{v} = \lambda \nabla g(\mathbf{x}) \text{ for some } \lambda \in \mathbb{R}\}$ defined by $\nabla g(\mathbf{x})$. To avoid the need for a parameter λ in the optimality condition, equivalently, we can say the projection of $\nabla f(\mathbf{x})$ onto the space $\mathcal{T}M(\mathbf{x})$, which consists of vector $\mathbf{v}$ that is orthogonal to $\nabla g(\mathbf{x})$, is 0. We introduce the projection norm onto $\mathcal{T}M(\mathbf{x})$ to quantify this "orthogonal" relation. The projection norm of M is defined as follows: for each $\phi \in \mathbb{R}^d$,

$$\|\phi\|_{\mathcal{T}M(\mathbf{x})} = \sup_{\mathbf{v} \in \mathcal{T}M(\mathbf{x})} \frac{\langle \mathbf{v}, \phi \rangle}{\|\mathbf{v}\|}. \tag{16}$$

It is easy to see that algebraic form (15) is equivalent to

$$\|\nabla f(\mathbf{x})\|_{\mathcal{T}M(\mathbf{x})} = 0. \tag{17}$$

We term this form (17) the geometric form as it stems from the geometric perspective outlined above.

The optimality condition (17) can be easily extended to Wasserstein space. To extend the geometric form (17) of KKT condition to Wasserstein space, we need to extend the notion of "tangent vector" first. As explained in Section 3.2 (particularly in Lemma 3), for an absolutely continuous curve $\mu(\cdot) : (0, 1) \to (\mathcal{P}_2(\mathbb{R}^d), W_2)$, there exists a time-dependent velocity field $v(\cdot, \cdot) : (0, 1) \times \mathbb{R}^d \to \mathbb{R}^d$ such that the continuity equation (8) holds in weak sense. The "tangent vector" to this curve $\mu(\cdot)$ at time t is the instantaneous velocity field $v(t, \cdot) : \mathbb{R}^d \to \mathbb{R}^d$ (velocity

field at time t). Hence, it is natural to define "tangent space" by the continuity equation (8) as follows. Define the tangent space at $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_K) \in \mathcal{P}_2(\mathbb{R}^d)^{\otimes K}$ as

$$\begin{aligned}\mathcal{TP}_{\pi,\mathbf{p}}(\boldsymbol{\mu}_0) \doteq\{&(\phi_1, \ldots, \phi_K) \in \mathcal{L}^2_{\boldsymbol{\mu}} : \exists \boldsymbol{\mu}(t) : [0,1] \to \mathcal{P}_{\pi,\mathbf{p}}, \text{ s.t. } \forall k \in [K], \\ &\tfrac{d}{dt}\mu_k(t)|_{t=0} = -\nabla \cdot (\mu_k(0)\phi_k); \boldsymbol{\mu}(0) = \boldsymbol{\mu}_0\}.\end{aligned} \tag{18}$$

The equation $\frac{d}{dt}\mu_k(t)|_{t=0} = -\nabla \cdot (\mu_k(0)\phi_k)$ holds in the weak sense; that is, for all $f \in C_c^\infty(\mathbb{R}^d)$,

$$\frac{d}{dt}\int_{\mathbb{R}^d} f(\mathbf{x}) d\mu_k(t, \mathbf{x})|_{t=0} = \int_{\mathbb{R}^d} \langle \nabla f(\mathbf{x}), \phi_k(\mathbf{x})\rangle d\mu_k(0, \mathbf{x}).$$

The *projection norm* of tangent space $\mathcal{TP}_{\pi,\mathbf{p}}(\boldsymbol{\mu})$ is

$$\|\boldsymbol{\phi}\|_{\mathcal{TP}_{\pi,\mathbf{p}}(\boldsymbol{\mu})} \doteq \sup_{u \in \mathcal{TP}_{\pi,\mathbf{p}}(\boldsymbol{\mu})} \frac{\langle u, \boldsymbol{\phi}\rangle_{\boldsymbol{\mu}}}{\|u\|_{\boldsymbol{\mu}}},$$

for all $\boldsymbol{\phi} \in \mathcal{L}^2_{\boldsymbol{\mu}}$. We prove that the optimal solution to Problem 2 satisfies the following condition.

Proposition 2 (Optimality condition for Problem 2)**.** Suppose $\boldsymbol{\mu}^*$ is the optimal solution to Problem 2. Then,

$$\|\nabla F_{\mathbf{p}}(\boldsymbol{\mu}^*)\|_{\mathcal{TP}_{\pi,\mathbf{p}}(\boldsymbol{\mu}^*)} = 0.$$

Similarly, for Problem 1, we define the following tanget space,

$$\begin{aligned}\mathcal{TP}_{\pi}(\boldsymbol{\mu}_0, \mathbf{p}_0) \doteq\{&(\boldsymbol{\phi}, \mathbf{v}) \in \mathcal{L}^2_{\boldsymbol{\mu}} \times \mathcal{L}^2 : \exists (\boldsymbol{\mu}(t), \mathbf{p}(t)) \in \mathcal{P}_{\pi}, \text{ s.t.} \\ &\boldsymbol{\mu}(t) \in \mathcal{P}_{\pi,\mathbf{p}(t)}, \boldsymbol{\mu}(0) = \boldsymbol{\mu}_0, \text{ and } \mathbf{p}(0) = \mathbf{p}_0; \\ &\forall k \in [K], \frac{d}{dt}\mu_k(t)|_{t=0} = -\nabla \cdot (\mu_k(0)\phi_k) \text{ and } \tfrac{d}{dt}p_k(t)|_{t=0} = v_k\},\end{aligned}$$

for all $(\boldsymbol{\mu}_0, \mathbf{p}_0) \in \mathcal{P}_2(\mathbb{R}^d)^{\otimes K} \times \mathbb{R}^K$. Note that this tangent space is not empty, since for any $(\boldsymbol{\mu}_0, \mathbf{p}_0)$, it contains the zero element $(\mathbf{p}, \mathbf{v}) = (\mathbf{0}, \mathbf{0})$.

The projection norm of $\mathcal{TP}_{\pi}(\boldsymbol{\mu}, \mathbf{p})$ is,

$$\|(\boldsymbol{\phi}, \mathbf{v})\|_{\mathcal{TP}_{\pi}(\boldsymbol{\mu},\mathbf{p})} \doteq \sup_{(u_1,u_2) \in \mathcal{TP}_{\pi}(\boldsymbol{\mu},\mathbf{p})} \frac{\langle u_1, \boldsymbol{\phi}\rangle_{\boldsymbol{\mu}} + \langle u_2, \mathbf{v}\rangle}{\|u_1\|_{\boldsymbol{\mu}} + \|u_2\|}$$

for all $(\boldsymbol{\phi}, \mathbf{v}) \in \mathcal{L}^2_{\boldsymbol{\mu}} \times \mathcal{L}^2_{\boldsymbol{\mu}}$. We prove that the optimal solution to Problem 1 satisfies the following condition.

Proposition 3. Suppose $(\boldsymbol{\mu}^*, \mathbf{p}^*)$ is the optimal solution to Problem 1. Then,

$$\|\nabla F\|_{\mathcal{TP}_{\pi}(\boldsymbol{\mu}^*,\mathbf{p}^*)} = 0.$$

As a remark, the main difficulty of proving Propositions 2 and 3 by contradiction is constructing feasible solutions that are strictly better. This is nontrivial since the tangent space is not invariant, so moving along an element in the tangent space may leave the feasible set, and one needs to "project" the infeasible solution back. All these procedures have to be done within the Wasserstein metric.

5 Constraint controlled gradient flow (CCGF) for optimal probability measure decomposition

We aim to develop a concept of gradient flow for Problems 1 and 2 in Wasserstein space. We first illustrate the notion of gradient flow in Euclidean space in Section 5.1. The purpose of elucidating this Euclidean gradient flow is to illustrate the main idea behind our flow design in Wasserstein space. Then, with the machinery introduced in Section 3, we extend Euclidean

gradient flow to Wasserstein space in Sections 5.2 and 5.3. We note that, in Section 6, we implement Wasserstein gradient flow via discretization. This yields an iterative algorithm.

5.1 CCGF in Euclidean space. In this subsection, we design a constraint-controlled gradient flow (CCGF) in Euclidean space to solve the following constrained minimization problem. If we implement the CCGF, we can get an iterative algorithm. The concept of CCGF has also inspired the design of particle-based algorithms to solve constrained sampling problems in Liu et al. (2021) and Zhang et al. (2022). While the CCGF is an intuitive idea, we could not find any other references to it, even in standard Euclidean space settings. Therefore, we will first consider the following finite-dimensional optimization problem:

Problem 3 (Finite-dimensional optimization problem)**.** Let $f : \mathbb{R}^n \to \mathbb{R}$ and $g : \mathbb{R}^n \to [0, \infty]$ be two smooth functions and solve

$$\begin{aligned} \min_{\mathbf{x}\in\mathbb{R}^n} \quad & f(\mathbf{x}) \\ \text{s.t.} \quad & g(\mathbf{x}) = 0, \end{aligned}$$

where $C_0 \doteq \{\mathbf{x} \in \mathbb{R}^n : g(\mathbf{x}) = 0\}$ be the feasible region.

We will impose necessary assumptions when they are needed in later development. We do not assume that f, g are convex or concave.

To solve Problem 3, a first thought is to use the projected gradient descent algorithm. Let $d : \mathbb{R}^n \times \mathbb{R}^n \to [0, \infty)$ be the Euclidean metric. The project gradient descent algorithm is given by

$$\begin{aligned} \mathbf{x}'_{k+1} &= \text{argmin}_{y\in\mathbb{R}^n} f(\mathbf{x}_k) + \langle \nabla f(\mathbf{x}_k), y - \mathbf{x}_k\rangle + \frac{1}{2} d(y, \mathbf{x}_k)^2 \\ x_{k+1} &= \Pi_{C_0}(\mathbf{x}'_{k+1}), \end{aligned}$$

where $\Pi_{C_0}(\mathbf{x}'_{k+1})$ is the projection of $\mathbf{x}'_{k+1}$ to the feasible set C_0. However, such an algorithm can be hard to implement for two reasons: (1) projections can be difficult to find, and (2) the projected point $\mathbf{x}_{k+1}$ can be far away from $\mathbf{x}'_{k+1}$, which makes it difficult to derive the gradient flow in the limit.

We propose the following "variational interpolation" approach to design our gradient flow to avoid these two challenges. The idea of the variational interpolation approach is as follows. We aim to design a vector field $\phi : \mathbb{R}^d \to \mathbb{R}^d$ to specify the "velocity" of movement at each position in $\mathbb{R}^d$. Accordingly, we call it a "velocity field." The purpose of our design is to guarantee that if a particle is moving in $\mathbb{R}^d$ according to this velocity field ϕ, it will eventurally converge to a feasible solution $\mathbf{x}^*$ (i.e., $g(\mathbf{x}^*) = 0$) which satisfies optimality condition (17). Recall that the trajectory of the particle is given by the solution to the ODE: $\dot{\mathbf{x}}(t) = \phi(\mathbf{x}(t))$. Hence, mathematically speaking, our challenge is to design ϕ such that the solution to the ODE $\dot{\mathbf{x}}(t) = \phi(\mathbf{x}(t))$ converges to a feasible solution to Problem 3 (i.e., $\lim_{t\to\infty} g(\mathbf{x}(t)) = 0$) satisfying optimality condition (17) in the end. To design such a velocity field ϕ, we start with a sequence of discrete points generated by an iterative scheme explained as follows.

Fixing any time step parameter $\tau > 0$, for some constraint-control parameter $\alpha \in (0, 1/\tau)$, we look for a sequence of points $(\mathbf{x}^\tau_k)_{k\in\mathbb{N}}$ defined through the following iterated scheme,

$$\mathbf{x}^\tau_{k+1} \doteq \text{argmin}_{\mathbf{x}\in C_{(1-\alpha\tau)g(\mathbf{x}^\tau_k)}} f(\mathbf{x}^\tau_k) + \langle \nabla f(\mathbf{x}^\tau_k), \mathbf{x} - \mathbf{x}^\tau_k\rangle + \frac{1}{2\tau} d(\mathbf{x}, \mathbf{x}^\tau_k)^2, \tag{19}$$

where $C_{(1-\alpha\tau)g(x^\tau_k)} = \{\mathbf{x} \in \mathbb{R}^n : g(\mathbf{x}) = (1-\alpha\tau)g(\mathbf{x}^\tau_k)\}$. We can interpret this sequence of points

as the positions of the trajectory $\mathbf{x}(t)$ at time points $t = 0, \tau, 2\tau, \ldots, k\tau, \ldots$; that is, $\mathbf{x}(k\tau) = \mathbf{x}_k^\tau$. Hence, the parameter τ represents the time length between two consecutive points $\mathbf{x}_k^\tau$ and $\mathbf{x}_{k+1}^\tau$.

If $g(x_k^\tau) = 0$, i.e., $\mathbf{x}_k^\tau \in C_0$, the $(k+1)$-st step produced by equation (19) is the same as the original projected gradient descent algorithm. If $g(\mathbf{x}_k^\tau) > 0$, i.e., $\mathbf{x}_k^\tau \notin C_0$, we try to find a solution $\mathbf{x}_{k+1}^\tau$ which is closer to the actual feasible region C_0 in the sense that $g(\mathbf{x}_{k+1}^\tau) = (1-\alpha\tau)g(\mathbf{x}_k^\tau)$, where $(1-\alpha\tau) \in (0,1)$.

Moreover, we can further simplify the equation (19) by replacing the nonlinear constraint $g(\mathbf{x}) = (1-\alpha\tau)g(\mathbf{x}_k^\tau)$ by its linear approximation. This linearization simplifies the problem and makes it easily solvable using Lagrangian-type arguments as shown in Lemma 7. It is clear that, as $\tau \to 0^+$, we have $\mathbf{x}_{k+1}^\tau \to \mathbf{x}_k^\tau$. Hence, if time step τ is small enough, the $(k+1)$-st step $\mathbf{x}_{k+1}^\tau$ should be close to the k-th step $\mathbf{x}_k^\tau$. Therefore, we can replace the function g in the left-hand side of the constraint defining $C_{(1-\alpha\tau)g(\mathbf{x}_k^\tau)}$ with its linear approximation $\tilde{g}(\mathbf{x}) = g(\mathbf{x}_k^\tau) + \langle \mathbf{x} - \mathbf{x}_k^\tau, \nabla g(\mathbf{x}_k^\tau)\rangle$. Thus, now we consider the sequence $(\mathbf{x}_k^\tau)_{k\in\mathbb{N}}$ generated by the following iterative scheme

$$\begin{aligned} \mathbf{x}_k^\tau \doteq \operatorname{argmin}_{\mathbf{x}\in\mathbb{R}^n} \quad & f(\mathbf{x}_k^\tau) + \langle \nabla f(\mathbf{x}_k^\tau), \mathbf{x} - \mathbf{x}_k^\tau\rangle + \frac{1}{2\tau} d(\mathbf{x}, \mathbf{x}_k^\tau)^2 \\ \text{s.t.} \quad & \tilde{g}(\mathbf{x}) = (1-\alpha\tau)g(\mathbf{x}_k^\tau), \end{aligned} \tag{20}$$

where the constraint is equivalent to $\alpha\tau g(\mathbf{x}_k^\tau) + \langle \nabla g(\mathbf{x}_k^\tau), \mathbf{x} - \mathbf{x}_k^\tau\rangle = 0$. We solve (20) with the Lagrangian method as follows.

Lemma 7. For any $\tau > 0$, at any iteration t, the optimal solution $\mathbf{x}_{k+1}^\tau$ to equation (20) is given by

$$\mathbf{x}_{k+1}^\tau = \mathbf{x}_k^\tau - \tau(\nabla f(\mathbf{x}_k^\tau) + \lambda^* \nabla g(\mathbf{x}_k^\tau)), \quad \lambda^* = \frac{-\langle \nabla g(\mathbf{x}_k^\tau), \nabla f(\mathbf{x}_k^\tau)\rangle + \alpha g(\mathbf{x}_k^\tau)}{\|\nabla g(\mathbf{x}_k^\tau)\|^2}.$$

Suppose a particle travels at a constant velocity from point $\mathbf{x}_k^\tau$ to point $\mathbf{x}_{k+1}^\tau$. Since the duration between two consecutive points is τ, the velocity ϕ_k^τ of moving from position $\mathbf{x}_k^\tau$ to position $\mathbf{x}_{k+1}^\tau$ is given by

$$\phi_k^\tau \doteq \frac{\mathbf{x}_{k+1}^\tau - \mathbf{x}_k^\tau}{\tau} = -(\nabla f(\mathbf{x}_k^\tau) + \lambda^* \nabla g(\mathbf{x}_k^\tau)).$$

This motivates us to consider the following design of the velocity field ϕ.

Definition 7 (Euclidean CCGF)**.** The *Euclidean constraint controlled gradient flow (CCGF)* is defined to be the solution to the ODE:

$$\dot{\mathbf{x}}(t) = \phi(\mathbf{x}(t)),$$

where the velocity field $\phi : \mathbb{R}^n \to \mathbb{R}^n$ is

$$\phi(\mathbf{x}) \doteq -(\nabla f(\mathbf{x}) + \lambda(x)\nabla g(\mathbf{x})) \text{ and } \lambda(\mathbf{x}) = \frac{-\langle \nabla g(\mathbf{x}), \nabla f(\mathbf{x})\rangle + \alpha g(\mathbf{x})}{\|\nabla g(\mathbf{x})\|^2}. \tag{21}$$

We show the convergence of the Euclidean CCGF in Definition 7 under the following assumptions.

Assumption 1. $\|\nabla g(\mathbf{x})\| > 0$ if $g(\mathbf{x}) \neq 0$.

Assumption 1 guarantees that $\lambda(\mathbf{x})$ in (21) is well-defined. We also need the following assumptions to prove convergence.

Assumption 2. (i) f is bounded from below by a real number $f_{\min}$;

(ii) there exists an $L > 0$ such that $\|\nabla f(\mathbf{x})\| \leq L$ for all $\mathbf{x} \in \mathbb{R}^n$;

(iii) (Polyak-Łojasiewicz condition) there exists a $\kappa > 0$ such that $\|\nabla g(\mathbf{x})\|^2 \geq \kappa g(\mathbf{x})$ for all $\mathbf{x} \in \mathbb{R}^n$.

As we discussed in Section 4, the KKT condition has a geometric form described in equation (17), using the tangent norm defined in equation (16). We show that, in the limit (i.e., as $t \to \infty$), the KKT condition (17) is satisfied by the Euclidean CCGF in Definition 7.

Theorem 1. Suppose Assumption 1 holds. Let $\mathbf{x}(t)$ be the Euclidean CCGF defined in Definition 7. Then,

(i) $g(\mathbf{x}(t)) = e^{-\alpha t} g(\mathbf{x}(0))$ for all $t > 0$;

(ii) If Assumption 2 also holds, then, for any $T > 0$,

$$\min_{t \leq T} \|\nabla f(\mathbf{x}(t))\|_{\mathcal{T}M(\mathbf{x})} \leq \frac{C}{\sqrt{T}},$$

where $C \doteq f(\mathbf{x}(0)) - f_{\min} + \frac{2g(\mathbf{x}(0))}{\kappa} + \frac{L}{\alpha\sqrt{\kappa}}\sqrt{g(\mathbf{x}(0))}$.

Theorem 1 shows that the CCGF $\mathbf{x}(t)$ in Definition 7 converges in the limit to a feasible solution (i.e., $\lim_{t\to\infty} g(\mathbf{x}(t)) = 0$) exponentially fast with speed controlled by the parameter α. Furthermore, the tangent norm of the gradient decreases to 0 at the speed of $O(\frac{1}{\sqrt{T}})$, i.e., the KKT condition (17) is satisfied in the limit as $t \to \infty$.

To implement the Euclidean CCGF approach in practice, we can select a short time length τ and implement the iterative scheme in Lemma 7. After enough iterations (i.e., k is large enough, approximating $t \to \infty$), the solution $\mathbf{x}_k^\tau$ will be close to feasible and optimal.

Remark 1. Another version of the CCGF can be derived with the constraint in (20) is replaced with $\tilde{g}(x) \leq (1 - \alpha\tau) g(\mathbf{x}_k^\tau)$. The corresponding CCGF remains largely the same, except $\lambda(\mathbf{x})$ is replaced by its positive part, i.e $\max\{\lambda(\mathbf{x}), 0\}$. Gradient flow approaches that use a structure analogous to $\max\{\lambda(\mathbf{x}), 0\}$ can be found in Liu et al. (2021), Zhang et al. (2022), but in a Wasserstein context, not in Euclidean space. Similar bounds as in Theorem 1 can be obtained, except one can only obtain $g(\mathbf{x}(t)) \leq e^{-\alpha t} g(\mathbf{x}(0))$. While this provides faster convergence to the feasible set, it can also leads to instability in the computation of $\lambda(x)$, since $\nabla g(\mathbf{x})$ appears in the denominator. Using the current version of CCGF resolves this issue.

5.2 Wasserstein CCGF for Problem 2.

In this subsection, we extend the notion of CCGF in Definition 7 to Wasserstein space to solve Problem 2.

For ease of development, from now on, we assume π is absolutely continuous with respect to Lebesgue measure; that is, we think of π as a probability density function.

Now, $F_{\mathbf{p}}$ in Problem 2 plays the role of the objective function f in Problem 3. We use Kullback-Leibler (KL) divergence to measure deviations from the feasibility constraint $\sum_{k\in[K]} p_k \mu_k = \pi$. Recall that the KL divergence between two densities is 0 if and only if the two densities are equal almost everywhere. Define $\bar{\mu} \doteq \sum_{k\in[K]} p_k \mu_k$. The constraint $g(x) = 0$ in Problem 3 becomes $\mathrm{KL}(\bar{\mu}\|\pi) = 0$ in Problem 2.

Since Problem 2 is defined on the product Wasserstein space $(\mathcal{P}_{2,ac}(\mathbb{R}^d), W_2)^{\otimes K}$, our approach is to design a velocity field $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_K)$ on $(\mathcal{P}_{2,ac}(\mathbb{R}^d), W_2)^{\otimes K}$ to solve Problem 2. Given any point $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_K) \in (\mathcal{P}_{2,ac}(\mathbb{R}^d), W_2)^{\otimes K}$, the corresponding velocity field $\boldsymbol{\phi}(\boldsymbol{\mu})$

should specify the "velocity of movement" at position $\boldsymbol{\mu}$. Recall that, as we discussed in Section 3.2, in a single Wasserstein space $(\mathcal{P}_{2,ac}(\mathbb{R}^d), W_2)$, the "velocity" of any trajectory at any position $\mu \in (\mathcal{P}_{2,ac}(\mathbb{R}^d), W_2)$ is a vector field $\phi(\mu)(\cdot) : \mathbb{R}^d \to \mathbb{R}^d$ on Euclidean space. Given such a velocity field, the trajectory of movement in Wasserstein space is the solution to the continuity equation (8). Hence, the velocity field $\boldsymbol{\phi}$ in the product Wasserstein space $(\mathcal{P}_{2,ac}(\mathbb{R}^d), W_2)^{\otimes K}$ should map any $\boldsymbol{\mu} \in (\mathcal{P}_{2,ac}(\mathbb{R}^d), W_2)^{\otimes K}$ to K vector fields $\boldsymbol{\phi}(\boldsymbol{\mu}) = (\phi_1(\boldsymbol{\mu}), \dots, \phi_K(\boldsymbol{\mu}))$, where each $\phi_k(\boldsymbol{\mu})(\cdot) : \mathbb{R}^d \to \mathbb{R}^d$ in $\mathcal{L}^2_{\mu_k}$ describes the velocity of $\mu_k \in (\mathcal{P}_{2,ac}(\mathbb{R}^d), W_2)$.

Problem 2 can be solved using our velocity field $\boldsymbol{\phi}$ by the following procedure. Starting from arbitrary initial densities $\boldsymbol{\mu}(0) = (\mu_1(0), \dots, \mu_K(0))$, the solution $\boldsymbol{\mu}(t) : \mathbb{R} \to (\mathcal{P}_{2,ac}(\mathbb{R}^d), W_2)^{\otimes K}$ of the following K continuity equations

$$\frac{d}{dt}\mu_k(t) = -\nabla \cdot (\phi_k(\boldsymbol{\mu}(t))\mu_k(t)), \quad \forall k \in [K] \tag{22}$$

converges to a "stationary point" of Problem 2 in the sense that

- $\lim_{t\to\infty} \mathrm{KL}(\bar{\mu}(t)\|\pi) = 0$ with $\bar{\mu}(t) \doteq \sum_{k\in[K]} p_k \mu_k(t)$;
- the optimality condition in Proposition 2 is approximately satisfied.

Following the Euclidean CCGF in Definition 7, we design the Wasserstein CCGF as follows.

Definition 8 (Wasserstein CCGF)**.** The *Wasserstein constraint controlled gradient flow (CCGF)* is the solution $\boldsymbol{\mu}(t) : \mathbb{R} \to (\mathcal{P}_{2,ac}(\mathbb{R}^d), W_2)^{\otimes K}$ to the K continuity equations (22), where the velocity field $\phi_k(\boldsymbol{\mu})(\cdot) : \mathbb{R}^d \to \mathbb{R}^d$ is

$$\phi_k(\boldsymbol{\mu}) = -(\nabla_{\mu_k} F_{\mathbf{p}}(\boldsymbol{\mu}) + \lambda(\boldsymbol{\mu})\nabla_{\mu_k}\mathrm{KL}(\bar{\mu}\|\pi)), \quad \forall k \in [K]$$

with

$$\lambda(\boldsymbol{\mu}) = \frac{-\langle \nabla_{\boldsymbol{\mu}} F_{\mathbf{p}}(\boldsymbol{\mu}), \nabla_{\boldsymbol{\mu}}\mathrm{KL}(\bar{\mu}\|\pi)\rangle_{\boldsymbol{\mu}} + \alpha \mathrm{KL}(\bar{\mu}\|\pi)}{\|\nabla_{\boldsymbol{\mu}}\mathrm{KL}(\bar{\mu}\|\pi)\|^2_{\boldsymbol{\mu}}}.$$

By Lemmas 4 and 5, we have

$$\begin{aligned} \phi_k(\boldsymbol{\mu}) &= -(p_k \nabla L(\mu_k) + \lambda(\boldsymbol{\mu}) p_k (s_{\bar{\mu}} - s_\pi))) \in \mathcal{L}^2_{\mu_k}, \quad \forall k \in [K] \text{ and} \\ \lambda(\boldsymbol{\mu}) &= \frac{-\sum_{k\in[K]} p_k \langle \nabla L(\mu_k), p_k(s_{\bar{\mu}} - s_\pi)\rangle_{\mu_k} + \alpha \mathrm{KL}(\bar{\mu}\|\pi)}{\sum_{k\in[K]} \|p_k(s_{\bar{\mu}} - s_\pi)\|^2_{\mu_k}} \in \mathbb{R}, \end{aligned} \tag{23}$$

where for any $\mathbf{x} \in \mathbb{R}^d$, $\nabla L(\mu_k)(\mathbf{x}), s_{\bar{\mu}}(\mathbf{x}), s_\pi(\mathbf{x})$ are vectors in $\mathbb{R}^d$ given by

$$\begin{aligned} \nabla L(\mu_k)(\mathbf{x}) &= \int_{\mathbb{R}^d} (\nabla_1 \ell(\mathbf{x}, \mathbf{z}) + \nabla_2 \ell(\mathbf{z}, \mathbf{x})) d\mu_k(\mathbf{z}), \\ s_{\bar{\mu}}(\mathbf{x}) &= \nabla \log \bar{\mu}(\mathbf{x}) \text{ and } s_\pi(\mathbf{x}) = \nabla \log \pi(\mathbf{x}). \end{aligned}$$

We make the following assumptions similar to Assumptions 1 and 2 in Euclidean space.

Assumption 3. (i) the kernel ℓ is bounded, i.e., for some $\ell_{\max} > 0$, $|\ell(\mathbf{x}, \mathbf{y})| < \ell_{\max}$ for all $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$;

(ii) the gradient of kernel ℓ is bounded, i.e., for some $L_{\max} > 0$, $\|\nabla \ell(\mathbf{x}, \mathbf{y})\| < L_{\max}$ for all $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$.

(iii) π follows κ-log Soblev inequality, i.e., for some $\kappa > 0$, $\|s_\nu - s_\pi\|^2_\nu \geq \kappa \mathrm{KL}(\nu\|\pi)$ for all $\nu \in \mathcal{P}_{2,ac}(\mathbb{R}^d)$.

It is known that π with bounded support or is strongly-log concave satisfies the κ-log Sobolev inequality. We prove the following convergence result for the Wasserstein CCGF defined in Definition 8.

Theorem 2. Suppose Assumption 3 holds. Let $\boldsymbol{\mu}(t)$ be the Wasserstein CCGF in Definition 8. Then,

(i) KL-divergence between $\bar{\mu}(t)$ and π decreases exponentially, i.e., $\mathrm{KL}(\bar{\mu}(t)\|\pi) = e^{-\alpha t}\mathrm{KL}(\bar{\mu}(0)\|\pi)$;

(ii) optimality condition in Proposition 2 is approximately satisfied in the sense that for $T > 0$,

$$\min_{t \le T} \|\nabla_{\boldsymbol{\mu}} F_{\mathbf{p}}(\boldsymbol{\mu}(t))\|_{\mathcal{TP}_{\pi,\mathbf{p}}(\boldsymbol{\mu}(t))} \le \frac{C}{\sqrt{T}},$$

where

$$C \doteq \tfrac{1}{\sqrt{K}}(F_{\mathbf{p}}(\boldsymbol{\mu}(0)) + \ell_{\max} + \frac{4\alpha L_{\max}\sqrt{K}}{p_{\min}\sqrt{\kappa}}\sqrt{\mathrm{KL}(\bar{\mu}(0)\|\pi)} + \alpha\frac{1}{p_{\min}\kappa}\mathrm{KL}(\bar{\mu}(0)\|\pi)),$$

with $p_{\min} = \min_{k\in[K]} p_k$.

Theorem 2 shows that the Wasserstein CCGF $\boldsymbol{\mu}(t)$ in Definition 8 converges in the limit to a feasible solution (i.e., $\lim_{t\to\infty}\mathrm{KL}(\bar{\mu}(t)\|\pi) = 0$) exponentially fast with speed controlled by the parameter α. Furthermore, the tangent norm of the Wasserstein gradient decreases to 0 at the speed of $O(\frac{1}{\sqrt{T}})$, i.e., the KKT condition in Proposition 2 is satisfied in the limit as $t \to \infty$. We discuss an algorithmic implementation of the Wasserstein CCGF in Section 6.

5.3 Wasserstein CCGF for Problem 1. In this subsection, we extend the Wasserstein CCGF algorithm to solve Problem 1. We must also design a velocity field for the extra decision variable p_k.

The decision variable $(\boldsymbol{\mu}, \mathbf{p}) = ((\mu_1, \ldots, \mu_K), (p_1, \ldots, p_K))$ of Problem 1 is defined in the space $(\mathcal{P}_{2,ac}(\mathbb{R}^d), W_2)^{\otimes K} \times \mathbb{R}_+^K$. Similar to Section 5.2, we design a velocity field $\boldsymbol{\phi}$ in the space $(\mathcal{P}_{2,ac}(\mathbb{R}^d), W_2)^{\otimes K}$ for $\boldsymbol{\mu}$ and a velocity field $\mathbf{v}$ in the space $\mathbb{R}_+^K$ for $\mathbf{p}$. Specifically, given any $(\boldsymbol{\mu}, \mathbf{p}) \in (\mathcal{P}_{2,ac}(\mathbb{R}^d), W_2)^{\otimes K} \times \mathbb{R}_+^K$, $\boldsymbol{\phi}(\boldsymbol{\mu}, \mathbf{p}) = (\phi_1(\boldsymbol{\mu}, \mathbf{p}), \ldots, \phi_K(\boldsymbol{\mu}, \mathbf{p}))$ is a vector of velocity fields, where each $\phi_k(\boldsymbol{\mu}, \mathbf{p})(\cdot) : \mathbb{R}^d \to \mathbb{R}^d \in \mathcal{L}^2_{\mu_k}$ describes the velocity of decision variable μ_k in Wasserstein space; and $\mathbf{v}(\boldsymbol{\mu}, \mathbf{p})(\cdot) : \mathbb{R}_+^K \to \mathbb{R}_+^K \in \mathcal{L}^2$ is a velocity field of the decision variable $\mathbf{p} \in \mathbb{R}_+^K$.

Our chosen velocity fields $(\boldsymbol{\phi}, \mathbf{v})$ should solve Problem 1 in the following sense. Starting with initial densities $\boldsymbol{\mu}(0) = (\mu_1(0), \ldots, \mu_K(0))$ and initial weights $\mathbf{p}(0) = (p_1(0), \ldots, p_K(0))$, the solution of the following system of partial differential equations,

$$\begin{aligned} \frac{d}{dt}\mu_k(t) &= -\nabla \cdot (\phi_k(\boldsymbol{\mu}(t), \mathbf{p}(t))\mu_k(t)) \\ \frac{d}{dt}p_k(t) &= v_k(\boldsymbol{\mu}(t), \mathbf{p}(t)) \end{aligned} \tag{24}$$

for $k \in [K]$ converges to a "stationary point" of Problem 1 in the sense that

- $\sum_{k\in[K]} p_k(t) = 1$, and $p_k(t) \ge 0$, for all $k \in [K]$ and all $t \ge 0$;
- $\lim_{t\to\infty}\mathrm{KL}(\bar{\mu}(t)\|\pi) = 0$ with $\bar{\mu}(t) = \sum_{k\in[K]} p_k(t)\mu_k(t)$;
- the optimality condition in Proposition 3 is approximately satisfied.

To see this, we first need gradients of the objective function $F(\boldsymbol{\mu}, \mathbf{p}) = \sum_{k\in[K]} p_k L(\mu_k) + \theta/p_k^\beta$ and the constraint $\mathrm{KL}(\bar{\mu}\|\pi)$. The Wasserstein gradient of both $F(\boldsymbol{\mu}, \mathbf{p})$ and $\mathrm{KL}(\bar{\mu}\|\pi)$ with respect to $\boldsymbol{\mu}$ remains the same as given by Lemmas 4 and 5. The following lemma gives the gradients with respect to $\mathbf{p}$.

Lemma 8. The gradients $\nabla_{\mathbf{p}} F(\boldsymbol{\mu}, \mathbf{p}) \in \mathbb{R}^d$ and $\nabla_{\mathbf{p}} \mathrm{KL}(\bar{\mu}\|\pi) \in \mathbb{R}^d$ are

$$\nabla_{\mathbf{p}} F(\boldsymbol{\mu}, \mathbf{p})_k = L(\mu_k) - \frac{\theta\beta}{p_k^{\beta+1}},$$
$$\nabla_{\mathbf{p}} \mathrm{KL}(\boldsymbol{\mu}\|\pi)_k = \int_{\mathbb{R}^d} \mu_k(x) \log \frac{\bar{\mu}(x)}{\pi(x)} dx + 1.$$

We define the following objects to extend the Wasserstein CCGF for Problem 2. Recall in Section 3, given any $\boldsymbol{\mu} \in \mathcal{P}_{2,ac}(\mathbb{R}^d)^{\otimes K}$, we define inner product $\langle \cdot, \cdot, \rangle_{\boldsymbol{\mu}}$ in the space $\mathcal{L}^2_{\boldsymbol{\mu}}(\mathbb{R}^d; \mathbb{R}^d)$ by $\langle f, g\rangle_{\boldsymbol{\mu}} = \sum_{k\in[K]} \int_{\mathbb{R}^d} \langle f_k(x), g_k(x)\rangle d\mu_k(x)$ for any $f, g \in \mathcal{L}^2_{\boldsymbol{\mu}}(\mathbb{R}^d; \mathbb{R}^d)$. Particularly, since $(\nabla_{\boldsymbol{\mu}} F(\boldsymbol{\mu}, \mathbf{p}), \nabla_{\boldsymbol{\mu}} \mathrm{KL}(\bar{\mu}\|\pi))$ are both in $\mathcal{L}^2_{\boldsymbol{\mu}}(\mathbb{R}^d; \mathbb{R}^d)$, we have

$$\langle \nabla_{\boldsymbol{\mu}} F(\boldsymbol{\mu}, \mathbf{p}), \nabla_{\boldsymbol{\mu}} \mathrm{KL}(\bar{\mu}\|\pi)\rangle_{\boldsymbol{\mu}} = \sum_{k\in[K]} \int_{\mathbb{R}^d} \langle \nabla_{\mu_k} F(\boldsymbol{\mu}, \mathbf{p}), \nabla_{\mu_k} \mathrm{KL}(\bar{\mu}\|\pi)\rangle d\mu_k.$$

To ensure that $\sum_{k\in[K]} p_k(t) = 1$ for all $t \geq 0$, we project the velocity field $\mathbf{v}(t)$ for $\mathbf{p}(t)$ so that $\sum_{k\in[K]} v_k = 0$. Let $\mathbf{1}$ denote the column vector $(1, \ldots, 1)^\top \in \mathbb{R}^K$. We define the projection operator $P : \mathbb{R}^K \to \mathbb{R}^K$ by

$$Pv = (\mathrm{id} - \frac{1}{K}\mathbf{1}\mathbf{1}^\top)v = v - \frac{\sum_{k\in[K]} v_k}{K}\mathbf{1}.$$

for $v \in \mathbb{R}^K$. This projection ensures that $\sum_{k\in[K]} Pv_k = 0$. With this projection operator, we can define an inner product and norm: given any $\boldsymbol{\mu} \in \mathcal{P}_2(\mathbb{R}^d)^{\otimes K}$ and $\mathbf{u} = (u_1, u_2), \mathbf{w} = (w_1, w_2)$ with $u_1, w_1 \in \mathcal{L}^2_{\boldsymbol{\mu}}(\mathbb{R}^d; \mathbb{R}^d)$ and $u_2, w_2 \in \mathbb{R}^K$,

$$\langle \mathbf{u}, \mathbf{w}\rangle_{\boldsymbol{\mu},P} = \langle u_1, w_1\rangle_{\boldsymbol{\mu}} + \langle Pu_2, Pw_2\rangle,$$
$$\|\mathbf{u}\|^2_{\boldsymbol{\mu},P} = \|u_1\|^2_{\boldsymbol{\mu}} + \|Pu_2\|^2,$$

where $\langle Pu_2, Pw_2\rangle$ and $\|Pu_2\|$ are the Euclidean inner product and norm, respectively.

In particular, since $\nabla F(\boldsymbol{\mu}, \mathbf{p}) = (\nabla_{\boldsymbol{\mu}} F(\boldsymbol{\mu}, \mathbf{p}), \nabla_{\mathbf{p}} F(\boldsymbol{\mu}, \mathbf{p}))$ and $\nabla \mathrm{KL}(\bar{\mu}\|\pi) = (\nabla_{\boldsymbol{\mu}} \mathrm{KL}(\bar{\mu}\|\pi), \nabla_{\mathbf{p}} \mathrm{KL}(\bar{\mu}\|\pi))$ are both in $\mathcal{L}^2_{\boldsymbol{\mu}} \times \mathbb{R}^K$, we have

$$\langle \nabla F(\boldsymbol{\mu}, \mathbf{p}), \nabla \mathrm{KL}(\bar{\mu}\|\pi)\rangle_{\boldsymbol{\mu},P} = \langle \nabla_{\boldsymbol{\mu}} F(\boldsymbol{\mu}, \mathbf{p}), \nabla_{\boldsymbol{\mu}} \mathrm{KL}(\bar{\mu}\|\pi))\rangle_{\boldsymbol{\mu}} + \langle P\nabla_{\mathbf{p}} F(\boldsymbol{\mu}, \mathbf{p}), P\nabla_{\mathbf{p}} \mathrm{KL}(\bar{\mu}\|\pi)\rangle,$$
$$\|\nabla \mathrm{KL}(\bar{\mu}\|\pi)\|^2_{\boldsymbol{\mu},P} = \|\nabla_{\boldsymbol{\mu}} \mathrm{KL}(\bar{\mu}\|\pi)\|^2_{\boldsymbol{\mu}} + \|P\nabla_{\mathbf{p}} \mathrm{KL}(\bar{\mu}\|\pi)\|^2.$$

Definition 9 (Wasserstein CCGF with dynamic weights)**.** The *Wasserstein constraint controlled gradient flow with dynamic weights* is defined to be the solution $\boldsymbol{\mu}(t) : \mathbb{R} \to (\mathcal{P}_{2,ac}(\mathbb{R}^d), W_2)^{\otimes K}$ and $\mathbf{p}(t) : \mathbb{R} \to \mathbb{R}^K$ to the system (24) of PDEs, where the velocity fields $\phi_k(\boldsymbol{\mu}, \mathbf{p})(\cdot) : \mathbb{R}^d \to \mathbb{R}^d$ and $\mathbf{v}(\boldsymbol{\mu}, \mathbf{p})(\cdot) : \mathbb{R}^K \to \mathbb{R}^K$ are, for all $k \in [K]$,

$$\begin{aligned} \phi_k(\boldsymbol{\mu}, \mathbf{p}) &= -(\nabla_{\mu_k} F(\boldsymbol{\mu}, \mathbf{p}) + \lambda(\boldsymbol{\mu}, \mathbf{p})\nabla_{\mu_k} \mathrm{KL}(\bar{\mu}\|\pi)), \\ v_k(\boldsymbol{\mu}, \mathbf{p}) &= -P(\nabla_{\mathbf{p}} F(\boldsymbol{\mu}, \mathbf{p})_k + \lambda(\boldsymbol{\mu}, \mathbf{p})\nabla_{\mathbf{p}} \mathrm{KL}(\bar{\mu}\|\pi)_k), \text{ and} \\ \lambda(\boldsymbol{\mu}, \mathbf{p}) &= \frac{-\langle \nabla F(\boldsymbol{\mu}, \mathbf{p}), \nabla \mathrm{KL}(\bar{\mu}\|\pi)\rangle_{\boldsymbol{\mu},P} + \alpha \mathrm{KL}(\bar{\mu}\|\pi)}{\|\nabla \mathrm{KL}(\bar{\mu}\|\pi)\|^2_{\boldsymbol{\mu},P}}. \end{aligned} \tag{25}$$

We prove the following convergence result for Wasserstein CCGF with dynamic weights defined in Definition 9.

Theorem 3. Suppose Assumption 3 holds. Let $(\boldsymbol{\mu}(t), \mathbf{p}(t))$ be the Wasserstein CCGF with dynamic weights defined in Definition 9. Then,

(i) $\sum_{k\in[K]} p_k(t) = 1$ and $\mathbf{p}(t) \geq p_{\min}$ m-almost surely, for some $p_{\min} > 0$;

(ii) the KL-divergence between $\bar{\mu}(t)$ and π decreases exponentially, i.e., $\mathrm{KL}(\bar{\mu}(t)\|\pi) = e^{-\alpha t}\mathrm{KL}(\bar{\mu}(0)\|\pi)$;

(iii) optimality condition in Proposition 3 is approximately satisfied in the sense that for all $T > 0$,

$$\min_{t \le T} \|\nabla F(\boldsymbol{\mu}(t), \mathbf{p}(t))\|_{\mathcal{T}\mathcal{P}_\pi(\boldsymbol{\mu}(t), \mathbf{p}(t))} \le \frac{C}{\sqrt{T}},$$

where

$$C \doteq F(0) + \ell_{\max} + \frac{4}{\alpha p_{\min}\sqrt{\kappa}}\sqrt{\mathrm{KL}(0)}\sqrt{L_{\max}^2 K^2 + K(\ell_{\max} + \frac{\theta\beta}{p_{\min}^{\beta+1}})^2} + \frac{1}{p_{\min}\kappa}\mathrm{KL}(0).$$

Theorem 3 shows that the Wasserstein CCGF with dynamic weights $(\boldsymbol{\mu}(t), \mathbf{p}(t))$ in Definition 9 (i) guarantees that $\mathbf{p}(t) \ge 0$ and $\sum_{k\in[K]} p_k(t) = 1$ almost surely for all $t > 0$; (ii) converges in the limit to a feasible solution (i.e., $\lim_{t\to\infty} \mathrm{KL}(\bar{\mu}(t)\|\pi) = 0$) exponentially fast with speed controlled by the parameter α. Furthermore, the tangent norm of the Wasserstein gradient decreases to 0 at the speed of $O(\frac{1}{\sqrt{T}})$, i.e., the KKT condition in Proposition 3 is satisfied in the limit as $t \to \infty$. An algorithmic implementation of this flow is discussed in Section 6.

6 Implementation of Wasserstein CCGF

In this section, we implement Definitions 8 and 9 in the settings described by Examples 1 and 2. In particular, we conduct a detailed analysis of the league design problem described in Example 1.

6.1 Implementation of Wasserstein flow in Definition 8. We begin with Definition 8. Recall that we aim to generate a curve of probability densities $\boldsymbol{\mu}(\cdot) : [0, T] \to \mathcal{P}_{2,ac}(\mathbb{R}^d)^{\otimes K}$ (i.e., $\boldsymbol{\mu}(t) = (\mu_1(t), \ldots, \mu_K(t))$) to solve Problem 2. By Theorem 2, this is achieved by solving the system of K continuity equations (22) given by our velocity field design $\boldsymbol{\phi}(\boldsymbol{\mu}) = (\phi_1(\boldsymbol{\mu}), \ldots, \phi_K(\boldsymbol{\mu}))$ in Definition 8. Implementing this Wasserstein CCGF involves two challenges. First, each $\mu_k(t)$ is infinite-dimensional and non-parametric, thus hard to represent. Second, we need to decide $\mu_k(t)$ for each t in a continuous interval $[0, T]$. We address these challenges via the "particles method". Specifically, we use a large population of particles sampled from $\mu_k(t)$ to represent the distribution with density $\mu_k(t)$ and discretize the time horizon $[0, T]$ by considering time steps $t = 0, 1, 2, \ldots$ to get an iterative algorithm.

For any curve $\boldsymbol{\mu}(\cdot) : [0, T] \to \mathcal{P}_{2,ac}(\mathbb{R}^d)^{\otimes K}$, at each $t \in [0, T]$, let $\phi_k(t, \cdot) : \mathbb{R}^d \to \mathbb{R}^d$ denote the velocity field $\phi_k(\boldsymbol{\mu}(t))$ given in Definition 8. For each $k \in [K]$, we sample N particles from an initial probability density $\mu_k(0)$. If all particles move together according to velocity field ϕ_k (i.e., the trajectory of each particle initially located at $\mathbf{x}_0$ is given by the solution of $\dot{\mathbf{x}}(t) = \phi_k(t, \mathbf{x}(t))$ with $\mathbf{x}(0) = \mathbf{x}_0$), then by Lemma 2, the curve $\mu_k(t)$ solves the continuity equation (22) together with the velocity field ϕ_k.

To move these N particles according to velocity field $\phi_k(t)$, we discretize the interval by considering time step $t = 0, 1, 2, \ldots$ and apply the following iterative scheme: for particle located at $\mathbf{x}_t$ at time step t, we generate its position $\mathbf{x}_{t+1}$ at time step $t + 1$ by

$$\mathbf{x}_{t+1} = \mathbf{x}_t + \eta\phi_k(t, \mathbf{x}_t),$$

for some small step size $\eta > 0$; that is, at each discrete time step t, we move the position of particle located at $\mathbf{x}_t$ along the direction given by $\phi_k(t, \mathbf{x}_t)$ by a small step. If this step size η is small enough, this discrete sequence $\{\mathbf{x}_t\}_{t\in\mathbb{N}}$ should approximate the original continuous curve $\mathbf{x}(t)$. Therefore, we get the following particle algorithm (Algorithm 1).

One final question remains: how can we utilize the output? We need the probability density

Algorithm 1 Wasserstein CCGF particle algorithm

Require: Input the number of iterations T, the step size $\eta > 0$, the number K of sub-populations, and k initial probability distributions $\mu_k(0)$.

1: For each $k \in [K]$, randomly sample a population of N particles $X_k(0) \doteq (\mathbf{x}_k^i(0))_{i=1}^N$ with $\mathbf{x}_k^i(0) \in \mathbb{R}^d$ according to $\mu_k(0)$. Similarly, define $X_k(t) \doteq (\mathbf{x}_k^i(t))_{i=1}^N$ for all $t \geq 0$.
2: **for** each time step $t = 1, \ldots, T$ **do**
3: **for** each sub-population $k = 1, \ldots, K$ **do**
4: **for** each particle $i = 1, \ldots, N$ **do**
5: Compute the velocity $\phi_k(t, \mathbf{x}_k^i(t)) \in \mathbb{R}^d$ according to equation (23) under Definition 8.
6: Update the particle with: $\mathbf{x}_k^i(t+1) = \mathbf{x}_k^i(t) + \eta \phi_k(t, \mathbf{x}_k^i(t))$.
7: **end for**
8: **end for**
9: **end for**
10: Output $X_k(T)$ for all $k \in [K]$.

function μ_k for each $k \in [K]$. We can approximate this density function μ_k with the population X_k of particles. If one is interested in computing the integral $\int_{\mathbb{R}^d} f(\mathbf{x}) d\mu_k(\mathbf{x})$ for some function $f : \mathbb{R}^d \to \mathbb{R}$ with respect to μ_k, this integral can be approximated by

$$\int_{\mathbb{R}^d} f(\mathbf{x}) d\mu_k(\mathbf{x}) \approx \frac{1}{N} \sum_{i=1}^N f(\mathbf{x}_k^i).$$

The density function value $\mu_k(\mathbf{x})$ at any point $\mathbf{x} \in \mathbb{R}^d$ can also be approximated by the method of "kernel density estimation". Let $\psi : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ be

$$\psi(\mathbf{x}, \mathbf{y}) \doteq \frac{1}{\sqrt{2\pi}} e^{-\frac{\|\mathbf{x}-\mathbf{y}\|^2}{2}}.$$

Then, the density $\mu_k(\mathbf{x})$ can be approximated by

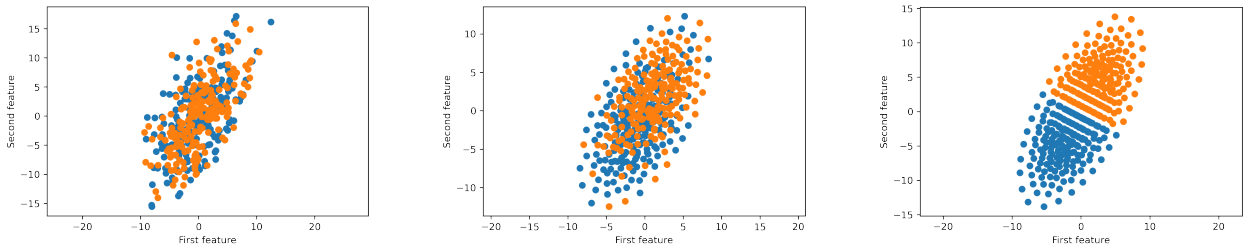
$$\mu_k(\mathbf{x}) \approx \frac{1}{N} \sum_{i=1}^N \psi(\mathbf{x}, \mathbf{x}_k^i), \quad \forall \mathbf{x} \in \mathbb{R}^d.$$

In the following example, we demonstrate how Algorithm 1 works.

Example 5. Let $\pi \in \mathcal{P}_{2,ac}(\mathbb{R}^2)$ be the density function of bivariate normal distribution $\mathcal{N}(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 4^2, 24\rho \\ 24\rho, 6^2 \end{bmatrix})$ with $\rho = 0.6$. In Problem 2, suppose we want to decompose π into two densities μ_1 and μ_2 with equal weights $p_1 = p_2 = 1/2$. Consider the loss

$$L(\mu) = \int_{\mathbb{R}^2} \int_{\mathbb{R}^2} \|\mathbf{x} - \mathbf{y}\|^2 d\mu(\mathbf{x}) d\mu(\mathbf{y}).$$

We implement Algorithm 1 to solve this instance of Problem 2 with $N = 200$ particles. Figure 1 shows the movement of particles in our velocity field design. As shown in Figure 1(a), for each $k = 1, 2$ we sample 200 particles to represent μ_k. Each colored dot is one particle and the 200 blue (resp. orange) particles represent probability density μ_1 (resp. μ_2). Then, we implement Algorithm 1 to compute the velocity of each particle (line 5 in Algorithm 1) and move each particle along the velocity by a short step (line 6 in Algorithm 1). After 100 such iterations, the 400 particles are arranged in an ellipse, as depicted in Figure 1(b). This elliptical shape is expected since the underlying distribution π is a bivariate normal. The contour plot of a bivariate normal density resembles an ellipse, and these points are intended to represent densities



(a) Initial positions of particles. (b) Particles at $t = 100$. (c) Final positions of particles.

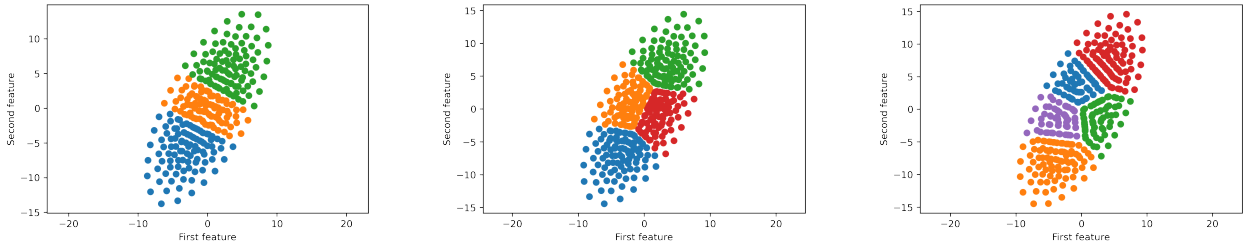
Figure 1: Movement of particles in Wasserstein CCGF.

μ_1 and μ_2, with $\mu_1/2 + \mu_2/2 = \pi$. However, particles of different colors are still intermingled after 100 iterations. As we demonstrated in Corollary 1, particles of different colors should separate from each other. This indeed occurs after a sufficient number of iterations, as illustrated in Figure 1(c). Particles are separated into two disjoint parts, each representing the support of one density μ_k. The area with denser particles has a higher density value. Additionally, particles of both colors are arranged in an ellipse, illustrating the contour plot of the bivariate normal density π. The final positions of the particles illustrate how the distribution π is optimally decomposed into μ_1 and μ_2.

Example 6 (Multiple classes)**.** We can also decompose the probability density $\pi \in \mathcal{P}_{2,ac}(\mathbb{R}^2)$ into multiple ($K > 2$) probability densities. Let $\pi \in \mathcal{P}_{2,ac}(\mathbb{R}^2)$ be the density function of the bivariate normal distribution as given in Example 5. Let σ_{X_1}, σ_{X_2} and ρ be the standard deviation of feature one, feature two, and the correlation between feature one and two, respectively. Let $\mu_{X_1} = \mu_{X_2} = 0$. In Problem 2, suppose we want to decompose π into K densities $\mu_1, \ldots, \mu_K$ with equal weights $p_k = 1/K$ for all $k \in [K]$. Consider the loss function L given by Example 2 with $W = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$. Figure 2 shows the decomposition found by Algorithm 1 under different setups. Again, each population of particles of one color represents one density μ_k for $k \in [K]$. The positions of particles of one color resemble the contour plot of the corresponding density function; that is, an area with denser particles has a larger density value. According to Corollary 1, the populations of particles of different colors should be disjoint, as illustrated in Figure 2. Additionally, we observe that the optimal decompositions may not always appear intuitive. For example, the decompositions shown in Figure 2(b), (c), and (f) are not immediately obvious. By contrast, Figure 2(a), (d), and (e) are more natural-looking partitions of the space due to their simple geometric nature.

We use the following natural decompositions as benchmarks to show the efficacy of the decompositions discovered by Algorithm 1.

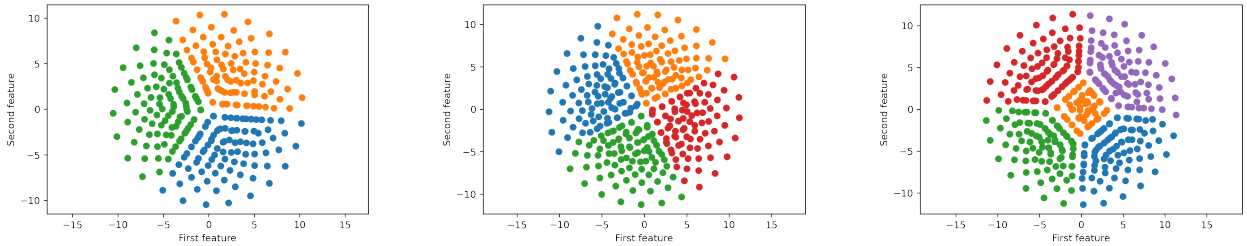
Example 7 (Comparision with natural decomposition)**.** Consider the settings described in Example 6. We consider the following simple ways (Figure 3) using parallel slices to decompose the underlying distribution π into K densities $\mu_1, \ldots, \mu_K$. Recall that we require $\sum_{k=1}^{K} \mu_k/K = \pi$. Compared with the alternative decompositions shown in Figure 3, the decompositions discovered by Algorithm 1 shown in Figure 2 improves the objective value (i.e., Problem 2 with variance



(a) $\sigma_{X_1} = 4, \sigma_{X_2} = 6, \rho = 0.6$ and $K = 3$

(b) $\sigma_{X_1} = 4, \sigma_{X_2} = 6, \rho = 0.6$ and $K = 4$

(c) $\sigma_{X_1} = 4, \sigma_{X_2} = 6, \rho = 0.6$ and $K = 5$



(d) $\sigma_{X_1} = 5, \sigma_{X_2} = 5, \rho = 0$ and $K = 3$

(e) $\sigma_{X_1} = 5, \sigma_{X_2} = 5, \rho = 0$ and $K = 4$

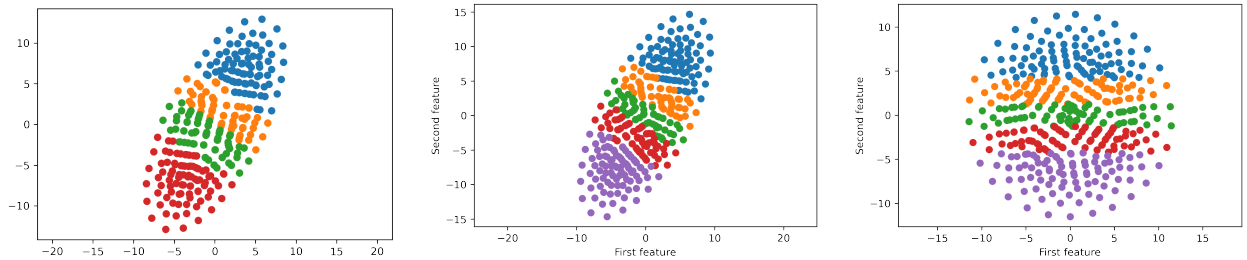
(f) $\sigma_{X_1} = 5, \sigma_{X_2} = 5, \rho = 0$ and $K = 5$

Figure 2: Decomposition with multiple groups.

loss described in Example 2) by 7.90%, 12.33% and 47.67%, respectively.

In certain scenarios, decision-makers need to divide the users into several groups that are homogeneous in one feature and diverse in other features. For example, one might want to group people with similar interests but have diverse demographic backgrounds.

Example 8. We can also maximize the similarity of feature one and maximize the diversity of feature two simultaneously. This can be achieved by considering the loss function L given by Example 2 with $W = \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix}$. In this case, our algorithm minimizes the variance of feature one minus the variance of feature two. Let $\pi \in \mathcal{P}_{2,ac}(\mathbb{R}^2)$ be the density function of a mixed bivariate normal distribution, i.e., $\pi = 1/3\pi_1 + 1/3\pi_2 + 1/3\pi_3$, where π_1, π_2, π_3 are densities of nor-

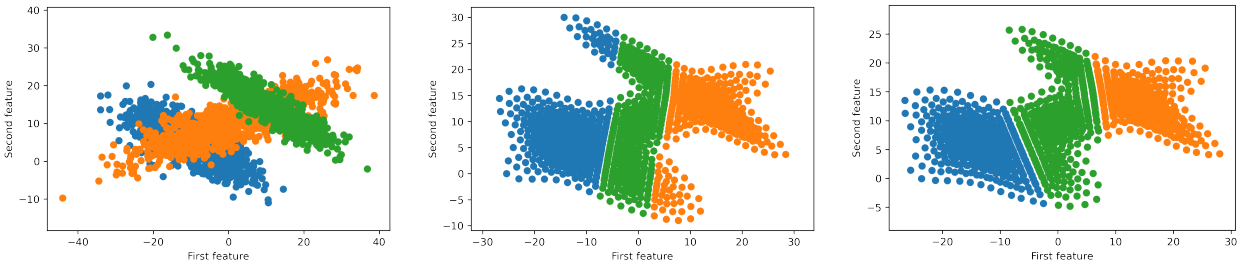


(a) $\sigma_{X_1} = 4, \sigma_{X_2} = 6, \rho = 0.6$ and $K = 4$

(b) $\sigma_{X_1} = 4, \sigma_{X_2} = 6, \rho = 0.6$ and $K = 5$

(c) $\sigma_{X_1} = 5, \sigma_{X_2} = 5, \rho = 0$ and $K = 5$

Figure 3: Results of Example 7



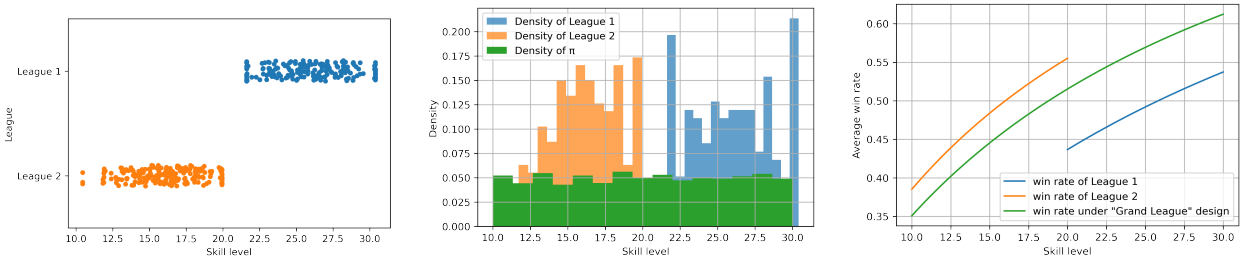
(a) Particles sampled from π. (b) Decomposition with W. (c) Decomposition when W is the identity matrix.

Figure 4: Results of Example 8

mal distributions with parameters $(\mu_{X_1}, \mu_{X_1}, \sigma_{X_1}, \sigma_{X_1}, \rho) = (-10, 5, 6, 5, -0.7), (0, 10, 7, 5, 0.8)$ and $(10, 15, 8, 5, -0.9)$, respectively. Figure 4(a) shows particles sampled from π. Figure 4(b) shows the decomposition generated by Algorithm 1. The result in Figure 4(b) is not surprising since, with this weight matrix W, we aim to minimize the dissimilarity of the first feature and maximize the diversity of the second feature. Notice that the regions in Figure 4(b) are "narrow" horizontally but "tall" vertically. For comparison, Figure 4(c) shows the decomposition when W is the identity matrix, which aims to define regions that are both "narrow" and "short". In this case, we minimize the variance of features one and two.

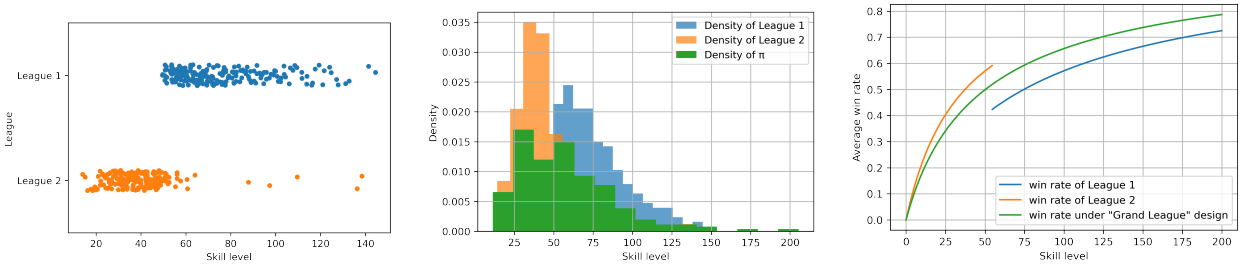
6.2 Case study: League design with Elo scores. In this subsection, we explore the league design problem in Example 1. Recall that the distribution loss function L for this problem is defined in (3).

Example 9 (π is uniform distribution)**.** Let $\pi \in \mathcal{P}_{2,ac}(\mathbb{R})$ be the density function of a uniform distribution over interval $[10, 30]$. We implement Algorithm 1 to solve this instance of Problem 2 with $N = 200$ particles. Our benchmark is the "Grand League" design, where all players are placed in the same league. This essentially means there is no specific league design in place. Figure 5(a) shows the final positions of particles given by Algorithm 1. Blue (resp. orange) particles in league 1 (resp. league 2) represent the density μ_1 (resp. μ_2). Figure 5(b) shows the histogram of μ_1 (in blue) and μ_2 (in orange), which are basically the densities of uniform distributions over $[10, 20]$ and $[20, 30]$, respectively. This means the optimal league design is to have a "novice" league (for players with skill levels in $[10, 20]$) and a "veteran" league (for players with skill levels in $[20, 30]$). This result is not surprising. Intuitively, since π is uniform distribution and we require $p_1 = p_2 = 1/2$, to minimize the Elo loss, μ_1 and μ_2 should also be uniform distributions supported on intervals with length 10. Figure 5(c) shows the average win rate of each skill level. The green curve represents the average win rate of each skill level under the "Grand League" design; that is, for players with skill level x, the average win rate is $\int_{y\in[10,30]} \frac{x}{x+y} d\pi(y)$. The blue curve represents the average win rate of each skill level in league 1 (veteran league), i.e., for players with skill level $x \in [20, 30]$, the average win rate is $\int_{y\in[20,30]} \frac{x}{x+y} d\mu_1(y)$. The orange curve represents the average win rate of each skill level in league 2 (novice league), i.e., for players with skill level $x \in [10, 20]$, the average win rate is $\int_{y\in[10,20]} \frac{x}{x+y} d\mu_2(y)$. Compared with the "Grand League" design, players in the novice league



(a) Final positions of particles (b) Histogram of μ_1, μ_2 and π (c) Average win rate

Figure 5: Results of Example 9



(a) Final positions of particles (b) Histogram of μ_1, μ_2 and π (c) Average win rate before and after grouping

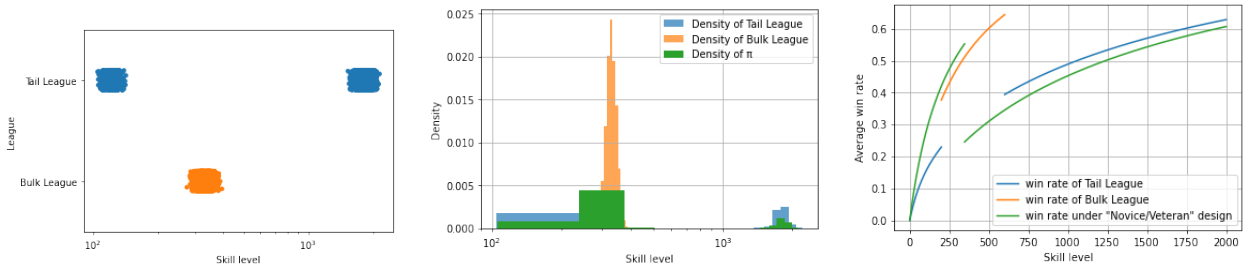
Figure 6: Results of Example 10

have higher win rates, and players in the veteran league have lower win rates. This is generally considered desirable as it reduces frustration for novice players and increases the challenge for veteran players. Particularly, notice that players in the veteran league have roughly the same win rates as the "high-end" players (in $[12.5, 20]$) in the novice league.

Example 10 (π is lognormal distribution)**.** Let $\pi \in \mathcal{P}_{2,ac}(\mathbb{R})$ be the density function of a lognormal distribution with parameter $(4, 0.5^2)$. As shown in the Figure 6(b), the density is single-peaked, indicated by the green color. We implement Algorithm 1 to solve this instance of Problem 2 with $N = 200$ particles. Again, we use the "Grand League" design as our benchmark. Figure 6(a) shows the final positions of particles given by Algorithm 1. Blue (resp. orange) particles in league 1 (resp. league 2) represent the density μ_1 (resp. μ_2). Figure 6(b) shows the histogram of μ_1 (in blue) and μ_2 (in orange). The optimal league design still consists of a novice league (league 2) and a veteran league (league 1). Figure 6(c) shows the average win rate of each skill level, which are computed in the same way as shown in Example 9. Compared with the "Grand League" design, players in the novice league have higher win rates, and players in the veteran league have lower win rates.

However, the "novice/veteran" league design is not always optimal. As we showed in Example 3, it is possible that the support of optimal solution is not convex in pairs. As shown below, this can happen if π is a mixed lognormal distribution with three peaks.

Example 11 (π is mixed lognormal distribution)**.** Suppose π is the density of a mixed log-



(a) Final positions of particles

(b) Histogram of μ_1, μ_2 and π

(c) Average win rate under different designs

Figure 7: Resuts of Example 11

normal distribution, i.e., $\pi = 0.1\pi_1 + 0.6\pi_2 + 0.3\pi_3$, where π_1, π_2, π_3 are densities of lognormal distributions with parameters $(4.8, 0.05^2), (5.8, 0.05^2)$, and $(7.5, 0.05^2)$. In this example, we decompose π into two probability densities μ_1 and μ_2 with weights $p_1 = 0.4$ and $p_2 = 0.6$. We use the "novice/veteran" league design as our benchmark. The novice league (resp. veteran league) consists of players with skill levels in $[0, 347]$ (resp. $[347, 2000]$). This boundary is chosen to guarantee that the weight of the novice league is 0.4 as required. Figure 7(a) shows one outcome of Algorithm 1. Blue particles form a "tail league" representing density μ_1 while orange particles form a "bulk league" representing density μ_2. Figure 7(b) shows the histogram of μ_1 (in blue) and μ_2 (in orange). Note that the optimal league design consists of a tail and bulk league. Compared with the "novice/veteran" league design, the Elo loss (i.e., the objective value in Problem 2 with Elo loss function in Example 1) is reduced by 7.49%. Figure 7(c) shows the average win rate of each skill level, which are computed in the same way as shown in Example 9. The green curve represents the win rate of each skill level under the "novice/veteran" league design. The "tail/bulk" league design significantly decreases the win rate of novice players (those with skill levels in the range of 0 to 347), markedly increasing the win rate of veteran players (those with skill levels greater than 347). In practice, this outcome is often considered undesirable as it may lead to frustration among low-skill players because they are paired with high-skill players and are likely to lose. This is an illustration of the failure of the convex in pairs condition.

6.3 Implementation of Wasserstein flow in Definition 9. In this subsection, we implement Definition 9 to solve Problem 1 in the setting described by Examples 1 and 2. In Definition 9, we also need to update the weight p_k of each density μ_k. Again, we discreterize the time and flow to get the following implementable algorithm (Algorithm 2).

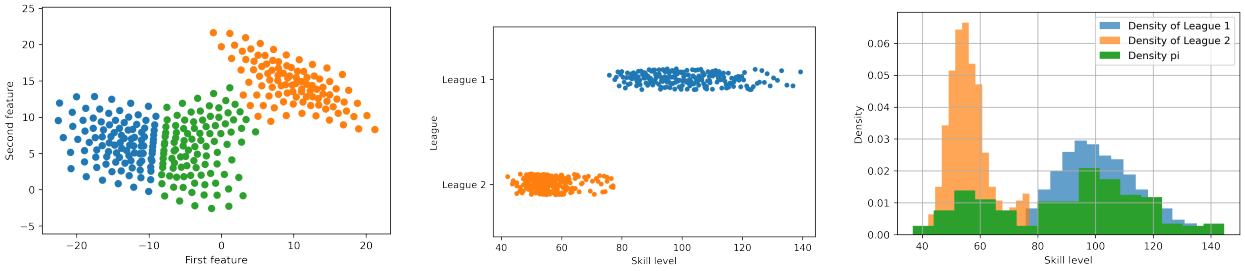
In the following example, we show that it is not always optimal to have equal weights.

Example 12 (WCCGF with dynamic weights)**. (i)** We revisit the setting described by Example 8 with W equal to the identity matrix where the result is shown in Figure 4(c). The output of Algorithm 2 is shown in Figure 8(a) with weight of blue particle-population being 0.2881, green particle-population being 0.3087, and orange particle-population being 0.4322. In comparison, the weights in Example 8 are fixed at $1/3$ for all populations. Compared with the result in Example 8, solution of Algorithm 2 reduces the objective value (computed by (2)) by

Algorithm 2 Wasserstein CCGF particles algorithm with dynamic weights

Require: Input number of iterations T, step size $\eta, \eta_2 > 0$, number K of sub-populations, and initial distributions $\mu_k(0)$.

For each $k \in [K]$, randomly sample a population of N particles $X_k(0) \doteq (\mathbf{x}_k^i(0))_{i=1}^N$ with $\mathbf{x}_k^i(0) \in \mathbb{R}^d$ from the initial distributions $\mu_k(0)$. Similarly, define $X_k(t) \doteq (\mathbf{x}_k^i(t))_{i=1}^N$ for all $t \geq 0$.
Initialize $p_k(0)$ for $k \in [K]$.
for each time step $t = 1, \ldots, T$ **do**
  for each sub-population $k = 1, \ldots, K$ **do**
    for each particle $i = 1, \ldots, N$ **do**
      Compute the velocity $\phi_k(t, \mathbf{x}_k^i(t)) \in \mathbb{R}^d$ according to equation (25) in Definition 9.
      Update the particle with: $\mathbf{x}_k^i(t+1) = \mathbf{x}_k^i(t) + \eta \phi_k(t, \mathbf{x}_k^i(t))$.
    end for
    Compute the velocity $v_k(t, p_k(t)) \in \mathbb{R}$ according to equation (25).
    Update the weight with $p_k(t+1) = p_k(t) + \eta_2 v_k(t, p_k(t))$.
  end for
end for
Output $X_k(T)$ and $p_k(T)$ for all $k \in [K]$.



(a) Final positions of particles (b) Final positions of particles (c) Histogram of μ_1, μ_2 and π

Figure 8: Resuts of Example 12

21%, reduces the weighted variance loss (computed by (6)) by 37.15%, and increases the total weight loss (computed by $\sum_{k=1}^K 1/p_k$) by 2.12%. Thus, a slight increase in the weight loss leads to a notable reduction in the weighted distribution loss and objective value. We can observe that, compared with Figure 4(c), orange particle population in Figure 8(a) occupied a larger area, while the blue particle population shrinks. These observations are also reflected in the change of their corresponding weights.

(ii) Suppose π is the density of a mixed lognoraml distribution defined by $\pi = 0.3\pi_1 + 0.7\pi_2$, where π_1 and π_2 are densities of lognormal distributions with (scale, shape) parameters given by $(4, 0.1)$ and $(4.6, 0.15)$. The density π is depicted by green color in Figure 8(c). The output of Algorithm 2 is shown in Figure 8(b) with weight of orange particle-population being 0.3303 and blue particle-population being 0.6697. Algorithm 2 creates a "novice/veteran" league design with 33.03% players being categorized as "novice", and 66.97% players being categorized as "veteran". Figure 8(c) shows the histogram of μ_1 and μ_2. Compared with the 50%-50% design,

league design of Algorithm 2 reduces the objective value (computed by (2)) by 32.85%, reduces the weighted Elo loss (computed by (6)) by 37.25%, and increases the total weight loss (computed by $0.0001 \sum_{k=1}^{K} 1/p_k$) by 13.02%. This 0.0001 guarantees that the weight loss is in the same scale as the distribution loss.

7 Conclusion

In the paper we have provided approaches to solving the optimal probability measure decomposition problem. The approaches use fresh ideas from Wasserstein gradient flow to handle a constrained problem. The numerical implementation uses a particle-based method that provides insights into our motivating examples.

Of course, there are opportunities for future work to extend our approach. On the application side, one could study the league design problem while also considering the spending habits of players as another dimension to designing leagues. For example, players who are big spenders in a game could be given higher priority for winning to keep them highly engaged with the game. This would require the loss function and applying our methodology to this new setting. Many similar extended applications are possible.

On the theory side, one could consider adapting our algorithm to tackle additional constraints for the nature of sub-population decomposition. For example, one could consider requiring sub-populations to have similar moments. This would require developing new optimality conditions and, adapting new optimality conditions and adapting the gradient flow algorithms accordingly.

References

Sharad Agarwal and Jacob R Lorch. Matchmaking for online games and other latency-sensitive P2P systems. In *Proceedings of the ACM SIGCOMM 2009 Conference on Data Communication*, pages 315–326, 2009.

Priyank Agrawal, Eric Balkanski, Vasilis Gkatzelis, Tingting Ou, and Xizhi Tan. Learning-augmented mechanism design: Leveraging predictions for facility location. *Mathematics of Operations Research*, 2023.

Luigi Ambrosio, Nicola Gigli, and Giuseppe Savaré. *Gradient Flows: In Metric Spaces and in the Space of Probability Measures*. Springer, 2005.

Dimitris Bertsimas, Melvyn Sim, and Meilin Zhang. Adaptive distributionally robust optimization. *Management Science*, 65(2):604–618, 2019.

Johannes Blömer, Sascha Brauer, and Kathrin Bujna. A complexity theoretical study of fuzzy K-means. *ACM Transactions on Algorithms (TALG)*, 16(4):1–25, 2020.

Yan Chen, Peter Cramton, John A List, and Axel Ockenfels. Market design, human behavior, and management. *Management Science*, 67(9):5317–5348, 2021.

Sinho Chewi. *Log-concave Sampling*. Book draft available at https://chewisinho.github.io, 2023.

Sinho Chewi, Tyler Maunu, Philippe Rigollet, and Austin J Stromme. Gradient descent algorithms for Bures-Wasserstein barycenters. In *Conference on Learning Theory*, pages 1276–1304. PMLR, 2020.

Lenaic Chizat and Francis Bach. On the global convergence of gradient descent for over-parameterized models using optimal transport. *Advances in Neural Information Processing Systems*, 31, 2018.

Xiaowu Dai and Michael I Jordan. Learning strategies in decentralized matching markets under uncertain preferences. *Journal of Machine Learning Research*, 22(260):1–50, 2021.

Charles Henry Edwards. *Advanced Calculus of Several Variables*. Courier Corporation, 1994.

Arpad E Elo and Sam Sloan. The rating of chess players: Past and present. 1978.

Lawrence Craig Evans. *Measure Theory and Fine Properties of Functions*. Routledge, 2018.

Yannick Francillette, Lylia Abrouk, and Abdelkader Gouaich. A players clustering method to enhance the players' experience in multi-player games. In *Proceedings of CGAMES'2013 USA*, pages 229–234. IEEE, 2013.

Yaofang Hu, Wanjie Wang, and Yi Yu. Graph matching beyond perfectly-overlapping erdős–rényi random graphs. *Statistics and Computing*, 32(1):19, 2022.

Abiodun M Ikotun, Absalom E Ezugwu, Laith Abualigah, Belal Abuhaija, and Jia Heming. K-means clustering algorithms: A comprehensive review, variants analysis, and advances in the era of big data. *Information Sciences*, 622:178–210, 2023.

Daeyoung Kim and Bruce G Lindsay. Empirical identifiability in finite mixture models. *Annals of the Institute of Statistical Mathematics*, 67:745–772, 2015.

Bar Light and Gabriel Y Weintraub. Mean field equilibrium: Uniqueness, existence, and comparative statics. *Operations Research*, 70(1):585–605, 2022.

Xingchao Liu, Xin Tong, and Qiang Liu. Sampling with trusthworthy constraints: A variational gradient framework. *Advances in Neural Information Processing Systems*, 34:23557–23568, 2021.

David G Luenberger. *Optimization by Vector Space Methods*. John Wiley & Sons, 1997.

James MacQueen. Some methods for classification and analysis of multivariate observations. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, volume 1, pages 281–297, 1967.

Justin Manweiler, Sharad Agarwal, Ming Zhang, Romit Roy Choudhury, and Paramvir Bahl. Switchboard: A matchmaking system for multiplayer mobile games. In *Proceedings of the 9th International Conference on Mobile Systems, Applications, and Services*, pages 71–84, 2011.

Geoffrey J McLachlan, Sharon X Lee, and Suren I Rathnayake. Finite mixture models. *Annual Review of Statistics and its Application*, 6:355–378, 2019.

Rad Niazadeh, Negin Golrezaei, Joshua Wang, Fransisca Susan, and Ashwinkumar Badanidiyuru. Online learning via offline greedy algorithms:: Applications in market design and optimization. *Management Science*, 69(7):3797–3817, 2023.

Adil Salim, Anna Korba, and Giulia Luise. The Wasserstein proximal gradient algorithm. *Advances in Neural Information Processing Systems*, 33:12356–12366, 2020.

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

Filippo Santambrogio. {Euclidean, metric, and Wasserstein} gradient flows: An overview. *Bulletin of Mathematical Sciences*, 7:87–154, 2017.

Andrey Simonov, Raluca M Ursu, and Carolina Zheng. Suspense and surprise in media product design: Evidence from Twitch. *Journal of Marketing Research*, 60(1):1–24, 2023.

Henry Teicher. Identifiability of finite mixtures. *The Annals of Mathematical Statistics*, pages 1265–1269, 1963.

Cédric Villani. *Optimal Transport: Old and New*. Springer, 2009.

Sidney J Yakowitz and John D Spragins. On the identifiability of finite mixtures. *The Annals of Mathematical Statistics*, 39(1):209–214, 1968.

Mohammad Zhalechian, Esmaeil Keyvanshokooh, Cong Shi, and Mark P Van Oyen. Online resource allocation with personalized learning. *Operations Research*, 70(4):2138–2161, 2022.

Ruiyi Zhang, Changyou Chen, Chunyuan Li, and Lawrence Carin. Policy optimization as Wasserstein gradient flows. In *International Conference on Machine Learning*, pages 5737–5746. PMLR, 2018.

Ruqi Zhang, Qiang Liu, and Xin Tong. Sampling in constrained domains with orthogonal-space variational gradient descent. *Advances in Neural Information Processing Systems*, 35:37108–37120, 2022.

Appendix: Omitted proofs

A.1 Proofs of Section 2

A.1.1 Proof of Proposition 1.

Proof. Consider a $\mathbf{x}_0$ and a unit norm vector $\mathbf{v} \in \mathbb{R}^d$ such that $\langle \mathbf{v}, \nabla^2_{1,2}\ell(\mathbf{x}_0, \mathbf{x}_0)\rangle < 0$. For the sake of contradiction, suppose for some $\delta, c > 0$ and $i, j \in [K]$, $\mathbf{x}_0 \in S_i(\delta, c) \cap S_j(\delta, c)$. Without loss of generality, we assume $x_0 = 0$, $i = 1$, and $j = 2$. Let B_0 be the ball of radius $r < \delta$ centered at x_0.

Define two half-balls

$$B_1 = \{\mathbf{x} \in B_0 : \langle \mathbf{v}, \mathbf{x}\rangle > 0\}, \quad B_2 = \{\mathbf{x} \in B_0 : \langle \mathbf{v}, \mathbf{x}\rangle < 0\}.$$

Define $q(\mathbf{x}) = 1_{B_1}(\mathbf{x}) - 1_{B_2}(\mathbf{x})$.

For $\epsilon > 0$ such that $c - \frac{\epsilon}{p_1} > 0$ and $c - \frac{\epsilon}{p_2} > 0$, construct the following densities

$$\mu_1^+(\mathbf{x}) = \mu_1(\mathbf{x}) - \frac{\epsilon}{p_1}q(\mathbf{x}), \quad \mu_2^+(\mathbf{x}) = \mu_2(x) + \frac{\epsilon}{p_2}q(\mathbf{x}), \quad \mu_k^+(\mathbf{x}) = \mu_k(\mathbf{x}), \quad k \geq 3.$$

$$\mu_1^-(\mathbf{x}) = \mu_1(\mathbf{x}) + \frac{\epsilon}{p_1}q(\mathbf{x}), \quad \mu_2^-(\mathbf{x}) = \mu_2(\mathbf{x}) - \frac{\epsilon}{p_2}q(\mathbf{x}), \quad \mu_k^-(\mathbf{x}) = \mu_k(\mathbf{x}), \quad k \geq 3.$$

Note that both $\{\mu_k^+\}_{k\in[K]}$ and $\{\mu_k^-\}_{k\in[K]}$ are feasible to Problem 1. To see this, by the choice of ϵ, $\mu_1^+(\mathbf{x}) > 0$ and $\mu_2^-(\mathbf{x}) > 0$ for any $\mathbf{x} \in \mathbb{R}^d$. Also, by the construction, $\{\mu_k^+\}_{k\in[K]}$ and $\{\mu_k^-\}_{k\in[K]}$ are two decomposition of π.

We have

$$\begin{aligned} p_1(L(\mu_1^+) - L(\mu_1)) &= -\epsilon \int\int_{B_0\times B_0} \ell(\mathbf{x},\mathbf{y})(q(\mathbf{x})\mu_1(\mathbf{y}) + q(\mathbf{y})\mu_1(\mathbf{x}))d\mathbf{x}d\mathbf{y} + \frac{\epsilon^2}{p_1}\int\int_{B_0\times B_0} \ell(\mathbf{x},\mathbf{y})q(\mathbf{x})q(\mathbf{y})d\mathbf{x}d\mathbf{y} \\ &= -2\epsilon \int\int_{B_0\times B_0} \ell(\mathbf{x},\mathbf{y})q(\mathbf{x})\mu_1(\mathbf{y})d\mathbf{x}d\mathbf{y} + \frac{\epsilon^2}{p_1}\int\int_{B_0\times B_0} \ell(\mathbf{x},\mathbf{y})q(\mathbf{x})q(\mathbf{y})d\mathbf{x}d\mathbf{y}, \end{aligned}$$

where the second equality is because $\ell : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ is symmetric. Likewise,

$$p_2(L(\mu_2^+) - L(\mu_2)) = 2\epsilon \int\int_{B_0\times B_0} \ell(\mathbf{x},\mathbf{y})q(\mathbf{x})\mu_2(\mathbf{y})d\mathbf{x}d\mathbf{y} + \frac{\epsilon^2}{p_2}\int\int_{B_0\times B_0} \ell(\mathbf{x},\mathbf{y})q(\mathbf{x})q(\mathbf{y})d\mathbf{x}d\mathbf{y}.$$

So we have

$$\begin{aligned} \sum_{k\in[K]} p_k(L(\mu_k^+) - L(\mu_k)) = 2\epsilon &\int\int_{B_0\times B_0} \ell(\mathbf{x},\mathbf{y})(\mu_2(\mathbf{y}) - \mu_1(\mathbf{y}))q(\mathbf{x})d\mathbf{x}d\mathbf{y} \\ &+ \epsilon^2(\frac{1}{p_1} + \frac{1}{p_2})\int\int_{B_0\times B_0} \ell(\mathbf{x},\mathbf{y})q(\mathbf{x})q(\mathbf{y})d\mathbf{x}d\mathbf{y}. \end{aligned}$$

Likewise, we have

$$\begin{aligned} \sum_{k\in[K]} p_k(L(\mu_k^-) - L(\mu_k)) = -2\epsilon &\int\int_{B_0\times B_0} \ell(\mathbf{x},\mathbf{y})(\mu_2(\mathbf{y}) - \mu_1(\mathbf{y}))q(\mathbf{x})d\mathbf{x}d\mathbf{y} \\ &+ \epsilon^2(\frac{1}{p_1} + \frac{1}{p_2})\int\int_{B_0\times B_0} \ell(\mathbf{x},\mathbf{y})q(\mathbf{x})q(\mathbf{y})d\mathbf{x}d\mathbf{y}. \end{aligned}$$

Suppose $\int\int_{B_0\times B_0} \ell(\mathbf{x},\mathbf{y})(\mu_2(\mathbf{y}) - \mu_1(\mathbf{y}))q(\mathbf{x})d\mathbf{x}d\mathbf{y} \neq 0$. Then, for ϵ small enough, either $\sum_{k\in[K]} p_k L(\mu_k^+)$ or $\sum_{k\in[K]} p_k L(\mu_k^-)$ is smaller than $\sum_{k\in[K]} p_k L(\mu_k)$, which contradicts the optimality of $\{\mu_k\}_{k\in[K]}$.

In the case that

$$\int\int_{B_0\times B_0} \ell(\mathbf{x},\mathbf{y})(\mu_2(\mathbf{y}) - \mu_1(\mathbf{y}))q(\mathbf{x})d\mathbf{x}d\mathbf{y} = 0,$$

we check the term

$$\int_{B_0\times B_0} \ell(\mathbf{x},\mathbf{y})q(\mathbf{x})q(\mathbf{y})d\mathbf{x}d\mathbf{y}. \tag{A.1}$$

We apply Talyor expansion as follows.

$$\ell(\mathbf{x},\mathbf{y}) = \ell(0,0) + (\mathbf{x},\mathbf{y})^\top \nabla\ell(0,0) + \frac{1}{2}(\mathbf{x},\mathbf{y})^\top \nabla^2\ell(0,0)(\mathbf{x},\mathbf{y}) + O(\|(\mathbf{x},\mathbf{y})\|^3).$$

Therefore, in the integral (A.1), the constant term becomes

$$\int\int_{B_0\times B_0} \ell(0,0)q(\mathbf{x})q(\mathbf{y})d\mathbf{x}d\mathbf{y} = \ell(0,0)\int_{B_0} q(\mathbf{x})d\mathbf{x}\int_{B_0} q(\mathbf{y})d\mathbf{y} = 0$$

where $\int_{B_0} q(\mathbf{x})d\mathbf{x} = 0$.

Note that $f(\mathbf{x},\mathbf{y}) \doteq (\mathbf{x},\mathbf{y})^\top \nabla\ell(0,0)q(\mathbf{x})q(\mathbf{y})$ is an odd function on $\mathbb{R}^d \times \mathbb{R}^d$, i.e., $f(\mathbf{x},\mathbf{y}) = -f(-\mathbf{x},-\mathbf{y})$. In the integral (A.1), the first order term is

$$\int\int_{B_0\times B_0} (\mathbf{x},\mathbf{y})^\top \nabla\ell(0,0)q(\mathbf{x})q(\mathbf{y})d\mathbf{x}d\mathbf{y} = 0.$$

Next, we deal with the second-order term. For each $\mathbf{x} \in B_0$, define $\mathbf{x}^- \doteq 2\langle \mathbf{x},\mathbf{v}\rangle\mathbf{v} - \mathbf{x}$, which is mirrow image of $\mathbf{x}$ with respect to the hyperplane $\{\mathbf{z} \in \mathbb{R}^d : \mathbf{z}^\top\mathbf{v} = 0\}$. Note that the distribution of $\mathbf{x}^-$ and $\mathbf{x}$ are identical under $q(\mathbf{x})$. Denote

$$\nabla^2\ell(0,0) = \begin{bmatrix} H_1 & H_2 \\ H_2^T & H_3 \end{bmatrix}.$$

In the integral (A.1), the second order term is

$$\begin{aligned}
&\int\int_{B_0\times B_0} (\mathbf{x},\mathbf{y})^\top \nabla^2\ell(0,0)(\mathbf{x},\mathbf{y})q(\mathbf{x})q(\mathbf{y})d\mathbf{x}d\mathbf{y} \\
&= \int\int_{B_0\times B_0} x^\top H_1 x q(\mathbf{x})q(\mathbf{y})d\mathbf{x}d\mathbf{y} + 2\int\int_{B_0\times B_0} \mathbf{x}^\top H_2\mathbf{y}q(\mathbf{x})q(\mathbf{y})d\mathbf{x}d\mathbf{y} + \int\int_{B_0\times B_0} \mathbf{y}^\top H_3\mathbf{y}q(\mathbf{y})q(\mathbf{x})d\mathbf{x}d\mathbf{y} \\
&= 2\int\int_{B_0\times B_0} \mathbf{x}^\top H_2\mathbf{y}q(\mathbf{x})q(\mathbf{y})d\mathbf{x}d\mathbf{y} \\
&= \int\int_{B_0\times B_0} (\mathbf{x}+\mathbf{x}^-)^\top H_2 y q(\mathbf{x})q(\mathbf{y})d\mathbf{x}d\mathbf{y} \\
&= \frac{1}{2}\int\int_{B_0\times B_0} (\mathbf{x}+\mathbf{x}^-)^\top H_2(y+y^-)q(\mathbf{x})q(\mathbf{y})d\mathbf{x}d\mathbf{y} \\
&= 2\int\int_{B_0\times B_0} \mathbf{v}^\top\mathbf{x}(\mathbf{v}^\top H_2\mathbf{v})\mathbf{v}^\top\mathbf{y}q(\mathbf{x})q(\mathbf{y})d\mathbf{x}d\mathbf{y} \\
&= 2(\mathbf{v}^\top H_2\mathbf{v})\int\int_{B_0\times B_0} (\mathbf{v}^\top\mathbf{x})(\mathbf{v}^\top\mathbf{y})(1_{\mathbf{v}^\top\mathbf{x}>0} - 1_{\mathbf{v}^\top\mathbf{x}<0})(1_{\mathbf{v}^\top\mathbf{y}>0} - 1_{\mathbf{v}^\top\mathbf{y}<0})dxdy \\
&= 2(\mathbf{v}^\top H_2\mathbf{v})(\int_{B_0} (\mathbf{v}^\top\mathbf{x})(1_{\mathbf{v}^\top\mathbf{x}>0} - 1_{\mathbf{v}^\top\mathbf{x}<0})d\mathbf{x})^2 = -O((rm(B_0))^2),
\end{aligned}$$

where the second equality is because

$$\int\int_{B_0\times B_0} \mathbf{x}^\top H_1\mathbf{x}q(\mathbf{x})q(\mathbf{y})d\mathbf{x}d\mathbf{y} = (\int_{B_0} \mathbf{x}^\top H_1\mathbf{x}q(\mathbf{x})d\mathbf{x})(\int_{B_0} q(\mathbf{y})d\mathbf{y}) = 0,$$

since $\int_{B_0} q(\mathbf{y})d\mathbf{y} = 0$; the third equality is because

$$\int\int_{B_0\times B_0} \mathbf{x}^\top H_2\mathbf{y}q(\mathbf{x})q(\mathbf{y})d\mathbf{x}d\mathbf{y} = \int\int_{B_0\times B_0} (\mathbf{x}^-)^\top H_2\mathbf{y}q(\mathbf{x})q(\mathbf{y})d\mathbf{x}d\mathbf{y};$$

and the last equality is because $v^\top H_2 v < 0$ by our condition.

Plug in this estimate and symmetry of B_1, B_2, we find

$$\int\int_{B_0\times B_0} \ell(\mathbf{x},\mathbf{y})q(\mathbf{x})q(\mathbf{y})d\mathbf{x}d\mathbf{y} = -|O(r^2)m(B_0)^2| + O(r^3)m(B_0)^2.$$

Thus, for r small enough,

$$\sum_{k\in[K]} p_k(L(\mu_k^-) - L(\mu_k)) = \int\int_{B_0\times B_0} \ell(\mathbf{x},\mathbf{y})q(\mathbf{x})q(\mathbf{y})d\mathbf{x}d\mathbf{y} < 0,$$

which again contradicts the optimality of $\{\mu_i\}_{i\in[K]}$. □

A.1.2 Proof of Corollary 1.

Proof. **Variance loss:** Let $\ell(\mathbf{x},\mathbf{y}) = \langle \mathbf{x}-\mathbf{y}, W(\mathbf{x}-\mathbf{y})\rangle$. The gradient and Hessian matrix are given by

$$\nabla\ell(\mathbf{x},\mathbf{y}) = 2W[\mathbf{x}-\mathbf{y}, \mathbf{y}-\mathbf{x}], \quad \nabla^2\ell(\mathbf{x},\mathbf{y}) = 2\begin{bmatrix} W & -W \\ -W & W \end{bmatrix}.$$

Thus, for any $\mathbf{x}_0$ and unit norm vector $\mathbf{v}$, $\langle \mathbf{v}, \nabla^2_{1,2}\ell(\mathbf{x}_0,\mathbf{x}_0)\mathbf{v}\rangle = -2W$. By Proposition 1, as long as not all entries in W are non-positive (equivalently, W is not negative semi-definite), the optimal densities have no overlap for their (δ, c)-support. In particular, if $W = \mathrm{id}$, i.e., $\ell(\mathbf{x},\mathbf{y}) = \|\mathbf{x}-\mathbf{y}\|^2$, then the optimal densities have no overlap for their (δ, c)-interior densities.

Elo loss: Let $\ell(x,y) = \frac{x^2+y^2}{(x+y)^2}$. We note that

$$\partial_1\ell = \frac{2xy-2y^2}{(x+y)^3}, \quad \partial_2\ell = \frac{2xy-2x^2}{(x+y)^3},$$

$$\nabla^2\ell = \frac{1}{(x+y)^4}\begin{bmatrix} 8y^2-4xy & 2(x^2+y^2-4xy) \\ 2(x^2+y^2-4xy) & 8x^2-4xy \end{bmatrix}.$$

In this case, $d = 1$ and the only unit norm vector is $v = 1$, thus

$$\nabla^2\ell(x,x) = \frac{1}{4x^2}\begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix}.$$

Hence, $\langle v, \nabla^2_{1,2}\ell(x,x)v\rangle < 0$. By Proposition 1, the optimal densities have no overlap for their (δ, c)- support. □

A.1.3 Proof of claim in Example 3.

Proof. Suppose we decompose π into two disbritutions μ_1 and μ_2, i.e., $K = 2$ in Problem 2. Consider the following decomposition:

$$\mu_1 = \frac{1}{2}\delta_{a-1} + \frac{1}{2}\delta_a, \quad \mu_2 = \frac{1}{2}\delta_{a+1} + \frac{1}{2}\delta_a.$$

Omitting the constant $-\frac{1}{4}$, the Elo loss of this decomposition is given by

$$\begin{aligned}
&\frac{1}{2}(\frac{1}{2}\cdot P(\text{equal match}) + \frac{2a^2-2a+1}{(2a-1)^2}\cdot P(\text{unequal match})) + \frac{1}{2}(\frac{1}{2}\cdot P(\text{equal match}) \\
&+ \frac{2a^2+2a+1}{(2a+1)^2}\cdot P(\text{unequal match})) \\
&= \frac{1}{4}(\frac{1}{2}+\frac{1}{2}+\frac{1}{2(2a-1)^2}) + \frac{1}{4}(\frac{1}{2}+\frac{1}{2}+\frac{1}{2(2a+1)^2}) \\
&= \frac{1}{2}(1+\frac{1}{4(2a-1)^2}+\frac{1}{4(2a+1)^2}),
\end{aligned}$$

where the second equality is due to $P(\text{equal match}) = P(\text{unequal match}) = \frac{1}{2}$ and $\frac{2a^2-2a+1}{(2a-1)^2} = \frac{1}{2} + \frac{1}{2(2a-1)^2}$.

Meanwhile, consider the following bulk/tail decomposition:

$$\mu_1 = \frac{1}{2}\delta_{a-1} + \frac{1}{2}\delta_{a+1}, \quad \mu_2 = \delta_a.$$

Omitting the constant $-\frac{1}{4}$, its Elo loss is

$$\begin{aligned}
&\frac{1}{2}(\frac{1}{2}\cdot P(\text{equal match}) + \frac{2a^2+2}{(2a)^2}\cdot P(\text{unequal match})) + \frac{1}{4} \\
=&\frac{1}{4}(\frac{1}{2} + \frac{1}{2} + \frac{2}{4a^2}) + \frac{1}{4} \\
=&\frac{1}{2}(1 + \frac{1}{4a^2}).
\end{aligned}$$

Then, when $a > 1$ is close to 1,

$$\frac{1}{4(2a-1)^2} + \frac{1}{4(2a+1)^2} \geq \frac{1}{4a^2}.$$

So, the second plan, which is nonconvex, yields a lower Elo loss. □

A.1.4 Proof of Lemma 1.

Proof. Let $(\mu_1, \ldots, \mu_K)$ be the optimal solution to Problem 2 with variance loss. For the sake of contradiction, suppose, for some $\delta, c > 0$, there exists $\mathbf{x}_0, \mathbf{x}_1 \in S_1(\delta, c)$, $\lambda \in (0,1)$ such that $\mathbf{z} \doteq \lambda\mathbf{x}_0 + (1-\lambda)\mathbf{x}_1 \in S_2(\delta, c)$. Without loss of generality, assume $\mathbf{x}_0 = 0$. Let $B_0, B_1, B_{\mathbf{z}}$ be open ball of radius δ centered at $\mathbf{x}_0, \mathbf{x}_1$, and $\mathbf{z}$, respectively. By the definition of $S_1(\delta, c)$ and $S_2(\delta, c)$, we know that $\mu_1(\mathbf{x}) > c\ \forall x \in B_0 \cup B_1$ and $\mu_2(\mathbf{x}) > c$ for all $\mathbf{x} \in B_{\mathbf{z}}$.

Consider the following alternative solution $(\mu_1', \ldots, \mu_K')$ obtained by

$$\mu_1'(\mathbf{x}) = \mu_1(\mathbf{x}) - \frac{c}{p_1}(1_{B_0}(\mathbf{x}) - 1_{B_z}(\mathbf{x})), \quad \mu_2'(\mathbf{x}) = \mu_2(\mathbf{x}) - \frac{c}{p_2}(1_{B_{\mathbf{z}}}(\mathbf{x}) - 1_{B_0}(\mathbf{x})),$$

and $\mu_k' = \mu_k$ for all $k > 2$.

Let $\mathbf{m}_i$ (resp. $\mathbf{m}_i'$) be the mean of probability measure μ_i (resp. μ_i') consisting of means in d-dimensions of the underlying features. Since $\mathbf{m}_1' \neq \mathbf{m}_1$ is the mean of μ_1', we have

$$L(\mu_1') = \int_{\mathbb{R}^d} (\mathbf{x} - m_1')^\top W(\mathbf{x} - \mathbf{m}_1') d\mu_1'(\mathbf{x}) < \int_{\mathbb{R}^d} (\mathbf{x} - \mathbf{m}_1)^\top W(\mathbf{x} - \mathbf{m}_1) d\mu_1'(\mathbf{x}).$$

Therefore,

$$\begin{aligned}
p_1(L(\mu_1) - L(\mu_1')) &> c\int_{B_0} (\mathbf{x} - \mathbf{m}_1)^\top W(\mathbf{x} - \mathbf{m}_1)dx - c\int_{B_z} (\mathbf{x} - \mathbf{m}_1)^\top W(\mathbf{x} - \mathbf{m}_1)dx \\
&= c\int_{B_0} ((\mathbf{x} - \mathbf{m}_1)^\top W(\mathbf{x} - \mathbf{m}_1) - (\mathbf{x} + \mathbf{z} - \mathbf{m}_1)^\top W(\mathbf{x} + \mathbf{z} - \mathbf{m}_1))dx \\
&= c\int_{B_0} (-2(\mathbf{x} - \mathbf{m}_1)^\top W\mathbf{z} - \mathbf{z}^\top W\mathbf{z})d\mathbf{x} \\
&= cm(B_0)(2\mathbf{m}_1 - \mathbf{z})^\top W\mathbf{z}
\end{aligned}$$

Similarly,

$$p_2(L(\mu_2) - L(\mu_2')) > -c\int_{B_0} (\mathbf{x} - \mathbf{m}_2)^\top W(\mathbf{x} - m_2)d\mathbf{x} + c\int_{B_{\mathbf{z}}} (\mathbf{x} - \mathbf{m}_2)^\top W(\mathbf{x} - \mathbf{m}_2)dx = -cm(B_0)(2\mathbf{m}_2 - \mathbf{z})^\top W\mathbf{z}.$$

The change of objective value is given by

$$\Delta_1 \doteq \sum_{k=1}^{K} p_k(L(\mu_k) - L(\mu_k')) = p_1(L(\mu_1) - L(\mu_1')) + p_2(L(\mu_2) - L(\mu_2')) > 2cm(B_0)(\mathbf{m}_1 - \mathbf{m}_2)^\top W\mathbf{z}.$$

Next, we consider another swapping

$$\mu_1''(\mathbf{x}) = \mu_1(\mathbf{x}) - \frac{c}{p_1}(1_{B_1}(\mathbf{x}) - 1_{B_z}(\mathbf{x})), \quad \mu_2''(\mathbf{x}) = \mu_2(\mathbf{x}) - \frac{c}{p_2}(1_{B_z}(\mathbf{x}) - 1_{B_1}(\mathbf{x})),$$

and $\mu_k'' = \mu_k$ for all $k > 2$. We note that

$$\begin{aligned} p_1(L(\mu_1) - L(\mu_1'')) &> c\int_{B_1}(\mathbf{x} - m_1)^\top W(\mathbf{x} - \mathbf{m}_1)d\mathbf{x} - c\int_{B_z}(\mathbf{x} - \mathbf{m}_1)^\top W(x - \mathbf{m}_1)d\mathbf{x} \\ &= cm(B_0)(2\mathbf{m}_1 - \mathbf{z} + \mathbf{x}_1)^\top W(\mathbf{z} - \mathbf{x}_1), \end{aligned}$$

and

$$\begin{aligned} p_2(L(\mu_2) - L(\mu_2'')) &> -c\int_{B_1}(\mathbf{x} - \mathbf{m}_2)^\top W(\mathbf{x} - \mathbf{m}_2)d\mathbf{x} + c\int_{B_z}(\mathbf{x} - \mathbf{m}_2)^\top W(\mathbf{x} - \mathbf{m}_2)d\mathbf{x} \\ &= cm(B_0)(-2\mathbf{m}_2 + \mathbf{z} - \mathbf{x}_1)^\top W(\mathbf{z} - \mathbf{x}_1). \end{aligned}$$

The change of objective value is given by

$$\Delta_2 \doteq \sum_{k=1}^{K} p_k(L(\mu_k) - L(\mu_k'')) = p_1(L(\mu_1) - L(\mu_1'')) + p_2(L(\mu_2) - L(\mu_2'')) > 2cm(B_0)(\mathbf{z} - \mathbf{x}_1)^\top W(\mathbf{m}_1 - \mathbf{m}_2).$$

Note that $\lambda\Delta_1 + (1-\lambda)\Delta_2 > 2cm(B_0)(\mathbf{m}_1 - \mathbf{m}_2)^\top W(\mathbf{z} - (1-\lambda)\mathbf{x}_1) = 0$ since $\mathbf{z} - (1-\lambda)\mathbf{x}_1 = \lambda\mathbf{x}_0 = 0$, which indicates that either $\Delta_1 > 0$ or $\Delta_2 > 0$. In either case $(\mu_k)_{k\in[K]}$ is not optimal. □

A.2 Proofs for Section 3

A.2.1 Proof of Lemma 4.

Proof. Consider a curve $\mu(\cdot) : [0,1] \to \mathcal{P}_2(\mathbb{R}^d)$ with $\mu(0) = \mu$ and velocity $\phi : \mathbb{R}^d \to \mathbb{R}^d$ at time $t = 0$; that is, $\frac{d}{dt}\mu(t,\mathbf{x})|_{t=0} = -\nabla \cdot (\phi(\mathbf{x})\mu(0,\mathbf{x}))$. Then

$$\begin{aligned} \frac{d}{dt}L(\mu(t))|_{t=0} &= \tfrac{d}{dt}\int_{\mathbb{R}^d}\int_{\mathbb{R}^d}\ell(\mathbf{x},\mathbf{y})d\mu(t,\mathbf{x})d\mu(t,\mathbf{y})|_{t=0} \\ &= \int_{\mathbb{R}^d}\frac{d}{dt}\int_{\mathbb{R}^d}\ell(\mathbf{x},\mathbf{y})d\mu(t,\mathbf{x}))d\mu(t,y) + \int_{\mathbb{R}^d}\frac{d}{dt}\int_{\mathbb{R}^d}\ell(\mathbf{x},\mathbf{y})d\mu(t,\mathbf{y}))d\mu(t,\mathbf{x})|_{t=0} \\ &= -\int_{\mathbb{R}^d}\int_{\mathbb{R}^d}\ell(\mathbf{x},\mathbf{y})\nabla\cdot(\phi(\mathbf{x})\mu(\mathbf{x}))d(\mathbf{x})d\mu(\mathbf{y}) - \int_{\mathbb{R}^d}\int_{\mathbb{R}^d}\ell(\mathbf{x},\mathbf{y})\nabla\cdot(\phi(\mathbf{y})\mu(\mathbf{y}))d\mathbf{y}d\mu(\mathbf{x}) \\ &= \int_{\mathbb{R}^d}\int_{\mathbb{R}^d}\phi(\mathbf{x})^\top\nabla_1\ell(\mathbf{x},\mathbf{y})d\mu(\mathbf{x})d\mu(\mathbf{y}) + \int_{\mathbb{R}^d}\int_{\mathbb{R}^d}\phi(\mathbf{y})^\top\nabla_2\ell(\mathbf{x},\mathbf{y})d\mu(\mathbf{x})d\mu(\mathbf{y}) \\ &= \int_{\mathbb{R}^d}d\mathbf{x}\phi(\mathbf{x})^\top\left(\int_{\mathbb{R}^d}(\nabla_1\ell(\mathbf{x},\mathbf{z}) + \nabla_2\ell(\mathbf{z},\mathbf{x}))d\mu(\mathbf{z})\right). \end{aligned}$$

□

A.2.2 Proof of Lemma 5.

Proof. Let $\bar{\mu} = \sum_{k=1}^{K} p_k\mu_k$. For each $k \in [K]$, consider a curve $\mu_k(\cdot) : [0,1] \to \mathcal{P}_2(\mathbb{R}^d)$ with $\mu_k(0) = \mu_k$ and velocity $\phi_k : \mathbb{R}^d \to \mathbb{R}^d$ at time $t = 0$; that is, $\frac{d}{dt}\mu(t,\mathbf{x})_k|_{t=0} = -\nabla \cdot$

$(\phi_k(\mathbf{x})\mu_k(0,\mathbf{x}))$. We have

$$\begin{aligned}
\tfrac{d}{dt}\mathrm{KL}(\bar{\mu}(t)\|\pi)|_{t=0} &= \tfrac{d}{dt}\int_{\mathbb{R}^d} \bar{\mu}(t)\log\tfrac{\bar{\mu}(t)}{\pi}d\mathbf{x}|_{t=0} \\
&= \int (\tfrac{d}{dt}\bar{\mu}(t))\log\tfrac{\bar{\mu}(t)}{\pi}d\mathbf{x}|_{t=0} \\
&= \sum_{k\in[K]} p_k \int_{\mathbb{R}^d} (\tfrac{d}{dt}\mu_k(t))\log\tfrac{\bar{\mu}(t)}{\pi}d\mathbf{x}|_{t=0} \\
&= -\sum_{k\in[K]} p_k \int_{\mathbb{R}^d} \nabla\cdot(\phi_k\mu_k)\log\tfrac{\bar{\mu}}{\pi}d\mathbf{x} \\
&= \sum_{k\in[K]} p_k \int_{\mathbb{R}^d} \langle\phi_k(\mathbf{x}), s_{\bar{\mu}}(x) - s_\pi(\mathbf{x})\rangle\mu_k(\mathbf{x})d\mathbf{x} \\
&= \sum_{k\in[K]} \int_{\mathbb{R}^d} \langle\phi_k(\mathbf{x}), p_k(s_{\bar{\mu}}(x) - s_\pi(x))\rangle\mu_k(\mathbf{x})d\mathbf{x} \\
&= \sum_{k\in[K]} \langle\phi_k(\mathbf{x}), p_k(s_{\bar{\mu}}(\mathbf{x}) - s_\pi(\mathbf{x}))\rangle_{\mu_k},
\end{aligned}$$

which satisfies the definition of the Wasserstein gradient in equation (14). □

A.3 Proofs for Section 4

A.3.1 Proof of Proposition 2.

We start with several useful lemmas.

Lemma A.1. For any $\boldsymbol{\mu}\in\mathcal{P}_{\pi,\mathbf{p}}$,

$$\mathcal{T}\mathcal{P}_{\pi,p}(\boldsymbol{\mu}) \subseteq \mathcal{T}'\mathcal{P}_{\pi,p}(\boldsymbol{\mu}) \doteq \left\{\boldsymbol{\phi} : \sum_{k\in[K]} p_k\nabla\cdot(\mu_k\phi_k) = 0\right\}.$$

Proof. For any $\boldsymbol{\phi}\in\mathcal{T}\mathcal{P}_{\pi,p}(\boldsymbol{\mu})$, there exists a curve $\boldsymbol{\mu}(t)$ in $\mathcal{P}_{\pi,p}$ such that $\frac{d}{dt}\mu_k(t)|_{t=0} = -\nabla\cdot(\mu_k(0)\phi_k)$ and $\mu(0)=\boldsymbol{\mu}$. Since $\boldsymbol{\mu}(t)\in\mathcal{P}_{\pi,p}$, $\sum_{k\in[K]} p_k\mu_k(t) = \pi\ \forall t\in[0,1]$. This equality holds in the sense that for all $t\in[0,1]$,

$$\sum_{k\in[K]} p_k\int_{\mathbb{R}^d} f(\mathbf{x})d\mu_k(t,\mathbf{x}) = \int_{\mathbb{R}^d} f(\mathbf{x})d\pi(\mathbf{x}), \quad \forall f\in C_c^\infty(\mathbb{R}^d).$$

Note that the right-hand side does not depend on t. Hence, $\sum_{k\in[K]} p_k\frac{d}{dt}\int_{\mathbb{R}^d} f(\mathbf{x})d\mu_k(t,\mathbf{x}) = 0$, $\forall f\in C_c^\infty(\mathbb{R}^d)$ for all $t\in[0,1]$, which means $\frac{d}{dt}\sum_{k\in[K]} p_k\mu_k(t) = 0$ for all $t\in[0,1]$. Therefore, we have

$$\sum_{k\in[K]} p_k\nabla\cdot(\mu_k(0)\phi_k) = -\frac{d}{dt}\sum_{k\in[K]} p_k\mu_k(t)|_{t=0} = 0,$$

which means $\boldsymbol{\phi}\in\mathcal{T}'\mathcal{P}_{\pi,p}(\boldsymbol{\mu})$, thus completing the proof. □

Lemma A.2. Given a $\phi = \nabla\Psi$ with $\Psi\in C_c^\infty(\mathbb{R}^d)$, denote $T_t(\mathbf{x}) = \mathbf{x} + \phi(\mathbf{x})t$. Then for $t < t_0 \doteq 1/\|\nabla\phi\|_\infty$, T_t is a diffeomorphism of $\mathbb{R}^d$. Moreover, $\|T_t^{-1}(\mathbf{x}) - \mathbf{x}\| \le \|\phi\|_\infty t$ (for all t, not just $t < t_0$).

Proof. First, we show T_t is one-to-one. If $T_t(\mathbf{x}) = T_t(\mathbf{y})$ for some given $\mathbf{x}$ and $\mathbf{y}$, then we have

$$\|\mathbf{x}-\mathbf{y}\| = t\|\phi(\mathbf{x}) - \phi(\mathbf{y})\| \le t\|\nabla\phi\|_\infty\|\mathbf{x}-\mathbf{y}\|,$$

by the mean value theorem and the Cauchy-Schwartz inequality. Thus, if $t < 1/\|\nabla\phi\|_\infty$, we have $\mathbf{x} = \mathbf{y}$.

Next, we show that T_t is onto. It suffices to show for any $\mathbf{x}$, there exists a $\mathbf{z}$ such that $\mathbf{x} = \mathbf{z} + \phi(\mathbf{z})t$. Since $\Phi = \nabla\Psi$ and ψ is compactly supported, $T_t = \mathrm{id}$ outside the compact support. Within the compact support, define $f_{\mathbf{x},\mathbf{t}}(\mathbf{z}) = \mathbf{x} - \phi(\mathbf{z})t$. We show that $f_{\mathbf{x},t}$ is a contraction mapping. To see this, for any $\mathbf{z}_1 \neq \mathbf{z}_2$ in the compact support, $\|f_{\mathbf{x},t}(\mathbf{z}_1) - f_{\mathbf{x},t}(\mathbf{z}_2)\| = \|\phi(\mathbf{z}_1) - \phi(\mathbf{z}_2)\|t$. By the mean value theorem, we have $\|\phi(\mathbf{z}_1) - \phi(\mathbf{z}_2)\| \leq \|\nabla\phi\|_\infty \|\mathbf{z}_1 - \mathbf{z}_2\|$. This implies $\|f_{\mathbf{x},t}(\mathbf{z}_1) - f_{\mathbf{x},t}(\mathbf{z}_2)\| \leq \|\nabla\phi\|_\infty t\|\mathbf{z}_1 - \mathbf{z}_2\| < \|\mathbf{z}_1 - \mathbf{z}_2\|$ for $t < t_0$. Hence, $f_{\mathbf{x},t}$ is a contraction mapping. By the Banach fixed point theorem, $f_{\mathbf{x},t}$ admits a fixed point, i.e., there exists a $\mathbf{z}$ such that $\mathbf{x} - \phi(\mathbf{z})t = \mathbf{z}$. Hence, T_t is also onto.

Then, we show that T_t is differentiable with a differentiable inverse. The differential of T_t is given by

$$\nabla T_t(\mathbf{x}) = \mathrm{id} + t\nabla\phi(\mathbf{x}),$$

where id is the identity matrix. Then, using the inverse map rule, the differential of the inverse T_t^{-1} is given by

$$\nabla T_t^{-1}(\mathbf{x}) = (\mathrm{id} + t\nabla\phi(T_t^{-1}(\mathbf{x})))^{-1}.$$

Note that id and $\nabla\phi(T_t^{-1}(x))$ are both invertible. Recall that the set of invertible matrices is a convex cone. Hence, since $t > 0$, $\mathrm{id} + t\nabla\phi(T_t^{-1}(\mathbf{x}))$ is also invertible. This means that $\nabla T_t^{-1}(\mathbf{x})$ exists.

Thus, T_t is a diffeomorphism. Also, note that if we let $\mathbf{z} = T_t^{-1}(\mathbf{x})$, i.e., $\mathbf{x} = \mathbf{z} + \phi(\mathbf{z})t$, then $\|\mathbf{x} - \mathbf{z}\| \leq t\|\phi\|_\infty$. □

Proof of Proposition 2. Recall that, by definition, $\nabla F_{\mathbf{p}}$ is the unique element in $\mathcal{TP}_2^{\otimes K}(\boldsymbol{\mu}^*) = \bigotimes_{k\in[K]} \overline{\{\nabla\Psi : \Psi \in C_c^\infty(\mathbb{R}^d)\}}^{\mathcal{L}^2_{\mu_k}}$ such that for all $\boldsymbol{\mu}(\cdot) : [0,\infty) \to \mathcal{P}_2^{\otimes K}(\mathbb{R}^d)$ with $\boldsymbol{\mu}(0) = \boldsymbol{\mu}^*$ and velocity $\boldsymbol{\phi} \in \mathcal{TP}_2^{\otimes K}(\boldsymbol{\mu}^*)$,

$$\begin{aligned}\frac{d}{dt}F_{\mathbf{p}}(\boldsymbol{\mu}(t))|_{t=0} &= \langle \nabla F_{\mathbf{p}}, \phi\rangle_{\boldsymbol{\mu}^*} \\ &= \sum_{k\in[K]} \langle \nabla_{\mu_k} F_{\mathbf{p}}, \phi_k\rangle_{\mu_k^*} \\ &= \sum_{k\in[K]} \int \phi_k(x)^\top \nabla_{\mu_k} F_{\mathbf{p}}(x) d\mu_k^*(x).\end{aligned}$$

Since $F_{\mathbf{p}}(\boldsymbol{\mu}) = \sum_{k\in[K]} p_k L(\mu_k)$, $\nabla_{\mu_k} F_{\mathbf{p}}(\boldsymbol{\mu}) = p_k \nabla L(\mu_k)$.

For the sake of contradiction, suppose $\|\nabla F_{\mathbf{p}}(\boldsymbol{\mu}^*)\|_{\mathcal{TP}_{\pi,\mathbf{p}}(\boldsymbol{\mu}^*)} > 0$. Then, there exists a $\boldsymbol{\phi} \in \mathcal{TP}_2^{\otimes K}(\boldsymbol{\mu}^*)$ such that

(i) $\boldsymbol{\phi} = \nabla\Phi$ for some $\Phi \in C_c^\infty(\mathbb{R}^d)$;
(ii) $A \doteq \langle \nabla F_{\mathbf{p}}, \boldsymbol{\phi}\rangle_{\boldsymbol{\mu}^*} = \sum_{k\in[K]} p_k \langle \phi_k, \nabla L\rangle_{\mu_k^*} < 0$;
(iii) $\sum_{k\in[K]} p_k \nabla \cdot (\mu_k^* \phi_k) = 0$,

where the equality in claim (iii) holds in the sense that, for any $f \in C_c^\infty(\mathbb{R}^d)$,

$$\sum_{k\in[K]} p_k \int_{\mathbb{R}^d} \langle \nabla f(x), \phi_k(x)\rangle d\mu_k^*(x) = 0.$$

Claim (i) and (ii) hold because $\nabla F_{\mathbf{p}}$ is in the closure of $\{\nabla\Psi : \Psi \in C_c^\infty(\mathbb{R}^d)\}$. Claim (iii) holds by Lemma A.1 since $\boldsymbol{\phi} \in \mathcal{TP}_2^{\otimes K}(\boldsymbol{\mu}^*)$.

When such a $\boldsymbol{\phi}$ exists, we show that we can construct a solution to Problem 2 which is

strictly better than $\boldsymbol{\mu}^*$. Consider the following time-independent velocity field

$$\phi_k(t) = \phi_k, \quad \forall t \in [0,1].$$

Then, the map $T^{\phi}_{k,t}(\mathbf{x}) = \mathbf{x} + \phi_k(\mathbf{x})t$ defines a curve $\mu_k(t) \doteq T^{\phi}_{k,t}\sharp\mu^*_k$ for $t \in [0,1]$ in $\mathcal{P}_2(\mathbb{R}^d)$. Define $\bar{\mu}(t) = \sum_{k\in[K]} p_k\mu_k(t)$ in the sense that, for all $t \in [0,1]$, $\bar{\mu}(t)$ is a probability measure in $\mathcal{P}_2(\mathbb{R}^d)$ such that for all $f \in C^\infty_c(\mathbb{R}^d)$, $\int_{\mathbb{R}^d} f d\bar{\mu}(t) = \sum_{k\in[K]} p_k \int_{\mathbb{R}^d} f d\mu_k(t)$. In particular, $\bar{\mu}(0) = \pi$ since $\sum_{k\in[K]} p_k\mu_k(0) = \pi$.

Note that the curve $\boldsymbol{\mu} \doteq (\mu_1(t), \ldots, \mu_K(t))$ may not be in the feasible set $\mathcal{P}_{\pi,\mathbf{p}}$, i.e., $\bar{\mu}(t)$ may not equal to π. Hence, to get the "better solution" we are looking for, we need to "project" this $\boldsymbol{\mu}(t)$ back to the feasible set $\mathcal{P}_{\pi,\mathbf{p}}$. However, this "projection" is subtle since the underlying space is not a Hilbert space.

Below, we show the following facts: there exists a $C > 0$ such that for all $t \in [0,1]$,

(i) $F_{\mathbf{p}}(\boldsymbol{\mu}(t)) \le F_{\mathbf{p}}(\boldsymbol{\mu}^*) + At + \frac{1}{2}Ct^2$;

(ii) there exists an optimal transport map S_t from $\bar{\mu}(t)$ to π where

$$F_{\mathbf{p}}(S_t\sharp\boldsymbol{\mu}(t)) - F_{\mathbf{p}}(\boldsymbol{\mu}(t)) \le W_1(\bar{\mu}(t), \pi) \le \frac{1}{2}Ct^2.$$

Given these two facts, note that $S_t\sharp\boldsymbol{\mu}(t) = (S_t\sharp\mu_1(t), \ldots, S_t\sharp\mu_K(t))$ is feasible, since

$$\sum_{k\in[K]} p_k S_t\sharp\mu_k(t) = S_t\sharp(\sum_{k\in[K]} p_k\mu_k(t)) = S_t\sharp\bar{\mu}(t) = \pi.$$

Combining these two inequalities, we have

$$F_{\mathbf{p}}(S\sharp\boldsymbol{\mu}(t)) \le F_{\mathbf{p}}(\boldsymbol{\mu}^*) + At + Ct^2.$$

Since $A < 0$ and $C > 0$, by setting t small enough, we show that $F_{\mathbf{p}}(S\sharp\boldsymbol{\mu}(t)) < F_{\mathbf{p}}(\boldsymbol{\mu}^*)$, which contradicts the optimality of $\boldsymbol{\mu}^*$.

Thus, it suffices to prove Claim (i) and (ii).

Claim (i): $F_{\mathbf{p}}(\boldsymbol{\mu}(t)) \le F_{\mathbf{p}}(\boldsymbol{\mu}^*) + At + \frac{1}{2}Ct^2$.

Observe that

$$\begin{aligned}
L(\mu_k(t)) - L(\mu_k(0)) &= \int_0^t \int_{\mathbb{R}^d} \langle \nabla L(\mu_k(s)), \phi_k\rangle d\mu_k(s,\mathbf{x})ds \\
&= \int_0^t \int_{\mathbb{R}^d}\int_{\mathbb{R}^d} \langle \nabla_1\ell(\mathbf{x},\mathbf{z}) + \nabla_2\ell(\mathbf{z},\mathbf{x}), \phi_k(\mathbf{x})\rangle d\mu_k(s,\mathbf{x})d\mu_k(s,\mathbf{z})ds \\
&= \int_0^t \int_{\mathbb{R}^d}\int_{\mathbb{R}^d} \langle \nabla_x\ell(T^{-1}_{k,s}x, T^{-1}_{k,s}\mathbf{z}) + \nabla_2\ell(T^{-1}_{k,s}\mathbf{z}, T^{-1}_{k,s}\mathbf{x}), \phi_k(T^{-1}_{k,s}\mathbf{x})\rangle d\mu^*_k(\mathbf{x})d\mu^*_k(\mathbf{z})ds,
\end{aligned}$$

where the first equality is due to the fundamental theorem of calculus, the second equality is due to Lemma 4, and the third equality is due to the definition of the pushforward measure.

Denote

$$Q_{k,s} \doteq \int_{\mathbb{R}^d}\int_{\mathbb{R}^d} \langle \nabla_1\ell(T^{-1}_{k,s}\mathbf{x}, T^{-1}_{k,s}\mathbf{z}) + \nabla_2\ell(T^{-1}_{k,s}\mathbf{z}, T^{-1}_{k,s}\mathbf{x}), \phi_k(T^{-1}_{k,s}\mathbf{x})\rangle d\mu^*_k(\mathbf{x})d\mu^*_k(\mathbf{z})$$

In particular, when $s = 0$,

$$Q_{k,0} = \int_{\mathbb{R}^d} \langle \nabla_1\ell(\mathbf{x},\mathbf{z}) + \nabla_2\ell(\mathbf{z},\mathbf{x}), \phi_k(\mathbf{x})\rangle d\mu^*_k(\mathbf{x})d\mu^*_k(\mathbf{z}).$$

Then note that $\|T^{-1}_{k,s}\mathbf{x} - \mathbf{x}\| \le Cs$, we have that $\|\nabla_2\ell(T^{-1}_{k,s}\mathbf{x}, T^{-1}_{k,s}\mathbf{z}) - \nabla_1\ell(T^{-1}_{k,s}\mathbf{x}, T^{-1}_{k,s}\mathbf{z})\| \le Cs(\|\mathbf{x}\| + \|\mathbf{z}\| + 1), \phi_k(T^{-1}_{k,s}\mathbf{x}) - \phi_k(\mathbf{x}) \le Cs$. Therefore

$$Q_{k,s} - Q_{k,0} \le Cs.$$

In summary, we have

$$L(\mu_k(t)) - L(\mu_k^*) - tQ_{k,0} \le \frac{1}{2}Ct^2.$$

Hence,

$$\begin{aligned} F_{\mathbf{p}}(\boldsymbol{\mu}(t)) &= \sum_{k\in[K]} p_k L(\mu_k(t)) \\ &\le \sum_{k\in[K]} p_k L(\mu_k^*) + t\sum_{k\in[K]} p_k Q_{k,0} + \frac{1}{2}Ct^2 \\ &= F_{\mathbf{p}}(\boldsymbol{\mu}^*) + At + \frac{1}{2}Ct^2. \end{aligned}$$

Claim (ii): There exists an optimal transport map S_t from $\bar{\mu}(t)$ to π and

$$F_{\mathbf{p}}(S_t\sharp\boldsymbol{\mu}(t)) - F_{\mathbf{p}}(\boldsymbol{\mu}(t)) \le W_1(\bar{\mu}(t),\pi) \le \frac{1}{2}Ct^2.$$

By the duality formula of Kantorovich-Rubinstein distance (Villani 2009), there is a Lipschitz continuous function f with Lipschitz constant no larger than 1 such that

$$\begin{aligned} W_1(\bar{\mu}(t),\pi) &= \sup_{f'\in\text{Lip-1}} \mathbb{E}_{\bar{\mu}(t)}[f'] - \mathbb{E}_{\pi}[f'] = \mathbb{E}_{\bar{\mu}(t)}[f] - \mathbb{E}_{\pi}[f] = \mathbb{E}_{\bar{\mu}(t)}[f] - \mathbb{E}_{\bar{\mu}(0)}[f] \\ &= \int_0^t \sum_{k\in[K]} p_k \int_{\mathbb{R}^d} \langle \nabla f(\mathbf{x}), \phi_k(\mathbf{x})\rangle d\mu_k(s,\mathbf{x})ds, \end{aligned}$$

where the last equality is by the definition of the distributional solution of the continuity equation with boundary conditions.

Define $R_{k,s} \doteq \int \langle \nabla T_{k,s}^{-1}(\mathbf{x}) \nabla f(T_{k,s}^{-1}\mathbf{x}), \phi_k(\mathbf{x})\rangle d\mu_k^*(\mathbf{x})$. There exists a constant $C > 0$ such that for any $s \in [0,t]$,

$$\begin{aligned} \int_{\mathbb{R}^d} \langle \nabla f(\mathbf{x}), \phi_k(\mathbf{x})\rangle d\mu_k(s,\mathbf{x}) &= \int_{\mathbb{R}^d} \langle \nabla f(T_{k,s}^{-1}\mathbf{x}), \phi_k(T_{k,s}^{-1}\mathbf{x})\rangle d\mu_k^*(\mathbf{x}) \\ &= \int_{\mathbb{R}^d} \langle \nabla f(T_{k,s}^{-1}\mathbf{x}), \phi_k(T_{k,s}^{-1}x) - \nabla T_{k,s}^{-1}(\mathbf{x})\phi_k(\mathbf{x})\rangle d\mu_k^*(\mathbf{x}) + R_{k,s} \\ &\le \int_{\mathbb{R}^d} \|\phi_k(T_{k,s}^{-1}\mathbf{x}) - \nabla T_{k,s}^{-1}(\mathbf{x})\phi_k(\mathbf{x})\| d\mu_k^*(\mathbf{x}) + R_{k,s} \\ &\le \int_{\mathbb{R}^d} (\|\phi_k(T_{k,s}^{-1}\mathbf{x}) - \phi_k(\mathbf{x})\| + \|T_{k,s}^{-1}(\mathbf{x}) - I\|\|\phi_k(\mathbf{x})\|) d\mu_k^*(\mathbf{x}) + R_{k,s} \\ &\le Cs + R_{k,s}. \end{aligned}$$

Because $\sum_{k\in[K]} p_k \nabla\cdot(\mu_k^*\phi_k) = 0$, we have

$$\begin{aligned} \sum p_k R_{k,s} &= \int_{\mathbb{R}^d} \sum_k p_k \langle \nabla T_{k,s}^{-1}(\mathbf{x}) \nabla f(T_{k,s}^{-1}\mathbf{x}), \phi_k(\mathbf{x})\rangle d\mu_k(\mathbf{x}) \\ &= \int_{\mathbb{R}^d} \sum_k p_k \langle \nabla g(\mathbf{x}), \phi_k(\mathbf{x})\rangle d\mu_k^*(\mathbf{x}) = 0, \end{aligned}$$

where $g(\mathbf{x}) = f(T_{k,s}^{-1}\mathbf{x})$.

Plugging in these results, we have

$$W_1(\bar{\mu}(t),\pi) = \mathbb{E}_{\bar{\mu}(t)}[f] - \mathbb{E}_{\bar{\mu}(0)}[f] = \int_0^t \sum p_k \int_{\mathbb{R}^d} \langle \nabla f_k(\mathbf{x}), \phi_k(\mathbf{x})\rangle d\mu_k(s,\mathbf{x})ds \le \int_0^t Csds = \frac{1}{2}Ct^2.$$

Since $\bar{\mu}(t)$ is absolutely continuous (as we will show later), there exists an optimal transport map $S_t : \mathbb{R}^d \to \mathbb{R}^d$ from $\bar{\mu}(t)$ to $\pi = \bar{\mu}$ for the 1-Wasserstein distance $W_1(\bar{\mu}(t),\pi)$; that is,

$$S_t\sharp\bar{\mu}(t) = \pi, \quad W_1(\bar{\mu}(t),\pi) = \int_{\mathbb{R}^d} \|S_t(\mathbf{x}) - \mathbf{x}\| d\bar{\mu}(t,\mathbf{x}).$$

We show that $\bar{\mu}(t)$ is absolutely continuous. Recall $\mu_k(t) = T_{k,t}\sharp\mu_k^*$, where $T_{k,t} : \mathbb{R}^d \to \mathbb{R}^d$ is a diffeomorphism for small t by Lemma A.2. Since $T_{k,t}^{-1}$ is continuously differentiable, $T_{k,t}^{-1}$ satisfies the Luzin N property (Evans 2018), which claims that, $\lambda(T_{k,t}^{-1}(B)) = 0$ if $\lambda(B) = 0$ for any measurable set B, where λ is the Lebesgue measure. Hence, for any measurable set B with 0 measure, $\mu_k(t)(B) = \mu_k^*(T_{k,t}^{-1}(B)) = 0$, since μ_k^* is absolutely continuous.

Recall that, $L(\mu) = \int_{\mathbb{R}^d}\int_{\mathbb{R}^d} \ell(\mathbf{x},\mathbf{y})d\mu(\mathbf{x})d\mu(\mathbf{y})$ for some function $\ell : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ such that

$$|\ell(\mathbf{z},\mathbf{x}) - \ell(\mathbf{z},\mathbf{y})| \le \|\mathbf{x}-\mathbf{y}\|$$

for all $\mathbf{z} \in \mathbb{R}^d$. Define $h(\mathbf{y}) = \int_{\mathbb{R}^d} \ell(\mathbf{x},\mathbf{y})d\mu(\mathbf{x})$. Then, h is Lipschitz continuous with Lipschitz constant less than one because for every $\mathbf{x},\mathbf{y} \in \mathbb{R}^d$,

$$|h(\mathbf{x}) - h(\mathbf{y})| = \left|\int_{\mathbb{R}^d} \ell(\mathbf{z},\mathbf{x}) - \ell(\mathbf{z},\mathbf{y})d\mu(\mathbf{z})\right| \le \int_{\mathbb{R}^d} |\ell(\mathbf{z},\mathbf{x}) - \ell(\mathbf{z},\mathbf{y})|\, d\mu(\mathbf{z}) \le \|\mathbf{x}-\mathbf{y}\|.$$

This implies

$$\begin{aligned}
F_{\mathbf{p}}(S_t\sharp\boldsymbol{\mu}(t)) - F_{\mathbf{p}}(\boldsymbol{\mu}(t)) &= \sum_{k\in[K]} p_k \int_{\mathbb{R}^d} h dS_t\sharp\mu_k(t) - \sum_{k\in[K]} p_k \int_{\mathbb{R}^d} h d\mu_k(t) \\
&= \int_{\mathbb{R}^d} h d(\sum_{k\in[K]} p_k S_t\sharp\mu_k(t)) - \int_{\mathbb{R}^d} h d(\sum_{k\in[K]} p_k\mu_k(t)) \\
&= \int_{\mathbb{R}^d} h d\pi - \int_{\mathbb{R}^d} h d\bar{\mu}(t) \\
&\le \sup_{f'\in\text{Lip-1}} \mathbb{E}_{\bar{\mu}(t)}[f'] - \mathbb{E}_{\pi}[f'] = W_1(\pi,\bar{\mu}) \le \frac{1}{2}Ct^2,
\end{aligned}$$

where the first equality is due to the definition of objective function. □

A.3.2 Proof of Proposition 3. The proof is similar to that of Proposition 2 in the previous section. We need the following lemma.

Lemma A.3. For any $(\boldsymbol{\mu},\mathbf{p}) \in \mathcal{P}_\pi$,

$$\mathcal{TP}_\pi(\boldsymbol{\mu},\mathbf{p}) \subseteq \mathcal{T}'\mathcal{P}_\pi(\boldsymbol{\mu},\mathbf{p}) \doteq \left\{(\boldsymbol{\phi},\mathbf{v}) : \sum_{k\in[K]} p_k\nabla\cdot(\mu_k\phi_k) = \sum_{k\in[K]} v_k\mu_k, \sum_{k\in[K]} v_k = 0\right\}.$$

Proof. For any $(\boldsymbol{\phi},\mathbf{v}) \in \mathcal{TP}_\pi(\boldsymbol{\mu},\mathbf{p})$, there exists a curve $(\boldsymbol{\mu}(t),\mathbf{p}(t)) \in \mathcal{P}_\pi$ such that, $\boldsymbol{\mu}(0) = \boldsymbol{\mu}$, $\mathbf{p}(0) = \mathbf{p}$, $\boldsymbol{\mu}(t) \in \mathcal{P}_{\pi,\mathbf{p}(t)}$, $\frac{d}{dt}\mu_k(t)|_{t=0} = -\nabla\cdot(\mu_k\phi_k)$ and $\frac{d}{dt}p_k(t)|_{t=0} = v_k\ \forall k \in [K]$. Particularly, we have for all $t \in [0,1]$, $\sum_{k\in[K]} p_k(t)\mu_k(t) = \pi$ in the sense that, for all $f \in C_c^\infty(\mathbb{R}^d)$,

$$\sum_{k\in[K]} p_k(t)\int_{\mathbb{R}^d} f(\mathbf{x})d\mu_k(t,\mathbf{x}) = \int_{\mathbb{R}^d} f(\mathbf{x})d\pi(x).$$

Since the right hand side does not depend on t, we have for all $f \in C_c^\infty(\mathbb{R}^d)$,

$$\frac{d}{dt}\sum_{k\in[K]} p_k(t)\int_{\mathbb{R}^d} f(\mathbf{x})d\mu_k(t,\mathbf{x}) = 0.$$

Hence, for all $f \in C_c^\infty(\mathbb{R}^d)$,

$$\begin{aligned}
\frac{d}{dt}\sum_{k\in[K]} p_k(t)\int_{\mathbb{R}^d} f(\mathbf{x})d\mu_k(t,\mathbf{x})|_{t=0} &= \sum_{k\in[K]} (\tfrac{d}{dt}p_k(t))\int_{\mathbb{R}^d} f(\mathbf{x})d\mu_k(t,\mathbf{x})|_{t=0} + p_k(t)\frac{d}{dt}\int_{\mathbb{R}^d} f(\mathbf{x})d\mu_k(t,\mathbf{x})|_{t=0} \\
&= \sum_{k\in[K]} v_k\int_{\mathbb{R}^d} f(\mathbf{x})d\mu_k(\mathbf{x}) + p_k\int_{\mathbb{R}^d}\langle\nabla f(\mathbf{x}),\phi_k(\mathbf{x})\rangle d\mu_k(\mathbf{x}) = 0.
\end{aligned}$$

where the second equality holds because $\frac{d}{dt}\mu_k(t)|_{t=0} = -\nabla\cdot(\mu_k\phi_k)$ and $\frac{d}{dt}p_k(t)|_{t=0} = v_k$.

This is the definition of the following equation

$$\sum_{k\in[K]} v_k\mu_k = \sum_{k\in[K]} p_k\nabla\cdot(\mu_k\phi_k).$$

Similarly, since $\sum_{k\in[K]} p_k(t) = 1$ for all t,

$$\frac{d}{dt}(\sum_{k\in[K]} p_k(t)) = 0,$$

which implies, at $t = 0$,

$$\sum_{k\in[K]} \frac{d}{dt}p_k(0) = \sum_{k\in[K]} v_k = 0. \qquad \square$$

Proof of Proposition 3. Consider the following objective function $F(\boldsymbol{\mu}, \mathbf{p}) = \sum_{k\in[K]}(p_k L(\mu_k) + \frac{\theta}{p_k^\beta})$. The Wasserstein gradient is given by

$$\nabla F(\boldsymbol{\mu}, \mathbf{p}) = (\nabla_{\boldsymbol{\mu}} F(\boldsymbol{\mu}, \mathbf{p}), \nabla_{\mathbf{p}} F(\boldsymbol{\mu}, \mathbf{p})) = (p_1\nabla L(\mu_1), \ldots, p_K\nabla L(\mu_k), L(\mu_1) - \frac{\beta\theta}{p_1^{\beta+1}}, \ldots, L(\mu_K) - \frac{\beta\theta}{p_K^{\beta+1}}).$$

For any tangent vector $(\boldsymbol{\phi}, \mathbf{v}) \in \mathcal{TP}_\pi(\boldsymbol{\mu}, \mathbf{p})$,

$$\langle \nabla F(\boldsymbol{\mu}, \mathbf{p}), (\boldsymbol{\phi}, \mathbf{v})\rangle_{\boldsymbol{\mu}} = \sum_{k\in[K]} p_k\langle \nabla L(\mu_k), \phi_k\rangle_{\mu_k} + \sum_{k\in[K]} v_k(L(\mu_k) - \frac{\beta\theta}{p_k^{\beta+1}}),$$

where $\langle \nabla L(\mu_k), \phi_k\rangle_{\mu_k} = \int_{\mathbb{R}^d} \nabla L(\mu_k)(\mathbf{x})^\top \phi_k(\mathbf{x}) d\mu_k(\mathbf{x})$.

For the sake of contradiction, assume $\|\nabla F\|_{\mathcal{TP}_\pi(\boldsymbol{\mu}^*, \mathbf{p}^*)} > 0$. Then, with the same reason as in the proof of Proposition 2, there exists tangent vector $(\boldsymbol{\phi}, \mathbf{v}) \in \mathcal{TP}_\pi(\boldsymbol{\mu}^*, \mathbf{p}^*)$ such that

(i) $\phi_k \in C_c^\infty$;
(ii) $\sum_{k\in[K]} p_k^*\nabla\cdot(\mu_k^*\phi_k) = \sum_{k\in[K]} v_k\mu_k^*$;
(iii) $A \doteq \sum_{k\in[K]} v_k(L(\mu_k^*) - \frac{\beta\theta}{(p_k^*)^{\beta+1}}) + p_k^*\langle \nabla L(\mu_k^*), \phi_k\rangle_{\mu_k^*} < 0$.

We construct a new feasible solution with this tangent vector as follows. Consider the following time-independent velocity field

$$\phi_k(t) = \phi_k, \quad \forall t \in [0, 1].$$

Then, the map $T_{k,t}^{\boldsymbol{\phi}} = x + \phi_k(x)t$ and $p_k(t) = p_k^* + v_k t$ define a curve $(\boldsymbol{\mu}(t), \mathbf{p}(t))$ by $\mu_k(t) \doteq T_{k,t}^{\boldsymbol{\phi}}\sharp\mu_k^*$ for $t \in [0, 1]$. Define $\bar{\mu}(t) = \sum_{k\in[K]} p_k(t)\mu_k(t)$ in the sense that, $\forall t \in [0, 1]$, $\bar{\mu}(t)$ is a probability measure in $\mathcal{P}_2(\mathbb{R}^d)$ such that for all $f \in C_c^\infty(\mathbb{R}^d)$, $\int_{\mathbb{R}^d} f(x)d\bar{\mu}(t) = \sum_{k\in[K]} p_k(t)\int_{\mathbb{R}^d} f(x)d\mu_k(t)$. In particular, $\bar{\mu}(0) = \pi$ since $\sum_{k\in[K]} p_k(0)\mu_k(0) = \pi$.

Note that the curve $(\boldsymbol{\mu}(t), \mathbf{p}(t))$ may not be in the feasible set $\mathcal{P}_\pi$, i.e., $\bar{\mu}(t)$ may not equal to π. Hence, to get the "better solution" we are looking for, we need to "project" this $(\boldsymbol{\mu}(t), \mathbf{p}(t))$ back to the feasible set $\mathcal{P}_\pi$.

We show the following facts: there exists a $C > 0$ such that for all $t \in [0, 1]$,

(i) $F(\boldsymbol{\mu}(t), \mathbf{p}(t)) \le F(\boldsymbol{\mu}^*, \mathbf{p}^*) + At + \frac{1}{2}Ct^2$;
(ii) there exists an optimal transport map S_t from $\bar{\mu}(t)$ to π and

$$F(S_t\sharp\boldsymbol{\mu}(t), \mathbf{p}(t)) - F(\boldsymbol{\mu}(t), \mathbf{p}(t)) \le W_1(\bar{\mu}(t), \pi) \le \frac{1}{2}Ct^2.$$

Assuming that (i) and (ii) are true, note that $S_t\sharp\boldsymbol{\mu}(t) = (S_t\sharp\mu_1(t), \ldots, S_t\sharp\mu_K(t))$ is feasible, since

$$\sum_{k\in[K]} p_k(t)S_t\sharp\mu_k(t) = S_t\sharp(\sum_{k\in[K]} p_k(t)\mu_k(t)) = S_t\sharp\bar{\mu}(t) = \pi.$$

Moreover, since $\sum_{k\in[K]} v_k = 0$, we have $\sum_{k\in[K]} p_k(t) = \sum_{k\in[K]} p_k^* + v_k = 1$. Since $p_k^* > 0$,

$p_k(t) > 0$ for small t. Hence, $\mathbf{p}(t)$ is also feasible.

Combining inequalities in statements (i) and (ii), we have

$$F(S\sharp\boldsymbol{\mu}(t), \mathbf{p}(t)) \leq F(\boldsymbol{\mu}^*, \mathbf{p}^*) + At + Ct^2.$$

Since $A < 0$ and $C > 0$, by setting small enough t, we show that $F(S\sharp\boldsymbol{\mu}(t), \mathbf{p}(t)) < F(\boldsymbol{\mu}^*, \mathbf{p}^*)$, which contradicts the optimality of $\boldsymbol{\mu}^*$.

We prove Claim (i) and (ii) as follows.

Claim (i): $F(\boldsymbol{\mu}(t), \mathbf{p}(t)) \leq F(\boldsymbol{\mu}^*, \mathbf{p}^*) + At + \frac{1}{2}Ct^2$**.**

With the same computation as in Proposition 2,

$$L(\mu_k(t)) - L(\mu_k(0)) = \int_0^t \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} \langle \nabla_1 \ell(T_{k,s}^{-1}\mathbf{x}, T_{k,s}^{-1}\mathbf{z}) + \nabla_2 \ell(T_{k,s}^{-1}\mathbf{z}, T_{k,s}^{-1}\mathbf{x}), \phi_k(T_{k,s}^{-1}\mathbf{x})\rangle d\mu_k^*(\mathbf{x}) d\mu_k^*(\mathbf{z}) ds$$

and we define

$$Q_{k,s} \doteq \int \langle \nabla_1 \ell(T_{k,s}^{-1}\mathbf{x}, T_{k,s}^{-1}\mathbf{z}) + \nabla_2 \ell(T_{k,s}^{-1}z, T_{k,s}^{-1}\mathbf{x}), \phi_k(T_{k,s}^{-1}\mathbf{x})\rangle d\mu_k^*(\mathbf{x}) d\mu_k^*(\mathbf{z}).$$

In particular, when $s = 0$,

$$Q_{k,0} = \int \langle \nabla_1 \ell(\mathbf{x}, \mathbf{z}) + \nabla_2 \ell(\mathbf{z}, \mathbf{x}), \phi_k(\mathbf{x})\rangle d\mu_k^*(\mathbf{x}) d\mu_k^*(\mathbf{z}).$$

Note that $Q_{k,0} = \langle \nabla L(\mu_k^*), \phi_k \rangle_{\mu_k^*}$. Also, note that $\|T_{k,s}^{-1}\mathbf{x} - \mathbf{x}\| \leq C_1 s$, we have that $\|\nabla_1 \ell(T_{k,s}^{-1}\mathbf{x}, T_{k,s}^{-1}\mathbf{z}) - \nabla_2 \ell(T_{k,s}^{-1}\mathbf{x}, T_{k,s}^{-1}\mathbf{z})\| \leq C_2 s(\|\mathbf{x}\| + \|\mathbf{z}\| + 1), \phi_k(T_{k,s}^{-1}\mathbf{x}) - \phi_k(\mathbf{x}) \leq C_3 s$. Therefore

$$Q_{k,s} - Q_{k,0} \leq C_4 s.$$

In summary, we have

$$L(\mu_k(t)) - L(\mu_k^*) - tQ_{k,0} \leq \frac{1}{2}C_5 t^2. \tag{A.2}$$

Also, note that

$$\frac{\theta}{p_k^\beta(t)} \leq \frac{\theta}{(p_k^*)^\beta} - \frac{\theta\beta}{(p_k^*)^{\beta+1}} v_k t + \frac{1}{2}C_6 t^2. \tag{A.3}$$

This is because $\frac{\theta}{p_k^\beta(t)}$ has Lipschitz continuous derivative on $(0, 1)$.

Hence,

$$\begin{aligned}
F(\boldsymbol{\mu}(t), \mathbf{p}(t)) &= \sum_{k\in[K]} (p_k(t) L(\mu_k(t)) + \frac{\theta}{p_k(t)^\beta}) \\
&\leq \sum_{k\in[K]} p_k(t) L(\mu_k^*) + t \sum_{k\in[K]} p_k(t) Q_{k,0} + \sum_{k\in[K]} (\frac{\theta}{(p_k^*)^\beta} - \frac{\theta\beta}{(p_k^*)^{\beta+1}} v_k t) + \frac{1}{2}C_7 t^2 \\
&= \sum_{k\in[K]} (p_k^* L(\mu_k^*) + \frac{\theta}{(p_k^*)^\beta}) + t(\sum_{k\in[K]} v_k (L(\mu_k^*) - \frac{\theta\beta}{(p_k^*)^{\beta+1}}) + \sum_{k\in[K]} p_k^* Q_{k,0}) + t^2 \sum_{k\in[K]} v_k Q_{k,0} + \frac{1}{2}C_7 t^2 \\
&= F(\boldsymbol{\mu}^*, \mathbf{p}^*) + At + \frac{1}{2}Ct^2,
\end{aligned}$$

where the second line (or the first inequality) is because of the bounds (A.2) and (A.3), the third line is from the definition $p_k(t) = p_k^* + v_k t$, and the last line is by definition of A.

Claim (ii): there exists an optimal transport map S_t from $\bar{\mu}(t)$ to π and

$$F(S_t\sharp\boldsymbol{\mu}(t), \mathbf{p}(t)) - F(\boldsymbol{\mu}(t), \mathbf{p}(t)) \leq W_1(\bar{\mu}(t), \pi) \leq \frac{1}{2}Ct^2.$$

By the duality formula of Kantorovich-Rubinstein distance (Villani 2009), there is a Lipschitz

continuous function f with Lipschitz constant no larger than 1 such that

$$W_1(\bar{\mu}(t),\pi) = \sup_{f'\in\text{Lip-1}} \mathbb{E}_{\bar{\mu}(t)}[f'] - \mathbb{E}_{\pi}[f'] = \mathbb{E}_{\bar{\mu}(t)}[f] - \mathbb{E}_{\pi}[f] = \mathbb{E}_{\bar{\mu}(t)}[f] - \mathbb{E}_{\bar{\mu}(0)}[f]$$

$$= \int_0^t \sum_{k\in[K]} v_k \int_{\mathbb{R}^d} f(\mathbf{x}) d\mu_k(s,\mathbf{x})ds + \int_0^t \sum_{k\in[K]} p_k(s) \int_{\mathbb{R}^d} \langle \nabla f(\mathbf{x}), \phi_k(\mathbf{x})\rangle d\mu_k(s,\mathbf{x})ds.$$

Define $R_{k,s} \doteq \int_{\mathbb{R}^d} \langle \nabla T_{k,s}^{-1}(\mathbf{x}) \nabla f(T_{k,s}^{-1}\mathbf{x}), \phi_k(\mathbf{x})\rangle d\mu_k^*(\mathbf{x})$. There exists a constant $C > 0$ such that for any $s \in [0,t]$,

$$\begin{aligned}
\int_{\mathbb{R}^d} \langle \nabla f(\mathbf{x}), \phi_k(\mathbf{x})\rangle d\mu_k(s,\mathbf{x}) &= \int_{\mathbb{R}^d} \langle \nabla f(T_{k,s}^{-1}\mathbf{x}), \phi_k(T_{k,s}^{-1}\mathbf{x})\rangle d\mu_k^*(\mathbf{x}) \\
&= \int_{\mathbb{R}^d} \langle \nabla f(T_{k,s}^{-1}\mathbf{x}), \phi_k(T_{k,s}^{-1}\mathbf{x}) - \nabla T_{k,s}^{-1}(\mathbf{x})\phi_k(\mathbf{x})\rangle d\mu_k^*(\mathbf{x}) + R_{k,s} \\
&\le \int_{\mathbb{R}^d} \|\phi_k(T_{k,s}^{-1}\mathbf{x}) - \nabla T_{k,s}^{-1}(\mathbf{x})\phi_k(\mathbf{x})\| d\mu_k^*(x) + R_{k,s} \\
&\le \int_{\mathbb{R}^d} (\|\phi_k(T_{k,s}^{-1}\mathbf{x}) - \phi_k(\mathbf{x})\| + \|T_{k,s}^{-1}(\mathbf{x}) - I\|\|\phi_k(\mathbf{x})\|) d\mu_k^*(\mathbf{x}) + R_{k,s} \\
&\le Cs + R_{k,s}.
\end{aligned}$$

Because $\sum_{k\in[K]} p_k^* \nabla \cdot (\mu_k^* \phi_k) = \sum_{k\in[K]} v_k \mu_k^*$, we have

$$\begin{aligned}
&\sum_{k\in[K]} v_k \int_{\mathbb{R}^d} f(\mathbf{x}) d\mu_k(s,\mathbf{x}) + \sum p_k(s) R_{k,s} \\
&= \sum_{k\in[K]} \int_{\mathbb{R}^d} v_k f(T_{k,s}^{-1}(\mathbf{x})) d\mu_k^*(\mathbf{x}) + \int_{\mathbb{R}^d} \sum_k p_k(s) \langle \nabla T_{k,s}^{-1}(\mathbf{x}) \nabla f(T_{k,s}^{-1}\mathbf{x}), \phi_k(\mathbf{x})\rangle d\mu_k^*(\mathbf{x}) \\
&= \int_{\mathbb{R}^d} \sum_{k\in[K]} v_k g(\mathbf{x}) + p_k(s) \langle \nabla g(\mathbf{x}), \phi_k(\mathbf{x})\rangle d\mu_k^*(\mathbf{x}) \\
&= \int_{\mathbb{R}^d} \sum_{k\in[K]} v_k g(\mathbf{x}) + p_k^* \langle \nabla g(\mathbf{x}), \phi_k(\mathbf{x})\rangle d\mu_k^*(\mathbf{x}) + s \int_{\mathbb{R}^d} v_k \langle \nabla g(\mathbf{x}), \phi_k(\mathbf{x})\rangle d\mu_k^*(\mathbf{x}) \le Cs,
\end{aligned}$$

where $g(\mathbf{x}) = f(T_{k,s}^{-1}\mathbf{x})$.

Plugging in these results, we have

$$W_1(\bar{\mu}(t),\pi) = \mathbb{E}_{\bar{\mu}(t)}[f] - \mathbb{E}_{\bar{\mu}(0)}[f] = \int_0^t \sum_{k\in[K]} p_k(s) \int_{\mathbb{R}^d} \langle \nabla f_k(\mathbf{x}), \phi_k(\mathbf{x})\rangle d\mu_k(s,\mathbf{x})ds \le \int_0^t Csds = \frac{1}{2}Ct^2.$$

Since $\bar{\mu}(t)$ is absolutely continuous, there exists an optimal transport map $S_t : \mathbb{R}^d \to \mathbb{R}^d$ from $\bar{\mu}(t)$ to $\pi = \bar{\mu}$ for the 1-Wasserstein distance $W_1(\bar{\mu}(t),\pi)$; that is,

$$S_t \sharp \bar{\mu}(t) = \pi, \quad W_1(\bar{\mu}(t),\pi) = \int_{\mathbb{R}^d} \|S_t(\mathbf{x}) - \mathbf{x}\| d\bar{\mu}(t,\mathbf{x}).$$

With the same reason as in the proof of Proposition 2,

$$F(S_t \sharp \boldsymbol{\mu}(t), \mathbf{p}(t)) - F(\boldsymbol{\mu}(t), \mathbf{p}(t)) \le W_1(\pi,\bar{\mu}) \le \frac{1}{2}Ct^2. \qquad \square$$

A.4 Proofs of Section 5

We will frequently apply the following Grönwall's inequality in the proofs. This inequality is key to obtaining the exponential decrease of the Kullback-Leibler divergence in Section 5.

Lemma A.4 (Grönwall's inequality)**.** Let β and u be real-valued continuous functions defined on the closed interval $[a,b]$. If u is differentiable on (a,b) and satisfies $u'(t) \le \beta(t)u(t)$ for all

$t \in (a,b)$ then $u(t) \le u(a)e^{\int_a^t \beta(s)ds}$ for all $t \in [a,b]$.

A.4.1 Proof of Lemma 7.

Proof. Denote the optimal value of (20) by V^*,i.e.,

$$\begin{aligned} V^* \doteq \min_{\mathbf{x}\in\mathbb{R}^n} \quad & f(\mathbf{x}_k^\tau) + \langle \nabla f(\mathbf{x}_k^\tau), \mathbf{x} - \mathbf{x}_k^\tau \rangle + \frac{1}{2\tau} d(\mathbf{x}, \mathbf{x}_k^\tau)^2 \\ \text{s.t.} \quad & \tilde{g}(\mathbf{x}) = (1-\alpha h) g(\mathbf{x}_k^\tau). \end{aligned}$$

Note that this is a convex optimization problem with one affine inequality constraint. Thus, it satisfies the weak Slater's condition, which implies strong duality; that is,

$$V^* = \max_{\lambda\in\mathbb{R}} \min_{\mathbf{x}\in\mathbb{R}^n} f(\mathbf{x}_k^\tau) + \langle \nabla f(\mathbf{x}_k^\tau) + \lambda \nabla g(\mathbf{x}_k^\tau), x - \mathbf{x}_k^\tau \rangle + \frac{1}{2\tau}\|\mathbf{x} - \mathbf{x}_k\|^2 + \lambda\alpha\tau g(\mathbf{x}_k^\tau),$$

where the right-hand side is the optimal value of the Lagrangian dual problem.

Since the objective function of the right-hand side is a quadratic function in x, the solution $(\mathbf{x}_{k+1}^\tau, \lambda^*)$ is given by

$$\mathbf{x}_{k+1}^\tau = \mathbf{x}_k^\tau - \tau(\nabla f(\mathbf{x}_k^\tau) + \lambda^* \nabla g(\mathbf{x}_k^\tau)), \quad \lambda^* = \frac{-\langle \nabla g(\mathbf{x}_k^\tau), \nabla f(\mathbf{x}_k^\tau)\rangle + \alpha g(\mathbf{x}_k^\tau)}{\|\nabla g(\mathbf{x}_k^\tau)\|^2}. \qquad \square$$

A.4.2 Proof of Theorem 1.

Proof. **(i):**

$$\begin{aligned} \tfrac{d}{dt} g(\mathbf{x}(t)) &= \langle \mathbf{x}'(t), \nabla g(\mathbf{x}(t))\rangle = \langle \phi(\mathbf{x}(t)), \nabla g(\mathbf{x}(t))\rangle \\ &= -\langle \nabla f(\mathbf{x}(t)), \nabla g(\mathbf{x}(t))\rangle - \lambda(t)\|\nabla g(\mathbf{x}(t))\|^2 \\ &= -\alpha g(\mathbf{x}(t)). \end{aligned}$$

By Grönwall's inequality, we prove statement (i).

(ii): We start with the derivative of $f(\mathbf{x}(t))$.

$$\begin{aligned} \tfrac{d}{dt} f(\mathbf{x}(t)) &= \langle \mathbf{x}'(t), \nabla f(\mathbf{x}(t))\rangle = \langle \phi(\mathbf{x}(t)), \nabla f(\mathbf{x}(t))\rangle \\ &= -\|\nabla f(\mathbf{x}(t))\|^2 - \lambda(t)\langle \nabla g(\mathbf{x}(t)), \nabla f(\mathbf{x}(t))\rangle, \end{aligned} \tag{A.4}$$

where the first line is by chain's rule and the second line is due to the definition of ϕ.

Integrating both parts of (A.4) and using the fundamental theorem of calculus, we have

$$f(\mathbf{x}(T)) - f(\mathbf{x}(0)) = \int_0^T \frac{d}{dt} f(\mathbf{x}(t))dt = -\int_0^T \|\nabla f(\mathbf{x}(t)) + \lambda(t)\nabla g(\mathbf{x}(t))\|^2 + \alpha\lambda(t) g(\mathbf{x}(t))dt.$$

This implies

$$\begin{aligned} \int_0^T \|\nabla f(\mathbf{x}(t)) + \lambda(t)\nabla g(\mathbf{x}(t))\|^2 dt &= f(\mathbf{x}(0)) - f(\mathbf{x}(T)) + \int_0^T \alpha\lambda(t) g(\mathbf{x}(t))dt \\ &\le f(\mathbf{x}(0)) - f_{\min} + \int_0^T \frac{|\langle \nabla g(\mathbf{x}(t)), \nabla f(\mathbf{x}(t))\rangle| g(\mathbf{x}(t))}{\|\nabla g(\mathbf{x}(t))\|^2} + \frac{\alpha g(\mathbf{x}(t))^2}{\|\nabla g(\mathbf{x}(t))\|^2} dt \\ &\le f(\mathbf{x}(0)) - f_{\min} + \int_0^T L\sqrt{\frac{g(\mathbf{x}(t))}{\kappa}} + \frac{\alpha g(\mathbf{x}(t))}{\kappa} dt \\ &\le f(\mathbf{x}(0)) - f_{\min} + \frac{2g(\mathbf{x}(0))}{\kappa} + \frac{L}{\alpha\sqrt{\kappa}}\sqrt{g(\mathbf{x}(0))}, \end{aligned}$$

where the first inequality is because

$$\lambda(t) = \frac{-\langle \nabla f(\mathbf{x}(t)), \nabla g(\mathbf{x}(t))\rangle + \alpha g(\mathbf{x}(t))}{\|\nabla g(\mathbf{x}(t))\|^2} \le \frac{|\langle \nabla f(\mathbf{x}(t)), \nabla g(\mathbf{x}(t))\rangle| + \alpha g(\mathbf{x}(t))}{\|\nabla g(\mathbf{x}(t))\|^2};$$

the second inequality is by Assumption 2, and the last inequality is by statement (i), i.e., $g(\mathbf{x}(t)) \le e^{-\alpha t} g(\mathbf{x}(0))$.

Since the upper bound we have just shown does not depend on the time index T, we have for all $T > 0$ that

$$\min_{t\le T} \|\nabla f(\mathbf{x}(t)) + \lambda(\mathbf{x}(t))\nabla g(\mathbf{x}(t))\|^2 \le \frac{C}{T},$$

for some constant $C > 0$. Define $M \doteq \{\mathbf{x} \in \mathbb{R}^d : g(\mathbf{x}) = 0\}$ and $\mathcal{T}M(x) \doteq \{\mathbf{v} \in \mathbb{R}^d : \mathbf{v}^\top \nabla g(\mathbf{x}) = 0\}$. Then, for all $\mathbf{v} \in \mathcal{T}M(\mathbf{x})$, $\mathbf{v}^\top \nabla f(\mathbf{x}(t)) = \mathbf{v}^\top (f(\mathbf{x}(t)) + \lambda(t)\nabla g(\mathbf{x}(t)))$. Hence,

$$\begin{aligned}\min_{t\le T} \|\nabla f(\mathbf{x}(t))\|^2_{\mathcal{T}M_{g(\mathbf{x}(t))}} &= \min_{t\le T} \left(\sup_{v\in\mathcal{T}M(\mathbf{x})} \frac{\mathbf{v}^\top \nabla f(\mathbf{x}(t))}{\|\mathbf{v}\|}\right)^2 = \min_{t\le T} \left(\sup_{\mathbf{v}\in\mathcal{T}M(\mathbf{x})} \frac{v^\top (f(\mathbf{x}(t)) + \lambda(t)\nabla g(\mathbf{x}(t)))}{\|\mathbf{v}\|}\right)^2 \\ &\le \min_{t\le T} \|\nabla f(\mathbf{x}(t)) + \lambda(\mathbf{x}(t))\nabla g(\mathbf{x}(t))\|^2 \le \frac{C}{T}.\end{aligned}$$

□

A.4.3 Proof of Theorem 2. We first show the following direct consequences of Assumption 3 that will be useful later.

Lemma A.5. Suppose Assumption 3 holds. Then,

(i) $\|\nabla_{\mu_k} \mathrm{KL}(\bar{\mu}\|\pi)\|^2_{\bar{\mu}} > 0$ if $\mathrm{KL}(\bar{\mu}\|\pi) > 0$;
(ii) $\|L(\mu)\| \le \ell_{\max}$, $\forall \mu \in \mathcal{P}_{2,ac}(\mathbb{R}^d)$;
(iii) $F_{\mathbf{p}}(\boldsymbol{\mu}) > -\ell_{\max}$, $\forall \boldsymbol{\mu} \in \mathcal{P}^{\otimes K}_{2,ac}$;
(iv) $\|\nabla L(\mu)\|_\mu \le 2L_{\max}\sqrt{K}$, $\forall \mu \in \mathcal{P}_{2,ac}(\mathbb{R}^d)$;
(v) $\frac{\mathrm{KL}(\bar{\mu}\|\pi)}{\|\nabla_{\boldsymbol{\mu}} \mathrm{KL}(\bar{\mu}\|\pi)\|_{\boldsymbol{\mu}}} \le \frac{\sqrt{\mathrm{KL}(\bar{\mu}\|\pi)}}{p_{\min}\sqrt{\kappa}}$.

Proof. **(i)** This is because $\|\nabla_{\mu_k} \mathrm{KL}(\bar{\mu}\|\pi)\|^2_{\bar{\mu}} = p_k^2 \|s_{\bar{\mu}} - s_\pi\|^2_{\bar{\mu}} > 0$ if $\bar{\mu} \neq \pi$.
(ii) This is because $|L(\mu)| = |\int_{\mathbb{R}^d}\int_{\mathbb{R}^d} \ell(\mathbf{x},\mathbf{y}) d\mu(\mathbf{x}) d\mu(\mathbf{y}) \le \int_{\mathbb{R}^d}\int_{\mathbb{R}^d} |\ell(\mathbf{x},\mathbf{y})| d\mu(\mathbf{x}) d\mu(\mathbf{y}) \le \ell_{\max}$ for all $\mu \in \mathcal{P}_{2,ac}(\mathbb{R}^d)$.
(iii) $F_{\mathbf{p}}(\boldsymbol{\mu}) = \sum_{k=1}^K p_k L(\mu_k) > -\ell_{\max}$.
(iv) For all $\mathbf{x} \in \mathbb{R}^d$,

$$\begin{aligned}\|\nabla L(\mu)(\mathbf{x})\|^2 &= \|\int_{\mathbb{R}^d} \nabla_1 \ell(\mathbf{x},\mathbf{z}) + \nabla_2 \ell(\mathbf{z},\mathbf{x}) d\mu(\mathbf{z})\|^2 \\ &= \sum_{i=1}^{2d} (\int_{\mathbb{R}^d} \nabla_1 \ell(\mathbf{x},\mathbf{z})_i + \nabla_2 \ell(\mathbf{z},\mathbf{x})_i d\mu(\mathbf{z}))^2 \\ &\le \sum_{i=1}^{2d} \int_{\mathbb{R}^d} (\nabla_1 \ell(\mathbf{x},\mathbf{z})_i + \nabla_2 \ell(\mathbf{z},\mathbf{x})_i)^2 d\mu(\mathbf{z}) \\ &\le \sum_{i=1}^{2d} \int (|\nabla_1 \ell(\mathbf{x},\mathbf{z})_i| + |\nabla_2 \ell(\mathbf{z},\mathbf{x})_i|)^2 d\mu(z) \\ &\le 4L^2_{\max} K.\end{aligned}$$

Hence, $\|\nabla L(\mu)\|^2_\mu = \int_{\mathbb{R}^d} \|\nabla L(\mu)(\mathbf{x})\|^2 d\mu(\mathbf{x}) \le 4L^2_{\max} K$.

(v)

$$\begin{aligned}
&\frac{\mathrm{KL}(\bar{\mu}\|\pi)}{\sqrt{\sum_{k\in[K]}\|p_k(s_{\bar{\mu}}-s_\pi)\|^2_{\mu_k}}}\\
&\leq \frac{\mathrm{KL}(\bar{\mu}(t)\|\pi)}{p_{\min}\sqrt{\sum_{k\in[K]}p_k\|(s_{\bar{\mu}}-s_\pi)\|^2_{\mu_k}}}\\
&= \frac{\mathrm{KL}(\bar{\mu}\|\pi)}{p_{\min}\sqrt{\|(s_{\bar{\mu}}-s_\pi)\|^2_{\bar{\mu}}}}\\
&\leq \tfrac{1}{p_{\min}\sqrt{\kappa}}\sqrt{\mathrm{KL}(\bar{\mu}\|\pi)},
\end{aligned} \tag{A.5}$$

where $p_{\min}=\min_{k\in[K]}p_k$; the equality is because

$$\sum_{k\in[K]}p_k\|(s_{\bar{\mu}}-s_\pi)\|^2_{\mu_k}=\sum_{k\in[K]}p_k\int_{\mathbb{R}^d}\|(s_{\bar{\mu}}-s_\pi)\|^2d\mu_k(\mathbf{x})=\int_{\mathbb{R}^d}\|(s_{\bar{\mu}}-s_\pi)\|^2d\left(\sum_{k\in[K]}p_k\mu_k(\mathbf{x})\right)=\|(s_{\bar{\mu}}-s_\pi)\|^2_{\bar{\mu}};$$

and the last line is by the statement (iii) in Assumption 3. □

Proof of Theorem 2. **(i)** We show that $\frac{d}{dt}\mathrm{KL}(\bar{\mu}(t)\|\pi)\leq-\alpha\mathrm{KL}(\bar{\mu}(t)\|\pi)$. Then, by Grönwall's inequality, we have

$$\mathrm{KL}(\bar{\mu}(t)\|\pi)\leq e^{-\alpha t}\mathrm{KL}(\mu(0)\|\pi).$$

First note that

$$\begin{aligned}
\frac{d}{dt}\mathrm{KL}(\bar{\mu}(t)\|\pi)&=\frac{d}{dt}\int_{\mathbb{R}^d}\bar{\mu}(t,\mathbf{x})\log\frac{\bar{\mu}(t,\mathbf{x})}{\pi(\mathbf{x})}dx\\
&=\int_{\mathbb{R}^d}\frac{d}{dt}[\bar{\mu}(t,\mathbf{x})\log\frac{\bar{\mu}(t,\mathbf{x})}{\pi(\mathbf{x})}]dx\\
&=\int_{\mathbb{R}^d}[\frac{d}{dt}\bar{\mu}(t,\mathbf{x})]\cdot\log\frac{\bar{\mu}(t,\mathbf{x})}{\pi(\mathbf{x})}+\bar{\mu}(t,\mathbf{x})\cdot\frac{d}{dt}[\log\frac{\bar{\mu}(t,\mathbf{x})}{\pi(\mathbf{x})}]dx\\
&=\int_{\mathbb{R}^d}[\frac{d}{dt}\bar{\mu}(t,\mathbf{x})]\cdot\log\frac{\bar{\mu}(t,\mathbf{x})}{\pi(\mathbf{x})}+\bar{\mu}(t,\mathbf{x})\cdot\frac{\frac{d}{dt}\bar{\mu}(t,\mathbf{x})}{\bar{\mu}(t,\mathbf{x})}d\mathbf{x}\\
&=\int_{\mathbb{R}^d}[-\sum_{k\in[K]}p_k\nabla\cdot(\phi_k\mu_k(t))]\cdot\log\frac{\bar{\mu}(t,\mathbf{x})}{\pi(\mathbf{x})}-\sum_{k\in[K]}p_k\nabla\cdot(\phi_k\mu_k(t))d\mathbf{x}\\
&=\sum_{k\in[K]}p_k\int_{\mathbb{R}^d}\langle\nabla\log\frac{\bar{\mu}(t,\mathbf{x})}{\pi(\mathbf{x})},\phi_k(\mathbf{x})\rangle d\mu_k(t,\mathbf{x})\\
&=\sum_{k\in[K]}p_k\langle s_{\bar{\mu}(t)}-s_\pi,\phi_k\rangle_{\mu_k(t)}\\
&=-\lambda(\boldsymbol{\mu}(t))\sum_{k\in[K]}\|p_k(s_{\bar{\mu}(t)}-s_\pi)\|^2_{\mu_k(t)}-p_k\langle p_k(s_{\bar{\mu}(t)}-s_\pi),\nabla L(\mu_k(t))\rangle_{\mu_k(t)}\\
&=-\alpha\mathrm{KL}(\bar{\mu}(t)\|\pi).
\end{aligned}$$

The sixth equality follows by integration by parts and the fact that the integral of the divergence of a vector field vanishing at infinity is 0 by the divergence theorem. The seventh equality is by the definition of $s_{\bar{\mu}}$ and s_π. The eighth equality is due to equation (23). We get the last equality by inserting the formula of $\lambda(\boldsymbol{\mu}(t))$ into equation (23).

(ii)

$$\begin{aligned}
\frac{d}{dt}F_{\mathbf{p}}(\boldsymbol{\mu}(t)) &= \sum_{k\in[K]} p_k\langle \nabla L(\mu_k(t)), \phi_k\rangle_{\mu_k(t)} \\
&= \sum_{k\in[K]} p_k\langle \nabla L(\mu_k(t)), -p_k(\nabla L(\mu_k(t)) + \lambda(\boldsymbol{\mu}(t))(s_{\bar{\mu}(t)} - s_\pi))\rangle_{\mu_k(t)} \\
&= \sum_{k\in[K]} p_k\langle \nabla L(\mu_k(t)) + \lambda(\boldsymbol{\mu}(t))(s_{\bar{\mu}(t)} - s_\pi), -p_k(\nabla L(\mu_k(t)) + \lambda(\boldsymbol{\mu}(t))(s_{\bar{\mu}(t)} - s_\pi))\rangle_{\mu_k(t)} \\
&\quad - \sum_{k\in[K]} p_k\langle \lambda(\boldsymbol{\mu}(t))(s_{\bar{\mu}(t)} - s_\pi)), -p_k(\nabla L(\mu_k(t)) + \lambda(\boldsymbol{\mu}(t))(s_{\bar{\mu}(t)} - s_\pi))\rangle_{\mu_k(t)} \\
&= -\sum_{k\in[K]} p_k^2\|\nabla L(\mu_k(t)) + \lambda(\boldsymbol{\mu}(t))(s_{\bar{\mu}(t)} - s_\pi)\|^2_{\mu_k(t)} - \lambda(\boldsymbol{\mu}(t)) \sum_{k\in[K]} p_k\langle s_{\bar{\mu}(t)} - s_\pi, \phi_k(\boldsymbol{\mu}(t))\rangle_{\mu_k(t)} \\
&= -\sum_{k\in[K]} \|\phi_k(\boldsymbol{\mu}(t))\|^2_{\mu_k(t)} - \lambda(\boldsymbol{\mu}(t)) \sum_{k\in[K]} p_k\langle s_{\bar{\mu}(t)} - s_\pi, \phi_k(\boldsymbol{\mu}(t))\rangle_{\mu_k(t)}
\end{aligned}$$

The first equality follows from the chain rule. The second equality is by the formula (23) of $\boldsymbol{\phi}$. The last two lines are due to the formula (23) of $\boldsymbol{\phi}$.

By the formula for $\lambda(\boldsymbol{\mu}(t))$ in equation (23), we have

$$\sum_{k\in[K]} p_k\langle s_{\bar{\mu}(t)} - s_\pi, \phi_k(\boldsymbol{\mu}(t))\rangle_{\mu_k(t)} = -\alpha \mathrm{KL}(\bar{\mu}(t)\|\pi).$$

Hence,

$$\frac{d}{dt}F_{\mathbf{p}}(\boldsymbol{\mu}(t)) = -\sum_{k\in[K]} \|\phi_k(\boldsymbol{\mu}(t))\|^2_{\mu_k(t)} + \lambda(\boldsymbol{\mu}(t))\alpha \mathrm{KL}(\bar{\mu}(t)\|\pi).$$

Then, using the fundamental theorem of calculus, we have for any $T > 0$,

$$\begin{aligned}
F_{\mathbf{p}}(\boldsymbol{\mu}(T)) - F_{\mathbf{p}}(\boldsymbol{\mu}(0)) &= \int_0^T \frac{d}{dt}F_{\mathbf{p}}(\boldsymbol{\mu}(t))dt \\
&= \int_0^T -\sum_{k\in[K]} \|\phi_k(\boldsymbol{\mu}(t))\|^2_{\mu_k(t)} + \lambda(\boldsymbol{\mu}(t))\alpha \mathrm{KL}(\bar{\mu}(t)\|\pi)dt.
\end{aligned}$$

This implies

$$\begin{aligned}
\int_0^T \sum_{k\in[K]} \|\phi_k(\boldsymbol{\mu}(t))\|^2_{\mu_k(t)}dt &= F_{\mathbf{p}}(\boldsymbol{\mu}(0)) - F_{\mathbf{p}}(\boldsymbol{\mu}(T)) + \int_0^T \lambda(\boldsymbol{\mu}(t))\alpha \mathrm{KL}(\bar{\mu}(t)\|\pi)dt \\
&\le F_{\mathbf{p}}(\boldsymbol{\mu}(0)) + \ell_{\max} + \int_0^T \lambda(\boldsymbol{\mu}(t))\alpha \mathrm{KL}(\bar{\mu}(t)\|\pi)dt,
\end{aligned}$$

where the inequality is due to statement (iii) in Lemma A.5, i.e., $F_{\mathbf{p}}(\mu) \ge -l_{\max}$.

By inserting the formula for $\lambda(\boldsymbol{\mu}(t))$, we have

$$\begin{aligned}
\int_0^T \lambda(\boldsymbol{\mu}(t))\alpha \mathrm{KL}(\bar{\mu}(t)\|\pi)dt = -\alpha \int_0^T \frac{\sum_{k\in[K]}\langle p_k \nabla L(\mu_k), p_k(s_{\bar{\mu}(t)} - s_\pi)\rangle_{\mu_k(t)}}{\sum_{k\in[K]} \|p_k(s_{\bar{\mu}(t)} - s_\pi)\|^2_{\mu_k(t)}} \mathrm{KL}(\bar{\mu}(t)\|\pi)dt \\
+ \alpha^2 \int_0^T \frac{\mathrm{KL}(\bar{\mu}(t)\|\pi)^2}{\sum_{k\in[K]} \|p_k(s_{\bar{\mu}(t)} - s_\pi)\|^2_{\mu_k(t)}} dt
\end{aligned} \tag{A.6}$$

We bound these two terms in this equation separately. For the first term, we have

$$
\begin{aligned}
&-\alpha\int_0^T \frac{\sum_{k\in[K]}\langle p_k\nabla L(\mu_k), p_k(s_{\bar{\mu}(t)}-s_\pi)\rangle_{\mu_k(t)}}{\sum_{k\in[K]}\|p_k(s_{\bar{\mu}(t)}-s_\pi)\|^2_{\mu_k(t)}}\mathrm{KL}(\bar{\mu}(t)\|\pi)dt \\
&\le \alpha\int_0^T \frac{\sqrt{\sum_{k\in[K]}\|p_k\nabla L(\mu_k(t))\|^2_{\mu_k(t)}}}{\sqrt{\sum_{k\in[K]}\|p_k(s_{\bar{\mu}(t)}-s_\pi)\|^2_{\mu_k(t)}}}\mathrm{KL}(\bar{\mu}(t)\|\pi)dt,
\end{aligned} \tag{A.7}
$$

by Cauchy-Schwartz inequality.

With statement (iv) in Lemma A.5, we can bound the numerator as

$$
\sqrt{\sum_{k\in[K]}\|p_k\nabla L(\mu_k(t))\|^2_{\mu_k(t)}} \le 2L_{\max}\sqrt{K}. \tag{A.8}
$$

By statement (v) in Lemma A.5, we can bound the denominator as

$$
\int_0^T \frac{1}{\sqrt{\sum_{k\in[K]}\|p_k(s_{\bar{\mu}(t)}-s_\pi)\|^2_{\mu_k(t)}}}\mathrm{KL}(\bar{\mu}(t)\|\pi)dt \le \tfrac{1}{p_{\min}\sqrt{\kappa}}\int_0^T\sqrt{\mathrm{KL}(\bar{\mu}(t)\|\pi)}dt, \tag{A.9}
$$

Hence, equation (A.7) becomes

$$
\begin{aligned}
&-\alpha\int_0^T \frac{\sum_{k\in[K]}\langle p_k\nabla L(\mu_k), p_k(s_{\bar{\mu}(t)}-s_\pi)\rangle_{\mu_k(t)}}{\sum_{k\in[K]}\|p_k(s_{\bar{\mu}(t)}-s_\pi)\|^2_{\mu_k(t)}}\mathrm{KL}(\bar{\mu}(t)\|\pi)dt \\
&\le \alpha\int_0^T \frac{\sqrt{\sum_{k\in[K]}\|p_k\nabla L(\mu_k(t))\|^2_{\mu_k(t)}}}{\sqrt{\sum_{k\in[K]}\|p_k(s_{\bar{\mu}(t)}-s_\pi)\|^2_{\mu_k(t)}}}\mathrm{KL}(\bar{\mu}(t)\|\pi)dt \\
&\le \frac{\alpha 2L_{\max}\sqrt{K}}{p_{\min}\sqrt{\kappa}}\int_0^T\sqrt{\mathrm{KL}(\bar{\mu}(t)\|\pi)}dt \\
&\le \frac{\alpha 2L_{\max}\sqrt{K}}{p_{\min}\sqrt{\kappa}}\int_0^T\sqrt{e^{-\alpha t}\mathrm{KL}(\bar{\mu}(0)\|\pi)}dt \\
&\le \frac{4\alpha L_{\max}\sqrt{K}}{p_{\min}\sqrt{\kappa}}\sqrt{\mathrm{KL}(\bar{\mu}(0)\|\pi)},
\end{aligned}
$$

where the second inequality is by (A.8) and (A.9), the third inequality is by statement (i) of Theorem 2, and the last inequality is by integration.

Next, we bound the second term in (A.6) as follows.

$$
\begin{aligned}
&\alpha^2\int_0^T \frac{\mathrm{KL}(\bar{\mu}(t)\|\pi)^2}{\sum_{k\in[K]}\|p_k(s_{\bar{\mu}(t)}-s_\pi)\|^2_{\mu_k(t)}}dt \\
&\le \alpha^2\int_0^T \frac{\mathrm{KL}(\bar{\mu}(t)\|\pi)^2}{p_{\min}\|p_k(s_{\bar{\mu}(t)}-s_\pi)\|^2_{\bar{\mu}(t)}}dt \\
&\le \alpha^2\frac{1}{p_{\min}\kappa}\int_0^T\mathrm{KL}(\bar{\mu}(t)\|\pi)dt \\
&\le \alpha^2\frac{1}{\kappa}\mathrm{KL}(\bar{\mu}(0)\|\pi)\int_0^T e^{-\alpha t}dt \\
&\le \alpha\frac{1}{p_{\min}\kappa}\mathrm{KL}(\bar{\mu}(0)\|\pi),
\end{aligned}
$$

where the first inequality is obtained by the same way as in (A.9), the second inequality is due to statement (iii) of Assumption 3, and the third inequality is due to part (i) of Theorem 2.

To conclude,

$$\int_0^T \sum_{k\in[K]} \|\phi_k(\boldsymbol{\mu}(t))\|^2_{\mu_k(t)} dt \le F_{\mathbf{p}}(\boldsymbol{\mu}(0)) + \ell_{\max} + \frac{4\alpha L_{\max}\sqrt{K}}{p_{\min}\sqrt{\kappa}} \sqrt{\mathrm{KL}(\bar{\mu}(0)\|\pi)} + \alpha \frac{1}{p_{\min}\kappa} \mathrm{KL}(\bar{\mu}(0)\|\pi).$$

Let C denote the constant $F_{\mathbf{p}}(\boldsymbol{\mu}(0)) + \ell_{\max} + \frac{4\alpha L_{\max}\sqrt{K}}{p_{\min}\sqrt{\kappa}} \sqrt{\mathrm{KL}(\bar{\mu}(0)\|\pi)} + \alpha \frac{1}{p_{\min}\kappa} \mathrm{KL}(\bar{\mu}(0)\|\pi)$. This implies

$$\min_{t\le T} \sum_{k\in[K]} \|\phi_k(\boldsymbol{\mu}(t))\|^2_{\mu_k(t)} \le \frac{C}{T}.$$

For any unit-norm $v \in \mathcal{TP}_{\pi,\mathbf{p}}(\boldsymbol{\mu}(t))$, we have

$$\begin{aligned}\langle \nabla F_{\mathbf{p}}(\boldsymbol{\mu}(t)), v\rangle &= \sum_{k\in[K]} \langle p_k \nabla L(\mu_k(t)), v_k\rangle_{\mu_k(t)} = \sum_{k\in[K]} \langle p_k \nabla L(\mu_k(t)) + p_k \lambda(\boldsymbol{\mu}(t))(s_{\bar{\mu}(t)} - s_\pi), v_k\rangle_{\mu_k(t)} \\ &\quad - \sum_{k\in[K]} \langle p_k \lambda(\boldsymbol{\mu}(t))(s_{\bar{\mu}(t)} - s_\pi), v_k\rangle_{\mu_k(t)}.\end{aligned}$$

By Lemma A.1, $\sum_{k\in[K]} p_k \nabla \cdot (v_k \mu_k(t)) = 0$. Thus,

$$\sum_{k\in[K]} \langle p_k (s_{\bar{\mu}(t)} - s_\pi), v_k\rangle_{\mu_k(t)} = \sum_{k\in[K]} p_k \int_{\mathbb{R}^d} \langle \nabla \log \frac{\bar{\mu}(t,x)}{\pi(x)}, v_k\rangle d\mu_k(t,x) = 0,$$

where the second equality is due to $\sum_{k\in[K]} p_k \nabla \cdot (v_k \mu_k(t)) = 0$.

To conclude,

$$\begin{aligned}\langle \nabla F_{\mathbf{p}}(\boldsymbol{\mu}(t)), v\rangle &= \sum_{k\in[K]} \langle p_k \nabla L(\mu_k(t)) + p_k \lambda(\boldsymbol{\mu}(t))(s_{\bar{\mu}(t)} - s_\pi), v_k\rangle_{\mu_k(t)} \\ &\le \sum_{k\in[K]} \|p_k \nabla L(\mu_k(t)) + p_k \lambda(\boldsymbol{\mu}(t))(s_{\bar{\mu}(t)} - s_\pi)\|_{\mu_k(t)} \|v_k\|_{\mu_k(t)} \\ &\le \frac{1}{\sqrt{K}} \sum_{k\in[K]} \|\phi_k(\boldsymbol{\mu}(t))\|_{\boldsymbol{\mu}(t)}.\end{aligned}$$

Therefore,

$$\min_{t\le T} \|\nabla F_{\mathbf{p}}(\boldsymbol{\mu}(t))\|_{\mathcal{TP}_{\pi,\mathbf{p}}(\boldsymbol{\mu}(t))} \le \frac{C}{\sqrt{KT}}.$$ □

A.4.4 Proof of Theorem 3. We first show the following bounds that will be useful later.

Lemma A.6. Suppose Assumption 3 holds. Then for all $\boldsymbol{\mu} \in \mathcal{P}_{2,ac}(\mathbb{R}^d)^{\otimes K}, \mathbf{p} \in \mathbb{R}^{Kd}$,

$$\|\nabla F(\boldsymbol{\mu}, \mathbf{p})\|_{\boldsymbol{\mu},P} \le 2\sqrt{L^2_{\max} K^2 + K(\ell_{\max} + \tfrac{\theta\beta}{p_{\min}^{\beta+1}})^2}.$$

Proof. By definition,

$$\|\nabla F(\boldsymbol{\mu}, \mathbf{p})\|^2_{\boldsymbol{\mu},P} = \|\nabla_{\boldsymbol{\mu}} F(\boldsymbol{\mu}, \mathbf{p})\|^2_{\boldsymbol{\mu}} + \|P \nabla_{\mathbf{p}} F(\boldsymbol{\mu}, \mathbf{p})\|^2.$$

We bound these two terms separately.

$$\|\nabla_{\boldsymbol{\mu}} F(\boldsymbol{\mu}, \mathbf{p})\|^2_{\boldsymbol{\mu}} = \sum_{k=1}^K \|\nabla_{\mu_k} F(\boldsymbol{\mu}, \mathbf{p})\|^2_{\mu_k} = \sum_{k=1}^K \|p_k \nabla L(\mu_k)\|^2_{\mu_k} \le 4L^2_{\max} K^2,$$

where the last inequality is by statement (iv) in Lemma A.5.

$$\|P\nabla_{\mathbf{p}}F(\boldsymbol{\mu},\mathbf{p})\|^2 = \sum_{k=1}^K [P(L(\mu_k) - \tfrac{\theta\beta}{p_k^{\beta+1}})]^2 = \sum_{k=1}^K (L(\mu_k) - \tfrac{\theta\beta}{p_k^{\beta+1}} - \tfrac{1}{K}\sum_{j=1}^K (L(\mu_j) - \tfrac{\theta\beta}{p_j^{\beta+1}}))^2$$
$$\le \sum_{k=1}^K (|L(\mu_k)| + |\tfrac{\theta\beta}{p_k^{\beta+1}}| + \tfrac{1}{K}\sum_{j=1}^K (|L(\mu_j)| + |\tfrac{\theta\beta}{p_j^{\beta+1}}|))^2$$
$$\le \sum_{k=1}^K (\ell_{\max} + \tfrac{\theta\beta}{p_{\min}^{\beta+1}} + \ell_{\max} + \tfrac{\theta\beta}{p_{\min}^{\beta+1}})^2 = 4K(\ell_{\max} + \tfrac{\theta\beta}{p_{\min}^{\beta+1}})^2,$$

where the last line is by Lemma A.5. Combining these two bounds, we prove the lemma. □

Proof of Theorem 3. **(i):** For any k,

$$v_k = -P(\nabla_{\mathbf{p}}F + \lambda\nabla_{\mathbf{p}}\mathrm{KL}_k)$$
$$= -[L(\mu_k) - \tfrac{\beta}{p_k^{\beta+1}} + \lambda(\int \mu_k \log\tfrac{\bar\mu}{\pi}dx + 1) - \tfrac{1}{K}\sum_j (L(\mu_j) - \tfrac{\beta}{p_j^{\beta+1}} + \lambda(\int \mu_j \log\tfrac{\bar\mu}{\pi}dx + 1))]$$
$$= \tfrac{1}{K}\sum_j (L(\mu_j) - L(\mu_k)) - \tfrac{1}{K}\sum_j (\tfrac{\beta}{p_j^{\beta+1}} - \tfrac{\beta}{p_k^{\beta+1}}) + \tfrac{1}{K}\lambda\sum_j \int (\mu_k - \mu_j)\log\tfrac{\bar\mu}{\pi}dx).$$

Now suppose t_0 is the first t that $p_i(t) = p_{\min}$ for some i. Without loss of generality, assume $i = 1$. Then $\tilde p(t) = [p_2, \dots, p_K] \in S := \{\sum_{j=2}^K p_j = 1 - p_{\min}, p_j \ge p_{\min}\}$, which is a simplex. Observe that $f(x) = \sum x_j^{-(\beta+1)}$ is a convex function; therefore its maximal value is reached at a vertice of S, that is

$$\sum_{j=2}^K \frac{1}{p_j^{\beta+1}} \le \frac{1}{(1-(K-1)p_{\min})^{\beta+1}} + \frac{K-2}{p_{\min}^{\beta+1}}.$$

Next note that

$$\frac{d}{dt}\mathrm{KL}(\bar\mu(t)\|\pi) = \langle\nabla_{\boldsymbol{\mu}}\mathrm{KL}(\bar\mu(t)\|\pi), \boldsymbol{\phi}\rangle_{\boldsymbol{\mu}} + \langle\nabla_{\mathbf{p}}\mathrm{KL}(\bar\mu(t)\|\pi), \mathbf{v}\rangle = -\alpha\mathrm{KL}(\bar\mu(t)\|\pi) \le 0.$$

Moreover, we note that $\sum p_j\mu_j(x) = \bar\mu(x)$, so $0 \le \mu_j(x) \le \frac{1}{p_{\min}}\bar\mu(x)$. Therefore by the convexity of $g(y) = y\log\frac{y}{\pi(x)}$, $-\frac{\pi(x)}{e} \le g(y) \le \max\{\frac{\bar\mu(x)}{p_{\min}}\log\frac{\bar\mu(x)}{p_{\min}\pi(x)}, 0\}$

$$-\frac{\pi(x)}{e} \le \mu_j(x)\log\frac{\mu_j(x)}{\pi(x)} \le \max\{\frac{\bar\mu(x)}{p_{\min}}\log\frac{\bar\mu(x)}{p_{\min}\pi(x)}, 0\} \le \frac{\bar\mu(x)}{p_{\min}}\log\frac{\bar\mu(x)}{p_{\min}\pi(x)} + \frac{\pi(x)}{e}.$$

$$\mu_j(x)\log\frac{\bar\mu(x)}{\pi(x)} \le \max\{\frac{\bar\mu(x)}{p_{\min}}\log\frac{\bar\mu(x)}{\pi(x)}, 0\} \le \frac{\bar\mu(x)}{p_{\min}}\log\frac{\bar\mu(x)}{\pi(x)} + \frac{\pi(x)}{p_{\min}e}.$$

Since $-\frac{\bar\mu(x)}{p_{\min}}\log\frac{\bar\mu(x)}{\pi(x)} \le \frac{\pi(x)}{ep_{\min}}$. Likewise

$$\mu_j(x)\log\frac{\bar\mu(x)}{\pi(x)} \ge \min\{\frac{\bar\mu(x)}{p_{\min}}\log\frac{\bar\mu(x)}{\pi(x)}, 0\} \ge -\frac{\pi(x)}{ep_{\min}}.$$

Assuming $|L| \le L_{\max}$, we find

$$v_1 = \tfrac{1}{K}\sum_j (L(\mu_j) - L(\mu_1)) - \tfrac{1}{K}\sum_j (\tfrac{\beta}{p_j^{\beta+1}} - \tfrac{\beta}{p_1^{\beta+1}}) + \tfrac{1}{K}\lambda\sum_j \int_{\mathbb{R}^d} (\mu_1 - \mu_j)\log\tfrac{\bar\mu}{\pi}dx$$
$$\ge -2L_{\max} - \frac{\beta}{K(1-(K-1)p_{\min})^{\beta+1}} + \frac{\beta}{Kp_{\min}^{\beta+1}} + \tfrac{1}{K}\lambda\sum_j \int_{\mathbb{R}^d} (\mu_1 - \mu_j)\log\tfrac{\bar\mu}{\pi}dx$$
$$\ge -2L_{\max} - \frac{\beta}{K(1-(K-1)p_{\min})^{\beta+1}} + \frac{\beta}{Kp_{\min}^{\beta+1}} - \tfrac{1}{K}\lambda\sum_j \int_{\mathbb{R}^d} (\frac{\bar\mu(x)}{p_{\min}}\log\tfrac{\bar\mu}{\pi} + \frac{\pi(x)}{ep_{\min}})dx$$
$$\ge -C - C(1 + KL(\bar\mu(t)\|\pi))/p_{\min} + \frac{\beta}{Kp_{\min}^{\beta+1}}.$$

Hence, if $p_{\min}$ is small enough, we have $v_1 > 0$. This implies $p_1(t) < p_{\min}$ for some $t < t_0$. This contradicts the fact that t_0 is the first time that $p_i(t) = p_{\min}$ for some $i \in [K]$.

To show $\sum_{k=1}^K p_k(t) = 1$ for any $t > 0$, we note that

$$\sum_{k=1}^K \int_0^T \tfrac{d}{dt} p_k(t) dt = \int_0^T \sum_{k=1}^K v_k(\boldsymbol{\mu}(t), \mathbf{p}(t)) dt = 0,$$

where the first equality is by the construction of v_k and the second equality is due to the projection operator P we defined. Hence,

$$\sum_{k=1} p_k(T) - \sum_{k=1}^K p_k(0) = \sum_{k=1}^K \int_0^T \tfrac{d}{dt} p_k(t) dt = 0$$

for any $T > 0$. This implies $\sum_{k=1} p_k(T) = 1$ as long as $\sum_{k=1} p_k(0) = 1$.

(ii): Note that

$$\frac{d}{dt}\mathrm{KL}(\bar{\mu}(t)\|\pi) = \langle \nabla_{\boldsymbol{\mu}} \mathrm{KL}(\bar{\mu}(t)\|\pi), \boldsymbol{\phi} \rangle_{\boldsymbol{\mu}} + \langle \nabla_{\mathbf{p}} \mathrm{KL}(\bar{\mu}(t)\|\pi), \mathbf{v} \rangle = -\alpha \mathrm{KL}(\bar{\mu}(t)\|\pi),$$

where the first equality is by chain's rule and the second one is due to the formula of $\boldsymbol{\phi}, \mathbf{v}$ in equation (25). Then, by the Grönwall's inequality, we proved statement (ii).

(iii): For abrivation, in this proof, we use $F(t)$, $\mathrm{KL}(t)$, and $\lambda(t)$ to denote $F(\boldsymbol{\mu}(t), \mathbf{p}(t))$, $\mathrm{KL}(\bar{\mu}(t)\|\pi)$, and $\lambda(\boldsymbol{\mu}(t), \mathbf{p}(t))$, repsectively. In this proof, with a little abuse of notation, we define the inner product $\langle \nabla F(\boldsymbol{\mu}, \mathbf{p}), \nabla \mathrm{KL}(\bar{\mu}\|\pi) \rangle$ as

$$\langle \nabla_{\boldsymbol{\mu}} F(\boldsymbol{\mu}, \mathbf{p}), \nabla_{\boldsymbol{\mu}} \mathrm{KL}(\bar{\mu}\|\pi)) \rangle_{\boldsymbol{\mu}} + \langle \nabla_{\mathbf{p}} F(\boldsymbol{\mu}, \mathbf{p}), \nabla_{\mathbf{p}} \mathrm{KL}(\bar{\mu}\|\pi) \rangle,$$

where

$$\langle \nabla_{\boldsymbol{\mu}} F(\boldsymbol{\mu}, \mathbf{p}), \nabla_{\boldsymbol{\mu}} \mathrm{KL}(\bar{\mu}\|\pi)) \rangle_{\boldsymbol{\mu}} = \sum_{k \in [K]} \int_{\mathbb{R}^d} \langle \nabla_{\mu_k} F(\boldsymbol{\mu}, \mathbf{p}), \nabla_{\mu_k} \mathrm{KL}(\bar{\mu}\|\pi)) \rangle d\mu_k.$$

Now we calculate $\frac{d}{dt} F(t)$ as follows.

$$\begin{aligned}
\frac{d}{dt} F(t) &= \langle \nabla F(t), (\boldsymbol{\phi}, \mathbf{v}) \rangle \\
&= -\langle \nabla F(t), \nabla F(t) + \lambda(t) \nabla \mathrm{KL}(t) \rangle + \eta \langle \nabla_{\mathbf{p}} F(t), \mathbf{1} \rangle \\
&= -\langle \nabla F(t), \nabla F(t) + \lambda(t) \nabla \mathrm{KL}(t) \rangle + \frac{1}{K} \langle \nabla_{\mathbf{p}} F(t) + \lambda(t) \nabla_{\mathbf{p}} \mathrm{KL}(t), \mathbf{1} \rangle \langle \nabla_{\mathbf{p}} F(t), \mathbf{1} \rangle \\
&= -\langle \nabla_{\boldsymbol{\mu}} F(t), \nabla_{\boldsymbol{\mu}} F(t) + \lambda(t) \nabla_{\boldsymbol{\mu}} \mathrm{KL}(t) \rangle_{\boldsymbol{\mu}(t)} \\
&\quad - (\langle \nabla_{\mathbf{p}} F(t), \nabla_{\mathbf{p}} F(t) + \lambda(t) \nabla_{\mathbf{p}} \mathrm{KL}(t) \rangle - \frac{1}{K} \langle \nabla_{\mathbf{p}} F(t) + \lambda(t) \nabla_{\mathbf{p}} \mathrm{KL}(t), \mathbf{1} \rangle \langle \nabla_{\mathbf{p}} F(t), \mathbf{1} \rangle) \\
&= -(\langle \nabla_{\boldsymbol{\mu}} F(t), \nabla_{\boldsymbol{\mu}} F(t) + \lambda(t) \nabla_{\boldsymbol{\mu}} \mathrm{KL}(t) \rangle_{\boldsymbol{\mu}(t)} + \langle P \nabla_{\mathbf{p}} F(t), P(\nabla_{\mathbf{p}} F(t) + \lambda(t) \nabla_{\mathbf{p}} \mathrm{KL}(t)) \rangle) \\
&= -\langle \nabla F(t), \nabla F(t) + \lambda(t) \nabla \mathrm{KL}(t) \rangle_{\boldsymbol{\mu}(t), P} \\
&= -\|\nabla F(t) + \lambda(t) \nabla \mathrm{KL}(t)\|^2_{\boldsymbol{\mu}(t), P} + \lambda(t) \langle \nabla \mathrm{KL}(t), \nabla F(t) + \lambda(t) \nabla \mathrm{KL}(t) \rangle_{\boldsymbol{\mu}(t), P}.
\end{aligned}$$

By inserting the formula of $\lambda(t)$ given in equation (25), we observe that

$$\lambda(t) \langle \nabla \mathrm{KL}(t), \nabla F(t) + \lambda(t) \nabla \mathrm{KL}(t) \rangle_{\boldsymbol{\mu}(t), P} = \alpha \lambda(t) \mathrm{KL}(t).$$

Hence, we have

$$\frac{d}{dt} F(t) = -\|\nabla F(t) + \lambda(t) \nabla \mathrm{KL}(t)\|^2_{\boldsymbol{\mu}(t), P} + \alpha \lambda(t) \mathrm{KL}(t).$$

By the fundamental theorem of calculus,

$$\int_0^T \|\nabla F(t) + \lambda(t) \nabla \mathrm{KL}(t)\|^2_{\boldsymbol{\mu}(t), P} dt = F(0) - F(T) + \int_0^T \alpha \lambda(t) \mathrm{KL}(t) dt.$$

To bound $\int_0^T \|\nabla F(t) + \lambda(t)\nabla \mathrm{KL}(t)\|^2_{\boldsymbol{\mu}(t),P} dt$, we show the following bound for $\lambda(t)\mathrm{KL}(t)$:

$$\begin{aligned}
\lambda(t)\mathrm{KL}(t) &= -\frac{\langle \nabla F(t), \nabla \mathrm{KL}(t)\rangle_{\boldsymbol{\mu}(t),P}}{\|\nabla \mathrm{KL}(t)\|^2_{\boldsymbol{\mu}(t),P}}\mathrm{KL}(t) + \frac{\alpha \mathrm{KL}(t)^2}{\|\nabla \mathrm{KL}(t)\|^2_{\boldsymbol{\mu}(t),P}} \\
&\le \frac{|\langle \nabla F(t), \nabla \mathrm{KL}(t)\rangle_{\boldsymbol{\mu}(t),P}|}{\|\nabla \mathrm{KL}(t)\|^2_{\boldsymbol{\mu}(t),P}}\mathrm{KL}(t) + \frac{\alpha \mathrm{KL}(t)^2}{\|\nabla \mathrm{KL}(t)\|^2_{\boldsymbol{\mu}(t),P}} \\
&\le \frac{\|\nabla F(t)\|_{\boldsymbol{\mu}(t),P}}{\|\nabla \mathrm{KL}(t)\|_{\boldsymbol{\mu}(t),P}}\mathrm{KL}(t) + \frac{\alpha \mathrm{KL}(t)^2}{\|\nabla \mathrm{KL}(t)\|^2_{\boldsymbol{\mu}(t),P}} \\
&\le \frac{\|\nabla F(t)\|_{\boldsymbol{\mu}(t),P}}{\|\nabla_{\boldsymbol{\mu}} \mathrm{KL}(t)\|_{\boldsymbol{\mu}(t)}}\mathrm{KL}(t) + \frac{\alpha \mathrm{KL}(t)^2}{\|\nabla_{\boldsymbol{\mu}} \mathrm{KL}(t)\|^2_{\boldsymbol{\mu}(t)}} \\
&\le \frac{1}{\underline{p}\sqrt{\kappa}}\sqrt{\mathrm{KL}(t)}\|\nabla F(t)\|_{\boldsymbol{\mu}(t),P} + \frac{\alpha}{\underline{p}^2\kappa}\mathrm{KL}(t),
\end{aligned}$$

where the third line is due to Cauchy-Schwatz inequality, the fourth line is because $\|\nabla \mathrm{KL}(t)\|_{\boldsymbol{\mu}(t),P} \ge \|\nabla_{\boldsymbol{\mu}} \mathrm{KL}(t)\|_{\boldsymbol{\mu}(t)}$, and the last line is due to statement (v) in Lemma A.5 and statement (i) in Theorem 3 (i.e., $p_k(t) \ge \underline{p}\ \forall k \in [K], t \ge 0$).

By part (ii), $\mathrm{KL}(\bar{\mu}(t)\|\pi) = e^{-\alpha t}\mathrm{KL}(\bar{\mu}(0)\|\pi)$. Hence,

$$\lambda(t)\mathrm{KL}(t) \le \frac{1}{\underline{p}\sqrt{\kappa}}\sqrt{\mathrm{KL}(0)}\|\nabla F(t)\|_{\boldsymbol{\mu}(t),P} e^{-\frac{1}{2}\alpha t} + \frac{\alpha}{\underline{p}^2\kappa}\mathrm{KL}(0)e^{-\alpha t}.$$

By Lemma A.6, we have

$$\lambda(t)\mathrm{KL}(t) \le \frac{1}{\underline{p}\sqrt{\kappa}}\sqrt{\mathrm{KL}(0)}2\sqrt{L^2_{\max}K^2 + K(\ell_{\max} + \tfrac{\theta\beta}{\underline{p}^{\beta+1}})^2} e^{-\frac{1}{2}\alpha t} + \frac{\alpha}{\underline{p}^2\kappa}\mathrm{KL}(0)e^{-\alpha t}.$$

This implies

$$\int_0^T \lambda(t)\mathrm{KL}(t)dt \le \frac{4}{\alpha\underline{p}\sqrt{\kappa}}\sqrt{\mathrm{KL}(0)}\sqrt{L^2_{\max}K^2 + K(\ell_{\max} + \tfrac{\theta\beta}{\underline{p}^{\beta+1}})^2} + \frac{1}{\underline{p}\kappa}\mathrm{KL}(0).$$

Therefore,

$$\int_0^T \|\nabla F(t)+\lambda(t)\nabla \mathrm{KL}(t)\|^2_{\boldsymbol{\mu}(t),P} dt \le F(0)+\ell_{\max}+\frac{4}{\alpha\underline{p}\sqrt{\kappa}}\sqrt{\mathrm{KL}(0)}\sqrt{L^2_{\max}K^2 + K(\ell_{\max} + \tfrac{\theta\beta}{\underline{p}^{\beta+1}})^2}+\frac{1}{\underline{p}\kappa}\mathrm{KL}(0).$$

Note that the right-hand side is a constant that does not depend on T. Denote the upper bound by C. This implies

$$\min_{t\le T} \|\nabla F(t) + \lambda(t)\nabla \mathrm{KL}(t)\|^2_{\boldsymbol{\mu}(t),P} \le \frac{C}{T}.$$

Finally, we will show our statement (iii) as follows. Recall $\|\nabla F(\boldsymbol{\mu}^*, \mathbf{p}^*)\|_{\mathcal{TP}_\pi(\boldsymbol{\mu}^*,\mathbf{p}^*)}$ is defined as

$$\|\nabla F(\boldsymbol{\mu}(t), \mathbf{p}(t))\|_{\mathcal{TP}_\pi(\boldsymbol{\mu}(t),\mathbf{p}(t))} = \sup_{(\boldsymbol{\phi},\mathbf{v})\in\mathcal{TP}_\pi(\boldsymbol{\mu}(t),\mathbf{p}(t))} \frac{\langle \nabla_{\boldsymbol{\mu}} F(t), \boldsymbol{\phi}\rangle_{\boldsymbol{\mu}} + \langle \nabla_{\mathbf{p}} F(t), \mathbf{v}\rangle}{\|\boldsymbol{\phi}\|^2_{\boldsymbol{\mu}} + \|\mathbf{v}\|^2},$$

and the vector $\nabla F(t) + \lambda(t)\nabla \mathrm{KL}(t)$ is given by

$$\begin{aligned}
&(p_1(\nabla L(\mu_1(t)) + \lambda(s_{\bar{\mu}(t)} - s_\pi), \dots, p_K(\nabla L(\mu_K(t)) + \lambda(s_{\bar{\mu}(t)} - s_\pi), \\
&\nabla_{\mathbf{p}} F(t)_1 + \lambda(t)\langle \log\frac{\bar{\mu}(t)}{\pi} + 1, \mu_1(t)\rangle, \dots, \nabla_{\mathbf{p}} F(t)_K + \lambda(t)\langle \log\frac{\bar{\mu}(t)}{\pi} + 1, \mu_K(t)\rangle).
\end{aligned}$$

For any pair $(\boldsymbol{\phi}, \mathbf{v}) \in \mathcal{TP}_\pi(\boldsymbol{\mu}, \mathbf{p})$, by Lemma A.3, we have

$$\sum_{k\in[K]} p_k \nabla\cdot(\mu_k\phi_k) = \sum_{k\in[K]} v_k\mu_k, \quad \sum_{k\in[K]} v_k = 0.$$

Thus,

$$\langle \nabla_{\boldsymbol{\mu}} F(t), \boldsymbol{\phi} \rangle_{\boldsymbol{\mu}} + \langle \nabla_{\mathbf{p}} F(t), \mathbf{v} \rangle = \langle \nabla_{\boldsymbol{\mu}} F(t) + \lambda(t) \nabla_{\boldsymbol{\mu}} \mathrm{KL}(t), \boldsymbol{\phi} \rangle_{\boldsymbol{\mu}(t)} + \langle \nabla_{\mathbf{p}} F(t) + \lambda(t) \nabla_{\mathbf{p}} \mathrm{KL}(t), \mathbf{v} \rangle$$
$$- (\langle \lambda(t) \nabla_{\boldsymbol{\mu}} \mathrm{KL}(t), \boldsymbol{\phi} \rangle_{\boldsymbol{\mu}(t)} + \langle \lambda(t) \nabla_{\mathbf{p}} \mathrm{KL}(t), \mathbf{v} \rangle).$$

Note that

$$\langle \nabla_{\boldsymbol{\mu}} F(t) + \lambda(t) \nabla_{\boldsymbol{\mu}} \mathrm{KL}(t), \boldsymbol{\phi} \rangle_{\boldsymbol{\mu}(t)} + \langle \nabla_{\mathbf{p}} F(t) + \lambda(t) \nabla_{\mathbf{p}} \mathrm{KL}(t), \mathbf{v} \rangle = \langle \nabla F(t) + \lambda(t) \nabla \mathrm{KL}(t), (\boldsymbol{\phi}, \mathbf{v}) \rangle_{\boldsymbol{\mu}(t), P}.$$

To see this,

$$\langle P \nabla_{\mathbf{p}} F(t) + \lambda(t) \nabla_{\mathbf{p}} \mathrm{KL}(t), P\mathbf{v} \rangle = \langle \nabla_{\mathbf{p}} F(t) + \lambda(t) \nabla_{\mathbf{p}} \mathrm{KL}(t), \mathbf{v} \rangle - \frac{1}{K} \langle \nabla_{\mathbf{p}} F(t) + \lambda(t) \nabla_{\mathbf{p}} \mathrm{KL}(t), \mathbf{1} \rangle \langle \mathbf{v}, \mathbf{1} \rangle$$
$$= \langle \nabla_{\mathbf{p}} F(t) + \lambda(t) \nabla_{\mathbf{p}} \mathrm{KL}(t), \mathbf{v} \rangle,$$

because $\langle \mathbf{v}, \mathbf{1} \rangle = 0$. Hence,

$$\begin{aligned} &\langle \nabla_{\boldsymbol{\mu}} F(t) + \lambda(t) \nabla_{\boldsymbol{\mu}} \mathrm{KL}(t), \boldsymbol{\phi} \rangle_{\boldsymbol{\mu}(t)} + \langle \nabla_{\mathbf{p}} F(t) + \lambda(t) \nabla_{\mathbf{p}} \mathrm{KL}(t), \mathbf{v} \rangle \\ &= \langle \nabla_{\boldsymbol{\mu}} F(t) + \lambda(t) \nabla_{\boldsymbol{\mu}} \mathrm{KL}(t), \boldsymbol{\phi} \rangle_{\boldsymbol{\mu}(t)} + \langle P \nabla_{\mathbf{p}} F(t) + \lambda(t) \nabla_{\mathbf{p}} \mathrm{KL}(t), P\mathbf{v} \rangle \\ &= \langle \nabla F(t) + \lambda(t) \nabla \mathrm{KL}(t), (\boldsymbol{\phi}, \mathbf{v}) \rangle_{\boldsymbol{\mu}(t), P}. \end{aligned}$$

Moreover,

$$\langle \lambda(t) \nabla_{\boldsymbol{\mu}} \mathrm{KL}(t), \boldsymbol{\phi} \rangle_{\boldsymbol{\mu}(t)} + \langle \lambda(t) \nabla_{\mathbf{p}} \mathrm{KL}(t), \mathbf{v} \rangle = 0.$$

To see this,

$$\langle \nabla_{\boldsymbol{\mu}} \mathrm{KL}(t), \boldsymbol{\phi} \rangle_{\boldsymbol{\mu}(t)} = \sum_{k \in [K]} \langle \nabla_{\mu_k} \mathrm{KL}(t), \phi_k \rangle_{\mu_k(t)} = \sum_{k \in [K]} p_k \langle s_{\bar{\mu}(t)} - s_\pi, \phi_k \rangle_{\mu_k(t)}$$
$$= - \sum_{k \in [K]} v_k \langle \log \frac{\bar{\mu}(t)}{\pi} + 1, \mu_k(t) \rangle = - \langle \nabla_{\mathbf{p}} \mathrm{KL}(t), \mathbf{v} \rangle,$$

where the first equality is the definition for the inner product, the second equality is due to the formula of $\nabla_{\boldsymbol{\mu}} \mathrm{KL}(t)$, the third equality is because $\sum_{k \in [K]} p_k(t) \nabla \cdot (\mu_k(t) \phi_k) = \sum_{k \in [K]} v_k \mu_k(t)$, and the last equality is due to the definition of inner product and the formula of $\nabla_{\mathbf{p}} \mathrm{KL}(t)$.

To conclude, we get

$$\langle \nabla_{\boldsymbol{\mu}} F(t), \boldsymbol{\phi} \rangle_{\boldsymbol{\mu}} + \langle \nabla_{\mathbf{p}} F(t), \mathbf{v} \rangle = \langle \nabla F(t) + \lambda(t) \nabla \mathrm{KL}(t), (\boldsymbol{\phi}, \mathbf{v}) \rangle_{\boldsymbol{\mu}(t), P}$$
$$\leq \| \nabla F(t) + \lambda(t) \nabla \mathrm{KL}(t) \|_{\boldsymbol{\mu}(t), P} \| (\boldsymbol{\phi}, \mathbf{v}) \|_{\boldsymbol{\mu}(t), P},$$

where the second inequality is by Cauchy-Schwartz inequality. Also, note that

$$\| (\boldsymbol{\phi}, \mathbf{v}) \|_{\boldsymbol{\mu}(t), P} = \| \boldsymbol{\phi} \|_{\boldsymbol{\mu}}^2 + \| P\mathbf{v} \|^2 = \| \boldsymbol{\phi} \|_{\boldsymbol{\mu}}^2 + \| \mathbf{v} \|^2,$$

since $P\mathbf{v} = \mathbf{v}$ by the fact $\sum_{k \in [K]} v_k = 0$.

Hence, for any pair $(\boldsymbol{\phi}, \mathbf{v}) \in \mathcal{TP}_\pi(\boldsymbol{\mu}^*, \mathbf{p}^*)$,

$$\frac{\langle \nabla_{\boldsymbol{\mu}} F(t), \boldsymbol{\phi} \rangle_{\boldsymbol{\mu}} + \langle \nabla_{\mathbf{p}} F(t), \mathbf{v} \rangle}{\| \boldsymbol{\phi} \|_{\boldsymbol{\mu}}^2 + \| \mathbf{v} \|^2} \leq \| \nabla F(t) + \lambda(t) \nabla \mathrm{KL}(t) \|_{\boldsymbol{\mu}(t), P},$$

which implies

$$\| \nabla F(\boldsymbol{\mu}(t), \mathbf{p}(t)) \|_{\mathcal{TP}_\pi(\boldsymbol{\mu}(t), \mathbf{p}(t))} \leq \| \nabla F(t) + \lambda(t) \nabla \mathrm{KL}(t) \|_{\boldsymbol{\mu}(t), P},$$

and thus,

$$\min_{t \leq T} \| \nabla F(\boldsymbol{\mu}(t), \mathbf{p}(t)) \|_{\mathcal{TP}_\pi(\boldsymbol{\mu}(t), \mathbf{p}(t))} \leq \min_{t \leq T} \| \nabla F(t) + \lambda(t) \nabla \mathrm{KL}(t) \|_{\boldsymbol{\mu}(t), P} \leq \frac{C}{\sqrt{T}}. \qquad \square$$