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 $\mathbf{p} = (p_1, \ldots, p_K) \in \mathbb{R}^{\otimes K}$ 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

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

[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.

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

$$L(\mu) = \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} \langle \mathbf{x} - \mathbf{y}, W(\mathbf{x} - \mathbf{y}) \rangle d\mu(\mathbf{x}) d\mu(\mathbf{y}), \tag{4}$$

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

References

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.

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

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.

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

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.

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.

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