Supplementary material

1. Pointers of the proofs and experiments

Table 1

Paper	Supplementary material	Position of proof
Proposition 3.1	PROPOSITION 7	A.10
Theorem 3.2	THEOREM 3	
Theorem 3.3	THEOREM 4	A.14
Proposition 4.1	PROPOSITION 8	A.18
Theorem 4.2	THEOREM 7	
Theorem 4.3	THEOREM 8	A.19
Theorem 5.1	THEOREM 10	
$l_K \to l$ convergence rate	LEMMA 5	A.7

2. List of Notation

3. Inverse Multiobjective Optimization

3.1. Decision Making Problem with Multiple Objectives

Consider the following decision making problem with p (≥ 2) objective functions parameterized by θ:

$$\begin{array}{l} \min\limits_{\mathbf{x}\in\mathbb{R}^n} \{f_1(\mathbf{x},\theta), f_2(\mathbf{x},\theta), \ldots, f_p(\mathbf{x},\theta)\} \\ s.t.\ \ \mathbf{x} \in X(\theta). \end{array} \qquad \text{DMP}$$

ASSUMPTION 1. *Θ is a convex set. For each $\theta \in \Theta$, $\mathbf{f}(\mathbf{x},\theta)$ is convex in $\mathbf{x}$, i.e., $f_l(\mathbf{x})$ is convex on $X(\theta)$ for all $l \in [p]$. Here, $X(\theta)$ is also a convex set for each $\theta \in \Theta$.*

A common way to derive a Pareto optimal solution is to solve a problem with a single objective function constructed by the weighted sum of original functions, i.e., to solve the following problem Gass and Saaty (1955).

$$\begin{array}{l} \min\ w^T \mathbf{f}(\mathbf{x},\theta) \\ s.t.\ \ \mathbf{x} \in X(\theta) \end{array} \qquad \text{WP}$$

where $w = (w^1, \ldots, w^p)^T$ is the nonnegative weight vector in the $(p-1)$-simplex $\mathscr{W}_p \equiv \{w \in \mathbb{R}^p_+ : \mathbf{1}^T w = 1\}$.

PROPOSITION 1. *Let $\mathbf{x} \in S(w,\theta)$ be an optimal solution of WP. The following statements hold.*
(a) If $w \in \mathscr{W}_p^+$, then $\mathbf{x} \in X_P(\theta)$.
(b) If $\mathbf{x}$ is the unique optimal solution of WP, then $\mathbf{x} \in X_P(\theta)$.

PROPOSITION 2. *Given that DMP is convex and $\mathbf{x} \in X_P(\theta)$, there exists a weight vector $w \in \mathscr{W}_p$ such that $\mathbf{x}$ is an optimal solution to WP, i.e., $\mathbf{x} \in S(w,\theta)$.*

COROLLARY 1. *For a convex DMP,*

$$\bigcup_{w \in \mathscr{W}_p^+} S(w,\theta) \subseteq X_P(\theta) \subseteq \bigcup_{w \in \mathscr{W}_p} S(w,\theta).$$

3.2. Models for IMOP as an Unsupervised Learning Task

$$l(\mathbf{y},\theta) = \min_{\mathbf{x} \in X_P(\theta)} \|\mathbf{y}-\mathbf{x}\|_2^2, \qquad \text{loss function}$$

where $X_P(\theta)$ is the Pareto optimal set of DMP for a given θ.

Using (loss function), our inverse multiobjective optimization problem can be formulated as follows

$$\min_{\theta \in \Theta} M(\theta) \equiv \mathbb{E}\Big(l(\mathbf{y},\theta)\Big), \qquad \text{IMOP}$$

where $M(\theta)$ is also called the risk of the loss function $l(\mathbf{y},\theta)$ for the hypothesis θ.

Practically, θ can not be learned by directly solving IMOP as $\mathbb{P}_{\mathbf{y}}$ is not known a priori. Given available observations $\{\mathbf{y}_i\}_{i\in[N]}$, it is often the case that θ will be inferred through solving the following empirical risk minimizing problem:

$$\min_{\theta \in \Theta} M^N(\theta) \equiv \tfrac{1}{N} \sum_{i \in [N]} l(\mathbf{y}_i,\theta). \qquad \text{IMOP-EMP}$$

Nevertheless, one remaining challenge of using (loss function) is that there is no general approach to comprehensively and explicitly characterize the Pareto optimal set $X_P(\theta)$. One way is to introduce weight variable representing the appropriate weight and convert the (loss function) into

$$\min_{w \in \mathscr{W}_p, \mathbf{x} \in S(w,\theta)} \|\mathbf{y}-\mathbf{x}\|_2^2.$$

However, this approach might not be suitable for a data-driven study, since it results in a drastically complicated model, where every single observation requires one weight variable and the nonlinear term between it and θ is heavily involved. On the contrary, according to Corollary 1 and its following remarks, we adopt a sampling approach to generate $w_k \in \mathscr{W}_p$ for each $k \in [K]$ and approximate $X_P(\theta)$ as the union of their $S(w_k,\theta)$s. Then, by utilizing binary variables that

select an appropriate Pareto optimal solution from this union, the loss function is converted into the following *sampling based loss problem.*

$$\begin{aligned} l_K(\mathbf{y},\theta) = \quad & \min_{\mathbf{x}_k, z_k \in \{0,1\}} \ \|\mathbf{y} - \sum_{k\in[K]} z_k \mathbf{x}_k\|_2^2 \\ & \text{s.t.} \sum_{k\in[K]} z_k = 1, \ \mathbf{x}_k \in S(w_k,\theta). \end{aligned} \qquad \text{surrogate loss function}$$

As previously mentioned, we indeed do not have the explicit representation of $X_P(\theta)$. Through the sampling approach described in the last subsection, variants of IMOP using (surrogate loss function) can be easily defined. The following one is to reformulate IMOP with weight samples, which helps us perform theoretical analysis of the reformulation of IMOP-EMP.

$$\min_{\theta\in\Theta} M_K(\theta) \equiv \mathbb{E}\Big(l_K(\mathbf{y},\theta) \Big). \qquad \text{IMOP-WS}$$

Next, we provide the reformulation of IMOP-EMP with the (surrogate loss function). As it serves as the primary model for analysis and computation, we present its comprehensive form to facilitate our discussion and understanding.

$$\begin{aligned} \min_{\theta\in\Theta} \ & M_K^N(\theta) \equiv \tfrac{1}{N} \sum_{i\in[N]} \|\mathbf{y}_i - \sum_{k\in[K]} z_{ik}\mathbf{x}_k\|_2^2 \\ \text{s.t.} \ & \mathbf{x}_k \in S(w_k,\theta), && \forall k \in [K], \\ & \sum_{k\in[K]} z_{ik} = 1, && \forall i \in [N], \\ & z_{ik} \in \{0,1\}, && \forall i \in [N], \ k \in [K]. \end{aligned} \qquad \text{IMOP-EMP-WS}$$

4. Estimators' Risk Consistency and Generalization Bound

4.1. Risk Consistency of IMOP-EMP-WS

ASSUMPTION 2. *(i) The parameter set Θ is compact.*

(ii) For each $\theta \in \Theta$, $X(\theta)$ is compact, and has a nonempty relatively interior. Also, $X(\theta)$ is uniformly bounded. Namely, there exists $B > 0$ such that $\|\mathbf{x}\|_2 \le B$ for all $\mathbf{x} \in X(\theta)$ and $\theta \in \Theta$.

(iii) Functions $\mathbf{f}(\mathbf{x},\theta)$ and $\mathbf{g}(\mathbf{x},\theta)$ are continuous on $\mathbb{R}^n \times \Theta$.

(iv) $\mathbb{E}[\mathbf{y}^T\mathbf{y}] < +\infty$.

LEMMA 1. *Suppose Assumptions 1 - 2 hold. $X(\theta)$ is continuous on Θ.*

The continuity of $X(\theta)$ follows from its lower semicontinuity (l.s.c.) and upper semicontinuity (u.s.c.), both of which can be derived by using Hogan (1973) under our assumptions.

LEMMA 2. *Suppose Assumptions 1 - 2 hold. If $\mathbf{f}(\mathbf{x},\theta)$ is strictly convex in $\mathbf{x}$ for each $\theta \in \Theta$, then $X_P(\theta)$ is continuous on Θ.*

PROPOSITION 3 **(ULLN for $M^N(\theta)$ in N)**. *Under the same conditions of Lemma 2, $M^N(\theta)$ uniformly converges to $M(\theta)$ in N. That is,*

$$\sup_{\theta\in\Theta}|M^N(\theta)-M(\theta)|\xrightarrow{p}0.$$

PROPOSITION 4 **(ULLN for $M_K^N(\theta)$ in N)**. *Under the same conditions of Lemma 2, $M_K^N(\theta)$ uniformly converges to $M_K(\theta)$ in N. That is, $\forall K$,*

$$\sup_{\theta\in\Theta}|M_K^N(\theta)-M_K(\theta)|\xrightarrow{p}0.$$

Throughout the paper, we use $K_2\geq K_1$ to denote the set of weights $\{w_k\}_{k\in[K_1]}\subseteq\{w_k\}_{k\in[K_2]}$, and $K_2>K_1$ to denote the set of weights $\{w_k\}_{k\in[K_1]}\subsetneq\{w_k\}_{k\in[K_2]}$. Then, we depict the monotonicity of $\{M_K(\theta)\}$ and $\{M_K^N(\theta)\}$ in K for each $\theta\in\Theta$ in the following lemma.

LEMMA 3 **(Monotonicity of $\{M_K(\theta)\}$ and $\{M_K^N(\theta)\}$ in K)**. *We have the following:*
(a) *The sequence $\{M_K(\theta)\}$ is monotone decreasing in K for all $\theta\in\Theta$. Moreover, $\{M_K(\hat{\theta}_K)\}$ is monotone decreasing in K. Specially, $M_K(\hat{\theta}_K)\geq M(\theta^*)$.*
(b) *Given any $\{\mathbf{y}_i\}_{i\in[N]}$, the sequence $\{M_K^N(\theta)\}$ is monotone decreasing in K for all $\theta\in\Theta$. Moreover, $\{M_K^N(\hat{\theta}_K^N)\}$ is monotone decreasing in K. Specially, $M_K^N(\hat{\theta}_K^N)\geq M^N(\hat{\theta}^N)$.*

LEMMA 4. *Suppose Assumptions 1 - 2 hold. Suppose also that $\mathbf{f}(\mathbf{x},\theta)$ is strongly convex in $\mathbf{x}$ for each $\theta\in\Theta$, that is, $\forall l\in[p]$, $\exists\lambda_l>0$, $\forall\mathbf{x},\mathbf{y}\in\mathbb{R}^n$,*

$$f_l(\mathbf{y},\theta)\geq f_l(\mathbf{x},\theta)+\nabla f_l(\mathbf{x},\theta)^T(\mathbf{y}-\mathbf{x})+\frac{\lambda_l}{2}\|\mathbf{y}-\mathbf{x}\|_2^2.$$

Then, $\forall\theta\in\Theta$, $\forall w,w_0\in\mathscr{W}_p$,

$$\|S(w,\theta)-S(w_0,\theta)\|_2\leq\frac{2L}{\lambda}\|w-w_0\|_2,$$

where $L=\sqrt{p}\cdot\max_{l\in[p],\theta\in\Theta,\mathbf{x}\in X(\theta)}|f_l(\mathbf{x},\theta)|$ is a finite number, and $\lambda=\min_{l\in[p]}\{\lambda_l\}$.

LEMMA 5. *Under Assumptions 1 - 2, we have that $\forall\mathbf{y}\in\mathcal{Y},\forall\theta\in\Theta$,*

$$0\leq l_K(\mathbf{y},\theta)-l(\mathbf{y},\theta)\leq\frac{4(B+R)\zeta}{\lambda}\cdot\frac{\sqrt{2p}}{\Lambda-1},$$

where

$$K=\frac{(\Lambda+p-2)!}{(\Lambda-1)!(p-1)!},\zeta=\max_{l\in[p],\mathbf{x}\in X(\theta),\theta\in\Theta}|f_l(\mathbf{x},\theta)|.$$

Furthermore,

$$0\leq l_K(\mathbf{y},\theta)-l(\mathbf{y},\theta)\leq\frac{16e(B+R)\zeta}{\lambda}\cdot\frac{1}{K^{\frac{1}{p-1}}}.$$

PROPOSITION 5 **(Uniform convergence of $M_K(\theta)$ in K)**. *Under the same conditions of Lemma 4, $M_K(\theta)$ uniformly converges to $M(\theta)$ in K for $\theta \in \Theta$. That is,* $\sup_{\theta\in\Theta} |M_K(\theta) - M(\theta)| \longrightarrow 0$.

Next, we present a very mild assumption to bound random observations.

ASSUMPTION 3. *The support $\mathcal{Y}$ of the distribution $\mathbf{y}$ is contained within a ball of radius R almost surely, where $R < \infty$. That is,* $\mathbb{P}(\|\mathbf{y}\|_2 \leq R) = 1$.

PROPOSITION 6 **(Uniform convergence of $M_K^N(\theta)$ in K)**. *Suppose Assumptions 1 - 3 hold. If $\mathbf{f}(\mathbf{x}, \theta)$ is strongly convex in $\mathbf{x}$ for each $\theta \in \Theta$, then $M_K^N(\theta)$ uniformly converges to $M(\theta)$ in K for $\theta \in \Theta$ and N. That is, $\forall N$,* $\sup_{\theta\in\Theta} |M_K^N(\theta) - M^N(\theta)| \xrightarrow{p} 0$.

DEFINITION 1 (DOUBLE-INDEX CONVERGENCE). Let $\{X_{mn}\}$ be an array of double-index random variables. Let X be a random variable. If $\forall \delta > 0, \forall \epsilon > 0$, $\exists N$, s.t. $\forall m, n \geq N$, $\mathbb{P}(|X_{mn} - X| > \epsilon) < \delta$. Then X_{mn} is said to converge in probability to X (denoted by $X_{mn} \xrightarrow{P} X$).

PROPOSITION 7 **(Uniform convergence of $M_K^N(\theta)$ in N and K)**. *Under the same conditions of Proposition 6, $M_K^N(\theta)$ uniformly converges to $M(\theta)$ in N and K for all $\theta \in \Theta$. That is,*

$$\sup_{\theta\in\Theta} |M_K^N(\theta) - M(\theta)| \xrightarrow{P} 0.$$

We next show the risk consistency of the estimators. We denote Θ^* the set of parameters that minimizes the risk and refer to it as the optimal set. Namely, $\Theta^* = \{\theta^* \in \Theta : M(\theta^*) = \min_{\theta\in\Theta} M(\theta)\}$.

THEOREM 1 **(Consistency of IMOP-EMP)**. *Suppose Assumptions 1 - 2 hold. If $\mathbf{f}(\mathbf{x}, \theta)$ is strictly convex in $\mathbf{x}$ for each $\theta \in \Theta$, then $M(\hat{\theta}^N) \xrightarrow{p} M(\theta^*)$.*

THEOREM 2 **(Consistency of IMOP-WS)**. *Suppose Assumptions 1 - 2 hold. If $\mathbf{f}(\mathbf{x}, \theta)$ is strongly convex in $\mathbf{x}$ for each $\theta \in \Theta$, then $M(\hat{\theta}_K) \xrightarrow{P} M(\theta^*)$.*

THEOREM 3 **(Consistency of IMOP-EMP-WS)**. *Suppose Assumptions 1 - 3 hold. If $\mathbf{f}(\mathbf{x}, \theta)$ is strongly convex in $\mathbf{x}$ for each $\theta \in \Theta$, then $M(\hat{\theta}_K^N) \xrightarrow{P} M(\theta^*)$.*

4.2. Generalization Bound of IMOP-EMP-WS

DEFINITION 2 (RADEMACHER RANDOM VARIABLES). Random variables $\sigma_1, \ldots, \sigma_N$ are called *Rademacher random variables* if they are independent, identically distributed and $\mathbb{P}(\sigma_i = 1) = \mathbb{P}(\sigma_i = -1) = 1/2$ for $i \in [N]$.

Let $\mathcal{F}$ be a class of functions mapping from Z to $[a, b]$, and $Z_1, \ldots, Z_N$ be independent and identically distributed (i.i.d.) random variables on Z.

DEFINITION 3. The Rademacher complexity of $\mathcal{F}$ is

$$Rad_N(\mathcal{F}) = \frac{1}{N}\mathbb{E}\left[\sup_{f\in\mathcal{F}} \sum_{i\in[N]} \sigma_i f(Z_i)\right],$$

where the expectation is taken over σ and $Z_1, \ldots, Z_N$.

LEMMA 6. *Let $\mathcal{F}$ be a class of functions mapping from Z to $[a,b]$. Let $Z_1,\cdots,Z_N$ be i.i.d. random variables on Z. Then, for any $0<\delta<1$, with probability at least $1-\delta$, every $f\in\mathcal{F}$ satisfies*

$$\mathbb{E}[f(Z)] \le \frac{1}{N}\sum_{i\in[N]} f(Z_i) + 2Rad_N(\mathcal{F}) + (b-a)\sqrt{\frac{log(1/\delta)}{2N}}.$$

Given K and θ, we define a function $f(\cdot,\theta)$ by $f(\mathbf{y},\theta)=\min_{k\in[K]}\|\mathbf{y}-\mathbf{x}_k\|_2^2$, where $\mathbf{x}_k\in S(w_k,\theta)$ for all $k\in[K]$. Now consider the class of functions $\mathcal{F}=\{f(\cdot,\theta):\theta\in\Theta\}$. To bound the risk $\mathbb{E}[f(\mathbf{y},\theta)]$ using Lemma 6, we need to either compute the vaule of $Rad_N(\mathcal{F})$ or find an upper bound of it. Note that the computation of $Rad_N(\mathcal{F})$ involves solving a difficult optimization problem over $\mathcal{F}$. In contrast, obtaining a bound of $Rad_N(\mathcal{F})$ is relatively easier. Therefore, we seek to bound $Rad_N(\mathcal{F})$ in the following lemma.

LEMMA 7. *The Rademacher complexity of $\mathcal{F}$ is bounded by a function of sample size N,*

$$Rad_N\big(\mathcal{F}\big) \le \frac{K}{\sqrt{N}}\bigg(B^2+2BR\bigg).$$

THEOREM 4 **(Generalization bound)**. *Suppose Assumptions 1 - 3 hold. For any $0<\delta<1$, with probability at least $1-\delta$ with respect to the observations,*

$$M_K(\hat{\theta}_K^N) \le M_K^N(\hat{\theta}_K^N) + \frac{1}{\sqrt{N}}\bigg(2K(B^2+2BR)+(B+R)^2\sqrt{\log(1/\delta)/2}\bigg) \text{ for each } K.$$

5. Identifiability Analysis for IMOP

DEFINITION 4 (HAUSDORFF SEMI-DISTANCE). Let X and Y be two nonempty set. We define their *Hausdorff semi-distance* by

$$d_{sH}(X,Y) = \sup_{x\in X}\inf_{y\in Y} d(x,y).$$

LEMMA 8. *$d_{sH}(X,Y)=0$ if and only if $X\subseteq Y$.*

DEFINITION 5 (IDENTIFIABILITY). A DMP is said to be identifiable at $\theta\in\Theta$, if for all $\theta'\in\Theta\setminus\theta$,

$$d_{sH}(X_P(\theta),X_P(\theta'))>0.$$

5.0.1. Estimation Consistency of IMOP under Identifiability Let θ_0 be the underlying parameter of the DMP that generates the data. If DMP is identifiable at θ_0, and the data is not corrupted by noise, then $M(\theta)$ achieves its minimum uniquely at θ_0. We are now ready to state our result regarding the estimation consistency of $\hat{\theta}_K^N$.

THEOREM 5 **(Consistency of $\hat{\theta}_K^N$)**. *Suppose Assumptions 1 - 2 hold. Suppose also that $\mathbf{f}(\mathbf{x},\theta)$ is strongly convex in $\mathbf{x}$ for each $\theta\in\Theta$, and that $\forall\mathbf{y}\in\mathcal{Y},\mathbf{y}\in X_P(\theta_0)$. That is, there is no noise in the data. If DMP is identifiable at $\theta_0\in\Theta$, then $\hat{\theta}_K^N \xrightarrow{P} \theta_0$.*

DEFINITION 6 (BIJECTIVITY). A DMP is said to be bijective at $\theta \in \Theta$ if $X_P(\theta) = \bigcup_{w \in \mathscr{W}_p} S(w, \theta)$, $S(w, \theta)$ is single valued for w almost surely, and $\forall w_1, w_2 \in \mathscr{W}_p$, $w_1 \neq w_2$ implies $S(w_1, \theta) \neq S(w_2, \theta)$.

With a slight abuse of notation, we let $w_{\mathbf{y}}$ be the true weight for $\mathbf{y}$, and $w_{\mathbf{y}}^{NK}$ be the estimated weight for $\mathbf{y}$ given $\hat{\theta}_K^N$. More precisely, $w_{\mathbf{y}}^{NK} = \arg\min_{w_k : k \in [K]} \{l_K(\mathbf{y}, \hat{\theta}_K^N)\}$. The following theorem shows that the inferred preference converges in probability to the true preference if the DMP we investigate enjoys the identifiability and the bijectivity defined above.

THEOREM 6 **(Consistency of $w_{\mathbf{y}}^{NK}$)**. *Suppose the same conditions of Theorem 5 hold. If DMP is bijective at θ_0, then $\|w_{\mathbf{y}} - w_{\mathbf{y}}^{NK}\|_2 \xrightarrow{P} 0$ for $\mathbf{y} \in \mathcal{Y}$ almost surely.*

6. Connections between IMOP, Clustering and Manifold Learning

6.1. Connection bewteen IMOP and Clustering

K-means clustering aims to partition the observations into K clusters such that the average squared distance between each observation and its closest cluster centroid is minimized. Given observations $\{\mathbf{y}_i\}_{i \in [N]}$, a mathematical formulation of K-means clustering is presented in the following (Bagirov 2008, Aloise and Hansen 2009).

$$\begin{array}{lll} \min\limits_{\mathbf{x}_k, z_{ik}} & \frac{1}{N} \sum\limits_{i \in [N]} \|\mathbf{y}_i - \sum\limits_{k \in [K]} z_{ik} \mathbf{x}_k\|_2^2 & \\ \text{s.t.} & \sum\limits_{k \in [K]} z_{ik} = 1, & \forall i \in [N], \\ & \mathbf{x}_k \in \mathbb{R}^n,\ z_{ik} \in \{0, 1\}, & \forall i \in [N],\ k \in [K], \end{array} \qquad \text{K-means clustering}$$

where K is the number of clusters, and $\{\mathbf{x}_k\}_{k \in [K]}$ are the centroids of the clusters.

PROPOSITION 8. *Given any K-means clustering problem, we can construct an instance of IMOP-EMP-WS, such that solving the K-means clustering problem is equivalent to solving the instance of IMOP-EMP-WS.*

LEMMA 9 **(Aloise et al. (2009), Mahajan et al. (2012))**. *K-means clustering is NP-hard.*

One should distinguish K-means clustering problem from K-means algorithm (a.k.a. Lloyd's algorithm) Lloyd (1982), where the later one is a fast heuristic to solve the former problem. Indeed, K-means clustering problem is NP-hard to solve even for instances in the plane Mahajan et al. (2012), or $K = 2$ in general dimension Aloise et al. (2009).

THEOREM 7 **(NP-hardness of IMOP)**. *IMOP-EMP-WS is NP-hard to solve.*

6.2. Connection between IMOP and Manifold Learning

Given a set of high-dimensional observations $\{\mathbf{y}_i\}_{i \in [N]}$ in $\mathbb{R}^n$, manifold learning attempts to find an embedding set $\{\mathbf{x}_i\}_{i \in [N]}$ in a low-dimensional space $\mathbb{R}^d$ $(d < n)$, and the local manifold structure formed by $\{\mathbf{y}_i\}_{i \in [N]}$ is preserved in the embedded space Tenenbaum et al. (2000), Roweis and Saul

(2000), Saul and Roweis (2003). Formally, given a set of data points $\{\mathbf{y}_i\}_{i\in[N]}$, we are required to find a mapping $f:\mathbb{R}^d\to\mathbb{R}^n$ and another set of points $\{\mathbf{x}_i\}_{i\in[N]}$ in $\mathbb{R}^d$ such that

$$\mathbf{y}_i = f(\mathbf{x}_i)+\epsilon_i,\quad i\in[N], \tag{1}$$

where ϵ_i represents random noise.

THEOREM 8 **(Pareto manifold)**. *Under the same assumptions as Theorem 3, if* $\forall w_1,w_2\in\mathscr{W}_p$, $w_1\neq w_2$ *implies* $S(w_1,\theta)\neq S(w_2,\theta)$ *for each* $\theta\in\Theta$, *for each* $\theta\in\Theta$, *we have that the Pareto optimal set of DMP is a* $(p-1)$*-dimensional manifold.*

From this theorem, one can see that the Pareto optimal set of a DMP with two objectives is a piecewise continuous curve, and the Pareto optimal set of a DMP with three objectives is a piecewise continuous surface, etc.

THEOREM 9. *Suppose that both* $\mathbf{f}(\mathbf{x},\theta)$ *and* $\mathbf{g}(\mathbf{x},\theta)$ *are linear functions in* $\mathbf{x}$ *for all* $\theta\in\Theta$. *Then,* $X_P(\theta)$ *is a piecewise linear manifold that has dimension not exceeding* $p-1$ *for all* $\theta\in\Theta$.

Note that the feasible set for a multiobjective linear program is a polyhedron. Thus, one way to interpret Theorem 9 is that the Pareto optimal set of such a program consists of Pareto optimal faces of the polyhedron that are arc-wise connected. Therefore, the Pareto optimal set naturally has a piecewise linear structure and forms a manifold. Note that each piece might have different dimensions. In this case, the Pareto optimal set of a linear program is a special manifold that is the disjoint union of topological manifolds with different dimensions.

7. Solutions Approaches to IMOP-EMP-WS

7.1. Solving IMOP through a Clustering-type Approach

For each $k\in[K]$, we denote C_k the set of noisy decisions with $z_{ik}=1$ after solving IMOP-EMP-WS to optimal. That is, observations in C_k are closest to $\mathbf{x}_k$. Consequently, we partition $\{\mathbf{y}_i\}_{i\in[N]}$ into K clusters $\{C_k\}_{k\in[K]}$. Let $\overline{\mathbf{y}}_k=\frac{1}{|C_k|}\sum_{\mathbf{y}_i\in C_k}\mathbf{y}_i$ be the centroid of cluster C_k, and denote $Var(C_k)$ the variance of C_k. Through an algebraic calculation, we get

$$M_K^N(\theta)=\frac{1}{N}\sum_{i\in[N]}\|\mathbf{y}_i-\sum_{k\in[K]}z_{ik}\mathbf{x}_k\|_2^2=\frac{1}{N}\sum_{k\in[K]}|C_k|\bigg(\|\overline{\mathbf{y}}_k-\mathbf{x}_k\|_2^2+Var(C_k)\bigg). \tag{2}$$

We propose a procedure that alternately clusters the noisy decisions (assignment step) and find θ and $\{\mathbf{x}_k\}_{k\in[K]}$ (update step) until convergence. Given θ and $\{\mathbf{x}_k\}_{k\in[K]}$, the assignment step can be done easily as we discussed previously. Moreover, the update step can be established by solving the problem as follows.

$$\begin{aligned}
\min_{\theta,\mathbf{x}_{k'}}\ & \tfrac{1}{N}\sum_{k\in[K]}|C_k|\|\overline{\mathbf{y}}_k-\textstyle\sum_{k'\in[K]}z_{kk'}\mathbf{x}_{k'}\|_2^2 & & \\
\text{s.t. }\ & \mathbf{x}_{k'}\in S(w_{k'},\theta), & \forall k'\in[K], & \\
& \sum_{k'\in[K]}z_{kk'}=1, & \forall k\in[K], & \\
& z_{kk'}\in\{0,1\}, & \forall k\in[K],\ k'\in[K]. &
\end{aligned} \tag{Kmeans-IMOP}$$

The expectation-maximization (EM)-style algorithm is formally presented in the following.

Algorithm 1 Solving IMOP-EMP-WS through a Clustering-type Approach

Input: Noisy decisions $\{\mathbf{y}_i\}_{i\in[N]}$, weight samples $\{w_k\}_{k\in[K]}$.

1: **Initialization**: Partition $\{\mathbf{y}_i\}_{i\in[N]}$ into K clusters using K-means clustering. Calculate $\{\overline{\mathbf{y}}_k\}_{k\in[K]}$. Solve Kmeans-IMOP and get an initial estimation of θ and $\{\mathbf{x}_k\}_{k\in[K]}$.

2: **while** stopping criterion is not satisfied **do**

3: **Assignment step**: Assign each $\mathbf{y}_i$ to the closest $\mathbf{x}_k$ to form new clusters. Calculate their centroids $\{\overline{\mathbf{y}}_k\}_{k\in[K]}$.

4: **Update step**: Update θ and $\{\mathbf{x}_k\}_{k\in[K]}$ by solving Kmeans-IMOP.

5: **end while**

Output: An estimate of the parameter θ of DMP.

LEMMA 10. *Both the **Assignment step** and the **Update step** in Algorithm 1 decrease $M_K^N(\theta)$.*

THEOREM 10 **(Finite convergence)**. *Suppose there is an oracle to solve Kmeans-IMOP. Algorithm 1 converges to a (local) optimal solution of IMOP-EMP-WS in a finite number of iterations.*

Proof. Since there is at most K^N ways to partition $\{\mathbf{y}_i\}_{i\in[N]}$ into K clusters, the monotonically decreasing Algorithm 1 will eventually arrive at a (local) optimal solution in finite steps. □

7.2. An Enhanced Algorithm for Solving IMOP with Manifold Learning

Algorithm 2 An initialization with manifold learning

1: **Input:** Noisy decision $\{\mathbf{y}_i\}_{i\in[N]}$, evenly sampled weights $\{w_k\}_{k\in[K]}$.

2: Apply any nonlinear manifold learning algorithm: $\mathbf{y}_i \in \mathbb{R}^n \to \mathbf{x}_i \in \mathbb{R}^{p-1}, \forall i \in [N]$.

3: Group $\{\mathbf{x}_i\}_{i\in[N]}$ into K clusters by solving K-means clustering. Denote I_K the set of labels of $\{\mathbf{x}_i\}_{i\in[N]}$. Find the clusters $\{C_k\}_{k\in[K]}$ and centroids $\{\overline{\mathbf{y}}_k\}_{k\in[K]}$ of $\{\mathbf{y}_i\}_{i\in[N]}$ according to I_K.

4: Solve Kmeans-IMOP and get $\hat{\theta}$ and $\{\mathbf{x}_k\}_{k\in[K]}$.

5: Run Step 2 - 5 in Algorithm 1.

Output: An estimate of the parameter θ of DMP.

THEOREM 11. *Suppose there is an oracle to solve Kmeans-IMOP. Algorithm 2 converges to a (local) optimal solution of IMOP-EMP-WS in a finite number of iterations.*

8. Computational Experiments

8.1. Learning the Objective Functions of an MLP

Consider the following Tri-objective linear programming problem:

$$\begin{aligned} \min\ & \{-x_1, -x_2, -x_3\} \\ s.t.\ & x_1 + x_2 + x_3 \leq 5, \\ & x_1 + x_2 + 3x_3 \leq 9, \\ & x_1, x_2, x_3 \geq 0. \end{aligned}$$

In this example, there are two efficient faces, one is the triangle defined by vertices $(2, 4, 5)$, the other one is the tetragon defined by vertices $(1, 3, 5, 4)$ as shown by Figure 1.

We generate the data as follows. First, $N = 10000$ Pareto optimal points $\{\mathbf{x}_i\}_{i\in[N]}$ are uniformly sampled on faces $(2, 4, 5)$ and $(1, 3, 5, 4)$. Next, the observations $\{\mathbf{y}_i\}_{i\in[N]}$ are obtained by adding noise to each Pareto optimal point, where the noise has a jointly normal distribution with zero mean and 0.5^2 units identity covariance. Namely, $\mathbf{y}_i = \mathbf{x}_i + \epsilon_i$, where $\epsilon_i \sim \mathcal{N}(\mathbf{0}_3, 0.5^2\mathbf{I}_3)$ for each $i \in [N]$. We assume that the parameters to be learned are non-positive. In addition, we add the normalization constraints $\mathbf{1}^T\mathbf{c}_1 = -1, \mathbf{1}^T\mathbf{c}_2 = -1$ and $\mathbf{1}^T\mathbf{c}_3 = -1$ to prevent the arise of trivial solutions, such as $\mathbf{c}_1 = \mathbf{c}_2 = \mathbf{c}_3 = [0, 0, 0]^T$. Then, we uniformly choose the weights $\{w_k\}_{k\in[K]}$ such that $w_k \in \mathscr{W}_3$ for each $k \in [K]$. Here, we set $K = 81$.

Algorithms 1 - 2 are used to solve IMOP-EMP-WS. In Algorithm 1, we run K-means++ algorithm 10 times to find the best clustering results. Centroids of the $K = 81$ clusters are plotted in Figure 1b. In Algorithm 2, we use Kernel PCA (Schölkopf et al. 1997) to project the data into a 2-dimension space, and then apply K-means++ clustering algorithm to find $K = 81$ clusters. Centroids of the $K = 81$ clusters are plotted in Figure 1c. As shown in Figures 1b - 1c, Algorithm 2 provides the better estimation of the manifold before solving IMOP-EMP-WS than Algorithm 1. Nevertheless, both solve IMOP-EMP-WS as they all recover the true Pareto optimal set even with the initial estimation of the parameter in the Initialization step. Thus, we won't run the later steps in Algorithm 1. The estimating results using Algorithm 1 are $\hat{\mathbf{c}}_1 = [0, 0, -1]^T$, $\hat{\mathbf{c}}_2 = [-0.3333, -0.3333, -0.3333]^T$ and $\hat{\mathbf{c}}_3 = [-0.2871, -0.2871, -0.4258]^T$ and $\hat{\mathbf{c}}_3 = [-0.2871, -0.2871, -0.4258]^T$. The estimating results using Algorithm 2 are $\hat{\mathbf{c}}_1 = [-0.4, -0.4, -0.2]^T$, $\hat{\mathbf{c}}_2 = [-0.2, -0.2, -0.6]^T$ and $\hat{\mathbf{c}}_3 = [-0.3333, -0.3333, -0.3333]^T$.

8.2. Learning the Preferences and Constraints of an MQP

We consider the following multiobjective quadratic programming problem.

$$\begin{aligned} \min_{\mathbf{x}\in\mathbb{R}^2_+}\ & \begin{pmatrix} f_1(\mathbf{x}) = \frac{1}{2}\mathbf{x}^T Q_1 \mathbf{x} + \mathbf{c}_1^T\mathbf{x} \\ f_2(\mathbf{x}) = \frac{1}{2}\mathbf{x}^T Q_2 \mathbf{x} + \mathbf{c}_2^T\mathbf{x} \end{pmatrix} \\ s.t.\ & A\mathbf{x} \geq \mathbf{b}, \end{aligned}$$

where parameters of the objective functions and the constraints are

$$Q_1 = \begin{bmatrix} 1 & 0 \\ 0 & 2 \end{bmatrix}, \mathbf{c}_1 = \begin{bmatrix} 3 \\ 1 \end{bmatrix}, Q_2 = \begin{bmatrix} 2 & 0 \\ 0 & 1 \end{bmatrix}, \mathbf{c}_2 = \begin{bmatrix} -6 \\ -5 \end{bmatrix}, A = \begin{bmatrix} -3 & 1 \\ 0 & -1 \end{bmatrix}, \mathbf{b} = \begin{bmatrix} -6 \\ -3 \end{bmatrix}.$$

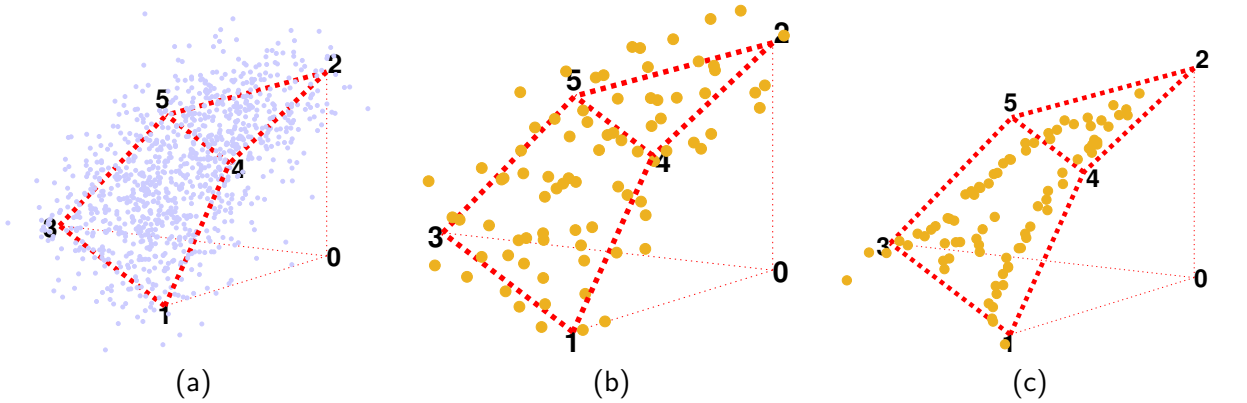

Figure 1 Learning the Objective Functions of a Tri-objective Linear Program Using $N = 10000$ Observations. (a) The Light Blue Dots Indicate the 1000 Observations Randomly Selected From the Data Set. Two Pareto Optimal Faces are the Triangle $(2, 4, 5)$, and the Tetragon $(1, 3, 5, 4)$. (b) Orange Dots Indicate the Centroids After Using K-means Clustering. (C) Orange Dots Indicate the Centroids after Using Kernel PCA and K-means Clustering.

8.2.1. Learning the Objective Functions In the experiments, suppose $\mathbf{c}_1$ and $\mathbf{c}_2$ are unknown, and the learner seeks to learn them given the noisy decisions. Assume that $\mathbf{c}_1$ and $\mathbf{c}_2$ are within range $[-10, 10]^2$. We generate the data in a way similar to the first set of experiments. The only difference is that each element of the noise has a uniform distribution supporting on $[-0.25, 0.25]$ with mean 0 for all $i \in [N]$.

We would like to use Algorithm 1 to solve large-scale IMOP-EMP-WS. We note that the SR approach can not handle cases when $N \geq 10$ and $K \geq 11$ in the **Update step**. Hence, the ADMM approach (Algorithm 3) is applied to solve Kmeans-IMOP. The stopping criterion for Algorithm 1 is that the maximum iteration number reaches five. In the **Initialization step**, we run K-means++ algorithm 50 times to find the best clustering results. When solving Kmeans-IMOP using ADMM, we partition the observations in such a way that each group has only one observation. We pick the penalty parameter $\rho = 0.5$ as the best out of a few trials. We use the initialization $\mathbf{c}_1^0 = \mathbf{c}_2^0 = \mathbf{v}_1^{t,0} = \mathbf{v}_2^{t,0} = \mathbf{0}_2$ for the iterations. The tolerances of the primal and dual residuals are set to be $\epsilon^{pri} = \epsilon^{dual} = 10^{-3}$. The termination criterion is that either the norms of the primal and dual residuals are smaller than 10^{-3} or the iteration number k reaches 50.

In Figure 2a, we report the prediction errors averaged over 10 repetitions of the experiments for different N and K. Here, we use an independent validation set that consists of 10^5 noisy decisions generated in the same way as the training data to compute the prediction error. We also calculate the prediction error using the true parameter and $M(\theta_{true}) = 0.022742$. More precisely, we evenly generate $K = 10^4$ weight samples and calculate the associated Pareto optimal solutions on the true Pareto optimal set. These Pareto optimal solutions are then used to find the prediction error of the true parameter. We observe that the prediction error has the trend to decrease to $M(\theta_{true})$ with

the increase of the data size N and weight sample size K. This makes lots of sense because IMOP-EMP-WS is risk consistent by Theorem 3 for this example. To further illustrate the performance of the algorithm, we plot the change of assignments versus iteration in the **Assignment step** over 10 repetitions of the experiments with $N = 5 \times 10^4, K = 21$ in Figure 2b. One can see the assignments become stable in 5 iterations, indicating the fast convergence of our algorithm. Also, we plot the estimated Pareto optimal set with $N = 5 \times 10^4, K = 21$ in the first repetition in Figure 2c. Here, $\hat{\mathbf{c}}_1 = [2.0023, 0.0454]^T$ and $\hat{\mathbf{c}}_2 = [-5.7197, -4.6949]^T$. They are not equal to the true parameters as this MQP is non-identifiable. However, our method still recovers the unknown parameters quite well as the estimated Pareto optimal set almost coincides with the true one.

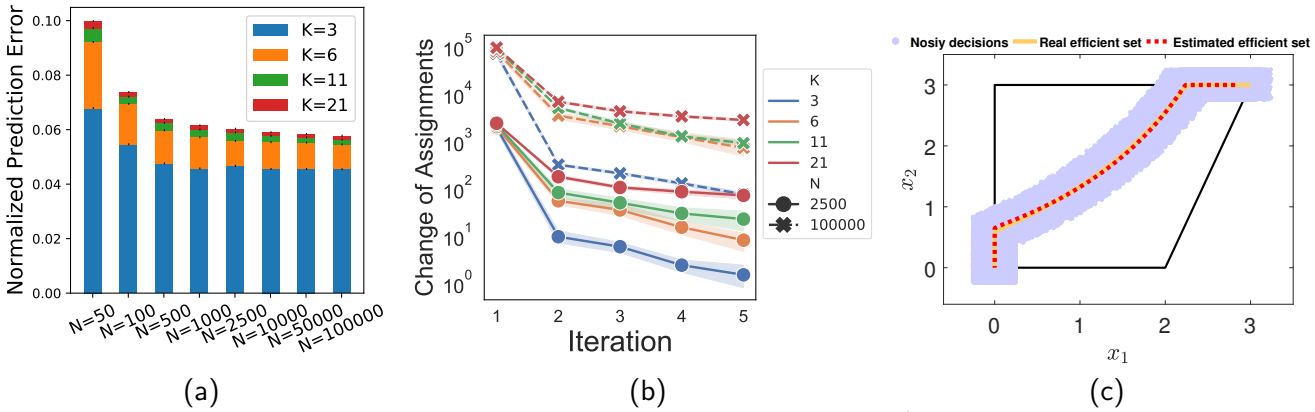

Figure 2 Learning the Objective Functions of an MQP. (a) Prediction Error $M(\hat{\theta}_K^N)$ for Different N and K. (b) The Dotted Yellow Line is the Error Bar Plot of the Change of the Assignments in Five Iterations over 10 Repetitions. (c) We Pick the First Repetition of the Experiments with $N = 5 \times 10^4$ and $K = 21$. Purple Dots Indicate the Data. The Estimated Pareto Optimal Set is Indicated by the Red Dotted Line. The Real Pareto Optimal Set is Shown by the Yellow Line.

We next compare the performance of Algorithm 1 and Algorithm 2. We find that the manifold learning based method generally performs better when the data has lots of noise. Specifically, in the third set of experiments, suppose $\mathbf{c}_1$ and $\mathbf{c}_2$ are unknown, and the learner seeks to learn them given the noisy decisions. Assume that $\mathbf{c}_1$ and $\mathbf{c}_2$ are within range $[-10, 10]^2$. We generate the data in a way similar to the previous two sets of experiments. The difference is that each element of the noise has a uniform distribution supporting on $[-1, 1]$ with mean 0 for all $i \in [N]$.

To further illustrate the performance of the two algorithms, we plot the centroids obtained in two algorithms when $N = 1000$ and $K = 6$ in Figure 3a, 3b, and 3c, respectively. Figures 3b and 3c shows clearly that the principal points (centroids) in Algorithm 2 almost lie on and recover the true Pareto optimal set, while centroids in Algorithm 1 lie around the true Pareto optimal set. This explains why Algorithm 2 would give us better estimation results. Also, we can see in Figures 3b and 3c that the estimated Pareto optimal set almost coincides with the true Pareto optimal set.

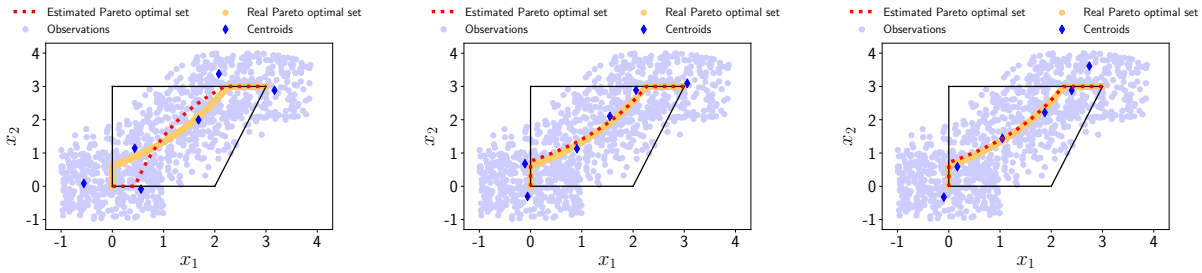

Figure 3 (a) Estimation Result for Algorithm 1. Purple Dots Indicate the Noisy Decisions. The Real Pareto Optimal Set is Shown by the Yellow Line. Blue Diamonds are the Centroids Obtained in the Initialization Step. The Estimated Pareto Optimal Set is Indicated by the Red Dotted Line. (b) Estimation Result for Algorithm 2 Using tSNE. Blue Diamonds are the Centroids Obtained through Manifold Learning and Clustering in Step 3. (c) Estimation Result for Algorithm 2 Using Factor Analysis. Blue Diamonds are the Centroids Obtained through Manifold Learning and Clustering in Step 3.

Table 2 11 sectors for the assets

Materials	Communication	Consumer Cyclical	Consumer Defensive	Energy	Financial Services
Healthcare	Industrial	Real Estate	Technology	Utilities	

8.3. Experiments with Real Data: Learning the Expected Returns

We consider various decisions arising from different investors in a stock market. Specifically, we consider a portfolio selection problem. The classical Markovitz mean-variance portfolio selection Markowitz (1952) is

$$\begin{aligned} \min_{\mathbf{x}} \ & \begin{pmatrix} f_1(\mathbf{x}) = -\mathbf{r}^T\mathbf{x} \\ f_2(\mathbf{x}) = \mathbf{x}^T Q\mathbf{x} \end{pmatrix} \\ s.t. \ & 0 \le x_i \le b_i, \forall i \in [n], \\ & \sum_{i=1}^{n} x_i = 1, \end{aligned}$$

where $\mathbf{r} \in \mathbb{R}^n_+$ is a vector of individual security expected returns, $Q \in \mathbb{R}^{n\times n}$ is the covariance matrix of securities returns, $\mathbf{x}$ is a portfolio specifying the proportions of capital to be invested in the different securities, and b_i is an upper bound put on the proportion of security $i \in [n]$.

Dataset: Stock price data is scraped from S&P 500 Index. Quarterly portfolio data is scraped from the mutual fund VHCAX (Vanguard Capital Opportunity Fund Admiral Shares) from March 2010 to December 2019. The assets are grouped into 11 sectors.

We first learn the average quarterly returns $\mathbf{r}$ for the 11 sectors from the portfolio data. We treat the learned returns as the market equilibrium returns. Note that the Black-Litterman model is an asset allocation approach that allows investment analysts to incorporate subjective views (based on investment analyst estimates) into market equilibrium returns. By blending analyst views and

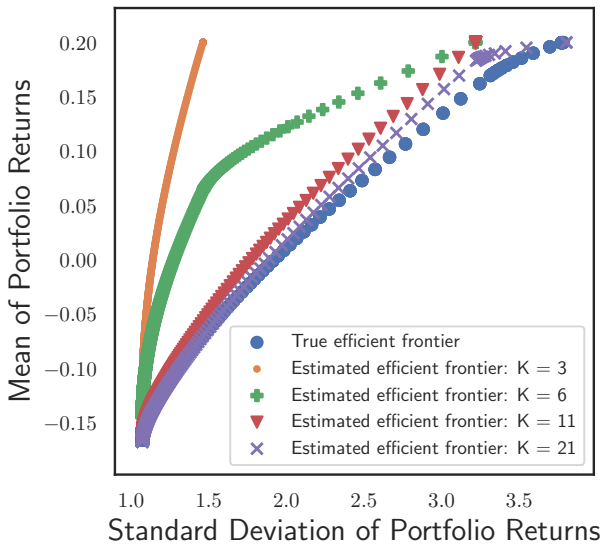

Figure 4 The efficient frontier and estimated efficient frontier.

equilibrium returns instead of relying only on historical asset returns, the Black-Litterman model provides a systematic way to estimate the mean of asset returns. Our market views of the 11 sectors: Energy, Healthcare, Technology and Utilities have higher returns than the equilibrium returns. We therefore add an additional return that follows the uniform distribution $U(0, 0.1)$ to each of the 4 sectors. We then seek to learn the blended expected return for each of the 4 sectors.

As we can see from Figure 4, the estimated efficient frontier almost coincides with the true efficient frontier, meaning that our model for IMOP indeed learns the expected returns well.

Appendix A: Omitted Proofs

A.1. Proof of Lemma 1

Proof. Since $\mathbf{g}(\mathbf{x}, \theta)$ is continuous and thus l.s.c. on $\mathbb{R}^n \times \Theta$ by ASSUMPTION 2, $X(\theta)$ is u.s.c. for each $\theta \in \Theta$ by Theorem 10 in Hogan (1973). From ASSUMPTION 1, we know that $\mathbf{g}(\mathbf{x}, \theta)$ is convex in $\mathbf{x}$ for each $\theta \in \Theta$. From ASSUMPTION 2, $X(\theta)$ has a nonempty relatively interior. Namely, there exists a $\bar{\mathbf{x}} \in \mathbb{R}^n$ such that $\mathbf{g}(\bar{\mathbf{x}}, \theta) < \mathbf{0}$. Then, $X(\theta)$ is l.s.c. for each $\theta \in \Theta$ by Theorem 12 in Hogan (1973). Hence, $X(\theta)$ is continuous on Θ. □

A.2. Proof of Lemma 2

Proof. First, we will show that $X_P(\theta)$ is u.s.c. on Θ. Since $\mathbf{f}(\mathbf{x}, \theta)$ is strictly convex in $\mathbf{x}$ for each $\theta \in \Theta$, the Pareto optimal set $X_P(\theta)$ coincides with the weakly Pareto optimal set. In addition, we know that $X(\theta)$ is continuous on Θ by Lemma 1. Also, note the pointed convex cone we use throughout this paper has the same meaning as the domination structure D in Tanino and Sawaragi (1980), and we set $D = \mathbb{R}^p_+$. To this end, we can readily verify that the sufficient conditions for upper semicontinuity in Theorem 7.1 of Tanino and Sawaragi (1980) are satisfied. Thus, $X_P(\theta)$ is u.s.c..

Next, we will show that $X_P(\theta)$ is l.s.c. on Θ. Theorem 7.2 of Tanino and Sawaragi (1980) provides the sufficient conditions for the lower semicontinuity of $X_P(\theta)$. All of these conditions are naturally satisfied under Assumptions 1 - 2 except the one that requires $\mathbf{f}(\mathbf{x},\theta)$ to be one-to-one, i.e., injective in $\mathbf{x}$. Next, we will show that the one-to-one condition can be safely replaced by the strict quasi-convexity of $\mathbf{f}(\mathbf{x},\theta)$ in $\mathbf{x}$.

Theorem 7.2 of Tanino and Sawaragi (1980) is a direct result of part (ii) in Lemma 7.2 of Tanino and Sawaragi (1980). To complete our proof, we only need to sightly modify the last part of the proof in Lemma 7.2. In what follows we will use notations in that paper.

Since strict convexity implies strict quasi-convexity, f is strictly quasi-convex. Suppose that $f(\bar{x},\hat{u}) = f(\hat{x},\hat{u})$ does not imply $\bar{x} = \hat{x}$. Let $z = \frac{\bar{x}+\hat{x}}{2}$. By the strict quasi-convexity of f, we have

$$f(z,\hat{u}) = f(\frac{\bar{x}+\hat{x}}{2},\hat{u}) < \max\{f(\bar{x},\hat{u}), f(\hat{x},\hat{u})\} = f(\hat{x},\hat{u}).$$

This contradicts the fact that $\hat{x} \in M(\hat{u})$, where $M(\hat{u})$ is the Pareto optimal set given $\hat{u}$. Hence, $\bar{x}$ must be equal to $\hat{x}$. The remain part of the proof is the same as that of Lemma 7.2. $\square$

A.3. Proof of Proposition 3

Proof. We apply Theorem 2 of Jennrich (1969) in our proof. We start by checking that the three conditions for using this theorem are satisfied. First, by Lemma 2, $X_P(\theta)$ is continuous. Then, applying Berge Maximum Theorem (Berge 1963) to IMOP-EMP implies that the empirical risk $M^N(\theta)$ is continuous. Second, by Assumption 2, Θ is a compact set. Third, $\forall \mathbf{y} \in \mathcal{Y}, \min_{\mathbf{x}\in X_P(\theta)} \|\mathbf{y}-\mathbf{x}\|_2^2 \le \|\mathbf{y}\|_2^2 + B^2 + 2B\|\mathbf{y}\|_2$ and the right-hand side is integrable with respect to $\mathbf{y}$ under Assumption 2. Consequently, all three conditions are satisfied and the proof is concluded. $\square$

A.4. Proof of Proposition 4

Proof. Similar to Proposition 3, the key step is to show the continuity of $M_K^N(\theta)$ in θ for each K. It suffices to show that $\bigcup_{k\in[K]} S(w_k,\theta)$ is continuous in θ for all K. First, let us establish the continuity of $S(w_k,\theta)$ in θ for each $k \in [K]$. Note that the feasible region $X(\theta)$ is irrelevant to w. Thus, applying the Berge Maximum Theorem (Berge 1963) to (WP) implies that $S(w_k,\theta)$ is upper semicontinuous in θ. Hence, $S(w_k,\theta)$ is continuous in θ as it is a single-valued set. Second, let us show the continuity of $\bigcup_{k\in[K]} S(w_k,\theta)$ in θ. By Propositions 2 and 4 of Hogan (1973), we know that a finite union of continuous sets, i.e., $\bigcup_{k\in[K]} S(w_k,\theta)$, is continuous in θ. Finally, applying Theorem 2 of Jennrich (1969) yields the uniform convergence of $M_K^N(\theta)$ to $M_K(\theta)$ in N. $\square$

A.5. Proof of Lemma 3

Proof. **(a)** Let $K_2 \ge K_1$. Under our setting, $K_2 \ge K_1$ implies $\{w_k\}_{k\in[K_1]} \subseteq \{w_k\}_{k\in[K_2]}$. By the definition of $l_K(\mathbf{y},\theta)$, we have $l_{K_1}(\mathbf{y},\theta) \ge l_{K_2}(\mathbf{y},\theta)$ for all $\mathbf{y} \in \mathcal{Y}$, and thus $M_{K_1}(\theta) \ge M_{K_2}(\theta)$ for all $\theta \in \Theta$. Therefore, $\{M_K(\theta)\}$ is monotone decreasing in K.

Recall the definition of $\hat{\theta}_K$, we know $\hat{\theta}_{K_2}$ minimizes $M_{K_2}(\theta)$. Therefore, $M_{K_2}(\hat{\theta}_{K_1}) \ge M_{K_2}(\hat{\theta}_{K_2})$. In addition, $M_{K_1}(\hat{\theta}_{K_1}) \ge M_{K_2}(\hat{\theta}_{K_1})$ by the first part of **(a)**. Consequently,

$$M_{K_1}(\hat{\theta}_{K_1}) \ge M_{K_2}(\hat{\theta}_{K_1}) \ge M_{K_2}(\hat{\theta}_{K_2}).$$

Therefore, $M_{K_1}(\hat{\theta}_{K_1}) \geq M_{K_2}(\hat{\theta}_{K_2})$ for $K_2 \geq K_1$.

Similarly, we can readily show that $M_K(\hat{\theta}_K) \geq M(\theta^*)$ by noting that

$$M_K(\hat{\theta}_K) \geq M(\hat{\theta}_K) \geq M(\theta^*).$$

The first inequality is a direct result of the first part of **(a)**; the second inequality follows from the fact that θ^* minimizes $M(\theta)$ by definition.

(b) Let $K_2 \geq K_1$. By the definition of $l_K(\mathbf{y}, \theta)$, we have $l_{K_1}(\mathbf{y}_i, \theta) \geq l_{K_2}(\mathbf{y}_i, \theta)$ for all $i \in [N]$, and thus $M^N_{K_1}(\theta) \geq M^N_{K_2}(\theta)$ for all $\theta \in \Theta$. Therefore, $\{M^N_K(\theta)\}$ is monotone decreasing in K.

Recall the definition of $\hat{\theta}^N_K$ in Table **??**, we know $\hat{\theta}^N_{K_2}$ minimizes $M^N_{K_2}(\theta)$. Therefore, $M^N_{K_2}(\hat{\theta}^N_{K_1}) \geq M^N_{K_2}(\hat{\theta}^N_{K_2})$. In addition, $M^N_{K_1}(\hat{\theta}^N_{K_1}) \geq M^N_{K_2}(\hat{\theta}^N_{K_1})$ by the first part of **(b)**. Consequently,

$$M^N_{K_1}(\hat{\theta}^N_{K_1}) \geq M^N_{K_2}(\hat{\theta}^N_{K_1}) \geq M^N_{K_2}(\hat{\theta}^N_{K_2}).$$

Hence, $M^N_{K_1}(\hat{\theta}^N_{K_1}) \geq M^N_{K_2}(\hat{\theta}^N_{K_2})$ for $K_2 \geq K_1$.

Finally, we can show $M^N_K(\hat{\theta}^N_K) \geq M^N(\hat{\theta}^N)$ by noting that $M^N_K(\hat{\theta}^N_K) \geq M^N(\hat{\theta}^N_K) \geq M^N(\hat{\theta}^N)$. □

A.6. Proof of Lemma 4

Proof. $\forall w \in \mathscr{W}_p$, one can readily check that $w^T\mathbf{f}(\cdot, \theta)$ is strongly convex for each θ and thus

$$w^T\mathbf{f}(\mathbf{y}, \theta) \geq w^T\mathbf{f}(\mathbf{x}, \theta) + \nabla w^T\mathbf{f}(\mathbf{x}, \theta)^T(\mathbf{y} - \mathbf{x}) + \frac{\lambda}{2}\|\mathbf{y} - \mathbf{x}\|_2^2.$$

Thus, the second-order growth condition holds for $w^T\mathbf{f}(\cdot, \theta)$ for all $\theta \in \Theta$. That is,

$$w^T\mathbf{f}(\mathbf{x}, \theta) \geq w^T\mathbf{f}(S(w, \theta), \theta) + \frac{\lambda}{2}\|(S(w, \theta) - \mathbf{x}\|_2^2. \tag{3}$$

In addition, $\forall w, w_0 \in \mathscr{W}_p$, we have

$$\begin{aligned} |w^T\mathbf{f}(\mathbf{x}, \theta) - w_0^T\mathbf{f}(\mathbf{x}, \theta)| &= |(w - w_0)^T\mathbf{f}(\mathbf{x}, \theta)| \\ &\leq \|w - w_0\|_2\|\mathbf{f}(\mathbf{x}, \theta)\|_2 \quad \text{(Cauchy-Schwarz inequality)} \\ &\leq L\|w - w_0\|_2. \end{aligned} \tag{4}$$

A.7. Proof of Lemma 5

Proof. By definition,

$$l_K(\mathbf{y}, \theta) - l(\mathbf{y}, \theta) = \min_{\mathbf{x} \in \bigcup_{k \in [K]} S(w_k, \theta)} \|\mathbf{y} - \mathbf{x}\|_2^2 - \min_{\mathbf{x} \in X_P(\theta)} \|\mathbf{y} - \mathbf{x}\|_2^2 \geq 0.$$

Let $\|\mathbf{y} - S(w^{\mathbf{y}}_K, \theta)\|_2^2 = \min_{\mathbf{x} \in \bigcup_{k \in [K]} S(w_k, \theta)} \|\mathbf{y} - \mathbf{x}\|_2^2$, and $\|\mathbf{y} - S(w_{\mathbf{y}}, \theta)\|_2^2 = \min_{\mathbf{x} \in X_P(\theta)} \|\mathbf{y} - \mathbf{x}\|_2^2$. Let $w^K_{\mathbf{y}}$ be the closest weight sample among $\{w_k\}_{k \in [K]}$ to $w_{\mathbf{y}}$. Then,

$$\begin{aligned} l_K(\mathbf{y}, \theta) - l(\mathbf{y}, \theta) &= \|\mathbf{y} - S(w^{\mathbf{y}}_K, \theta)\|_2^2 - \|\mathbf{y} - S(w_{\mathbf{y}}, \theta)\|_2^2 \\ &\leq \|\mathbf{y} - S(w^K_{\mathbf{y}}, \theta)\|_2^2 - \|\mathbf{y} - S(w_{\mathbf{y}}, \theta)\|_2^2 \\ &= \left(2\mathbf{y} - S(w^K_{\mathbf{y}}, \theta) - S(w_{\mathbf{y}}, \theta)\right)^T \left(S(w_{\mathbf{y}}, \theta) - S(w^K_{\mathbf{y}}, \theta)\right) \\ &\leq \|2\mathbf{y} - S(w^K_{\mathbf{y}}, \theta) - S(w_{\mathbf{y}}, \theta)\|_2 \|S(w_{\mathbf{y}}, \theta) - S(w^K_{\mathbf{y}}, \theta)\|_2 \\ &\leq 2(B + R)\|S(w_{\mathbf{y}}, \theta) - S(w^K_{\mathbf{y}}, \theta)\|_2 \\ &\leq \tfrac{4(B+R)\zeta\sqrt{p}}{\lambda} \cdot \|w_{\mathbf{y}} - w^K_{\mathbf{y}}\|_2, \end{aligned} \tag{5}$$

where $\zeta = \max_{l\in[p],\mathbf{x}\in X(\theta),\theta\in\Theta} |f_l(\mathbf{x},\theta)|$. The third inequality is due to Cauchy Schwarz inequality. Under Assumptions 1 - 2, we can apply Lemma 4 to yield the last inequality.

Next, we will show that $\forall w \in \mathscr{W}_p$, the distance between w and its closest weight sample among $\{w_k\}_{k\in[K]}$ is upper bounded by the function of K and p and nothing else. More precisely, we will show that

$$\sup_{w\in\mathscr{W}_p} \min_{k\in[K]} \|w - w_k\|_2 \leq \frac{\sqrt{2}}{\Lambda - 1}. \tag{6}$$

Here, Λ is the number of evenly spaced weight samples between any two extreme points of $\mathscr{W}_p$.

Note that $\{w_k\}_{k\in[K]}$ are evenly sampled from $\mathscr{W}_p$, and that the distance between any two extreme points of $\mathscr{W}_p$ equals to $\sqrt{2}$. Hence, the distances between any two neighboring weight samples are equal and can be calculated as the distance between any two extreme points of $\mathscr{W}_p$ divided by $\Lambda - 1$. Proof of (6) can be done by further noticing that the distance between any w and $\{w_k\}_{k\in[K]}$ is upper bounded by the distances between any two neighboring weight samples.

Combining (5) and (6) yields that

$$0 \leq l_K(\mathbf{y},\theta) - l(\mathbf{y},\theta) \leq \frac{4(B+R)\zeta}{\lambda} \cdot \frac{\sqrt{2p}}{\Lambda - 1}, \tag{7}$$

Then, we can prove that the total number of weight samples K and Λ has the following relationship:

$$K = \binom{\Lambda + p - 2}{p - 1} \tag{8}$$

Proof of (8) can be done by induction with respect to p. Obviously, (8) holds when $p = 2$ as $K = \Lambda$. Assume (8) holds for the $\leq p-1$ cases. For ease of notation, denote

$$K_p^\Lambda = \binom{\Lambda + p - 2}{p - 1}.$$

Then, for the p case, we note that the weight samples can be classified into two categories: $w_p = 0; w_p > 0$. For $w_p = 0$, the number of weight samples is simply K_{p-1}^Λ. For $w_p > 0$, the number of weight samples is $K_p^{\Lambda-1}$. Thus,

$$K = K_{p-1}^\Lambda + K_p^{\Lambda-1}. \tag{9}$$

Iteratively expanding $K_p^{\Lambda-1}$ through the same argument as (8) and using the fact that

$$\binom{n}{k} = \binom{n-1}{k-1} + \binom{n-1}{k},$$

we have

$$\begin{aligned} K &= K_{p-1}^\Lambda + K_p^{\Lambda-1} = K_{p-1}^\Lambda + K_{p-1}^{\Lambda-1} + K_p^{\Lambda-2} \\ &\vdots \\ &= K_{p-1}^\Lambda + K_{p-1}^{\Lambda-1} + \cdots + K_{p-1}^2 + K_p^1 \\ &= \binom{\Lambda+p-3}{p-2} + \binom{\Lambda+p-4}{p-2} + \cdots + \binom{p-1}{p-2} + \binom{p-1}{p-1} \\ &= \tfrac{(\Lambda+p-2)!}{(\Lambda-1)!(p-1)!} \end{aligned} \tag{10}$$

To this end, we complete the proof of (8).

Furthermore, we notice that

$$K = \frac{(\Lambda+p-2)!}{(\Lambda-1)!(p-1)!} \leq \frac{(\Lambda+p-2)^{p-1}}{(p-1)!} < \left(\frac{\Lambda+p-2}{p-1}\right)^{p-1} \cdot e^{p-1}.$$

Then, when $\Lambda \geq p (K \geq 2^{p-1})$, through simple algebraic calculation we have

$$\frac{e}{K^{\frac{1}{p-1}}} > \frac{p-1}{\Lambda+p-2} > \frac{1}{4} \cdot \frac{p}{\Lambda-1} \tag{11}$$

We complete the proof by combining (7) and (11) and noticing that $\sqrt{2p} \leq p$. $\square$

A.8. Proof of Proposition 5

Proof. Note that $\forall \theta \in \Theta$, $S(w,\theta)$ is single-valued due to the fact that $\mathbf{f}$ is strongly convex. $\forall \mathbf{y} \in \mathcal{Y}$, let $\mathbf{x}_\mathbf{y} \in X_P(\theta)$ be the nearest point to $\mathbf{y}$. By Proposition 2, there exists a $w_\mathbf{y} \in \mathscr{W}_p$ such that $\mathbf{x}_\mathbf{y} = S(w_\mathbf{y},\theta)$. Let $w_\mathbf{y}^K$ be the nearest one to $w_\mathbf{y}$ among the weight samples $\{w_k\}_{k\in[K]}$. Then,

$$\begin{aligned}
M_K(\theta) &= \mathbb{E}\Big(l_K(\mathbf{y},\theta)\Big) \\
&\le \mathbb{E}\Big(\|\mathbf{y} - S(w_\mathbf{y}^K,\theta)\|_2^2\Big) \\
&= \mathbb{E}\Big(\|\mathbf{y} - S(w_\mathbf{y},\theta)\|_2^2\Big) + \mathbb{E}\Big(\|S(w_\mathbf{y},\theta) - S(w_\mathbf{y}^K,\theta)\|_2^2\Big) \\
&\quad +2\mathbb{E}\Big(\big\langle \mathbf{y} - S(w_\mathbf{y},\theta), S(w_\mathbf{y},\theta) - S(w_\mathbf{y}^K,\theta)\big\rangle\Big) \\
&\le \mathbb{E}\Big(\|\mathbf{y} - S(w_\mathbf{y},\theta)\|_2^2\Big) + \mathbb{E}\Big(\|S(w_\mathbf{y},\theta) - S(w_\mathbf{y}^K,\theta)\|_2^2\Big) \\
&\quad +2\mathbb{E}\Big(\|\mathbf{y} - S(w_\mathbf{y},\theta)\|_2 \|S(w_\mathbf{y},\theta) - S(w_\mathbf{y}^K,\theta)\|_2\Big) \quad \text{(Cauchy Schwarz inequality)} \\
&= M(\theta) + \mathbb{E}\Big(\|S(w_\mathbf{y},\theta) - S(w_\mathbf{y}^K,\theta)\|_2^2\Big) \\
&\quad +2\mathbb{E}\Big(\|\mathbf{y} - S(w_\mathbf{y},\theta)\|_2 \|S(w_\mathbf{y},\theta) - S(w_\mathbf{y}^K,\theta)\|_2\Big),
\end{aligned} \tag{12}$$

where the first inequality is due to the fact that $l_K(\mathbf{y},\theta) = \min_{k\in[K]}\{\|\mathbf{y} - \mathbf{x}_k\|_2^2 : \mathbf{x}_k = S(w_k,\theta)\} \le \|\mathbf{y} - S(w_\mathbf{y}^K,\theta)\|_2^2$.

Let $A_K := \sup_{\mathbf{y}\in\mathcal{Y},\theta\in\Theta}\|S(w_\mathbf{y},\theta) - S(w_\mathbf{y}^K,\theta)\|_2$. Then,

$$\mathbb{E}\Big(\|S(w_\mathbf{y},\theta) - S(w_\mathbf{y}^K,\theta)\|_2^2\Big) \le A_K^2. \tag{13}$$

Moreover,

$$\begin{aligned}
\mathbb{E}\Big(\|\mathbf{y} - S(w_\mathbf{y},\theta)\|_2 \|S(w_\mathbf{y},\theta) - S(w_\mathbf{y}^K,\theta)\|_2\Big) &\le A_K \mathbb{E}\Big(\|\mathbf{y} - S(w_\mathbf{y},\theta)\|_2\Big) \\
&\le A_K \mathbb{E}\Big(\|\mathbf{y}\|_2 + \|S(w_\mathbf{y},\theta)\|_2\Big) \\
&\le A_K \mathbb{E}\Big(\|\mathbf{y}\|_2 + B\Big).
\end{aligned} \tag{14}$$

Note that $\mathbb{E}\Big(\|\mathbf{y}\|_2 + B\Big)$ in (14) is a finite number under our assumptions. Putting (13) and (14) into (12), and further noticing that $M_K(\theta) \ge M(\theta)$ by part (a) of Lemma 3, we have

$$0 \le M_K(\theta) - M(\theta) \le A_K\Big(A_K + 2B + 2\mathbb{E}\big(\|\mathbf{y}\|_2\big)\Big). \tag{15}$$

By (15), we will conclude the proof if we can show $A_K \longrightarrow 0$ in K. By Lemma 4,

$$A_K \le \frac{2L}{\lambda}\sup_{\mathbf{y}\in\mathcal{Y}}\|w_\mathbf{y} - w_\mathbf{y}^K\|_2. \tag{16}$$

(16) implies that we only need to show $\|w_\mathbf{y} - w_\mathbf{y}^K\|_2^2 \longrightarrow 0$ in K for any $\mathbf{y} \in \mathcal{Y}$. It suffices to show that given any $w \in \mathscr{W}_p$, the nearest w_k to w among $\{w_k\}_{k\in[K]}$ can be arbitrarily small as $K \to \infty$. This is readily satisfied since we evenly sample $\{w_k\}_{k\in[K]}$ from $\mathscr{W}_p$. $\square$

A.9. Proof of Proposition 6

Proof. We use notations here similar to those in Proposition 5. We have

$$\begin{aligned} M_K^N(\theta) &= \tfrac{1}{N}\sum_{i\in[N]}\min_{k\in[K]}\|\mathbf{y}_i-\mathbf{x}_k\|_2^2 \\ &\le \tfrac{1}{N}\sum_{i\in[N]}\|\mathbf{y}_i - S(w_{\mathbf{y}_i}^K,\theta)\|_2^2 \\ &= \tfrac{1}{N}\sum_{i\in[N]}\|\mathbf{y}_i - S(w_{\mathbf{y}_i},\theta)\|_2^2 + \tfrac{1}{N}\sum_{i\in[N]}\|S(w_{\mathbf{y}_i},\theta) - S(w_{\mathbf{y}_i}^K,\theta)\|_2^2 \\ &\quad + \tfrac{2}{N}\sum_{i\in[N]}\left\langle \mathbf{y}_i - S(w_{\mathbf{y}_i},\theta), S(w_{\mathbf{y}_i},\theta) - S(w_{\mathbf{y}_i}^K,\theta)\right\rangle \\ &\le \tfrac{1}{N}\sum_{i\in[N]}\|\mathbf{y}_i - S(w_{\mathbf{y}_i},\theta)\|_2^2 + \tfrac{1}{N}\sum_{i\in[N]}\|S(w_{\mathbf{y}_i},\theta) - S(w_{\mathbf{y}_i}^K,\theta)\|_2^2 \\ &\quad + \tfrac{2}{N}\sum_{i\in[N]}\|\mathbf{y}_i - S(w_{\mathbf{y}_i},\theta)\|_2\|S(w_{\mathbf{y}_i},\theta) - S(w_{\mathbf{y}_i}^K,\theta)\|_2 \quad \text{(Cauchy Schwarz inequality).} \end{aligned} \tag{17}$$

Moreover, by part (b) of Lemma 3, we have $M_K^N(\theta) - M^N(\theta) \ge 0$. To this end, through a similar argument as in the proof of Proposition 5, we have

$$0 \le M_K^N(\theta) - M^N(\theta) \le A_K\bigg(A_K + 2B + 2R\bigg), \tag{18}$$

where the last inequality follows from the fact that $\max_{i\in[N],\theta\in\Theta}\|S(w_{\mathbf{y}_i},\theta) - S(w_{\mathbf{y}_i}^K,\theta)\|_2 \le A_K$.

The remaining proof is exactly the same as that of Proposition 5. □

A.10. Proof of Proposition 7

Proof. $\forall\theta\in\Theta$, $|M_K^N(\theta) - M(\theta)| \xrightarrow{P} 0$ if and only if $\forall\delta>0, \forall\epsilon>0$, $\exists J$, s.t. $\forall N, K \ge J$,

$$\mathbb{P}(|M_K^N(\theta) - M(\theta)| > \epsilon) < \delta. \tag{19}$$

To prove the above statement, we first note that

$$\begin{aligned} \mathbb{P}(|M_K^N(\theta) - M(\theta)| > \epsilon) &= \mathbb{P}(|M_K^N(\theta) - M^N(\theta) + M^N(\theta) - M(\theta)| > \epsilon) \\ &\le \mathbb{P}(|M_K^N(\theta) - M^N(\theta)| + |M^N(\theta) - M(\theta)| > \epsilon) \\ &\le \mathbb{P}(|M_K^N(\theta) - M^N(\theta)| > \epsilon/2) + \mathbb{P}(|M^N(\theta) - M(\theta)| > \epsilon/2). \end{aligned} \tag{20}$$

For the first term on the last line of (20), by Proposition 6, $\exists K_1$, s.t. $\forall K \ge K_1$, $\forall N$,

$$\mathbb{P}(|M_K^N(\theta) - M^N(\theta)| > \epsilon/2) < \delta/2. \tag{21}$$

For the second term on the last line of (20), by Proposition 3, $\exists N_1$, s.t. $\forall N \ge N_1$,

$$\mathbb{P}(|M^N(\theta) - M(\theta)| > \epsilon/2) < \delta/2. \tag{22}$$

Now, let $J = \max\{N_1, K_1\}$. Putting (21) and (22) in (20), we have $\forall N, K \ge J$,

$$\mathbb{P}(|M_K^N(\theta) - M(\theta)| > \epsilon) < \delta. \tag{23}$$

Hence, we complete the proof. □

A.11. Proof of Theorem 1

Proof. Let $\theta^* \in \Theta^*$, and $\hat{\theta}^N \in \arg\min\{M^N(\theta) : \theta\in\Theta\}$. Then, $M(\hat{\theta}^N) - M(\theta^*) \ge 0$. Also,

$$\begin{aligned} M(\hat{\theta}^N) - M(\theta^*) &= M(\hat{\theta}^N) - M^N(\hat{\theta}^N) + M^N(\hat{\theta}^N) - M(\theta^*) \\ &\le M(\hat{\theta}^N) - M^N(\hat{\theta}^N) + M^N(\theta^*) - M(\theta^*) \\ &\le 2\sup_{\theta\in\Theta}|M^N(\theta) - M(\theta)|, \end{aligned}$$

where the first inequality follows the fact that $M^N(\hat{\theta}^N) \le M^N(\theta^*)$.

Hence, applying Proposition 3 yields that $M(\hat{\theta}^N) - M(\theta^*) \xrightarrow{p} 0$. □

A.12. Proof of Lemma 6

Proof. Let $\mathcal{G}$ be a class of functions g mapping from Z to $\mathbb{R}$, where

$$g(Z) = \frac{f(Z) - a}{b - a}. \tag{24}$$

Note that $g(Z) \in [0, 1]$. By Theorem 3.1 in Mohri et al. (2012), we have

$$\mathbb{E}[g(Z)] \leq \frac{1}{N} \sum_{i \in [N]} g(Z_i) + 2Rad_N(\mathcal{G}) + \sqrt{\frac{log(1/\delta)}{2N}}. \tag{25}$$

Using part 3 in Theorem 12 of Bartlett and Mendelson (2002), and the translation invariant property, i.e., $Rad_N(\mathcal{F} - a) = Rad_N(\mathcal{F})$, we have

$$Rad_N(\mathcal{G}) = Rad_N\left(\frac{\mathcal{F} - a}{b - a}\right) = \frac{Rad_N(\mathcal{F})}{b - a}. \tag{26}$$

Plugging (24) and (26) in (25) yields the main result. □

A.13. Proof of Lemma 7

Proof. By the definition of Rademacher complexity, we have

$$\begin{aligned}
Rad_N\big(\mathcal{F}\big) &= \tfrac{1}{N}\mathbb{E}\left[\sup_{f \in \mathcal{F}} \sum_{i \in [N]} \sigma_i f(\mathbf{y}_i, \theta)\right] \\
&= \tfrac{1}{N}\mathbb{E}\left[\sup_{\theta \in \Theta} \sum_{i \in [N]} \sigma_i \min_{k \in [K]} \|\mathbf{y}_i - \mathbf{x}_k\|_2^2\right] \\
&= \tfrac{1}{N}\mathbb{E}\left[\sup_{\theta \in \Theta} \sum_{i \in [N]} \sigma_i \min_{k \in [K]} \big(\|\mathbf{y}_i\|_2^2 - 2\langle \mathbf{y}_i, \mathbf{x}_k\rangle + \|\mathbf{x}_k\|_2^2\big)\right] \\
&= \tfrac{1}{N}\mathbb{E}\left[\sup_{\theta \in \Theta} \sum_{i \in [N]} \sigma_i \min_{k \in [K]} \big(-2\langle \mathbf{y}_i, \mathbf{x}_k\rangle + \|\mathbf{x}_k\|_2^2\big)\right].
\end{aligned}$$

Note the fact $\mathbb{P}(\|\mathbf{x}\|_2 \leq B) = 1$ by Assumption 2. Through a similar argument in statement (ii) of Lemma 4.3 in Biau et al. (2008), we get

$$\frac{1}{N}\mathbb{E}\left[\sup_{\theta \in \Theta} \sum_{i \in [N]} \sigma_i \min_{k \in [K]} \big(-2\langle \mathbf{y}_i, \mathbf{x}_k\rangle + \|\mathbf{x}_k\|_2^2\big)\right] \leq 2K\left(\frac{1}{N}\mathbb{E}\left[\sup_{\|\mathbf{x}\|_2 \leq B} \sum_{i \in [N]} \sigma_i \langle \mathbf{y}_i, \mathbf{x}\rangle\right] + \frac{B^2}{2\sqrt{N}}\right). \tag{27}$$

The first term on the right-hand side of (27) can be upper bounded in the following way:

$$\begin{aligned}
\tfrac{1}{N}\mathbb{E}\left[\sup_{\|\mathbf{x}\|_2 \leq B} \sum_{i \in [N]} \sigma_i \langle \mathbf{y}_i, \mathbf{x}\rangle\right] &= \tfrac{1}{N}\mathbb{E}\left[\sup_{\|\mathbf{x}\|_2 \leq B} \langle \sum_{i \in [N]} \sigma_i \mathbf{y}_i, \mathbf{x}\rangle\right] \\
&\leq \tfrac{1}{N}\mathbb{E} \sup_{\|\mathbf{x}\|_2 \leq B} \|\mathbf{x}\|_2 \| \sum_{i \in [N]} \sigma_i \mathbf{y}_i\|_2 \quad \text{(Cauchy-Schwarz inequality)} \\
&\leq \tfrac{B}{N}\mathbb{E}\| \sum_{i \in [N]} \sigma_i \mathbf{y}_i\|_2 \\
&\leq \tfrac{B}{N}\sqrt{\mathbb{E}\| \sum_{i \in [N]} \sigma_i \mathbf{y}_i\|_2^2} \quad \text{(Jensen's inequality)} \\
&= \tfrac{B}{N}\sqrt{N\mathbb{E}\|\mathbf{y}\|_2^2} \\
&\leq \tfrac{BR}{\sqrt{N}} \quad (\mathbb{P}(\|\mathbf{y}\|_2 \leq R) = 1).
\end{aligned} \tag{28}$$

Plugging the result of (28) in (27), we get the bound for the Rademacher complexity of $\mathcal{F}$. □

A.14. Proof of Theorem 4

Proof. We specialize Lemmas 6 and 7 to prove the theorem. Note that

$$0 \le f(\mathbf{y},\theta) = \min_{k\in[K]} \|\mathbf{y}-\mathbf{x}_k\|_2^2 \le (B+R)^2.$$

Let $a=0, b=(B+R)^2$ in Lemma 6. Then, combining the results in Lemmas 6 and 7 yields this theorem. □

A.15. Proof of Lemma 8

Proof. Sufficiency: $d_{sH}(X,Y)=0$ implies that $\inf_{y\in Y}\|x-y\|_2=0, \forall x\in X$. That is, $\exists y\in Y$, st. $x=y$. Hence, $X\subseteq Y$. Necessity: $X\subseteq Y$ implies that $\forall x\in X$, $\exists y\in Y$, s.t. $y=x$. Thus, $\inf_{y\in Y}\|x-y\|_2=0$. Therefore, $d_{sH}(X,Y)=0$. □

A.16. Proof of Theorem 5

Proof. First, we show that θ_0 minimizes $M(\theta)$ among Θ. This is readily true since $M(\theta_0)=0$ by noting that there is no noise in the data. By Theorem 3, a direct result is $M(\hat{\theta}_K^N) \xrightarrow{P} M(\theta_0)=0$. Second, we show that θ_0 is the unique solution that minimizes $M(\theta)$ among Θ. $\forall\theta'\in\Theta\setminus\theta$, $M(\theta)=\mathbb{E}_{\mathbf{y}\in X_P(\theta_0)}\big(\min_{\mathbf{x}\in X_P(\theta)}\|\mathbf{y}-\mathbf{x}\|_2^2\big)>0$ as $d_{sH}(X_P(\theta),X_P(\theta'))>0$. Consequently, we have $M(\theta)>M(\theta_0)=0$. Finally, since DMP is identifiable at θ_0, then $\forall\epsilon>0$, $\exists\delta>0$, s.t. $M(\theta)-M(\theta_0)>\delta$ for every θ with $d(\theta,\theta_0)>\epsilon$. Thus, the event $\{d(\hat{\theta}_K^N,\theta_0)>\epsilon\}$ is contained in the event $\{M(\hat{\theta}_K^N)-M(\theta_0)>\delta\}$. Namely, $\mathbb{P}(d(\hat{\theta}_K^N,\theta_0)>\epsilon)\le\mathbb{P}(M(\hat{\theta}_K^N)-M(\theta_0)>\delta)$. We complete the proof by noting that the probability of the right term converges to 0 as $M(\hat{\theta}_K^N)\xrightarrow{P} M(\theta_0)$. □

A.17. Proof of Theorem 6

Proof. First, note that

$$\begin{aligned}\|S(w_{\mathbf{y}}^{NK},\theta_0)-S(w_{\mathbf{y}},\theta_0)\|_2 &= \|S(w_{\mathbf{y}}^{NK},\theta_0)-S(w_{\mathbf{y}}^{NK},\hat{\theta}_K^N)+S(w_{\mathbf{y}}^{NK},\hat{\theta}_K^N)-S(w_{\mathbf{y}},\theta_0)\|_2\\ &\le \|S(w_{\mathbf{y}}^{NK},\theta_0)-S(w_{\mathbf{y}}^{NK},\hat{\theta}_K^N)\|_2+\|S(w_{\mathbf{y}}^{NK},\hat{\theta}_K^N)-S(w_{\mathbf{y}},\theta_0)\|_2.\end{aligned} \tag{29}$$

By Theorem 5, we have $\hat{\theta}_K^N \xrightarrow{P} \theta_0$. Note that $S(w,\theta)$ is continuous in $\theta\in\Theta$. By continuous mapping theorem, the first term in the last line of (29) $\|S(w_{\mathbf{y}}^{NK},\theta_0)-S(w_{\mathbf{y}}^{NK},\hat{\theta}_K^N)\|_2 \xrightarrow{P} 0$.

By the argument in the proof of Theorem 5, the second term in the last line of (29) $\|S(w_{\mathbf{y}}^{NK},\hat{\theta}_K^N)-S(w_{\mathbf{y}},\theta_0)\|_2 \xrightarrow{P} 0$ almost surely. Otherwise, $M(\hat{\theta}_K^N)=\mathbb{E}_{\mathbf{y}\in X_P(\theta_0)}\big(\min_{\mathbf{x}\in X_P(\hat{\theta}_K^N)}\|\mathbf{y}-\mathbf{x}\|_2^2\big)=\mathbb{E}_{\mathbf{y}\in X_P(\theta_0)}\|S(w_{\mathbf{y}}^{NK},\hat{\theta}_K^N)-S(w_{\mathbf{y}},\theta_0)\|_2^2>0$, and thus will not converge to $M(\theta_0)$.

Putting the above two results into (29) yields $\|S(w_{\mathbf{y}}^{NK},\theta_0)-S(w_{\mathbf{y}},\theta_0)\|_2 \xrightarrow{P} 0$ almost surely.

Next, note that $S(w,\theta_0)$ is continuous in w, and that $MOP(\theta_0)$ is bijective. Then, we have that $S(\cdot,\theta_0): \mathscr{W}_p \to X_P(\theta_0)$ is a one-to-one correspondence. Thus, $S(\cdot,\theta_0)$ is a homeomorphism by the inverse mapping theorem (Sutherland 2009), meaning that the inverse map $S^{-1}(\cdot,\theta_0): X_P(\theta_0)\to\mathscr{W}_p$ is also continuous. Therefore, $\|S(w_{\mathbf{y}}^{NK},\theta_0)-S(w_{\mathbf{y}},\theta_0)\|_2 \xrightarrow{P} 0$ implies that $\|w_{\mathbf{y}}-w_{\mathbf{y}}^{NK}\|_2 \xrightarrow{P} 0$ by the continuous mapping theorem. □

A.18. Proof of Proposition 8

Proof. Denote $\{\mathbf{y}_i\}_{i\in[N]}$ these points, denote K the number of clusters, and denote $\{\mathbf{x}_k\}_{k\in[K]}$ the set of optimal centroids. We then construct an equivalent instance of IMOP-EMP-WS as follows. Note that an IMOP is determined by a DMP, the parameter space of θ, and a set of weight samples $\{w_k\}_{k\in[K]}$.

First, let us consider the DMP whose objective functions are quadratic and feasible region is a ball in $\mathbb{R}^n$ that has the following form

$$\begin{aligned} \min_{\mathbf{x}\in\mathbb{R}^n} \quad & \begin{bmatrix} \frac{1}{2}\mathbf{x}^T\mathbf{x}+\mathbf{c}_1^T\mathbf{x} \\ \vdots \\ \frac{1}{2}\mathbf{x}^T\mathbf{x}+\mathbf{c}_K^T\mathbf{x} \end{bmatrix} \\ s.t. \quad & \|\mathbf{x}\|_2 \le 2\max_{i\in[N]}\|\mathbf{y}_i\|_2, \end{aligned} \tag{MQP}$$

where $\mathbf{c}_l\in\mathbb{R}^n$ for $l\in[K]$. The task of IMOP for MQP is to learn $\{\mathbf{c}_l\}_{l\in[K]}$ given $\{\mathbf{y}_i\}_{i\in[N]}$. Since the objective functions and the constraint are convex, MQP is a convex DMP.

Second, we let Θ be the parameter space that consists of $\mathbf{c}_l\in\mathbb{R}^n$ such that $\|\mathbf{c}_l\|_2\le\max_{i\in[N]}\|\mathbf{y}_i\|_2$ for each $l\in[K]$. One can readily check Θ defined in such a way is a convex and compact set.

Assume $\{\mathbf{x}_k\}_{k\in[K]}$ are the optimal centroids for $\{\mathbf{y}_i\}_{i\in[N]}$ in K-means clustering. Since each of the optimal centroid is the mean of the points in that cluster, we have $\mathbf{x}_k\le\max_{i\in[N]}\|\mathbf{y}_i\|_2$. Now, let $\mathbf{c}_k=-\mathbf{x}_k$, which is achievable because $\|\mathbf{c}_k\|_2\le\max_{i\in[N]}\|\mathbf{y}_i\|_2$ for each $k\in[K]$.

Third, we select $\{w_k\}_{k\in[K]}$ in the following way: the K weights are the K vertices of $\mathscr{W}_K$.

With a slight abuse of notation, let $\mathbf{x}_k=S(w_k,\mathbf{c}_1,\ldots,\mathbf{c}_K)$ be the optimal solution of WP for MQP for each $k\in[K]$. This mild abuse of notation allows us to express our results in a unified manner. It will be clear from context whether we are treating K-means clustering problem or IMOP-EMP-WS, and consequently be clear which definition of $\mathbf{x}_k$ we mean. Since the objective functions in MQP are strictly convex, each $\mathbf{x}_k$ is a Pareto optimal point by Proposition 1.

Now, we are ready to show the equivalence between K-means clustering and the constructed IMOP. Note that the only difference between the two problems is that $\{\mathbf{x}_k\}_{k\in[K]}$ for IMOP are restricted to be Pareto optimal points for some DMP, while no restriction is put on $\{\mathbf{x}_k\}_{k\in[K]}$ for K-means clustering. Thus, the optimal value of the K-means clustering provides a lower bound for the optimal value of the constructed IMOP. Then, it suffices to show that they have the same optimal value, which can be done by proving that the previously defined $\{\mathbf{c}_k\}_{k\in[K]}$ and the optimal centroids $\{\mathbf{x}_k\}_{k\in[K]}$ solve IMOP-EMP-WS.

Since the K weights are vertices of the simplex $\mathscr{W}_K$, each Pareto optimal point $\mathbf{x}_k$ corresponds to the unique optimal solution for one single objective optimization problem. More specifically,

$$\mathbf{x}_k=\arg\min_{\mathbf{x}}\{\frac{1}{2}\mathbf{x}^T\mathbf{x}+\mathbf{c}_k^T\mathbf{x}:\|\mathbf{x}\|_2\le 2\max_{i\in[N]}\|\mathbf{y}_i\|_2\}. \tag{30}$$

One can readily check that the previously defined $\mathbf{c}_k=-\mathbf{x}_k$ indeed make the optimal centroid $\mathbf{x}_k$ of the K-means clustering problem also an optimal solution in (30). It shows that the construed IMOP-EMP-WS is solved bt $\{\mathbf{c}_k\}_{k\in[K]}$ and the optimal centroids $\{\mathbf{x}_k\}_{k\in[K]}$.

To this end, we have shown that the optimal values of IMOP-EMP-WS we constructed and K-means clustering are indeed equal. Therefore, solving the IMOP-EMP-WS we constructed provides an optimal partition of $\{\mathbf{y}_i\}_{i\in[N]}$ for the K-means clustering. □

A.19. Proof of Theorem 8

Proof. Under our assumptions, for each Pareto optimal point $\mathbf{y}$, $\exists w \in \mathscr{W}_p$, s.t. $\mathbf{y} \in S(w,\theta)$ by PROPOSITION 2. Recall that $w \in \mathbb{R}^p_+, \mathbf{1}^T w = 1$, thus w lies in a $(p-1)$-d manifold with boundary. We show that the mapping $S(w,\theta): w \to \mathbf{y}$ for fixed θ is continuous in w in Lemma 4. Also, $\forall w_1, w_2 \in \mathscr{W}_p$, $w_1 \neq w_2$ implies $S(w_1,\theta) \neq S(w_2,\theta)$ for each $\theta \in \Theta$. Then, we have that $S(\cdot,\theta): \mathscr{W}_p \to X_P(\theta)$ is a one-to-one correspondence. Thus, $S(\cdot,\theta)$ is a homeomorphism by the inverse mapping theorem (Sutherland 2009), meaning that the inverse map $S^{-1}(\cdot,\theta): X_P(\theta) \to \mathscr{W}_p$ is also continuous. By the definition of manifold, the Pareto optimal set is a $(p-1)$-d manifold with boundary. □

A.20. Proof of Lemma 10

Proof. First, $M_K^N(\theta)$ decreases in the **Assignment step** since each $\mathbf{y}_i$ is assigned to the closest $\mathbf{x}_k$. So the distance $\mathbf{y}_i$ contributes to $M_K^N(\theta)$ decreases. Second, $M_K^N(\theta)$ decreases in the **Update step** because the new θ and $\{\mathbf{x}_k\}_{k\in[K]}$ are the ones for which $M_K^N(\theta)$ attains its minimum. □

Appendix B: Omitted Algorithms

B.1. ADMM for IMOP

IMOP-EMP-WS is closely related to the global consensus problem discussed heavily in Boyd et al. (2011), but with the important difference that IMOP-EMP-WS is a nonconvex problem. In order to use ADMM, we first partition $\{\mathbf{y}_i\}_{i\in[N]}$ equally into T groups, and denote $\{\mathbf{y}_i\}_{i\in[N_t]}$ the observations in t-th group. Then, we introduce a set of new variables $\{\theta^t\}_{t\in T}$, typically called local variables, and transform IMOP-EMP-WS equivalently to the following problem:

$$\begin{aligned} \min_{\theta\in\Theta,\theta^t\in\Theta} \ & \sum_{t\in T}\sum_{i\in[N_t]} l_K(\mathbf{y}_i,\theta^t) \\ \text{s.t.} \quad & \theta^t = \theta, \qquad\qquad \forall t \in [T]. \end{aligned} \tag{31}$$

ADMM for problem (31) can be derived directly from the augmented Lagrangian

$$L_\rho(\theta,\{\theta^t\}_{t\in[T]},\{\mathbf{v}^t\}_{t\in[T]}) = \sum_{t\in[T]}\left(\sum_{i\in[N_t]} l_K(\mathbf{y}_i,\theta^t) + <\mathbf{v}^t,\theta^t-\theta> + (\rho/2)\|\theta^t-\theta\|_2^2\right),$$

where $\rho>0$ is an algorithm parameter, $\mathbf{v}^t$ is the dual variable for the constraint $\theta^t=\theta$.

Let $\overline{\theta}^k = \frac{1}{|T|}\sum_{t\in T}\theta^{t,k}$. As suggested in Boyd et al. (2011), the primal and dual residuals are

$$r^k_{pri} = \left(\theta^{1,k}-\overline{\theta}^k,\ldots,\theta^{|T|,k}-\overline{\theta}^k\right), \quad r^k_{dual} = -\rho\left(\overline{\theta}^k-\overline{\theta}^{k-1},\ldots,\overline{\theta}^k-\overline{\theta}^{k-1}\right),$$

so their squared norms are

$$\|r^k_{pri}\|_2^2 = \sum_{t\in T}\|\theta^{t,k}-\overline{\theta}^k\|_2^2, \ \ \|r^k_{dual}\|_2^2 = |T|\rho^2\|\overline{\theta}^k-\overline{\theta}^{k-1}\|_2^2.$$

$\|r^k_{pri}\|_2^2$ is $|T|$ times the variance of $\{\theta^{t,k}\}_{t\in T}$, which can be interpreted as a natural measure of (lack of) consensus. Similarly, $\|r^k_{dual}\|_2^2$ is a measure of the step length. These suggest that a reasonable stopping criterion is that the primal and dual residuals must be small.

The resulting ADMM algorithm in scaled form is formally presented in the following.

Algorithm 3 ADMM for IMOP-EMP-WS

Input: Noisy decisions $\{\mathbf{y}_i\}_{i\in[N]}$, weight samples $\{w_k\}_{k\in[K]}$.

1: Set $k=0$ and initialize θ^0 and $\mathbf{v}^{t,0}$ for each $t\in T$.
2: **while** stopping criterion is not satisfied **do**
3: **for** $t\in[T]$ **do**
4: $\theta^{t,k+1}\leftarrow \arg\min_{\theta^t}\left\{\sum_{i\in N_t} l_K(\mathbf{y}_i,\theta^t)+(\rho/2)\|\theta^t-\theta^k+\mathbf{v}^{t,k}\|_2^2\right\}$.
5: **end for**
6: $\theta^{k+1}\leftarrow \frac{1}{|T|}\sum_{t\in T}\left(\theta^{t,k+1}+\mathbf{v}^{t,k}\right)$.
7: **for** $t\in[T]$ **do**
8: $\mathbf{v}^{t,k+1}\leftarrow \mathbf{v}^{t,k}+\theta^{t,k+1}-\theta^{k+1}$.
9: **end for**
10: $k\leftarrow k+1$.
11: **end while**

References

Aloise, Daniel, Amit Deshpande, Pierre Hansen, Preyas Popat. 2009. Np-hardness of euclidean sum-of-squares clustering. *Machine learning* **75**(2) 245–248.

Aloise, Daniel, Pierre Hansen. 2009. A branch-and-cut sdp-based algorithm for minimum sum-of-squares clustering. *Pesquisa Operacional* **29**(3) 503–516.

Bagirov, Adil M. 2008. Modified global k-means algorithm for minimum sum-of-squares clustering problems. *Pattern Recognition* **41**(10) 3192–3199.

Bartlett, Peter L, Shahar Mendelson. 2002. Rademacher and Gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research* **3**(Nov) 463–482.

Berge, Claude. 1963. *Topological Spaces: Including a Treatment of Multi-valued Functions, Vector Spaces, and Convexity*. Courier Corporation.

Biau, Gérard, Luc Devroye, Gábor Lugosi. 2008. On the performance of clustering in hilbert spaces. *IEEE Transactions on Information Theory* **54**(2) 781–790.

Boyd, Stephen, Neal Parikh, Eric Chu, Borja Peleato, Jonathan Eckstein. 2011. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends® in Machine Learning* **3**(1) 1–122.

Gass, Saul, Thomas Saaty. 1955. The computational algorithm for the parametric objective function. *Naval Research Logistics* **2**(1-2) 39–45.

Hogan, William W. 1973. Point-to-set maps in mathematical programming. *SIAM Review* **15**(3) 591–603.

Jennrich, Robert I. 1969. Asymptotic properties of non-linear least squares estimators. *The Annals of Mathematical Statistics* **40**(2) 633–643.

Lloyd, S. 1982. Least squares quantization in pcm. *IEEE Transactions on Information Theory* **28**(2) 129–137. doi:10.1109/TIT.1982.1056489.

Mahajan, Meena, Prajakta Nimbhorkar, Kasturi Varadarajan. 2012. The planar k-means problem is np-hard. *Theoretical Computer Science* **442** 13–21.

Markowitz, Harry. 1952. Portfolio selection. *The Journal of Finance* **7**(1) 77–91.

Mohri, Mehryar, Afshin Rostamizadeh, Ameet Talwalkar. 2012. *Foundations of Machine Learning*. MIT Press.

Roweis, Sam T, Lawrence K Saul. 2000. Nonlinear dimensionality reduction by locally linear embedding. *science* **290**(5500) 2323–2326.

Saul, Lawrence K, Sam T Roweis. 2003. Think globally, fit locally: unsupervised learning of low dimensional manifolds. *Journal of machine learning research* **4**(Jun) 119–155.

Schölkopf, Bernhard, Alexander Smola, Klaus-Robert Müller. 1997. Kernel principal component analysis. *International Conference on Artificial Neural Networks*. Springer, 583–588.

Sutherland, Wilson A. 2009. *Introduction to Metric and Topological Spaces*. Oxford University Press.

Tanino, T, Y Sawaragi. 1980. Stability of nondominated solutions in multicriteria decision-making. *Journal of Optimization Theory and Applications* **30**(2) 229–253.

Tenenbaum, Joshua B, Vin De Silva, John C Langford. 2000. A global geometric framework for nonlinear dimensionality reduction. *science* **290**(5500) 2319–2323.