Asymptotic Risk of Bézier Simplex Fitting

Akinori Tanaka
RIKEN AIP, Keio University
akinori.tanaka@riken.jp

Akiyoshi Sannai
RIKEN AIP, Keio University
akiyoshi.sannai@riken.jp

Ken Kobayashi
Fujitsu Laboratories LTD., RIKEN AIP, Tokyo Tech
ken-kobayashi@fujitsu.com

Naoki Hamada
Fujitsu Laboratories LTD., RIKEN AIP
hamada-naoki@fujitsu.com

Abstract

The Bézier simplex fitting is a novel data modeling technique which utilizes geometric structures of data to approximate the Pareto set of multi-objective optimization problems. There are two fitting methods based on different sampling strategies. The *inductive skeleton fitting* employs a stratified subsampling from skeletons of a simplex, whereas the *all-at-once fitting* uses a non-stratified sampling which treats a simplex as a single object. In this paper, we analyze the asymptotic risks of those Bézier simplex fitting methods and derive the optimal subsample ratio for the inductive skeleton fitting. It is shown that the inductive skeleton fitting with the optimal ratio has a smaller risk when the degree of a Bézier simplex is less than three. Those results are verified numerically under small to moderate sample sizes. In addition, we provide two complementary applications of our theory: a generalized location problem and a multi-objective hyper-parameter tuning of the group lasso. The former can be represented by a Bézier simplex of degree two where the inductive skeleton fitting outperforms. The latter can be represented by a Bézier simplex of degree three where the all-at-once fitting gets an advantage.

1 Introduction

Given functions $f_1, \ldots, f_M : X \to \mathbb{R}$ on a subset X of a Euclidean space $\mathbb{R}^N$, consider the multi-objective optimization problem

$$\begin{aligned} &\text{minimize} \;\; f(x) := (f_1(x), \ldots, f_M(x)) \\ &\text{subject to} \;\; x \in X (\subseteq \mathbb{R}^N) \end{aligned}$$

with respect to the Pareto ordering defined as follows:

$$x \prec y \overset{\text{def}}{\Longleftrightarrow} \forall i \left[f_i(x) \leq f_i(y) \right] \wedge \exists j \left[f_j(x) < f_j(y) \right].$$

The goal is to find the *Pareto set* and its image, called the *Pareto front*, which are denoted by

$$X^*(f) := \{ \, x \in X \mid \forall y \in X \left[y \not\prec x \right] \}$$

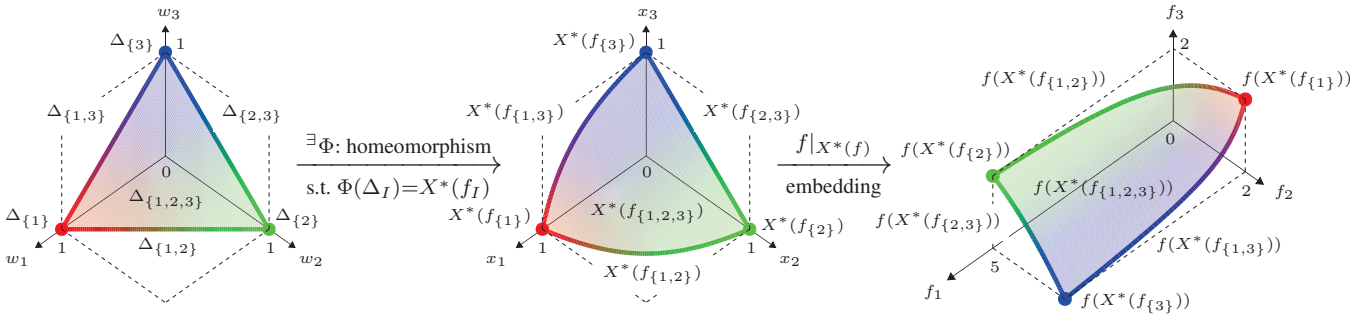
and

$$f(X^*(f)) := \left\{ \, f(x) \in \mathbb{R}^M \;\middle|\; x \in X^*(f) \, \right\},$$

respectively. Most numerical optimization approaches (e.g., goal programming (Miettinen 1999; Eichfelder 2008), evolutionary computation (Deb 2001; Zhang and Li 2007; Deb and Jain 2014), homotopy methods (Hillermeier 2001; Harada et al. 2007), Bayesian optimization (Hernandez-Lobato et al. 2016; Yang et al. 2019)) give a finite number of points as an approximation of the Pareto set or front. Since the Pareto set and front usually have an infinite number of points, such a point approximation cannot reveal the complete shapes of the Pareto set and front. In order to gain richer information, we consider in this paper a fitting problem of the Pareto set and front.

It is known that the Pareto set and front often have skeleton structures that can be used to enhance fitting accuracy. An M-objective problem is *simplicial* if the Pareto set and front are homeomorphic to an $(M-1)$-dimensional simplex and each m-dimensional subsimplex corresponds to the Pareto set of an $(m+1)$-objective subproblem for all $0 \leq m \leq M-1$ (see (Hamada et al. 2019) for precise definition and an example is shown in Figure 1). There are a lot of practical problems being simplicial: location problems (Kuhn 1967) and a phenotypic divergence model in evolutionary biology (Shoval et al. 2012) are shown to be simplicial, and an airplane design (Mastroddi and Gemma 2013) and a hydrologic modeling (Vrugt et al. 2003) have numerical solutions which imply those problems are simplicial. The Pareto set and front of any simplicial problem can be approximated with arbitrary accuracy by a Bézier simplex of an appropriate degree (Kobayashi et al. 2019). There are two fitting algorithms for Bézier simplices: the all-at-once fitting is a naïve extension of Borges-Pastva algorithm for Bézier curves (Borges and Pastva 2002), and the inductive skeleton fitting (Kobayashi et al. 2019) exploits the skeleton structure of simplicial problems discussed above.

An important problem class which is (generically) simplicial is strongly convex problems. It has been shown that many practical problems can be considered as strongly convex via appropriate transformations preserving the essential problem structure, i.e., the Pareto ordering and the topology (Hamada et al. 2019). For example, the multi-objective location problem (Kuhn 1967) becomes strongly convex by squaring each objective function. The resulting problem has



(a) Simplex Δ^2. (b) Pareto set $X^*(f)$. (c) Pareto front $f(X^*(f))$.

Figure 1: A simplicial problem $f = (f_1, f_2, f_3) : \mathbb{R}^3 \to \mathbb{R}^3$. An M-objective problem f is simplicial if the following conditions are satisfied: (i) there exists a homeomorphism $\Phi : \Delta^{M-1} \to X^*(f)$ such that $\Phi(\Delta_I) = X^*(f_I)$ for all $I \subseteq \{1, \ldots, M\}$; (ii) the restriction $f|_{X^*(f)} : X^*(f) \to \mathbb{R}^M$ is a topological embedding (and thus so is $f \circ \Phi : \Delta^{M-1} \to \mathbb{R}^M$).

a Pareto set that can be represented by a Bézier simplex of degree two (Hamada et al. 2019). As we will show in this paper, the group lasso (Yuan and Lin 2006) can be reformulated as a multi-objective simplicial problem. It has a twice-curving Pareto set that requires a Bézier simplex of degree three. The same reformulation can also be applied to a broad range of sparse modeling methods, including the (original) lasso (Tibshirani 1996), the fused lasso (Tibshirani et al. 2005), the smooth lasso (Hebiri and van de Geer 2011), and the elastic net (Zou and Hastie 2005). Since the required degree is observed to be problem-dependent, we need to understand the performance of the two Bézier simplex fittings with respect to the degree.

Moreover, use cases of the Bézier simplex fitting are not limited to post-optimal analysis. It can be applied to general data modeling problems as well. In the filed of evolutionary biology, (Shoval et al. 2012) showed that the phenotype of a species distributes like a curved simplex. Such a distribution can be modeled by a Bézier simplex for a better understanding of biological phenomena.

In this paper, we study the asymptotic risk of the two fitting methods of the Bézier simplex: the all-at-once fitting and the inductive skeleton fitting, and compare their performance with respect to the degree. While asymptotics on a Euclidean space (having no boundary) is well-studied, the Bézier simplex fitting is a regression method on a simplex (having a complex boundary, i.e., the skeleton), and its asymptotics have not been studied ever.

Our contributions are as follows:

- We have evaluated the asymptotic ℓ_2-risk, as the sample size tends to infinity, of two Bézier simplex fitting methods: the all-at-once fitting and the inductive skeleton fitting.
- In terms of minimizing the asymptotic risk, we have derived the optimal ratio of subsample sizes for the inductive skeleton fitting.
- We have shown when the inductive skeleton fitting with optimal ratio outperforms the all-at-once fitting when the degree of a Bézier simplex is two, whereas the all-at-once has an advantage at degree three.
- We have demonstrated that the location problem and the group lasso are transformed into strongly convex problems, and their Pareto sets and fronts are approximated by a Bézier simplex, which numerically verifies the asymptotic results.

The rest of this paper is organized as follows: Section 2 describes the problem definition. Section 3 analyzes the asymptotic risks of the all-at-once fitting and the inductive skeleton fitting. For the inductive skeleton fitting, the optimal subsample ratio in terms of minimizing the risk is derived. Those analyses are verified in Section 4 via numerical experiments. Section 5 concludes the paper and addresses future work.

2 Problem definition

Let M be a non-negative integer. The *standard* $(M-1)$*-simplex* is denoted by

$$\Delta^{M-1} = \left\{ (t_1, \ldots, t_M) \in \mathbb{R}^M \;\middle|\; \sum_{m=1}^{M} t_m = 1,\ t_m \geq 0 \right\}.$$

We define the *I-subsimplex* for an index set $I \subseteq \{1, \ldots, M\}$ by $\Delta_I^{M-1} = \{ (t_1, \ldots, t_M) \in \Delta^{M-1} \mid t_m = 0\ (m \notin I) \}$. In addition, the *$m$-skeleton* of Δ^{M-1} for an integer $0 \leq m \leq M-1$ is defined by

$$\Delta^{(m)} = \bigcup_{I \subseteq \{1,\ldots,M\} \text{ s.t. } |I|=m+1} \Delta_I^{M-1}.$$

2.1 Bézier simplex and its fitting methods

We denote the set of non-negative integers (including zero!) by $\mathbb{N}$. Let M, D, L be arbitrary integers in $\mathbb{N}$ and $\mathbb{N}_D^M := \{ (d_1, \ldots, d_M) \in \mathbb{N}^M \mid \sum_{m=1}^{M} d_m = D \}$. As shown in Figure 2, an $(M-1)$*-Bézier simplex of degree* D is a mapping $\boldsymbol{b} : \Delta^{M-1} \to \mathbb{R}^L$ determined by *control points* $\boldsymbol{p_d} \in \mathbb{R}^L$ $(\boldsymbol{d} \in \mathbb{N}_D^M)$ as follows:

$$\boldsymbol{b}(\boldsymbol{t}) := \sum_{\boldsymbol{d} \in \mathbb{N}_D^M} \binom{D}{\boldsymbol{d}} \boldsymbol{t}^{\boldsymbol{d}} \boldsymbol{p_d} \tag{1}$$

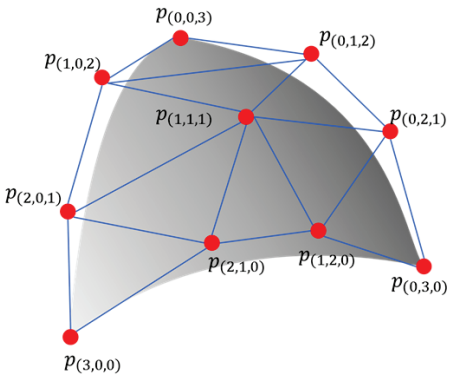


Figure 2: A Bézier simplex for $M = 3$, $D = 3$.

where $\binom{D}{\boldsymbol{d}} := \frac{D!}{d_1!d_2!\cdots d_M!}$ is a multinomial coefficient, and $\boldsymbol{t}^{\boldsymbol{d}} := t_1^{d_1} t_2^{d_2} \cdots t_M^{d_M}$ is a monomial (not vector) for each $\boldsymbol{t} := (t_1, \ldots, t_M) \in \Delta^{M-1}$ and $\boldsymbol{d} := (d_1, \ldots, d_M) \in \mathbb{N}_D^M$.

(Kobayashi et al. 2019) proposed two Bézier simplex fitting algorithms: the all-at-once fitting and the inductive skeleton fitting. They are different in not an only fitting algorithm but also sampling strategy. The all-at-once fitting requires a training set $S_N := \{ (\boldsymbol{t}_n, \boldsymbol{x}_n) \in \Delta^{M-1} \times \mathbb{R}^L \mid n = 1, \ldots, N \}$ and adjusts all control points at once by minimizing the OLS loss: $\frac{1}{N} \sum_{n=1}^{N} \|\boldsymbol{x}_n - \boldsymbol{b}(\boldsymbol{t}_n)\|^2$.

The inductive skeleton fitting, on the other hand, requires skeleton-wise sampled training sets $S_{N^{(m)}} := \{ (\boldsymbol{t}_n^{(m)}, \boldsymbol{x}_n^{(m)}) \in \Delta^{(m)} \times \mathbb{R}^L \mid n = 1, \ldots, N^{(m)} \}$ $(m = 0, \ldots, M-1)$. It also divides control points as $\boldsymbol{p}_{\boldsymbol{d}^{(m)}}$ such that $\boldsymbol{d}^{(m)}$ has exactly $m+1$ non-zero elements. Such $\boldsymbol{p}_{\boldsymbol{d}^{(m)}}$ determines the m-skeleton of a Bézier simplex. The inductive skeleton fitting inductively adjusts $\boldsymbol{p}_{\boldsymbol{d}^{(m)}}$ from $m = 0$ to $M-1$ by minimizing the OLS loss of the m-skeleton $\frac{1}{N^{(m)}} \sum_{n=1}^{N^{(m)}} \left\| \boldsymbol{x}_n^{(m)} - \boldsymbol{b}(\boldsymbol{t}_n^{(m)}) \right\|^2$.

2.2 The ℓ_2-risk

In this paper, we consider the following fitting problem: As Figure 3 illustrates, a sample point $(\boldsymbol{t}, \boldsymbol{x}) \in \Delta^{M-1} \times \mathbb{R}^L$ is taken from an unknown Bézier simplex $\boldsymbol{b} : \Delta^{M-1} \to \mathbb{R}^L$ with additive Gaussian noise $\boldsymbol{\varepsilon} \sim N(\boldsymbol{0}, \sigma^2 \boldsymbol{I})$, that is, $\boldsymbol{x} = \boldsymbol{b}(\boldsymbol{t}) + \boldsymbol{\varepsilon}$. For the all-at-once fitting, $S_N = \{ (\boldsymbol{t}_n, \boldsymbol{x}_n) \}$ follows the uniform distribution on the domain of the Bézier simplex: $\boldsymbol{t}_n \sim U(\Delta^{M-1})$ and $\boldsymbol{x}_n = \boldsymbol{b}(\boldsymbol{t}_n) + \boldsymbol{\varepsilon}_n$. For the inductive skeleton fitting, $S_{N^{(m)}} = \{ (\boldsymbol{t}_n^{(m)}, \boldsymbol{x}_n^{(m)}) \}$ follows the uniform distribution on the m-skeleton of the domain of the Bézier simplex: $\boldsymbol{t}_n^{(m)} \sim U(\Delta^{(m)})$ and $\boldsymbol{x}_n^{(m)} = \boldsymbol{b}(\boldsymbol{t}_n^{(m)}) + \boldsymbol{\varepsilon}_n^{(m)}$. A Bézier simplex estimated from S_N is denoted by $\hat{\boldsymbol{b}}(\boldsymbol{t}|S_N)$. For both method, we asymptotically evaluate the ℓ_2-risk below as $N \to \infty$.

$$R_N := \mathbb{E}_{S_N} \left[\mathbb{E}_{\boldsymbol{t} \sim U(\Delta^{M-1})} \left\| \boldsymbol{b}(\boldsymbol{t}) - \hat{\boldsymbol{b}}(\boldsymbol{t}|S_N) \right\|^2 \right]. \quad (2)$$

For the inductive skeleton fitting, we put $S_N = S_{N^{(0)}} \cup \cdots \cup S_{N^{(M-1)}}$ subject to $N = N^{(0)} + \cdots + N^{(M-1)}$.

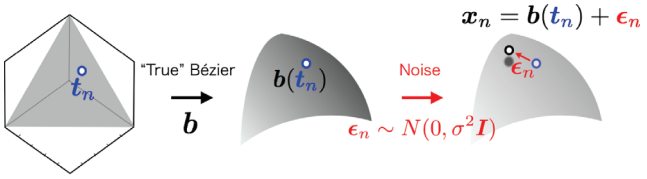


Figure 3: An illustration of taking a sample point on the true Bézier simplex with additive noise.

3 Asymptotic risk of Bézier simplex fitting

Let us first focus on the fact: the subtraction inside the ℓ_2-norm in (2) can be also written as a Bézier simplex:

$$\boldsymbol{b}(\boldsymbol{t}) - \hat{\boldsymbol{b}}(\boldsymbol{t}|S_N) = \sum_{A=1}^{\left|\mathbb{N}_D^M\right|} \binom{D}{\boldsymbol{d}_A} \boldsymbol{t}^{\boldsymbol{d}_A} \boldsymbol{p}'_{\boldsymbol{d}_A}. \quad (3)$$

Each A-th control point $\boldsymbol{p}'_{\boldsymbol{d}_A}$ of this Bézier simplex is defined by difference between the target control point $\boldsymbol{p}_{\boldsymbol{d}_A}$ and the model control point $\boldsymbol{p}_{\hat{\boldsymbol{d}}_A}(S_N)$, i.e. $\boldsymbol{p}'_{\boldsymbol{d}_A} = \boldsymbol{p}_{\boldsymbol{d}_A} - \hat{\boldsymbol{p}}_{\boldsymbol{d}_A}(S_N)$. To simplify the summation notation, we introduce a size $\left|\mathbb{N}_D^M\right| \times L$ matrix $\boldsymbol{P}$ composed by the l-th element of the A-th control point and a column vector $\boldsymbol{z}$:

$$(\boldsymbol{P})_{Al} = (\boldsymbol{p}'_{\boldsymbol{d}_A})_l, \ \boldsymbol{z}^\top = \left[\binom{D}{\boldsymbol{d}_1} \boldsymbol{t}^{\boldsymbol{d}_1}, \ldots, \binom{D}{\boldsymbol{d}_{\left|\mathbb{N}_D^M\right|}} \boldsymbol{t}^{\boldsymbol{d}_{\left|\mathbb{N}_D^M\right|}} \right]. \quad (4)$$

Note that $\boldsymbol{t}^{\boldsymbol{d}} = t_1^{d_1} t_2^{d_2} \cdots t_M^{d_M}$ is scalar, and $\boldsymbol{z}$ is a vector. Then, the Bézier simplex (3) can be represented by column vector $\boldsymbol{P}^\top \boldsymbol{z}$, and its squared norm is equal to $\left\|\boldsymbol{P}^\top \boldsymbol{z}\right\|^2 = \boldsymbol{z}^\top \boldsymbol{P}\boldsymbol{P}^\top \boldsymbol{z}$, or $\sum_{A,B} z_A z_B (\boldsymbol{P}\boldsymbol{P}^\top)_{AB}$ in component. The risk (2) is defined by an expectation value of this norm. $\mathbb{E}_{\boldsymbol{t}}$ only acts to $\boldsymbol{z}$ and $\mathbb{E}_{S_N}$ only acts to $\boldsymbol{P}$. Therefore, we arrive at

$$R_N = \sum_{\boldsymbol{d}_A, \boldsymbol{d}_B \in \mathbb{N}_D^M} \Sigma_{AB} \mathbb{E}_{S_N} \left[(\boldsymbol{P}\boldsymbol{P}^\top)_{AB} \right], \quad (5)$$

where

$$\Sigma_{AB} = \mathbb{E}_{\boldsymbol{t}}[z_A z_B] = \binom{D}{\boldsymbol{d}_A} \binom{D}{\boldsymbol{d}_B} \mathbb{E}_{\boldsymbol{t}}[\boldsymbol{t}^{\boldsymbol{d}_A + \boldsymbol{d}_B}]. \quad (6)$$

We can get closed form of Σ_{AB} by performing integral $\mathbb{E}_{\boldsymbol{t}}$ explicitly. The following theorem provides the result.

Theorem 1 *The matrix element Σ_{AB} is calculated by*

$$\Sigma_{AB} = \binom{2D + M - 1}{M - 1}^{-1} \binom{D}{\boldsymbol{d}_A} \binom{D}{\boldsymbol{d}_B} \binom{2D}{\boldsymbol{d}_A + \boldsymbol{d}_B}^{-1} \quad (7)$$

The proof is provided in the supplementary materials. [1]. The equation (5) means that the asymptotic value of the risk function depends only on a choice of the matrix $\boldsymbol{P}$.

[1] A longer version of this paper including appendix is available at https://arxiv.org/abs/1906.06924

3.1 All-at-once fitting

The matrix $\boldsymbol{P}$ determined by the all-at-once fitting algorithm, $\boldsymbol{P}_{\mathrm{AAO}}$, is minimizing the OLS loss:

$$\frac{1}{N}\sum_{n=1}^{N}\Big\|\underbrace{\boldsymbol{b}(\boldsymbol{t}_n)+\boldsymbol{\varepsilon}_n}_{\boldsymbol{x}_n}-\hat{\boldsymbol{b}}(\boldsymbol{t}_n)\Big\|^2=\frac{1}{N}\sum_{n=1}^{N}\left\|\boldsymbol{P}^\top\boldsymbol{z}_n+\boldsymbol{\varepsilon}_n\right\|^2$$
$$=\frac{1}{N}\left\|\boldsymbol{Z}\boldsymbol{P}+\boldsymbol{Y}\right\|_{\mathrm{F}}^2, \quad (8)$$

where $\|\cdot\|_{\mathrm{F}}$ is the Frobenius norm. Here, we introduced an $N\times\left|\mathbb{N}_D^M\right|$ matrix $\boldsymbol{Z}$ and an $N\times L$ matrix $\boldsymbol{Y}$:

$$\boldsymbol{Z}=\left[\boldsymbol{z}_1\boldsymbol{z}_2\cdots\boldsymbol{z}_N\right]^\top,\ \boldsymbol{Y}=\left[\boldsymbol{\varepsilon}_1\boldsymbol{\varepsilon}_2\cdots\boldsymbol{\varepsilon}_N\right]^\top. \quad (9)$$

Minimizing (8) is a traditional problem and we get $\boldsymbol{P}_{\mathrm{AAO}}=-\left(\boldsymbol{Z}^\top\boldsymbol{Z}\right)^{-1}\boldsymbol{Z}^\top\boldsymbol{Y}$. Note that $\boldsymbol{Z}$ includes N sample points on Δ^{M-1} and $\boldsymbol{Y}$ is a set of N noises on $\mathbb{R}^L$. These are all independent, so the expectation $\mathbb{E}_{S_N}$ can be factorized to $\mathbb{E}_{\boldsymbol{Z}}\mathbb{E}_{\boldsymbol{Y}}$.

Calculation of the asymptotics We need to calculate the expectation value of the matrix $\boldsymbol{P}_{\mathrm{AAO}}\boldsymbol{P}_{\mathrm{AAO}}^\top=\left(\boldsymbol{Z}^\top\boldsymbol{Z}\right)^{-1}\boldsymbol{Z}^\top\boldsymbol{Y}\boldsymbol{Y}^\top\boldsymbol{Z}\left(\boldsymbol{Z}^\top\boldsymbol{Z}\right)^{-1}$ over $\boldsymbol{Z}$ and $\boldsymbol{Y}$. As easily checked, $\mathbb{E}_{\boldsymbol{Y}}[\boldsymbol{Y}\boldsymbol{Y}^\top]=\sigma^2L\boldsymbol{I}_{N\times N}$, so we get

$$\mathbb{E}_{S_N}\left[\boldsymbol{P}_{\mathrm{AAO}}\boldsymbol{P}_{\mathrm{AAO}}^\top\right]=\sigma^2L\cdot\mathbb{E}_{\boldsymbol{Z}}\left[\left(\boldsymbol{Z}^\top\boldsymbol{Z}\right)^{-1}\right]. \quad (10)$$

Now, the matrix $\left(\boldsymbol{Z}^\top\boldsymbol{Z}\right)$ is an average over the sample:

$$\frac{1}{N}\left(\boldsymbol{Z}^\top\boldsymbol{Z}\right)_{AB}=\binom{D}{\boldsymbol{d}_A}\binom{D}{\boldsymbol{d}_B}\sum_{n=1}^{N}\frac{1}{N}\boldsymbol{t}_n^{\boldsymbol{d}_A+\boldsymbol{d}_B}, \quad (11)$$

and it converges to the matrix Σ_{AB} defined in (6) and (7) as $N\to\infty$ by using the law of large numbers: $\left(\boldsymbol{Z}^\top\boldsymbol{Z}\right)_{AB}\xrightarrow{p}N\Sigma_{AB}$. To substitute it to (10), however, we need to guarantee Σ_{AB} has the inverse matrix. We can show it by the following theorem.

Theorem 2 *Let $V_{M,D}$ be a vector space spanned by*

$$Z=\left\{\left.\frac{\boldsymbol{t}^{\boldsymbol{d}}}{\boldsymbol{d}!}\ \right|\ \boldsymbol{d}=(d_1,\ldots,d_M)\in\mathbb{N}_D^M\right\}.$$

Then the map

$$\begin{array}{ccc} L: V_{M,D}\times V_{M,D} & \longrightarrow & \mathbb{R} \\ \cup & & \cup \\ (P,Q) & \longmapsto & \int_{\Delta^{M-1}}P(\boldsymbol{t})Q(\boldsymbol{t})d\boldsymbol{t} \end{array},$$

is a non-degenerate bilinear form. Moreover, the matrix corresponding to this bilinear form is Σ_{AB} in (7)*. In particular, for any D, M, the matrix Σ_{AB} is non-singular.*

The precise proof is given in the supplementary materials. In summary, our formula for the asymptotic form of the risk for the all-at-once fitting is

$$R_N\xrightarrow{p}\frac{\sigma^2L}{N}\sum_{A,B}\Sigma_{AB}\Sigma_{AB}^{-1}=\frac{\sigma^2L}{N}\binom{D+M-1}{D}\quad(N\to\infty).$$

We can further simplify the result by using: $\sum_{AB}\Sigma_{AB}\Sigma_{AB}^{-1}=\left|\mathbb{N}_D^M\right|=\binom{D+M-1}{D}$, which is relatively easy to show (see the supplementary materials).

3.2 Inductive skeleton fitting

So far, we did not take any explicit order of the control point indices A. From now on, let us take a specific order

$$\boldsymbol{P}^\top=[\boldsymbol{P}^{(0)\top}\boldsymbol{P}^{(1)\top}\cdots\boldsymbol{P}^{(M-1)\top}], \quad (12)$$

where $\boldsymbol{P}^{(m)}$ is the submatrix of $\boldsymbol{P}$ composed by control points on $\Delta^{(m)}$. Similarly, we introduce an order of control point indices $\boldsymbol{d}_A$ as follows:

$$[\boldsymbol{d}_1^{(0)},\ldots,\boldsymbol{d}_{n_0}^{(0)},\boldsymbol{d}_1^{(1)},\ldots,\boldsymbol{d}_{n_1}^{(1)},\ldots,\boldsymbol{d}_1^{(M-1)},\ldots,\boldsymbol{d}_{n_{M-1}}^{(M-1)}],$$

where $\boldsymbol{d}_n^{(m)}$ is the n-th index of control points on the m-skeleton and n_m is the number of control points on the m-skeleton. The inductive skeleton fitting is described by an inductive procedure of determining control points matrices $\boldsymbol{P}^{(m)}$ from low $m=0,1,\ldots,M-1$. That is, first it fits the vertices of a Bézier simplex by moving the control points of the lowest dimension ($\boldsymbol{P}^{(0)}$); then, it fits the edges by moving the control points of the second lowest dimension ($\boldsymbol{P}^{(1)}$); this process goes on with increasing dimensions and finishes at the highest dimension ($\boldsymbol{P}^{(M-1)}$). In the m-th step, sample points $\boldsymbol{t}^{(m)}$ on $\Delta^{(m)}$ are given. The corresponding $\boldsymbol{z}^{(m)}$ defined in (4) has the following form:

$$\boldsymbol{z}^{(m)\top}=[\boldsymbol{z}^{(m)[0]\top}\boldsymbol{z}^{(m)[1]\top}\cdots\boldsymbol{z}^{(m)[m]\top}\boldsymbol{0}\cdots\boldsymbol{0}],\ \text{where}$$
$$\boldsymbol{z}^{(m)[k]\top}=\left[\binom{D}{\boldsymbol{d}_1^{(k)}}(\boldsymbol{t}^{(m)})^{\boldsymbol{d}_1^{(k)}},\ldots,\binom{D}{\boldsymbol{d}_{n_k}^{(k)}}(\boldsymbol{t}^{(m)})^{\boldsymbol{d}_{n_k}^{(k)}}\right],$$

because $(\boldsymbol{t}^{(m)})^{\boldsymbol{d}^{(k>m)}}$ includes $0^{d_\bullet^{(k)}\neq0}=0$ by definition. Thanks to these zeros, the OLS loss reduces as follows:

$$\sum_{n=1}^{N^{(m)}}\left\|(\boldsymbol{P}^{(m)})^T\boldsymbol{z}_n^{(m)[m]}+\sum_{k<m}(\boldsymbol{P}^{(k)})^T\boldsymbol{z}_n^{(m)[k]}+\boldsymbol{\varepsilon}_n^{(m)}\right\|^2$$
$$=\left\|\boldsymbol{Z}^{(m)[m]}\boldsymbol{P}^{(m)}+\sum_{k<m}\boldsymbol{Z}^{(m)[k]}\boldsymbol{P}^{(k)}+\boldsymbol{Y}^{(m)}\right\|_{\mathrm{F}}. \quad (13)$$

Each matrix is defined as follows.

$$\boldsymbol{Z}^{(m)[k]}=\left[\boldsymbol{z}_1^{(m)[k]}\cdots\boldsymbol{z}_{N^{(m)}}^{(m)[k]}\right]^\top,\boldsymbol{Y}^{(m)}=\left[\boldsymbol{\varepsilon}_1^{(m)}\cdots\boldsymbol{\varepsilon}_{N^{(m)}}^{(m)}\right]^\top.$$

In addition, we regard lower-dimensional control points already fixed, so the net objective control points are ones included in $\boldsymbol{P}^{(m)}$. By repeating similar procedure done in the all-at-once fitting, we can conclude $\boldsymbol{P}^{(m)}$ is determined as

$$\boldsymbol{P}_{\mathrm{ISK}}^{(m)}=-[(\boldsymbol{Z}^{(m)})^\top\boldsymbol{Z}^{(m)}]^{-1}(\boldsymbol{Z}^{(m)})^\top$$
$$\times\left(\boldsymbol{Y}^{(m)}+\sum_{k<m}\boldsymbol{Z}^{(m)[k]}\boldsymbol{P}_{\mathrm{ISK}}^{(k)}\right) \quad (14)$$

Calculation of the asymptotics From this expression, we get $\boldsymbol{P}_{\mathrm{ISK}}\boldsymbol{P}_{\mathrm{ISK}}{}^\top=\oplus_{i,j=0}^{M-1}\boldsymbol{P}_{\mathrm{ISK}}^{(i)}(\boldsymbol{P}_{\mathrm{ISK}}^{(j)})^\top$. The expected value of each (i,j)-term is needed to evaluate the risk (5). The following theorem provides us an algorithm to asymptotically calculate the expectation.

Theorem 3 *Let* $I_{\boldsymbol{d}} = \{\, i \mid d_i \neq 0 \,\} \subseteq \{\, 1, \ldots, M \,\}$ *and* $\boldsymbol{\Lambda}^{(m)[k]}$ *be an* $n_m \times n_k$ *matrix defined by*

$$(\boldsymbol{\Lambda}^{(m)[k]})_{\boldsymbol{d}^{(m)}\boldsymbol{d}^{(k)}} = 1_{I_{\boldsymbol{d}^{(m)}} \supseteq I_{\boldsymbol{d}^{(k)}}} \binom{M}{m+1}^{-1} \Sigma^{(m)}_{\boldsymbol{d}^{(m)}\boldsymbol{d}^{(k)}},$$

$$\textit{where } \Sigma^{(m)}_{\boldsymbol{d}_A\boldsymbol{d}_B} := \binom{2D+m}{m}^{-1} \binom{D}{\boldsymbol{d}_A}\binom{D}{\boldsymbol{d}_B}\binom{2D}{\boldsymbol{d}_A+\boldsymbol{d}_B}^{-1},$$

and $1_X = 1$ *if* X *is true, otherwise 0. Then we get the asymptotic submatrix* $\boldsymbol{X}^{(i)(j)}$

$$\mathbb{E}_{S_N}\left[\boldsymbol{P}^{(i)}_{\mathrm{ISK}}(\boldsymbol{P}^{(j)}_{\mathrm{ISK}})^\top\right] \xrightarrow{p} \boldsymbol{X}^{(i)(j)} := \sigma^2 L \sum_{\substack{m\le i\\ m\le j}} \sum_{\substack{m\le k_1<\cdots<k_\heartsuit<i\\ m\le l_1<\cdots<l_\spadesuit<j}} \frac{(-1)^{\heartsuit+\spadesuit}}{N^{(m)}} \boldsymbol{\Lambda}_\heartsuit \boldsymbol{\Lambda}_{(m)} \boldsymbol{\Lambda}_\spadesuit \quad (15)$$

where the summation runs for all possible increasing sequences $[k_1, \ldots, k_\heartsuit]$ *and* $[l_1, \ldots, l_\spadesuit]$*, and*

$$\begin{aligned}
&\boldsymbol{\Lambda}_\heartsuit = \boldsymbol{\Lambda}_{(i)}\boldsymbol{\Lambda}^{(i)[k_\heartsuit]}\boldsymbol{\Lambda}_{(k_\heartsuit)}\cdots\boldsymbol{\Lambda}^{(k_1)[m]},\\
&\boldsymbol{\Lambda}_\spadesuit = \boldsymbol{\Lambda}^{[m](l_1)}\cdots\boldsymbol{\Lambda}_{(l_\spadesuit)}\boldsymbol{\Lambda}^{[l_\spadesuit](j)}\boldsymbol{\Lambda}_{(j)},\\
&\boldsymbol{\Lambda}^{[k](m)} = (\boldsymbol{\Lambda}^{(m)[k]})^\top, \quad \boldsymbol{\Lambda}_{(m)} = (\boldsymbol{\Lambda}^{(m)[m]})^{-1}. \quad (16)
\end{aligned}$$

For the complete derivation, see the supplementary materials. The asymptotic form of the risk for the inductive skeleton fitting is, therefore, calculated by

$$R_{N^{(0)},\ldots,N^{(M-1)}} = \sum_{\boldsymbol{d}_A,\boldsymbol{d}_B\in\mathbb{N}^M_D} \Sigma_{AB}\Big(\bigoplus_{i,j=0}^{M-1}\boldsymbol{X}^{(i)(j)}\Big)_{AB}.$$

We found a candidate of closed-form with $M = 2$ risks as $(D-1)/N^{(1)} + 4/D(D+2)N^{(0)}$, but postpone deriving the closed-form of it for arbitrary (D, M) for future work. Instead, we show numerically computed risks in Table 1.

Table 1: Numerically computed asymptotic risks of the inductive skeleton fitting $R_{N^{(0)},\ldots,N^{(M-1)}}$ (M: the dimension of the Bézier simplex, D: the degree of the Bézier simplex, $N^{(m)}$: the sample size for the m-skeleton).

M	$D=2$	$D=3$
2	$\frac{1.00}{N^{(1)}} + \frac{0.50}{N^{(0)}}$	$\frac{2.00}{N^{(1)}} + \frac{0.27}{N^{(0)}}$
3	$\frac{3.00}{N^{(1)}} + \frac{0.38}{N^{(0)}}$	$\frac{1.00}{N^{(2)}} + \frac{3.54}{N^{(1)}} + \frac{0.15}{N^{(0)}}$
4	$\frac{5.14}{N^{(1)}} + \frac{0.46}{N^{(0)}}$	$\frac{5.33}{N^{(2)}} + \frac{4.70}{N^{(1)}} + \frac{0.17}{N^{(0)}}$
5	$\frac{7.14}{N^{(1)}} + \frac{0.64}{N^{(0)}}$	$\frac{13.33}{N^{(2)}} + \frac{6.67}{N^{(1)}} + \frac{0.21}{N^{(0)}}$
6	$\frac{8.93}{N^{(1)}} + \frac{0.82}{N^{(0)}}$	$\frac{24.24}{N^{(2)}} + \frac{9.74}{N^{(1)}} + \frac{0.26}{N^{(0)}}$
7	$\frac{10.50}{N^{(1)}} + \frac{1.02}{N^{(0)}}$	$\frac{37.10}{N^{(2)}} + \frac{13.84}{N^{(1)}} + \frac{0.31}{N^{(0)}}$
8	$\frac{11.87}{N^{(1)}} + \frac{1.21}{N^{(0)}}$	$\frac{51.17}{N^{(2)}} + \frac{18.73}{N^{(1)}} + \frac{0.37}{N^{(0)}}$

3.3 All-at-once vs. Inductive skeleton

Given a total sample size N, we can minimize the ISK-risk by finding the optimally-decoupled subsample sizes:

$$R_N := \min_{\substack{N^{(0)},\ldots,N^{(M-1)}\\ N=N^{(0)}+\cdots+N^{(M-1)}}} \left\{\, R_{N^{(0)},\ldots,N^{(M-1)}} \,\right\}. \quad (17)$$

We calculated optimal risks for all cases shown in Table 1 and compared them to the risks of the all-at-once fitting. Table 2 shows the results.

Table 2: Comparison of asymptotic risks of the all-at-once R^{AAO}_N vs. the inductive skeleton with the optimal subsample ratio R^{ISK}_N (M: the dimension of the Bézier simplex, D: the degree of the Bézier simplex, N: the sample size). The winner is shown in bold.

	$D=2$		$D=3$	
	R^{AAO}_N	R^{ISK}_N	R^{AAO}_N	R^{ISK}_N
$M=2$	$3.0/N$	$\mathbf{2.91}/N$	$4.0/N$	$\mathbf{3.73}/N$
$M=3$	$6.0/N$	$\mathbf{5.50}/N$	$\mathbf{10.0}/N$	$10.650/N$
$M=4$	$10.0/N$	$\mathbf{8.67}/N$	$\mathbf{20.0}/N$	$23.88/N$
$M=5$	$15.0/N$	$\mathbf{11.99}/N$	$\mathbf{35.0}/N$	$44.76/N$
$M=6$	$21.0/N$	$\mathbf{15.17}/N$	$\mathbf{56.0}/N$	$73.14/N$
$M=7$	$28.0/N$	$\mathbf{18.07}/N$	$\mathbf{84.0}/N$	$107.57/N$
$M=8$	$36.0/N$	$\mathbf{20.68}/N$	$\mathbf{120.0}/N$	$146.21/N$

As one can see, the optimum inductive skeleton fitting outperforms the all-at-once fitting for $D = 2$, but it is not always true for $D = 3$. On $D = 2$, in fact, we can show that the minimum value of the inductive skeleton always less than the asymptotic risk of the corresponding all-at-one fitting.

4 Numerical examples

We examine the empirical performances of the all-at-once fitting and the inductive skeleton fitting and verify the asymptotic risks derived in Section 3.1 over synthetic instances and multi-objective optimization instances. Experiment programs were implemented in Python 3.7.1 and run on a Windows 7 PC with an Intel Core i7-4790CPU (3.60 GHz) and 16 GB RAM[2].

4.1 Synthetic instances

We consider the fitting problem where the true Bézier simplex $\boldsymbol{b}(\boldsymbol{t})$ ($\boldsymbol{t} \in \Delta^{M-1}$) is an $(M-1)$-dimensional unit simplex on $\mathbb{R}^L$, and randomly generate N training points $\{(\boldsymbol{t}_n, \boldsymbol{x}_n)\}_{n=1}^N$ as $\boldsymbol{x}_n = \boldsymbol{b}(\boldsymbol{t}_n) + \boldsymbol{\varepsilon}_n$ ($\boldsymbol{\varepsilon}_n \sim N(\boldsymbol{0}, 0.1^2\boldsymbol{I})$). This synthetic instance is parameterized by a tuple (L, M, N). The detailed data generation processes are shown in the supplementary materials.

In this experiment, we estimated the Bézier simplex with degree $D = 2$ or 3, and compared the following three fitting methods:

all-at-once the all-at-once fitting (Section 3.1);

[2]The source code and library dependencies are provided in https://github.com/rafcc/aaai-20.1534.

inductive skeleton (non-optimal) the inductive skeleton fitting (Section 3.2) with $N^{(0)} = \cdots = N^{(M-1)} = N/M$, which does not provide the optimal value of the risk shown in Table 1;

inductive skeleton (optimal) the inductive skeleton fitting (Section 3.2) where $N^{(0)}, \ldots, N^{(M-1)}$ are determined by minimizing the risk shown in Table 1 under the constraints $\sum_{m=0}^{M-1} N^{(m)} = N$ and $N^{(m)} \geq 0$ $(m = 0, \ldots, M-1)$. The actual sample size $N^{(m)}$ for each (D, M) are shown in the supplementary materials.

When we calculated an approximation of the expected risk for each method, we randomly chose other 10000 parameters $\{\hat{\boldsymbol{t}}_n\}_{n=1}^{10000}$ from $U(\Delta^{M-1})$ as a test set and measured the mean squared error, $\mathrm{MSE} := \frac{1}{10000}\sum_{n=1}^{10000}\left\|\boldsymbol{b}(\hat{\boldsymbol{t}}_n) - \hat{\boldsymbol{b}}(\hat{\boldsymbol{t}}_n)\right\|^2$, where $\hat{\boldsymbol{b}}$ is the estimated Bézier simplex. This experiment was conducted with the following tuple (L, M, N) to observe how the empirical MSEs depend on L, M and N respectively:

- $N \in \{\, 250, 500, 1000, 2000 \,\}$ with $(L, M) = (100, 8)$,
- $M \in \{\, 3, 4, 5, 6, 7, 8 \,\}$ with $(L, N) = (100, 1000)$,
- $L \in \{\, 8, 25, 50, 100 \,\}$ with $(M, N) = (8, 1000)$,

For each (L, M, N) with $D \in \{\, 2, 3 \,\}$, we ran 20 trials and measured MSEs.

Owing to space limitation, we only present typical results here. The remaining results are provided in the supplementary materials. Figure 4 shows box plots of MSEs over 20 trials and our theoretical risks (5) and Table 1 for each $N \in \{\, 250, 500, 1000, 2000 \,\}$ with $(L, M) = (100, 8)$ and $D \in \{\, 2, 3 \,\}$. We observe that these figures empirically show that our theoretical risks are correct for both $D = 2$ and 3, and the gaps between the MSEs and the risks are sufficiently small at $N = 1000$. For both $D = 2$ and 3, the inductive skeleton (optimal) always achieved lower MSEs than that of the inductive skeleton (non-optimal). This result suggests the effectiveness of minimizing the risk (Table 2) with respect to the sample size of each skeleton. In addition, the inductive skeleton fitting (optimal) also outperformed the all-at-once fitting in the case of $D = 2$. This result also supports the discussion described in Section 3.3.

4.2 Multi-objective optimization instances

To investigate the relationship between the generalization performance and our theoretical risk, we provide two complementary instances of multi-objective optimization problems: a generalized location problem called `MED` (Harada, Sakuma, and Kobayashi 2006; Hamada et al. 2010) and a multi-objective hyper-parameter tuning of the group lasso (Yuan and Lin 2006) on the `Birthwt` dataset (Hosmer and Lemeshow 1989; Venables and Ripley 2002). The location problem has 3 objectives and 100 variables (that is $(M, L) = (3, 100)$). Its Pareto set/front can be represented by a Bézier simplex with degree $D = 2$. On the other hand, the group lasso has 3 objectives and 6 variables (that is $(M, L) = (3, 6)$). Its Pareto set/front cannot be represented with degree $D = 2$ but can be with $D = 3$ (see the supplementary materials). We will describe the details of problem settings in the subsequent sections.

A generalized location problem This problem is a generalization of the multi-objective location problem (Kuhn 1967) to a higher dimension:

$$\begin{aligned}&\text{minimize } f(x) = (f_1(x), f_2(x), f_3(x)) \text{ subject to } x \in \mathbb{R}^{100}\\ &\quad \text{where } f_m(x) = \|x - e_m\|^2 \quad (m = 1, \ldots, 3)\\ &\qquad e_1 = (1, 0, 0, 0, 0, \ldots, 0) \in \mathbb{R}^{100},\\ &\qquad e_2 = (0, 1, 0, 0, 0, \ldots, 0) \in \mathbb{R}^{100},\\ &\qquad e_3 = (0, 0, 1, 0, 0, \ldots, 0) \in \mathbb{R}^{100}.\end{aligned} \tag{18}$$

Note that this is a special case of the MED benchmark problem (Hamada et al. 2010). The MED problem is simplicial (Hamada 2017) and its Pareto set is known to be the convex hull of the minimizers of separate objective functions, i.e., the 2-simplex spanned by e_1, e_2, e_3. For each vertex, edge, face of this simplex, which is the Pareto set of each 1-, 2-, 3-objective subproblem, we generate a subsample according to the uniform distribution on it.

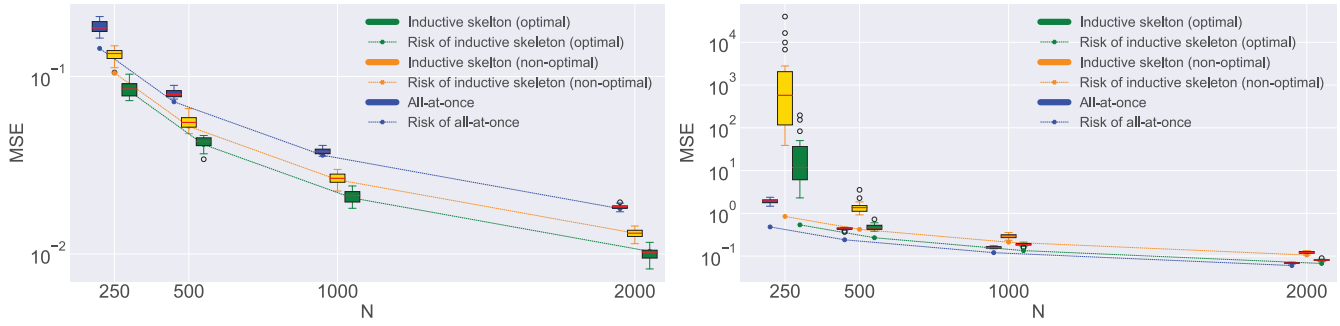
The group lasso We applied the Bézier simplex fittings to multi-objective hyper-parameter tuning of the group lasso. In this problem, we used the dataset, `Birthwt` in the R-package `MASS`, which contains 189 births at the Baystate Medical Centre, Springfield, Massachusetts during 1986 (Hosmer and Lemeshow 1989; Venables and Ripley 2002). From the dataset, we adopted six continuous features `age1`, `age2`, `age3`, `lwt1`, `lwt2`, `lwt3` as predictors and one continuous feature `bwt` as a response for regression analysis. Since the predictors are classified into two groups, `age` and `lwt`, the group lasso (Yuan and Lin 2006) was employed.

Put $N = 189$ and $M = 6$. Let A be an $N \times M$ matrix of observations of the predictors, $x \in \mathbb{R}^M$ be a row vector of the predictor coefficients to be estimated, separated into two groups $x_{\text{age}} = (x_1, x_2, x_3)^\top$ and $x_{\text{lwt}} = (x_4, x_5, x_6)^\top$, and $y \in \mathbb{R}^N$ be a row vector of observations of the response. The group lasso regressor is the solution to the following problem:

$$\begin{aligned}&\text{minimize } \frac{1}{2N}\|Ax - y\|^2 + \frac{\lambda}{\sqrt{3}}\left(\|x_{\text{age}}\| + \|x_{\text{lwt}}\|\right)\\ &\text{subject to } x \in \mathbb{R}^6\end{aligned} \tag{19}$$

where $\|\cdot\|$ is the Euclidean norm, and λ is a positive number to be tuned by users. This original form suffers from two drawbacks:

- Choosing an appropriate value for λ involves a grid search on an unbounded domain.
- Since two groups have physically different units of measurement, same weights are not always appropriate even if their values are normalized.



(a) $D = 2$ (b) $D = 3$

Figure 4: Sample size N vs. MSE with $(L, M) = (100, 8)$ (boxplots: empirical MSEs over 20 trials, lines: theoretical risks).

Table 3: MSE (avg. $\pm$ s.d. over 20 trials) for the Pareto sets of the location problem and the group lasso. The winners with significance level $p < 0.05$ are shown in bold.

(a) Location problem

D	N	All-at-once	Inductive-skeleton (optimal)
2	250	1.0246e+00 $\pm$ 1.7031e-03	**1.0227e+00 $\pm$ 2.3045e-03**
	500	1.0119e+00 $\pm$ 5.3916e-04	**1.0108e+00 $\pm$ 7.5912e-04**
	1000	1.0060e+00 $\pm$ 2.6182e-04	**1.0055e+00 $\pm$ 4.7063e-04**
	2000	1.0029e+00 $\pm$ 1.8406e-04	1.0028e+00 $\pm$ 2.7489e-04
3	250	**1.0430e+00 $\pm$ 3.3097e-03**	1.0458e+00 $\pm$ 4.4068e-03
	500	**1.0203e+00 $\pm$ 1.1845e-03**	1.0219e+00 $\pm$ 1.4369e-03
	1000	**1.0100e+00 $\pm$ 3.9162e-04**	1.0112e+00 $\pm$ 6.2094e-04
	2000	**1.0049e+00 $\pm$ 2.3927e-04**	1.0056e+00 $\pm$ 3.4550e-04

(b) Group lasso

D	N	All-at-once	Inductive-skeleton (optimal)
2	250	**2.9440e-04 $\pm$ 9.8629e-06**	1.6387e-03 $\pm$ 2.5544e-05
	500	**2.8576e-04 $\pm$ 3.3213e-06**	1.6213e-03 $\pm$ 1.8418e-05
	1000	**2.8395e-04 $\pm$ 2.3915e-06**	1.6133e-03 $\pm$ 1.4468e-05
	2000	**2.8219e-04 $\pm$ 1.3681e-06**	1.6110e-03 $\pm$ 8.2639e-06
3	250	**9.4367e-05 $\pm$ 8.2106e-06**	3.6896e-04 $\pm$ 1.0013e-05
	500	**8.7906e-05 $\pm$ 3.2759e-06**	3.6264e-04 $\pm$ 4.6550e-06
	1000	**8.6296e-05 $\pm$ 1.9045e-06**	3.5979e-04 $\pm$ 2.5247e-06
	2000	**8.5007e-05 $\pm$ 9.5520e-07**	3.5846e-04 $\pm$ 1.8742e-06

Instead, we consider each term in (19) as a separate objective function:

$$\text{minimize } f(x) = (f_1(x), f_2(x), f_3(x)) \text{ subject to } x \in \mathbb{R}^6 \tag{20}$$

where $f_1(x) = \|Ax - y\|^2$, $f_2(x) = \|x_{\text{age}}\|^2$, $f_3(x) = \|x_{\text{lwt}}\|^2$. Notice that the use of the squared norm in f_2 and f_3 does not change their solutions. It is easy to see that every objective function in (20) is convex but not strongly convex. We make them strongly convex by the following perturbation:

$$\tilde{f}_1 = f_1 + \varepsilon \|x\|^2,$$
$$\tilde{f}_2 = f_2 + \varepsilon \|x\|^2,$$
$$\tilde{f}_3 = f_3 + \varepsilon \|x\|^2$$

where ε is an arbitrarily small positive number (we set $\varepsilon = 10^{-4}$). Now the problem of minimizing a mapping $\tilde{f} = (\tilde{f}_1, \tilde{f}_2, \tilde{f}_3)$ is strongly convex. By (Hamada et al. 2019, Theorems 1.1 and 3.1), this problem is weakly simplicial and the mapping

$$x^*(w) = \arg\min_x \langle w,\ f(x) \rangle \tag{21}$$

is well-defined and continuous on Δ^2, satisfying $x^*(\Delta^2_I) = X^*(\tilde{f}_I)$ for all $I \subseteq \{1, 2, 3\}$.

Then, we obtained subsamples by solving (21) repeatedly with varying $w \in \Delta^2_I$ for each $I \subseteq \{1, 2, 3\}$. For each such I, the weight w was drawn from the uniform distribution on Δ^2_I and the problem (21) was solved by the steepest descent method.

The same idea can be applied to a broad range of sparse learning methods, including the original lasso (Tibshirani 1996), the fused lasso (Tibshirani et al. 2005), the smooth lasso (Hebiri and van de Geer 2011), and the elastic net (Zou and Hastie 2005). For those methods, their group-wise regularization terms can be considered as separate objectives, and the resulting problems would be many-objective (four-objective or more) where the all-at-once fitting will much outperform over the inductive skeleton fitting. We however remark that the bridge regression (lldiko E. Frank and Friedman 1993) is not the case since its regularization term using a nonconvex ℓ_p-norm (i.e., $p < 1$) cannot change into a strongly convex function via perturbations.

Data generation process and evaluation As we conducted in the previous experiments, we generated a training set and a test set on a Pareto set/front randomly. For the location problem, to evaluate the generalized performance for the noisy test data, we added the Gaussian noise ($N(\mathbf{0}, 0.1^2\boldsymbol{I})$) to each point of the training and test sets. Then, we fitted a Bézier simplex to the training set and eval-

uated the MSE between the estimated Bézier simplex and the test set. We changed the size of the training set from $N \in \{250, 500, 1000, 2000\}$. The size of the test set is 10000 and 1000 for the location problem and the group lasso, respectively. We repeated experiments 20 times for each (D, N).

Results and discussion Here, we show the results of fitting Pareto sets. The results of fitting Pareto fronts are provided in the supplementary materials. For each problem instance and method, the average and the standard deviation of the MSE are shown in Table 3. In the table, we highlighted the best score of MSE out of all-at-once fitting and inductive skeleton fitting (optimal) and added the results of one-sided Student's t-test[3] with significance level 0.05.

Table 3a shows results that the inductive skeleton (optimal) outperformed the all-at-once for $D = 2$, and the opposite results for $D = 3$. Note that these magnitude relationships of MSEs for the test data accord with those of the theoretical risks described in Table 2. Therefore, we found that the difference in MSEs can be derived from the risk of each fitting method.

Since the variance of added noise is relatively so large that the Pareto ordering of the training data points may be changed (see the scatter plots shown in the supplementary materials), this experimental setting is more challenging than those of real multi-objective optimization problems. Thus, the result of the location problem suggests that, even for a real problem, we are expected to see a significant difference in the generalized performance between the all-at-once and the inductive skeleton (optimal).

In case of the group lasso, on the other hand, Table 3b shows that the all-at-once was always better for both $D = 2$ and 3, and the differences are almost all significant. While our analysis assumes that the target hyper-surface to be fitted can be represented by a Bézier simplex, the Pareto set of the group lasso cannot for $D = 2$ but for $D = 3$. Therefore, we can see that the results for $D = 3$ that the all-at-once achieved better MSEs accords with our analysis.

From the above results, the validity of the analytic results is confirmed in practical situations.

5 Conclusion

In this paper, we have shown that the asymptotic ℓ_2-risk of the two Bézier simplex fitting methods developed previously: the all-at-once fitting and the inductive skeleton fitting. From our risk analysis, the optimal ratio of subsamples for the inductive skeleton fitting has been derived, which is useful for design of experiments to maximize the goodness of fit. We have discussed that superiority between the two fitting methods depends on the degree of a Bézier simplex to be fit: the inductive skeleton fitting with optimally-decoupled subsamples outperforms for degree two whereas the all-at-once fitting becomes the better for degree three, independent of the dimensionality of the Bézier simplex and its ambient space. The above theoretical results have been confirmed via numerical experiments under small to moderate sample sizes. We have demonstrated two applications of the analytic results in multi-objective optimization: a generalized location problem and a hyperparameter tuning of the group lasso.

As a remark for future work, we point out two important cases which the current theory does not cover. The first one is the case discussed in Section 4.2 that the true surface is not representable by a model. The second one is presented in the literature (Kobayashi et al. 2019). When the parameters of a Bézier simplex are not given in a sample and to be estimated as well as the control points, the inductive skeleton fitting outperforms the all-at-once fitting even if the Bézier simplex is of degree *three*. We believe that those cases would offer insightful examples to extend the scope of our theory.

[3]When we conducted a one-sided Student's t-test, we used a log transformation to MSEs in advance.

Acknowledgement

We wish to thank Prof. Shuji Yamamoto for making a number of valuable suggestions.

References

Borges, C. F., and Pastva, T. 2002. Total least squares fitting of Bézier and B-spline curves to ordered data. *Computer Aided Geometric Design* 19(4):275–289.

Deb, K., and Jain, H. 2014. An evolutionary many-objective optimization algorithm using reference-point-based nondominated sorting approach, part I: Solving problems with box constraints. *IEEE Transactions on Evolutionary Computation* 18(4):577–601.

Deb, K. 2001. *Multi-objective Optimization Using Evolutionary Algorithms*. New York, NY, USA: John Wiley & Sons, Inc.

Eichfelder, G. 2008. *Adaptive Scalarization Methods in Multiobjective Optimization*. Springer-Verlag, Berlin, Heidelberg.

Hamada, N.; Nagata, Y.; Kobayashi, S.; and Ono, I. 2010. Adaptive weighted aggregation: A multiobjective function optimization framework taking account of spread and evenness of approximate solutions. In *Proceedings of the 2010 IEEE Congress on Evolutionary Computation*, CEC 2010, 787–794.

Hamada, N.; Hayano, K.; Ichiki, S.; Kabata, Y.; and Teramoto, H. 2019. Topology of Pareto sets of strongly convex problems. *ArXiv e-prints*. http://arxiv.org/abs/1904.03615.

Hamada, N. 2017. Simple problems: The simplicial gluing structure of Pareto sets and Pareto fronts. In *Proceedings of the Genetic and Evolutionary Computation Conference Companion*, GECCO '17, 315–316. New York, NY, USA: ACM.

Harada, K.; Sakuma, J.; Kobayashi, S.; and Ono, I. 2007. Uniform sampling of local Pareto-optimal solution curves by Pareto path following and its applications in multi-objective GA. In *Proceedings of the Genetic and Evolutionary Computation Conference (GECCO)*, 813–820. New York, NY, USA: ACM.

Harada, K.; Sakuma, J.; and Kobayashi, S. 2006. Local search for multiobjective function optimization: Pareto descent method. In *Proceedings of the 8th Annual Conference on Genetic and Evolutionary Computation*, GECCO '06, 659–666. New York, NY, USA: ACM.

Hebiri, M., and van de Geer, S. 2011. The smooth-lasso and other $\ell_1 + \ell_2$-penalized methods. *Electron. J. Statist.* 5:1184–1226.

Hernandez-Lobato, D.; Hernandez-Lobato, J.; Shah, A.; and Adams, R. 2016. Predictive entropy search for multi-objective bayesian optimization. In Balcan, M. F., and Weinberger, K. Q., eds., *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, 1492–1501. New York, New York, USA: PMLR.

Hillermeier, C. 2001. *Nonlinear Multiobjective Optimization: A Generalized Homotopy Approach*, volume 25 of *International Series of Numerical Mathematics*. Birkhäuser Verlag, Basel, Boston, Berlin.

Hosmer, D. W., and Lemeshow, S. 1989. *Applied Logistic Regression*. New York: Wiley.

Kobayashi, K.; Hamada, N.; Sannai, A.; Tanaka, A.; Bannai, K.; and Sugiyama, M. 2019. Bézier simplex fitting: Describing pareto fronts of simplicial problems with small samples in multi-objective optimization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, 2304–2313.

Kuhn, H. W. 1967. On a pair of dual nonlinear programs. *Nonlinear Programming* 1:38–45.

lldiko E. Frank, and Friedman, J. H. 1993. A statistical view of some chemometrics regression tools. *Technometrics* 35(2):109–135.

Mastroddi, F., and Gemma, S. 2013. Analysis of Pareto frontiers for multidisciplinary design optimization of aircraft. *Aerosp. Sci. Technol.* 28(1):40–55.

Miettinen, K. M. 1999. *Nonlinear Multiobjective Optimization*, volume 12 of *International Series in Operations Research & Management Science*. Springer-Verlag, GmbH.

Shoval, O.; Sheftel, H.; Shinar, G.; Hart, Y.; Ramote, O.; Mayo, A.; Dekel, E.; Kavanagh, K.; and Alon, U. 2012. Evolutionary trade-offs, Pareto optimality, and the geometry of phenotype space. *Science* 336(6085):1157–1160.

Tibshirani, R.; Saunders, M.; Rosset, S.; Zhu, J.; and Knight, K. 2005. Sparsity and smoothness via the fused lasso. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 67(1):91–108.

Tibshirani, R. 1996. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)* 58(1):267–288.

Venables, W. N., and Ripley, B. D. 2002. *Modern Applied Statistics with S*. Springer, fourth edition.

Vrugt, J. A.; Gupta, H. V.; Bastidas, L. A.; Bouten, W.; and Sorooshian, S. 2003. Effective and efficient algorithm for multiobjective optimization of hydrologic models. *Water Resources Research* 39(8):1214–1232.

Yang, K.; Emmerich, M.; Deutz, A.; and Bäck, T. 2019. Multi-objective Bayesian global optimization using expected hypervolume improvement gradient. *Swarm and Evolutionary Computation* 44:945–956.

Yuan, M., and Lin, Y. 2006. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 68(1):49–67.

Zhang, Q., and Li, H. 2007. MOEA/D: A multiobjective evolutionary algorithm based on decomposition. *IEEE Transactions on Evolutionary Computation* 11(6):712–731.

Zou, H., and Hastie, T. 2005. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)* 67(2):301–320.