



Stochastic Gradient Descent for Parametric Hypersurface Representation of Pareto Set in Multi-Objective Optimization

Yasunari Hikima^{1,2}, Ken Kobayashi³, Akinori Tanaka⁴, Akiyoshi Sannai⁵, Naoki Hamada⁶

Fujitsu Limited¹, Kyushu University², Science Tokyo³, RIKEN AIP⁴, Kyoto University⁵, KLab Inc.⁶

Summary

1. Propose a multi-objective optimization method that **updates a hypersurface called Bézier simplex** with a modified stochastic gradient descent.
2. Provide a theoretical analysis that the proposed method **converges to optimal control points** under mild assumptions.
3. Can **efficiently find the control points** of a Bézier simplex that approximates the *Pareto set* for various MOO instances.

Background: Multi-objective Optimization

■ (Strongly convex) Multi-Objective Optimization (MOO)

$$\underset{\mathbf{x} \in \mathcal{X}(\subseteq \mathbb{R}^L)}{\text{minimize}} \quad \mathbf{f}(\mathbf{x}) := (f_1(\mathbf{x}), \dots, f_M(\mathbf{x}))^\top.$$

where f_1, \dots, f_M are assumed to be μ -strongly convex and ρ -smooth.

Goal: Find the set of Pareto optimal solutions, which is called *Pareto set*, and its image, called *Pareto front*, is given by

$$X^*(\mathbf{f}) := \{\mathbf{x} \in \mathcal{X} \mid f(\mathbf{y}) \not\leq f(\mathbf{x}) \text{ for all } \mathbf{y} \in \mathcal{X}\},$$

$$\mathbf{f}(X^*(\mathbf{f})) := \{\mathbf{f}(\mathbf{x}) \in \mathbb{R}^M \mid \mathbf{x} \in X^*(\mathbf{f})\}.$$

Theorem 1 (Mizota et al. '21). *Let $\mathcal{X} = \mathbb{R}^L$ and f_m be strongly convex for all $m \in [M]$. Then, the mapping $\text{argmin } \mathbb{E}(\mathbf{f})$ gives a continuous surjection onto $X^*(\mathbf{f})$.*

Preliminary: Bézier Simplex

Definition. Bézier simplex

The $(M-1)$ -dim Bézier simplex of degree D in \mathbb{R}^L is a map $\mathbf{b}: \Delta^{M-1} \rightarrow \mathbb{R}^L$ determined by *control points* $\mathbf{p}_d \subseteq \mathbb{R}^L$ ($d \in \mathbb{N}_D^M$)

$$\mathbf{b}(\mathbf{t} \mid \mathbf{P}) := \sum_{d \in \mathbb{N}_D^M} \binom{D}{d} \mathbf{t}^d \mathbf{p}_d.$$

$\mathbf{t} \in \Delta^{M-1}$: Parameter of the Bézier simplex

$\mathbf{P} = (\mathbf{p}_d)_{d \in \mathbb{N}_D^M}$: Control points

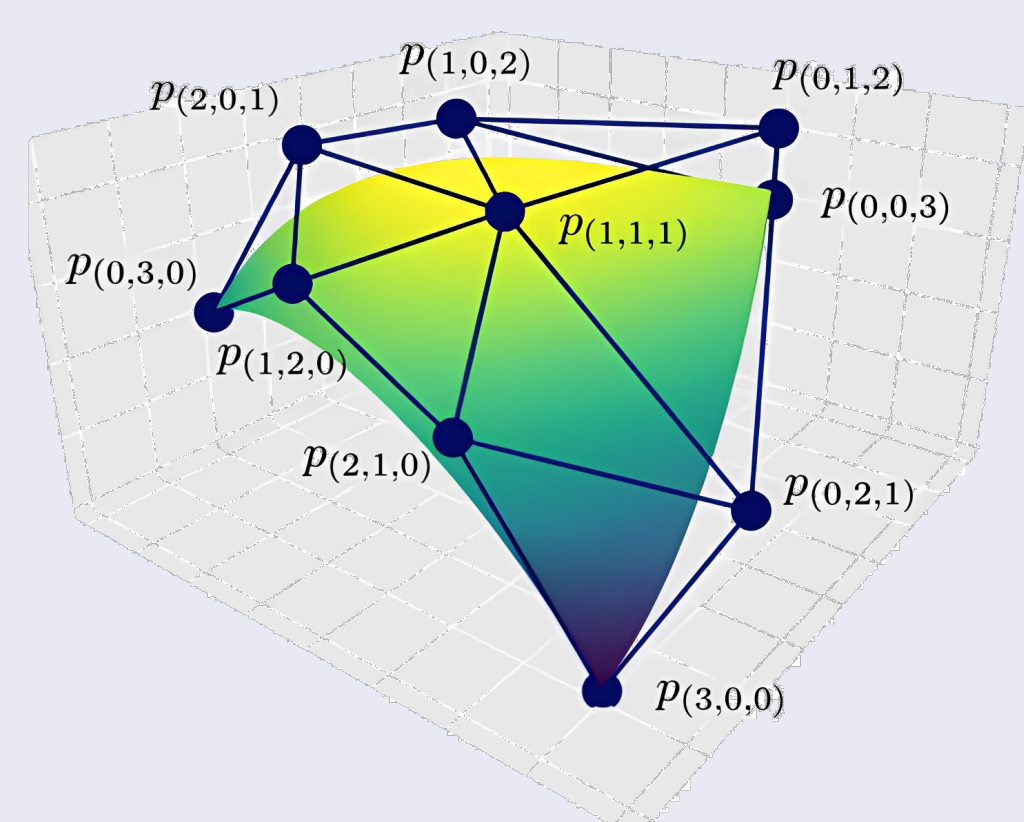


Fig: Bézier simplex with $D = 3$.

Theorem 2 (Kobayashi et al. '19). *Let $\phi: \Delta^{M-1} \rightarrow \mathbb{R}^M$ be a continuous map. There exists an infinite sequence of Bézier simplices $\mathbf{b}^{(i)}: \Delta^{M-1} \rightarrow \mathbb{R}^M$ such that*

$$\lim_{i \rightarrow \infty} \sum_{\mathbf{t} \in \Delta^{M-1}} |\phi(\mathbf{t}) - \mathbf{b}^{(i)}(\mathbf{t})| = 0.$$

→ Pareto set/front of any strongly convex MOO is approximated by some Bézier simplex in arbitrary precision

- Pareto sets/fronts are often curved simplices in practical MOO applications
 - Location problem (Kuhn '67), Phenotypic divergence model (Shoval et al. '12)
 - Hydrologic modeling (Vrugt et al. '03), Airplane design (Mastroddi & Gemma '13)

Scalarization-based Loss and SGD Method

■ Generalized loss for a Bézier simplex

$$\mathcal{L}_{\text{gen}}(\mathbf{P}) := \mathbb{E}_{\mathbf{t}} [\mathbf{t}^\top \mathbf{f}(\mathbf{b}(\mathbf{t} \mid \mathbf{P}))].$$

Similar losses have been studied in Pareto set learning (Lin et al., '20, Navon et al., '21, Chen & Kwok '24)

■ Minimize the empirical loss with $\{\mathbf{t}_i\}_{i=1}^n$

$$\underset{\mathbf{P}}{\text{minimize}} \quad \mathcal{L}(\mathbf{P}) := \frac{1}{n} \sum_{i=1}^n \mathbf{t}_i^\top \mathbf{f}(\mathbf{b}(\mathbf{t}_i \mid \mathbf{P})).$$

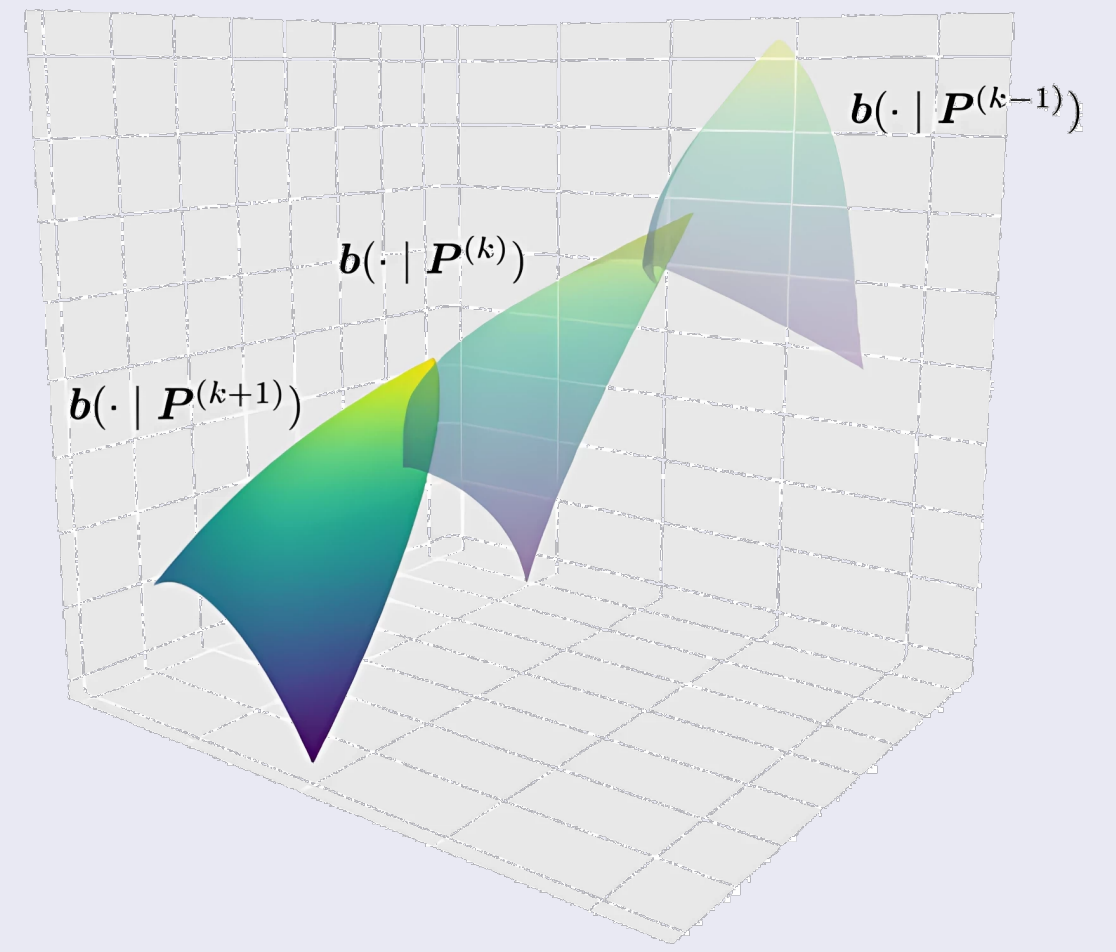


Fig: Update Bézier simplex.

■ Update \mathbf{P} with stochastic gradient descent

$$\mathbf{P}^{(k+1)} \in \underset{\mathbf{P}}{\text{argmin}} \left\{ \mathcal{L}(\mathbf{P}^{(k)}) + \left\langle \nabla \mathcal{L}_B(\mathbf{P}^{(k)}), \mathbf{P} - \mathbf{P}^{(k)} \right\rangle + \frac{1}{2\alpha_k} \mathbb{E}_{\mathbf{t}} \left[\left\| \mathbf{b}(\mathbf{t} \mid \mathbf{P}) - \mathbf{b}(\mathbf{t} \mid \mathbf{P}^{(k)}) \right\|_2^2 \right] \right\}.$$

- Objective function:

: Linear approximation of \mathcal{L} at $\mathbf{P}^{(k)}$ with a minibatch $B \subseteq [n]$
 : Average distance between $\mathbf{b}(\mathbf{t} \mid \mathbf{P})$ and $\mathbf{b}(\mathbf{t} \mid \mathbf{P}^{(k)})$

- By using [Tanaka et al. '20], we can obtain a closed form of the average distance with a positive definite matrix Σ :

$$\mathbb{E}_{\mathbf{t}} \left[\left\| \mathbf{b}(\mathbf{t} \mid \mathbf{P}) - \mathbf{b}(\mathbf{t} \mid \mathbf{P}^{(k)}) \right\|_2^2 \right] = \left\langle \left(\mathbf{P} - \mathbf{P}^{(k)} \right)^\top \Sigma, \left(\mathbf{P} - \mathbf{P}^{(k)} \right) \right\rangle.$$

Algorithm 1: Our Stochastic Gradient Descent

Initialization: Set $k \leftarrow 0$, $\mathbf{P}^{(k)} \leftarrow \mathbf{P}_0$.

1 Randomly draw n samples $\{\mathbf{t}_i\}_{i=1}^n \subseteq \Delta^{M-1}$ from $U(\Delta^{M-1})$.

2 **while** $k < K$ **do**

3 Choose $B \subseteq [n]$ uniformly at random.

4 Construct a gradient estimator as $\nabla \mathcal{L}_B(\mathbf{P}) = \frac{1}{|B|} \sum_{i \in B} \nabla \mathcal{L}_i(\mathbf{P})$.

5 Update $\mathbf{P}^{(k)}$ as $\mathbf{P}^{(k+1)} \leftarrow \mathbf{P}^{(k)} - \alpha_k \Sigma^{-1} \nabla \mathcal{L}_B(\mathbf{P}^{(k)})$.

6 **return** $\mathbf{P}^{(K)}$

Theoretical Analysis

1. Assume that the gradient noise $\sigma_{\mathcal{L}} = \sup_{\mathbf{P}^* \in \mathcal{P}^*} \mathbb{E} [\|\nabla \mathcal{L}(\mathbf{P}^*)\|^2]$ is finite.
2. Assume that \mathcal{L}_B is $\tilde{\rho}$ -expected smooth for any $B \in [n]$.

Theorem 3 (informal). *Let \mathbf{P}_0 be an initial control points of the Bézier simplex. Consider a sequence of control points $\{\mathbf{P}^{(k)}\}_{k \in \mathbb{N}}$ generated by Algorithm 1 with a stepsize $\alpha_k = \alpha < (2\tilde{\rho}\lambda_{\max}(\Sigma^{-1}))^{-1}$. Then, under the mild assumption of the sampling $\{\mathbf{t}_i\}_{i=1}^n$ from $U(\Delta^{M-1})$ we have*

$$\mathbb{E} \left[\left\| \mathbf{P}^{(k)} - \mathbf{P}^* \right\|_{\Sigma}^2 \right] \leq (1 - \alpha \bar{\mu} \lambda_{\min}(\Sigma^{-1}))^k \|\mathbf{P}_0 - \mathbf{P}^*\|_{\Sigma}^2 + \frac{2\alpha \cdot \text{cond}(\Sigma^{-1})}{\bar{\mu}} \sigma_{\mathcal{L}}.$$

Numerical Experiments

- Applied Algorithm 1 to a Group LASSO instance on Birthwt dataset (Venables and Ripley '12)
- Hyperparameter settings: $\mathbf{P}_0 = \mathbf{O}$, $\alpha_k = 0.05$ ($\forall k$), $K = 100$, $n = 100$, $m = 20$

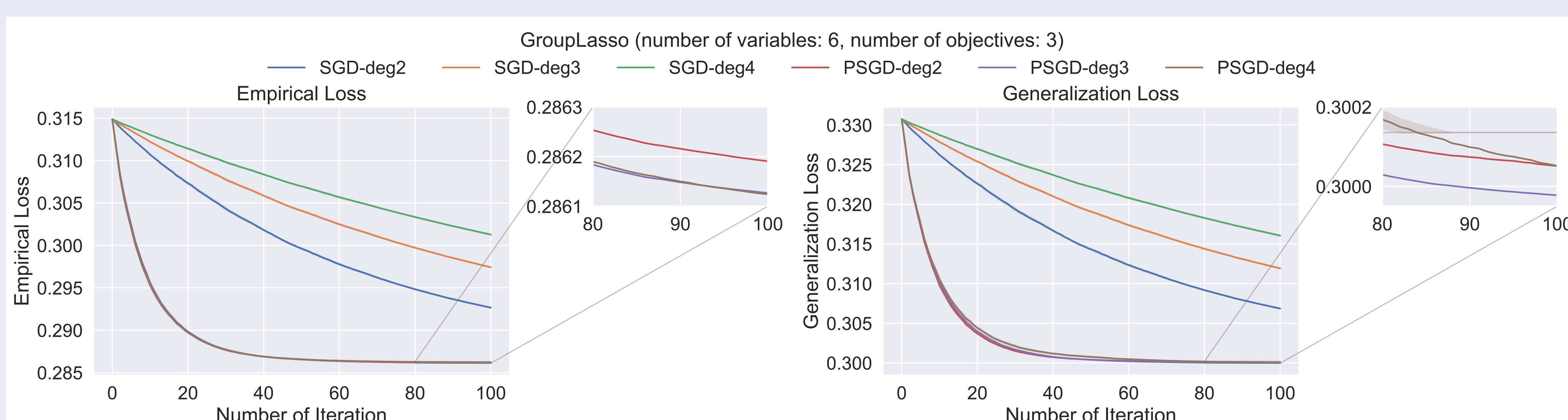


Fig: The learning curve for Group LASSO instance (left: training data, right: test data)

- Verify that the Pareto set can be well approximated by the proposed method

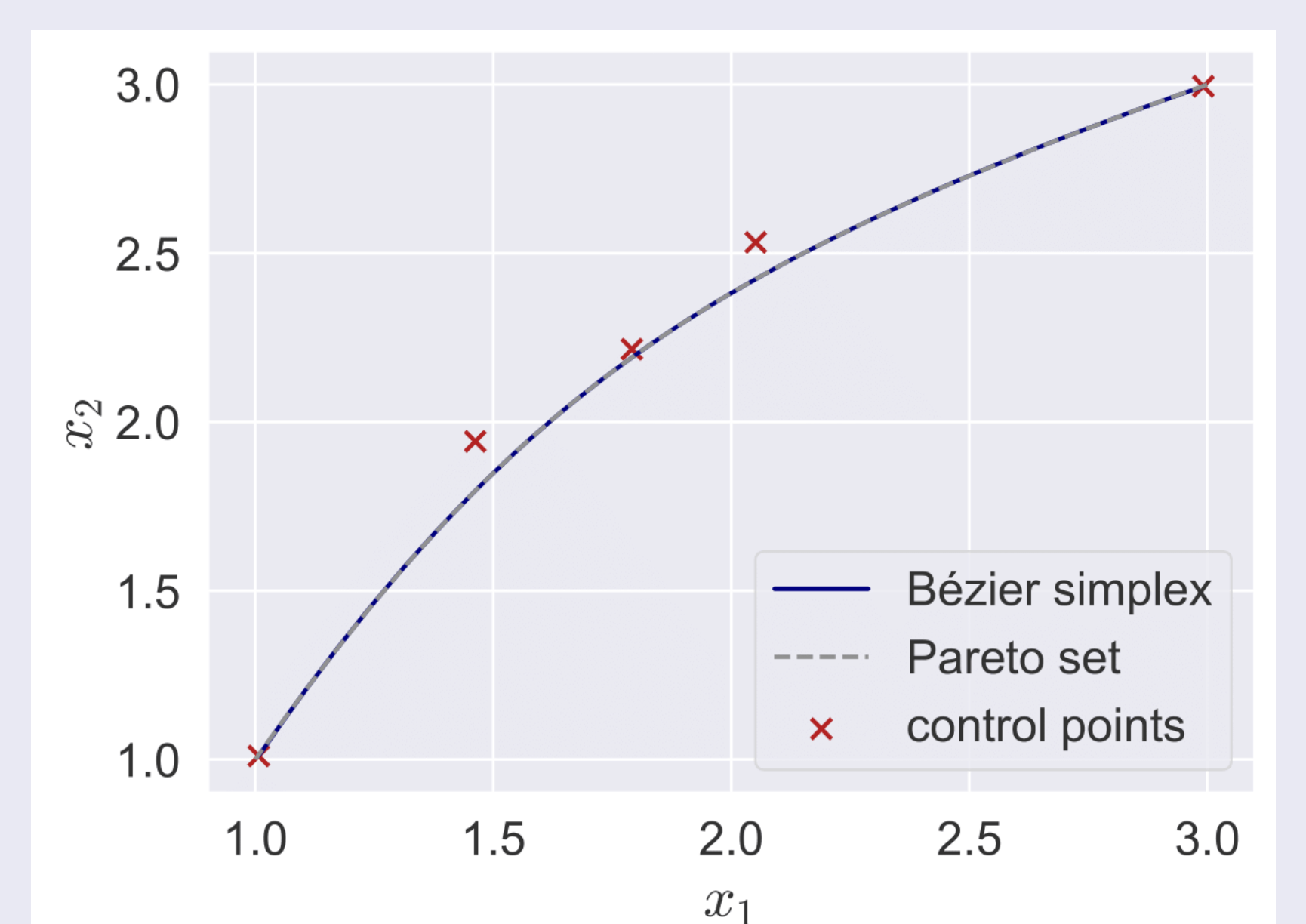


Fig: Approximated Pareto set