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

Yasunari Hikima[1,2], Ken Kobayashi[3], Akinori Tanaka[4], Akiyoshi Sannai[5], Naoki Hamada[6]

Fujitsu Limited[1], Kyushu University[2], Science Tokyo[3], RIKEN AIP[4], Kyoto University[5], 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)

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

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

$$X^*(\boldsymbol{f}) := \{\, \boldsymbol{x}\in\mathscr{X} \mid f(\boldsymbol{y}) \not\prec f(\boldsymbol{x}) \text{ for all } \boldsymbol{y}\in\mathscr{X} \,\},$$
$$\boldsymbol{f}(X^*(\boldsymbol{f})) := \{\, \boldsymbol{f}(\boldsymbol{x})\in\mathbb{R}^M \mid \boldsymbol{x}\in X^*(\boldsymbol{f}) \,\}.$$

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

Preliminary: Bézier Simplex

Definition. Bézier simplex

The $(M-1)$-dim Bézier simplex of degree D in $\mathbb{R}^L$ is a map $\boldsymbol{b}\colon \Delta^{M-1}\to\mathbb{R}^L$ determined by *control points* $\boldsymbol{p_d}\subseteq\mathbb{R}^L$ $(\boldsymbol{d}\in\mathbb{N}_D^M)$

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

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

$\boldsymbol{P}=(\boldsymbol{p_d})_{\boldsymbol{d}\in\mathbb{N}_D^M}$: Control points

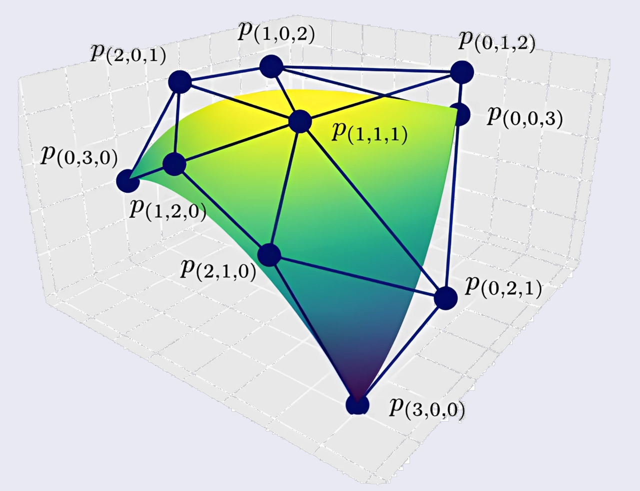

Fig: Bézier simplex with $D=3$.

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

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

→ 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

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

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 $\{\boldsymbol{t}_i\}_{i=1}^n$

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

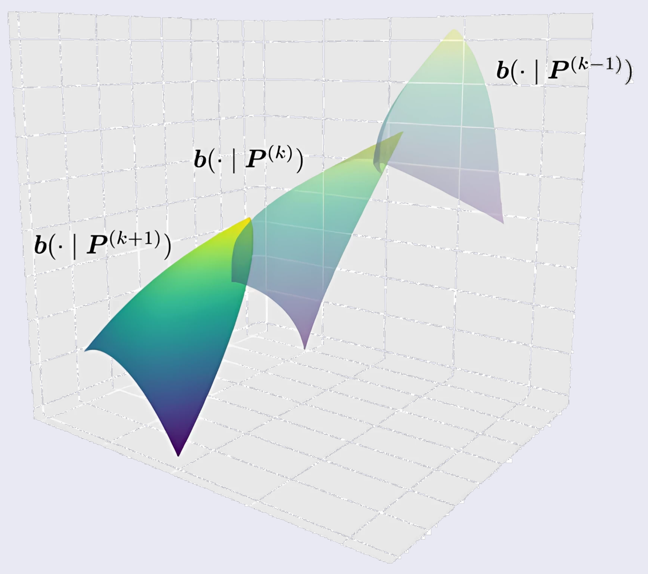

Fig: Update Bézier simplex.

■ Update $\boldsymbol{P}$ with stochastic gradient descent

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

- Objective function:
  - (red) : Linear approximation of $\mathscr{L}$ at $\boldsymbol{P}^{(k)}$ with a minibatch $B\subseteq[n]$
  - $+$ (green) : Average distance between $\boldsymbol{b}(\boldsymbol{t}\mid\boldsymbol{P})$ and $\boldsymbol{b}(\boldsymbol{t}\mid\boldsymbol{P}^{(k)})$
- By using [Tanaka et al. '20], we can obtain a closed form of the average distance with a positive definite matrix $\boldsymbol{\Sigma}$:

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

Algorithm 1: Our Stochastic Gradient Descent

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

Randomly draw n samples $\{\boldsymbol{t}_i\}_{i=1}^n\subseteq\Delta^{M-1}$ from $U(\Delta^{M-1})$.
while $k<K$ **do**
  Choose $B\subseteq[n]$ unifomly at random.
  Construct a gradient estimator as $\nabla\mathscr{L}_B(\boldsymbol{P})=\frac{1}{|B|}\sum_{i\in B}\nabla\mathscr{L}_i(\boldsymbol{P})$.
  Update $\boldsymbol{P}^{(k)}$ as $\boldsymbol{P}^{(k+1)}\leftarrow\boldsymbol{P}^{(k)}-\alpha_k\boldsymbol{\Sigma}^{-1}\mathscr{L}_B(\boldsymbol{P}^{(k)})$.
return $\boldsymbol{P}^{(K)}$

Theoretical Analysis

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

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

$$\mathbb{E}\left[\left\|\boldsymbol{P}^{(k)}-\boldsymbol{P}^*\right\|_{\boldsymbol{\Sigma}}^2\right]\le\left(1-\alpha\overline{\mu}\lambda_{\min}(\boldsymbol{\Sigma}^{-1})\right)^k\|\boldsymbol{P}_0-\boldsymbol{P}^*\|_{\boldsymbol{\Sigma}}^2+\frac{2\alpha\cdot\operatorname{cond}(\boldsymbol{\Sigma}^{-1})}{\overline{\mu}}\sigma_{\mathscr{L}}.$$

Numerical Experiments

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

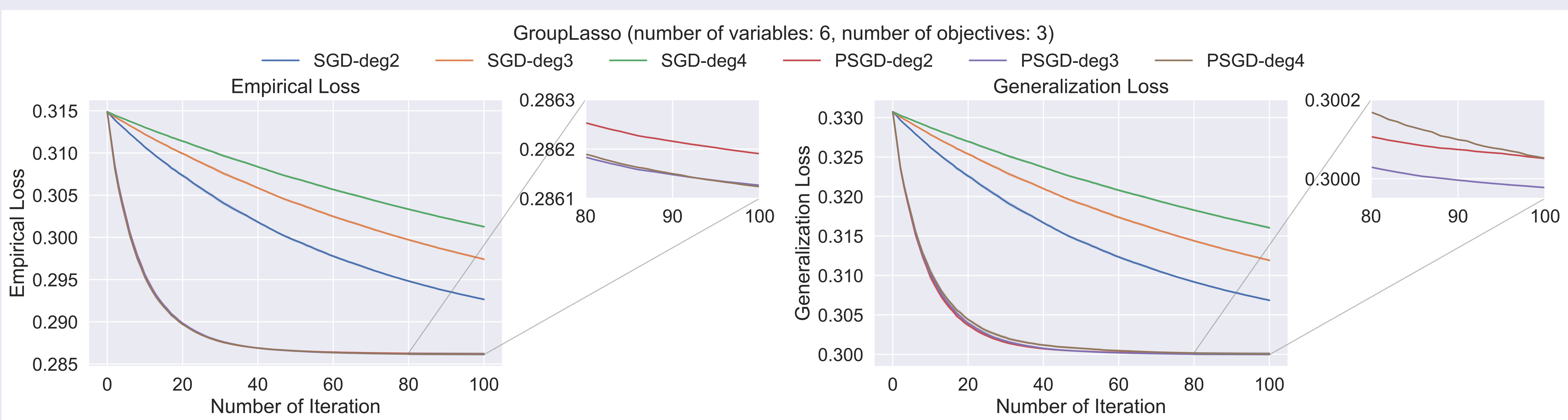

Fig: The learning curve for Group LASSO instance (left: training data, right: test data)

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

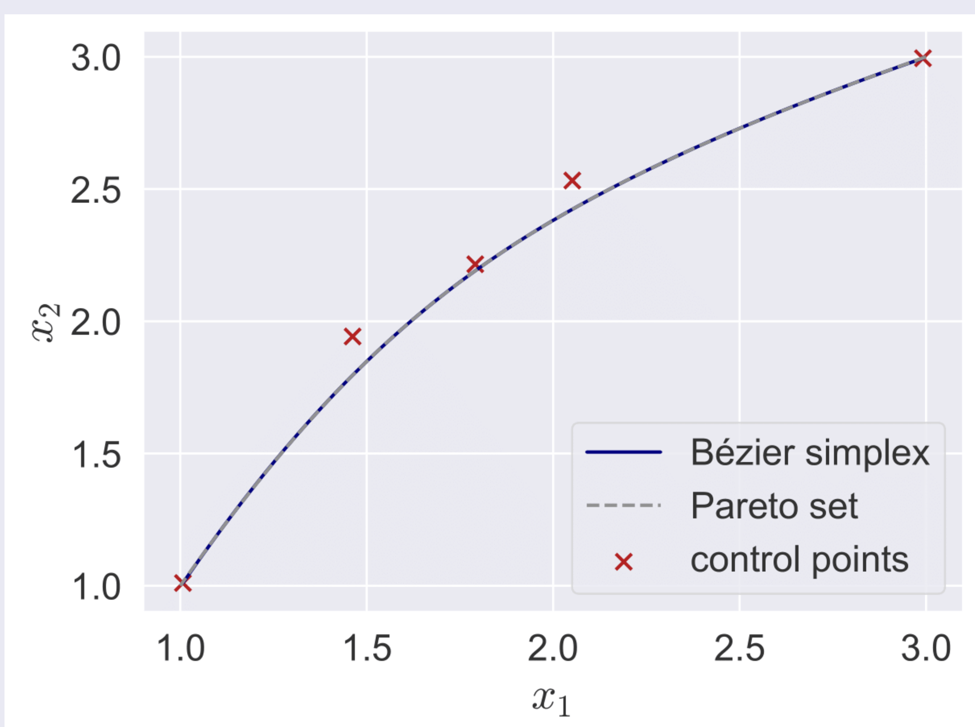

Fig: Approximated Pareto set