# Stochastic Gradient Descent for Bézier Simplex Representation of Pareto Sets in Multi-Objective Optimization

**Yasunari Hikima**
Fujitsu Limited
Kyushu University

**Ken Kobayashi**
Institute of Science Tokyo

**Akinori Tanaka**
RIKEN AIP
RIKEN iTHEMS
Keio University

**Akiyoshi Sannai**
Kyoto University

**Naoki Hamada**
KLab Inc.

## Abstract

Multi-objective optimization aims to find a set of solutions that achieve the best trade-off among multiple conflicting objective functions. While various multi-objective optimization algorithms have been proposed so far, most of them aim to find finite solutions as an approximation of the Pareto set, which may not adequately capture the entire structure of the Pareto set, especially when the number of variables is large. To overcome this limitation, we propose a method to obtain a parametric hypersurface representing the entire Pareto set instead of a finite set of points. Since the Pareto set of an $M$-objective optimization problem typically forms an $(M-1)$-dimensional simplex, we use a Bézier simplex as a model to express the Pareto set. We then develop a stochastic gradient descent-based algorithm that updates the Bézier simplex model toward the Pareto set, introducing a preconditioning matrix to enhance convergence. Our convergence analysis demonstrated that the proposed algorithm outperforms naive stochastic gradient descent in terms of convergence rate. Furthermore, we validate the effectiveness of our method through various multi-objective optimization problem instances, including real-world problems.

Proceedings of the 28[th] International Conference on Artificial Intelligence and Statistics (AISTATS) 2025, Mai Khao, Thailand. PMLR: Volume 258. 

## 1 INTRODUCTION

Given multiple objective functions denoted as $f_1, \ldots, f_M \colon \mathscr{X} \to \mathbb{R}$ on a subset $\mathscr{X}$ of the subspace of $L$-dimensional Euclidean space $\mathbb{R}^L$, consider the following multi-objective optimization problem:

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

The goal is to find the *Pareto set* and its image, called the *Pareto front*, which are defined as follows:

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

Here, the symbol $\prec$ stands for *Pareto ordering* which is defined as follows:

$$\begin{aligned} &\boldsymbol{f}(\boldsymbol{x}) \prec \boldsymbol{f}(\boldsymbol{y}) \\ &\overset{\text{def}}{\iff} f_m(\boldsymbol{x}) \leq f_m(\boldsymbol{y}) \text{ for all } m = 1, \ldots, M \\ &\quad \text{and } f_m(\boldsymbol{x}) < f_m(\boldsymbol{y}) \text{ for some } m = 1, \ldots, M. \end{aligned}$$

Many algorithms have been proposed to find a single solution on the Pareto set or a finite set of solutions to approximate the Pareto set, say, evolutionary computation (Deb, 2001; Zhang and Li, 2007; Deb and Jain, 2014), homotopy approaches (Hillermeier, 2001; Harada et al., 2007), Bayesian optimization (Hernandez-Lobato et al., 2016; Yang et al., 2019; Daulton et al., 2022, 2024), and gradient-based methods (Miettinen, 1999; Ehrgott, 2005; Fliege et al., 2018; Tanabe et al., 2018; Lin et al., 2024).

A finite point representation of the Pareto set suffers from the curse of dimensionality as the dimensionality of the Pareto set and Pareto front of "almost all" $M$-objective optimization problems is $M-1$ (see (Wan, 1977, 1978) for rigorous statement). To deal with this issue, several approaches to obtaining a parametric hypersurface representation of the Pareto set of multi-objective optimization problems. Lin et al. (2022) has proposed a learning-based method to approximate the

Pareto set with a surrogate model. Kobayashi et al. (2019); Tanaka et al. (2020, 2021) focused on a widely-observed structure of Pareto sets, i.e., simplicial problems, and proposed a post-optimization process that fits a hypersurface to the Pareto set with the Bézier simplex model. Sannai et al. (2022) extended this approach to a multi-objective optimization method that sequentially updates the Bézier simplex model. However, a theoretical guarantee of whether the resulting hypersurface can approximate a Pareto set or not has not been explored.

In this paper, we propose a multi-objective optimization method to obtain a hypersurface representing the Pareto set with a theoretical guarantee of its convergence properties. Inspired by the existing studies (Kobayashi et al., 2019; Tanaka et al., 2020, 2021; Sannai et al., 2022), we consider the Bèzier simplex model to represent the Pareto set and develop a stochastic gradient descent (SGD) algorithm that updates the Bèzier simplex toward the Pareto set. Also, we provide a theoretical analysis of the convergence properties of the proposed algorithm. The contributions of this paper are summarized as follows:

(i) We propose a multi-objective optimization method that updates the parameters of Bèzier simplex model with a stochastic gradient descent (SGD) approach to obtain a hypersurface representing the Pareto set. Unlike the ordinary SGD, we incorporate a scaling matrix into the update rule, which enhances the convergence speed.

(ii) We provide a theoretical analysis of the convergence properties of the proposed algorithm. The loss function that we introduced is shown to be strongly convex with respect to the parameters of Bézier simplex under mild conditions. We prove that the proposed algorithm achieves a faster convergence speed than the standard stochastic gradient descent method.

(iii) We demonstrate the effectiveness of the proposed algorithm through numerical experiments on various multi-objective optimization instances, including real-world problems. Compared with the algorithm based on the naive stochastic gradient descent method, our algorithm achieved faster convergence results.

**Notation** We use lowercase bold letters for vectors and uppercase bold letters for matrices. For a positive integer $M$, we define $[M] \coloneqq \{1, \dots, M\}$. For a real-valued symmetric matrix $\boldsymbol{A}$, $\lambda_{\min}(\boldsymbol{A})$ and $\lambda_{\max}(\boldsymbol{A})$ represent the minimum and maximum eigenvalues of $\boldsymbol{A}$, respectively. For a positive definite matrix $\boldsymbol{A}$, we define its condition number as $\mathrm{cond}(\boldsymbol{A}) \coloneqq \lambda_{\max}(\boldsymbol{A})/\lambda_{\min}(\boldsymbol{A})$. For real-valued matrices $\boldsymbol{A}, \boldsymbol{B} \in \mathbb{R}^{m\times n}$, we define the standard inner product as $\langle \boldsymbol{A}, \boldsymbol{B}\rangle \coloneqq \mathrm{Tr}(\boldsymbol{A}^\top \boldsymbol{B}) = \sum_{ij} A_{ij}B_{ij}$. Given a positive definite matrix $\boldsymbol{\Sigma} \in \mathbb{R}^{m\times m}$, we define its associated inner product as $\langle \boldsymbol{A}, \boldsymbol{B}\rangle_{\boldsymbol{\Sigma}} \coloneqq \langle \boldsymbol{A}, \boldsymbol{\Sigma}\boldsymbol{B}\rangle$ for $\boldsymbol{A}, \boldsymbol{B} \in \mathbb{R}^{m\times n}$. For these inner products, we define their associated norm as $\|\boldsymbol{A}\|_{\mathsf{F}} \coloneqq \sqrt{\langle \boldsymbol{A}, \boldsymbol{A}\rangle}$ and $\|\boldsymbol{A}\|_{\boldsymbol{\Sigma}} \coloneqq \sqrt{\langle \boldsymbol{A}, \boldsymbol{A}\rangle_{\boldsymbol{\Sigma}}}$ for all $\boldsymbol{A} \in \mathbb{R}^{m\times n}$.

## 2 PRELIMINARIES

This section provides the preliminaries to describe the algorithm and theoretical analysis.

### 2.1 Probability simplex

Let $M$ be a positive integer. We consider the set of probability distributions $\boldsymbol{t} \in \mathbb{R}^M$ over $[M]$. The set of probability distributions over $[M]$ is equal to the simplex as follows:

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

Let $C(\mathscr{X})$ be the space of continuous functions over $\mathscr{X}$, and we define the function $F\colon [M] \to C(\mathscr{X})$ by $F(m) = f_m$. Then, we have the expectation function

$$\begin{array}{rccc} \mathbb{E}(\boldsymbol{f})\colon & \Delta^{M-1} & \longrightarrow & C(\mathscr{X}) \\ & \cup & & \cup \\ & \boldsymbol{t} & \longmapsto & \mathbb{E}_{\boldsymbol{t}}(F) \end{array}.$$

Furthermore, if $f_m$ is strongly convex for all $m \in [M]$, then the following function is well-defined:

$$\begin{array}{rccc} \operatorname{argmin}\mathbb{E}(\boldsymbol{f})\colon & \Delta^{M-1} & \longrightarrow & \mathscr{X} \\ & \cup & & \cup \\ & \boldsymbol{t} & \longmapsto & \operatorname{argmin}\mathbb{E}_{\boldsymbol{t}}(F) \end{array}.$$

Note that $\mathbb{E}_{\boldsymbol{t}}(F) = \sum_m t_m f_m$ follows from the definition. $\mathbb{E}_{\boldsymbol{t}}(F)$ corresponds to the sum of a function chosen continuously along $\boldsymbol{t}$ from $\boldsymbol{f}$. As a direct consequence of Mizota et al. (2021, Theorem 2), the mapping $\operatorname{argmin}\mathbb{E}(\boldsymbol{f})$ gives a continuous surjection onto $X^*(\boldsymbol{f})$ if $f_m$ is strongly convex for all $m \in [M]$.

**Theorem 2.1.** *Let $\mathscr{X} = \mathbb{R}^L$ and $f_m\colon \mathscr{X} \to \mathbb{R}$ strongly convex for all $m \in [M]$. Then, the mapping $\operatorname{argmin}\mathbb{E}(\boldsymbol{f})$ gives a continuous surjection onto $X^*(\boldsymbol{f})$.*

### 2.2 Bézier Simplex

Let $\mathbb{N} \coloneqq \{0, 1, 2, \dots\}$ and $D$ be a positive integer. We define

$$\mathbb{N}_D^M \coloneqq \left\{ (d_1, \dots, d_M)^\top \in \mathbb{N}^M \;\middle|\; \sum_{m=1}^{M} d_m = D \right\}.$$

For $\boldsymbol{t} \coloneqq (t_1, \ldots, t_M)^\top \in \Delta^{M-1}$ and $\boldsymbol{d} \coloneqq (d_1, \ldots, d_M)^\top \in \mathbb{N}_D^M$, we denote by $\boldsymbol{t}^{\boldsymbol{d}}$ a monomial $t_1^{d_1} \ldots t_M^{d_M}$. The Bézier simplex of degree $D$ in $\mathbb{R}^L$ with control points $\{\boldsymbol{p}_{\boldsymbol{d}}\}_{\boldsymbol{d} \in \mathbb{N}_D^M} \subseteq \mathbb{R}^L$ is a map $\boldsymbol{b}\colon \Delta^{M-1} \to \mathbb{R}^L$, which is defined by

$$\boldsymbol{b}(\boldsymbol{t} \,|\, \boldsymbol{P}) \coloneqq \sum_{\boldsymbol{d} \in \mathbb{N}_D^M} \binom{D}{\boldsymbol{d}} \boldsymbol{t}^{\boldsymbol{d}} \boldsymbol{p}_{\boldsymbol{d}}, \tag{1}$$

where $\binom{D}{\boldsymbol{d}} \coloneqq \frac{D!}{\prod_{i=1}^M d_i!}$ is a multinomial coefficient and $\boldsymbol{P} \in \mathbb{R}^{|\mathbb{N}_D^M| \times L}$ represents a matrix of control points, which is defined as

$$\boldsymbol{P} \coloneqq \begin{pmatrix} (\boldsymbol{p}_1)_1 & (\boldsymbol{p}_1)_2 & \cdots & (\boldsymbol{p}_1)_L \\ (\boldsymbol{p}_2)_1 & (\boldsymbol{p}_2)_2 & \cdots & (\boldsymbol{p}_2)_L \\ \vdots & \vdots & \ddots & \vdots \\ (\boldsymbol{p}_{|\mathbb{N}_D^M|})_1 & (\boldsymbol{p}_{|\mathbb{N}_D^M|})_2 & \cdots & (\boldsymbol{p}_{|\mathbb{N}_D^M|})_L \end{pmatrix}.$$

Additionally, introducing a vector $\boldsymbol{z}(\boldsymbol{t})$ defined by

$$\boldsymbol{z}(\boldsymbol{t}) \coloneqq \left( \binom{D}{\boldsymbol{d}_1} \boldsymbol{t}^{\boldsymbol{d}_1}, \ldots, \binom{D}{\boldsymbol{d}_{|\mathbb{N}_D^M|}} \boldsymbol{t}^{\boldsymbol{d}_{|\mathbb{N}_D^M|}} \right)^\top \in \mathbb{R}^{|\mathbb{N}_D^M|},$$

the Bézier simplex given in Eq. (1) is represented as $\boldsymbol{b}(\boldsymbol{t} \,|\, \boldsymbol{P}) = \boldsymbol{P}^\top \boldsymbol{z}(\boldsymbol{t})$. It is known that a Bézier simplex is a universal approximator of continuous functions:

**Theorem 2.2** (Kobayashi et al. (2019, Theorem 1))**.** *Let $\boldsymbol{\phi} : \Delta^{M-1} \to \mathbb{R}^M$ be a continuous map. There exists an infinite sequence of Bézier simplices $\boldsymbol{b}^{(i)} : \Delta^{M-1} \to \mathbb{R}^M$ such that*

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

By Theorems 2.1 and 2.2, the Pareto set and the Pareto front of any strongly convex problem is approximated by some Bézier simplex in arbitrary precision. This fact motivates the use of Bézier simplices to approximate the Pareto set.

### 2.3 Stochastic Gradient Descent

Let $d > 0$ be a positive integer. For a finite number of functions $\ell_1, \ldots, \ell_n$, we consider minimizing the following finite sum of functions:

$$\min_{\boldsymbol{x} \in \mathbb{R}^d} \ell(\boldsymbol{x}) \coloneqq \frac{1}{n} \sum_{i=1}^n \ell_i(\boldsymbol{x}). \tag{2}$$

Let $\mathscr{X}^* \subseteq \mathbb{R}^d$ be the set of optimal solutions to the unconstrained optimization problem (2), and we assume that $\mathscr{X}^*$ is nonempty, i.e., $\mathscr{X}^* \coloneqq \arg\min_{\boldsymbol{x}} \ell(\boldsymbol{x}) \neq \emptyset$.

For a standard SGD algorithm, the gradient is estimated with a minibatch $B \subseteq \{1, \ldots, n\}$ as

$$\boldsymbol{g}(\boldsymbol{x}) = \frac{1}{|B|} \sum_{i \in B} v_i \nabla \ell_i(\boldsymbol{x}), \tag{3}$$

where $\boldsymbol{v} = (v_1, \ldots, v_{|B|})^\top \in \mathbb{R}_{\geq 0}^{|B|}$ is a random sampling vector satisfying $\mathbb{E}[v_i] = 1$ for $i = 1, \ldots, |B|$. One can show that the gradient estimator given as Eq. (3) is unbiased since we have $\mathbb{E}[\boldsymbol{g}(\boldsymbol{x})] = \frac{1}{|B|} \sum_{i \in B} \mathbb{E}[v_i] \nabla \ell_i(\boldsymbol{x}) = \nabla \ell(\boldsymbol{x})$.

Following Shamir and Zhang (2013); Gower et al. (2019, 2021), define the gradient noise as follows.

**Definition 2.3** (Gradient noise)**.** For an unbiased gradient estimator $\boldsymbol{g}(\boldsymbol{x})$ of $\ell(\boldsymbol{x})$, the gradient noise is defined by

$$\sigma_\ell \coloneqq \sup_{\boldsymbol{x}^* \in \mathscr{X}^*} \mathbb{E}\left[ \|\boldsymbol{g}(\boldsymbol{x}^*)\|^2 \right].$$

The analysis of the convergence property of the SGD algorithm is based on the Expected Smoothness (ES) condition, which is stated as follows.

**Definition 2.4** (Expected Smoothness)**.** For an unbiased gradient estimator $\boldsymbol{g}(\boldsymbol{x})$ of $\ell(\boldsymbol{x})$, we say that the expected smoothness holds or $\boldsymbol{g} \in \mathrm{ES}(\rho)$ if

$$\frac{1}{2\rho} \mathbb{E}\left[ \|\boldsymbol{g}(\boldsymbol{x}) - \boldsymbol{g}(\boldsymbol{x}^*)\|^2 \right] \leq \ell(\boldsymbol{x}) - \ell(\boldsymbol{x}^*)$$

holds for any $\boldsymbol{x} \in \mathbb{R}^d$.

## 3 OUR FRAMEWORK

This section provides our proposed method for multi-objective optimization with the Bézier simplex. Our method is motivated by the scalarization of the multi-objective optimization problem to a single-objective optimization problem.

Define the general loss function over all possible $\boldsymbol{t} \in \Delta^{M-1}$ as follows:

$$\mathscr{L}_{\mathrm{gen}}(\boldsymbol{P}) \coloneqq \mathbb{E}_{\boldsymbol{t} \sim U(\Delta^{M-1})}\left[ \boldsymbol{t}^\top \boldsymbol{f}(\boldsymbol{b}(\boldsymbol{t} \,|\, \boldsymbol{P})) \right],$$

where the symbol $U(\cdot)$ denotes a uniform distribution on the argumant space. Notice that the loss function of this type has been considered in previous research on Pareto set learning (Navon et al., 2020; Lin et al., 2021; Chen and Kwok, 2024). Evaluating the expectation over the entire simplex is intractable; thus we draw finite samples from the uniform distribution $U(\Delta^{M-1})$, and define the following empirical loss function over the empirical distribution:

$$\mathscr{L}(\boldsymbol{P}) = \frac{1}{n} \sum_{i=1}^n \boldsymbol{t}_i^\top \boldsymbol{f}(\boldsymbol{b}(\boldsymbol{t}_i \,|\, \boldsymbol{P})), \tag{4}$$

where $\{\boldsymbol{t}_i\}_{i=1}^n \subseteq \Delta^{M-1}$ is a set of $n$ samples drawn independently from the uniform distribution $U(\Delta^{M-1})$. For a minibatch $B \subseteq [n]$, let us define $\mathscr{L}_B(\boldsymbol{P}) = \frac{1}{|B|} \sum_{i \in B} \boldsymbol{t}_i^\top \boldsymbol{f}(\boldsymbol{b}(\boldsymbol{t}_i \,|\, \boldsymbol{P}))$ and its gradient estimator as

$\nabla\mathscr{L}_B(\boldsymbol{P})$. At the $k$th iteration, the standard SGD updates the current control points $\boldsymbol{P}^{(k)}$ as $\boldsymbol{P}^{(k+1)} \leftarrow \boldsymbol{P}^{(k)} - \alpha_k \nabla\mathscr{L}_B\boldsymbol{P}^{(k)}$ where $\alpha_k > 0$ is a stepsize. This update rule is derived from the following optimization problem:

$$\boldsymbol{P}^{(k+1)} \in \arg\min_{\boldsymbol{P}} \left\{ \mathscr{L}(\boldsymbol{P}^{(k)}) + \left\langle \nabla\mathscr{L}_B(\boldsymbol{P}^{(k)}), \boldsymbol{P} - \boldsymbol{P}^{(k)} \right\rangle + \frac{1}{2\alpha_k n} \sum_{i=1}^{n} \left\| \boldsymbol{P} - \boldsymbol{P}^{(k)} \right\|_{\mathsf{F}}^2 \right\},$$

While the above update rule can be seen as natural, the above update rule does not consider the geometric structure of the Bézier simplex. Thus, we alternatively consider updating the control points $\boldsymbol{P}^{(k)}$ as an optimal solution to the following optimization problem:

$$\begin{aligned} &\boldsymbol{P}^{(k+1)} \\ &\in \arg\min_{\boldsymbol{P}} \left\{ \mathscr{L}(\boldsymbol{P}^{(k)}) + \left\langle \nabla\mathscr{L}_B(\boldsymbol{P}^{(k)}), \boldsymbol{P} - \boldsymbol{P}^{(k)} \right\rangle \right. \\ &\quad \left. + \frac{1}{2\alpha_k} \mathbb{E}_{\boldsymbol{t}\sim U(\Delta^{M-1})} \left[ \left\| \boldsymbol{b}(\boldsymbol{t} \mid \boldsymbol{P}) - \boldsymbol{b}(\boldsymbol{t} \mid \boldsymbol{P}^{(k)}) \right\|_2^2 \right] \right\}. \end{aligned} \tag{5}$$

The last term of the objective function is a regularization term that prevents the Bézier simplex from being too far apart from the previous Bézier simplex in the variable space. Here, the regularization term in Eq. (5) is represented as

$$\begin{aligned} &\mathbb{E}_{\boldsymbol{t}\sim U(\Delta^{M-1})} \left[ \left\| \boldsymbol{b}(\boldsymbol{t} \mid \boldsymbol{P}) - \boldsymbol{b}(\boldsymbol{t} \mid \boldsymbol{P}^{(k)}) \right\|_2^2 \right] \\ &= \mathbb{E}_{\boldsymbol{t}\sim U(\Delta^{M-1})} \left[ \left\| \left( \boldsymbol{P} - \boldsymbol{P}^{(k)} \right)^\top \boldsymbol{z}(\boldsymbol{t}) \right\|_2^2 \right] \\ &= \left\langle \mathbb{E}_{\boldsymbol{t}} \left[ \boldsymbol{z}(\boldsymbol{t}) \boldsymbol{z}(\boldsymbol{t})^\top \right], \left( \boldsymbol{P} - \boldsymbol{P}^{(k)} \right) \left( \boldsymbol{P} - \boldsymbol{P}^{(k)} \right)^\top \right\rangle. \end{aligned}$$

Define $\boldsymbol{\Sigma} \in \mathbb{R}^{|\mathbb{N}_D^M| \times |\mathbb{N}_D^M|}$ as

$$\boldsymbol{\Sigma} \coloneqq \mathbb{E}_{\boldsymbol{t}\sim U(\Delta^{M-1})} \left[ \boldsymbol{z}(\boldsymbol{t}) \boldsymbol{z}(\boldsymbol{t})^\top \right].$$

It has shown that the matrix $\boldsymbol{\Sigma} \in \mathbb{R}^{|\mathbb{N}_D^M| \times |\mathbb{N}_D^M|}$ is invertible (Tanaka et al., 2020, Theorem 3) and its $(i, j)$ component is given as follows:

$$\boldsymbol{\Sigma}_{ij} = \frac{(2D)!(M-1)!}{(2D+M-1)!} \binom{D}{\boldsymbol{d}_i} \binom{D}{\boldsymbol{d}_j} \binom{2D}{\boldsymbol{d}_i + \boldsymbol{d}_j}^{-1}.$$

Since the matrix $\boldsymbol{\Sigma}$ is positive definite, the objective function in Eq. (5) is strongly convex, and its optimal solution is unique. Thus, $\boldsymbol{P}^{(k+1)}$ can be obtained as the solution to the following linear system:

$$\nabla\mathscr{L}_B(\boldsymbol{P}^{(k)}) + \left\langle \boldsymbol{P} - \boldsymbol{P}^{(k)}, \mathbb{E}_{\boldsymbol{t}} \left[ \boldsymbol{z}(\boldsymbol{t}) \boldsymbol{z}(\boldsymbol{t})^\top \right] \right\rangle = \boldsymbol{O}.$$

**Algorithm 1:** SGD for Parametric Hypersurface Representation of Pareto Set of Multi-Objective Optimization

**Input:** Initial control points $\boldsymbol{P}_0$, sample size $n$, batch size $m$, and max iterations $K > 0$
**Output:** Optimal control points $\boldsymbol{P}$.
**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 $U(\Delta^{M-1})$, and define the empirical loss function as $\mathscr{L}(\boldsymbol{P}) = \frac{1}{n}\sum_{i=1}^n \boldsymbol{t}_i^\top \boldsymbol{f}(\boldsymbol{b}(\boldsymbol{t} \mid \boldsymbol{P}))$.
**while** $k < K$ **do**
  Choose $B \subseteq [n]$ of $|B| = m$ 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** $\boldsymbol{P}^{(K)}$.

By solving the above equation, the update rule of the control points is represented as follows:

$$\boldsymbol{P}^{(k+1)} = \boldsymbol{P}^{(k)} - \alpha_k \boldsymbol{\Sigma}^{-1} \nabla\mathscr{L}_B(\boldsymbol{P}^{(k)}), \tag{6}$$

where $\alpha_k > 0$ is a predetermined step size at the $k$th iteration. Note that instead of explicitly computing $\boldsymbol{\Sigma}^{-1}$, performing the Cholesky decomposition of $\boldsymbol{\Sigma}$ once allows us to efficiently calculate $\boldsymbol{P}^{(k+1)}$ in Eq. (6). The computational complexity of the Cholesky decomposition is $\mathrm{O}(|\mathbb{N}_D^M|^3)$, but it needs to be performed only once, not in each iteration.

We present a pseudocode of a stochastic gradient descent algorithm for the empirical loss function in Algorithm 1.

## 4 CONVERGENCE ANALYSIS

In this section, we provide the convergence analysis of the proposed algorithm. For the space limitation, all the proofs are provided in the supplementary materials. To discuss the convergence analysis, we make the following mild assumptions on the objective functions and gradient noise, which are often used in the analysis of stochastic optimization algorithms (Gower et al., 2019, 2021; Gorbunov et al., 2022).

**Assumption 4.1.** All the objective functions $f_1, \ldots, f_M$ are $\mu$-strongly convex and $\rho$-smooth.

**Assumption 4.2.** The gradient noise $\sigma_{\mathscr{L}}$ is finite.

In addition, we put an assumption on the set of parameters $\{\boldsymbol{t}_i\}_{i=1}^n$ used for the empirical loss function.

**Assumption 4.3.** The matrix $\boldsymbol{Z} \coloneqq (\boldsymbol{z}(\boldsymbol{t}_1), \ldots, \boldsymbol{z}(\boldsymbol{t}_n))$ is full row rank.

**Lemma 4.4.** *Under Assumption 4.3, let $\overline{\sigma} > 0$ be the smallest singular value of $\boldsymbol{Z}$. Then, the empirical loss function $\mathscr{L}$ defined by Eq. (4) is strongly convex with constant $\overline{\mu} \coloneqq \mu\overline{\sigma}^2/n$.*

Under the assumptions above, we can show that the proposed method achieves the following bounds, which supports the empirical loss function decreasing. In what follows, the symbol $\boldsymbol{P}^*$ denotes the minimizer of the empirical loss $\mathscr{L}$ defined in Eq. (4).

**Theorem 4.5.** *Assume that Assumptions 4.1, 4.2 and 4.3 hold, and there exists $\tilde{\rho}$ such that the gradient estimator holds $\mathscr{L}_B \in \mathrm{ES}(\tilde{\rho})$ for any $B \subseteq [n]$. Consider a sequence of control points $\{\boldsymbol{P}^{(k)}\}_{k\in\mathbb{N}}$ generated by Algorithm 1 with a stepsize $\alpha_k \leq \min\{(4\widetilde{\rho}\lambda_{\max}(\boldsymbol{\Sigma}^{-1}))^{-1}, (\overline{\mu}\lambda_{\min}(\boldsymbol{\Sigma}^{-1}))^{-1}\}$ and an initial control point $\boldsymbol{P}_0$. Then, for any $K > 0$, we have*

$$\begin{aligned}&\mathbb{E}\left[\mathscr{L}(\widetilde{\boldsymbol{P}}^{(K)}) - \mathscr{L}(\boldsymbol{P}^*)\right]\\ &\leq \frac{1 - \alpha_0\overline{\mu}\lambda_{\min}(\boldsymbol{\Sigma}^{-1})}{\sum_{k=0}^{K-1}\alpha_k}\|\boldsymbol{P}_0 - \boldsymbol{P}^*\|_{\boldsymbol{\Sigma}}^2\\ &\quad + \frac{2\lambda_{\max}(\boldsymbol{\Sigma}^{-1})\sum_{k=0}^{K-1}\alpha_k^2}{\sum_{k=0}^{K-1}\alpha_k}\sigma_{\mathscr{L}},\end{aligned}$$

*where $\widetilde{\boldsymbol{P}}^{(K)}$ is the weighted average of control points $\boldsymbol{P}^{(k)}$ for $k = 0, 1, \ldots, K-1$ which is defines as*

$$\widetilde{\boldsymbol{P}}^{(K)} \coloneqq \sum_{k=0}^{K-1} q_{K,k}\boldsymbol{P}^{(k)},$$
$$\text{with}\quad q_{K,k} \coloneqq \frac{\alpha_k(1 - 2\alpha_k\tilde{\rho})}{2\sum_{\kappa=0}^{K-1}\alpha_\kappa(1 - 2\alpha_\kappa\tilde{\rho})}.$$

Compared to the convergence analysis of the standard SGD algorithm (Garrigos and Gower, 2023, Theorem 5.3), we obtain a better upper bound for $\mathbb{E}\left[\mathscr{L}(\widetilde{\boldsymbol{P}}^{(K)}) - \mathscr{L}(\boldsymbol{P}^*)\right]$ in Algorithm 1. The following corollary establishes the convergence rate with a specific stepsize, including constant stepsize and diminishing stepsize.

**Corollary 4.6.** *Consider the situation as in Theorem 4.5. Let $\alpha_k = \alpha \leq \min\{(4\widetilde{\rho}\lambda_{\max}(\boldsymbol{\Sigma}^{-1}))^{-1}, (\overline{\mu}\lambda_{\min}(\boldsymbol{\Sigma}^{-1}))^{-1}\}$. Then, for any $K > 0$, we have*

$$\begin{aligned}&\mathbb{E}\left[\mathscr{L}(\overline{\boldsymbol{P}}^{(K)}) - \mathscr{L}(\boldsymbol{P}^*)\right]\\ &\leq \frac{1 - \alpha\overline{\mu}\lambda_{\min}(\boldsymbol{\Sigma}^{-1})}{\alpha K}\|\boldsymbol{P}_0 - \boldsymbol{P}^*\|_{\boldsymbol{\Sigma}}^2\\ &\quad + 2\alpha\lambda_{\max}(\boldsymbol{\Sigma}^{-1})\sigma_{\mathscr{L}} = \mathrm{O}\left(\frac{1}{K}\right),\end{aligned}$$

*where $\overline{\boldsymbol{P}}^{(K)} \coloneqq \frac{1}{K}\sum_{k=1}^{K}\boldsymbol{P}^{(k)}$. Additionally, if we choose a stepsize as $\alpha_k = \frac{\alpha_0}{\sqrt{k+1}}$ for $\alpha_0 \leq \min\{(4\widetilde{\rho}\lambda_{\max}(\boldsymbol{\Sigma}^{-1}))^{-1}, (\overline{\mu}\lambda_{\min}(\boldsymbol{\Sigma}^{-1}))^{-1}\}$, then we have*

$$\begin{aligned}&\mathbb{E}\left[\mathscr{L}(\overline{\boldsymbol{P}}^{(K)}) - \mathscr{L}(\boldsymbol{P}^*)\right]\\ &\leq \frac{5(1 - \alpha_0\overline{\mu}\lambda_{\min}(\boldsymbol{\Sigma}^{-1}))}{4\alpha_0\sqrt{K}}\|\boldsymbol{P}_0 - \boldsymbol{P}^*\|_{\boldsymbol{\Sigma}}^2\\ &\quad + \frac{5\alpha_0\ln(K+1)}{\sqrt{K}}\sigma_{\mathscr{L}} = \mathrm{O}\left(\frac{\ln K}{\sqrt{K}}\right).\end{aligned}$$

Regarding the convergence property for the control points of the Bézier simplex generated by Algorithm 1, we can achieve the following convergence results.

**Theorem 4.7.** *Assume that Assumptions 4.1, 4.2 and and 4.3 hold, and there exists $\tilde{\rho}$ such that the gradient estomator holds $\mathscr{L}_B \in \mathrm{ES}(\tilde{\rho})$ for any $B \subseteq [n]$. 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 constant stepsize $\alpha_k = \alpha < \left(2\tilde{\rho}\lambda_{\max}(\boldsymbol{\Sigma}^{-1})\right)^{-1}$. Then, we have*

$$\begin{aligned}&\mathbb{E}\left[\left\|\boldsymbol{P}^{(k)} - \boldsymbol{P}^*\right\|_{\boldsymbol{\Sigma}}^2\right]\\ &\leq \left(1 - \alpha\overline{\mu}\lambda_{\min}(\boldsymbol{\Sigma}^{-1})\right)^k\|\boldsymbol{P}_0 - \boldsymbol{P}^*\|_{\boldsymbol{\Sigma}}^2\\ &\quad + \frac{2\alpha\,\mathrm{cond}(\boldsymbol{\Sigma}^{-1})}{\overline{\mu}}\sigma_{\mathscr{L}}.\end{aligned}$$

The following corollary provides the number of minimum iterations to achieve the given tolerance.

**Corollary 4.8.** *Consider the situation as in Theorem 4.7. Let $\alpha = \min\{\frac{\overline{\mu}\varepsilon}{4\,\mathrm{cond}(\boldsymbol{\Sigma}^{-1})\sigma_{\mathscr{L}}}, \frac{1}{4\widetilde{\rho}\lambda_{\max}(\boldsymbol{\Sigma}^{-1})}\}$. Then, for any $\varepsilon > 0$, we can guarantee that $\left\|\boldsymbol{P}^{(k)} - \boldsymbol{P}^*\right\|_{\boldsymbol{\Sigma}}^2 \leq \varepsilon$ provided that*

$$\begin{aligned}k \geq \max&\left\{\frac{1}{\varepsilon}\frac{4\,\mathrm{cond}(\boldsymbol{\Sigma}^{-1})\sigma_{\mathscr{L}}}{\overline{\mu}^2}, \frac{4\widetilde{\rho}\lambda_{\max}(\boldsymbol{\Sigma}^{-1})}{\overline{\mu}}\right\}\\ &\times \ln\left(\frac{2\|\boldsymbol{P}_0 - \boldsymbol{P}^*\|_{\boldsymbol{\Sigma}}^2}{\varepsilon}\right).\end{aligned}$$

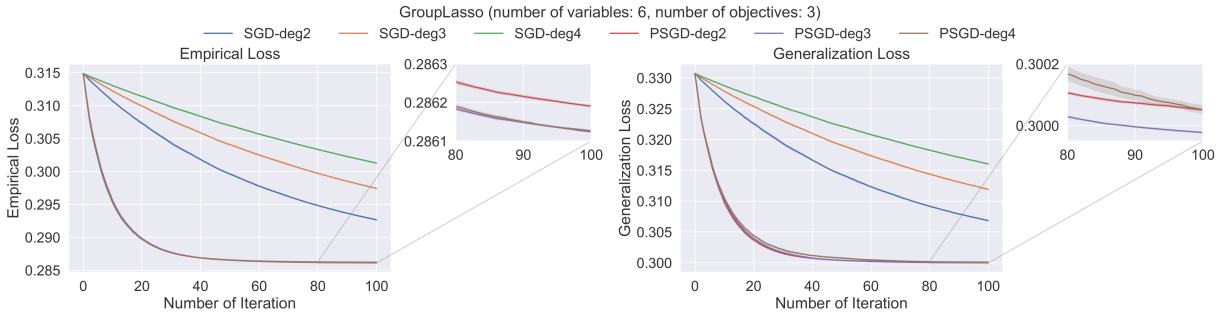
# 5 EXPERIMENTS

In this section, we investigate the performance of the proposed algorithm on several multi-objective optimization instances[1].

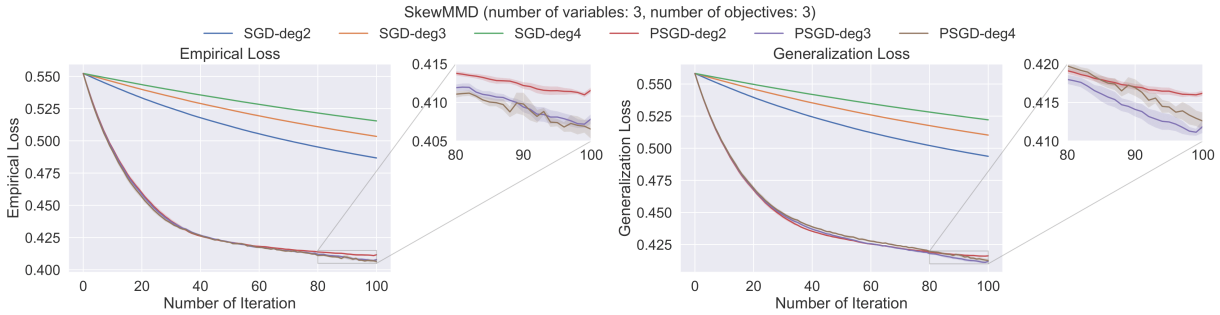
## 5.1 Experimental Setup

The primal objectives of the experiments are (i) to investigate the convergence behavior of the proposed

[1]The source code is available at `https://github.com/hikimay/sgd-bf`.




(a) Group LASSO




(b) Skew-MMD

Figure 1: The history of the empirical loss and the generalization loss for the group Lasso and skew-MMD.

algorithm, (ii) to verify that the proposed algorithm can well approximate the Pareto optimal set and front, and (iii) to compare the generalization performance of the proposed algorithm with the naive method.

As a baseline, we compare the proposed algorithm with the algorithm based on the naive stochastic gradient descent method, which updates the control points as $\boldsymbol{P}^{(k+1)} \leftarrow \boldsymbol{P}^{(k)} - \alpha_k \mathscr{L}_B(\boldsymbol{P}^{(k)})$ in the line 6 of Algorithm 1. We refer to the naive method as SGD, and the proposed method as PSGD. For both methods, we set the maximum number of iterations as $K = 100$, the stepsize as $\alpha_k = 0.05$ for all $k \in \{0, 1, \ldots, K-1\}$, the number of samples as $n = 100$, and the number of batchsize as $m = 20$. We set the degree of the Bézier simplex as $D \in \{2, 3, 4\}$. For each problem instance, we repeat the experiments five times with different samples from the uniform simplex. In all experiments, we set the initial control points as $\boldsymbol{P}_0 = \boldsymbol{O}$, that is, all control points are set to be an origin. To evaluate the generalization performance, we generated 10,000 samples from a uniform distribution on $\Delta^{M-1}$, which were not used for the algorithm. All the experiments were implemented in Python 3.12 and performed on macOS Sonoma with Apple M1 Pro CPU and 16 GB memory.

### 5.2 Problem Descriptions

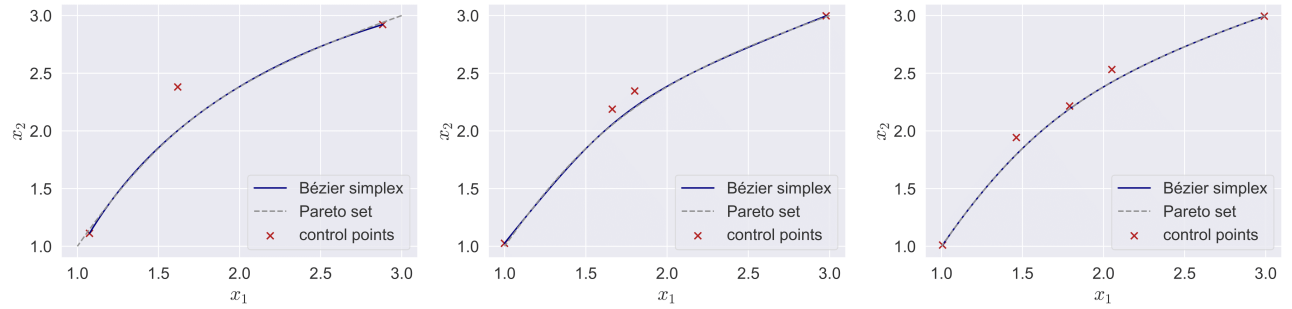
To investigate the convergence behavior, we solve the group LASSO (Yuan and Lin, 2006) on the `Birthwt` dataset (Hosmer Jr et al., 2013; Venables and Ripley, 2002), and skew-MMD problem (Harada et al., 2006; Hamada et al., 2010).

The problem statement for the group LASSO on the `Birthwt` dataset is described as follows. In this experiment, we adopted six continuous features $\boldsymbol{x} \in \mathbb{R}^6$ which consists of $\boldsymbol{x}_{\mathrm{age}} \in \mathbb{R}^3$ and $\boldsymbol{x}_{\mathrm{lwt}} \in \mathbb{R}^3$ as predictors and one continuous target $y \in \mathbb{R}$ as a response. According to Tanaka et al. (2020, Section. 4), the group LASSO regressor for the `Birthwt` dataset can be obtained by the solution to the following optimization problem:

$$\underset{\boldsymbol{x} \in \mathbb{R}^6}{\text{minimize}} \ \frac{1}{N}\|\boldsymbol{A}\boldsymbol{x} - \boldsymbol{y}\|_2^2 + \frac{\lambda}{3}(\|\boldsymbol{x}_{\mathrm{age}}\|_2 + \|\boldsymbol{x}_{\mathrm{lwt}}\|_2)$$

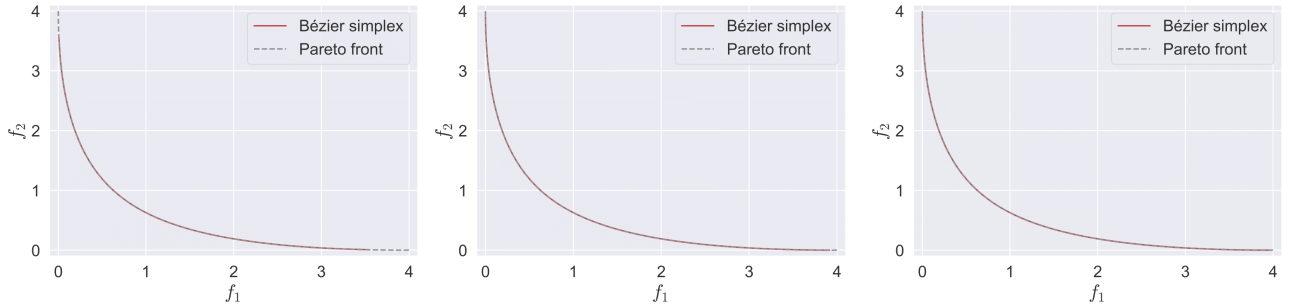
where $\boldsymbol{A}$ is a observation matrix of size $N \times M$ where $N = 189, M = 6$, and $\lambda > 0$ is a regularization parameter to be predetermined by users. One can reformulate the above optimization problem as the following multi-objective optimization problem:

$$\underset{\boldsymbol{x} \in \mathbb{R}^6}{\text{minimize}} \ \left(\widetilde{f}_1(\boldsymbol{x}), \widetilde{f}_2(\boldsymbol{x}), \widetilde{f}_3(\boldsymbol{x})\right)^\top$$
$$\text{where} \quad \widetilde{f}_1(\boldsymbol{x}) \coloneqq \frac{1}{N}\|\boldsymbol{A}\boldsymbol{x} - \boldsymbol{y}\|_2^2 + \widetilde{\varepsilon}\|\boldsymbol{x}\|_2^2,$$




(a) $D = 2$ (b) $D = 3$ (c) $D = 4$

Figure 2: Approximated Pareto set of the problem described in Eq. (10). The orange cross points are the control points of the Bézier simplex obtained by the proposed method.




(a) $D = 2$ (b) $D = 3$ (c) $D = 4$

Figure 3: Approximated Pareto front of the problem described in Eq. (10).

$$\widetilde{f}_2(\boldsymbol{x}) \coloneqq \|\boldsymbol{x}_{\mathrm{age}}\|_2^2 + \widetilde{\varepsilon}\|\boldsymbol{x}\|_2^2,$$
$$\widetilde{f}_3(\boldsymbol{x}) \coloneqq \|\boldsymbol{x}_{\mathrm{lwt}}\|_2^2 + \widetilde{\varepsilon}\|\boldsymbol{x}\|_2^2.$$

In the above formulation, $\widetilde{\varepsilon} > 0$ is a small perturbation parameter to make the original objective functions strongly convex. We set $\widetilde{\varepsilon} = 10^{-4}$ in the experiments.

The **skew-MMD** is a multi-objective optimization problem with $L$ number of variables and $M$ number of objective functions. This problem includes a generalized fitting problem that encompasses various statistical methods, such as sparse modelling, generalized linear models, transfer learning, and robust estimation.[2] For real matrices $\boldsymbol{A}_1, \ldots, \boldsymbol{A}_M$, real vectors $\boldsymbol{c}_1, \ldots, \boldsymbol{c}_M$, and positive real numbers $p_1, \ldots, p_M$, the skew-MMD is defined as follows:

$$\begin{aligned} &\underset{\boldsymbol{x}\in\mathbb{R}^L}{\text{minimize}} \ (f_1(\boldsymbol{x}), \ldots, f_M(\boldsymbol{x}))^\top \\ &\text{where} \quad f_i(\boldsymbol{x}) \coloneqq \|\boldsymbol{A}_i(\boldsymbol{x} - \boldsymbol{c}_i)\|_2^{p_i} \quad (i = 1, \ldots, M). \end{aligned}$$

In this experiments, we set $L = 3$, $M = 3$, and $\boldsymbol{A}_i$, $\boldsymbol{c}_i$, $p_i$ are defined as follows:

$$\begin{aligned} \boldsymbol{A}_1 &\coloneqq \mathrm{diag}(3/5, 4/5, 4/5), \\ \boldsymbol{A}_2 &\coloneqq \mathrm{diag}(4/5, 3/5, 4/5), \\ \boldsymbol{A}_3 &\coloneqq \mathrm{diag}(4/5, 4/5, 3/5), \\ \boldsymbol{c}_i &\coloneqq \boldsymbol{e}_i \quad (i = 1, \ldots, 3), \\ p_i &\coloneqq \exp\left(\frac{2(i-1)}{M-1} - 1\right) \quad (i = 1, \ldots, 3). \end{aligned}$$

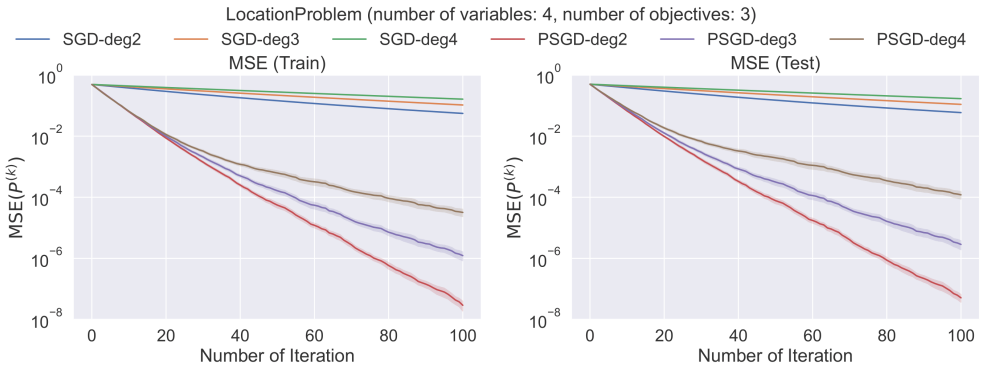
Next, to verify that the Pareto optimal set and front can be approximated by the proposed algorithm, we consider the problem instance from Bergou et al. (2020), which is defined as follows:

$$\begin{aligned} &\underset{\boldsymbol{x}\in\mathbb{R}^2}{\text{minimize}} \ (f_1(\boldsymbol{x}), f_2(\boldsymbol{x}))^\top \\ &\text{where} \quad f_1(\boldsymbol{x}) \coloneqq (x_1 - 1)^2 + (x_1 - x_2)^2, \\ &\qquad\quad f_2(\boldsymbol{x}) \coloneqq (x_2 - 3)^2 + (x_1 - x_2)^2. \end{aligned}$$

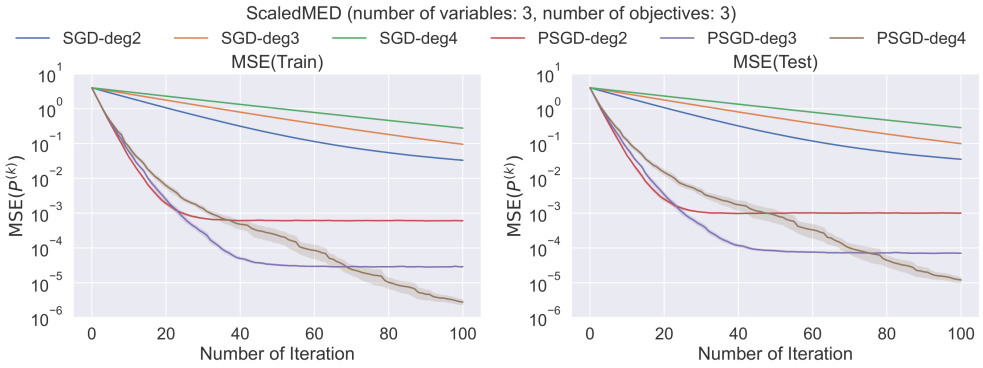
Note that the optimal solution to the scalarization function with $\boldsymbol{t} = (t_1, t_2)^\top \in \Delta^1$ is given as follows:

$$x_1^*(\boldsymbol{t}) = \frac{t_1^2 + t_1 - 3}{t_1^2 - t_1 - 1}, \quad x_2^*(\boldsymbol{t}) = \frac{3t_1^2 - t_1 - 3}{t_1^2 - t_1 - 1}.$$

[2] In fact, the group LASSO applied to the `Birthwt` dataset discussed in this paper is recovered by the skew-MMD with the following parameters: $\boldsymbol{A}_1 = \boldsymbol{A}$, $\boldsymbol{A}_2 = \mathrm{diag}(1, 1, 1, 0, 0, 0)$, $\boldsymbol{A}_3 = \mathrm{diag}(0, 0, 0, 1, 1, 1)$, $\boldsymbol{c}_1 = \boldsymbol{A}^+\boldsymbol{y}$, $\boldsymbol{c}_2 = \boldsymbol{0}$, $\boldsymbol{c}_3 = \boldsymbol{0}$, $p_1 = 2$, $p_2 = 1$, $p_3 = 1$.




(a) Location Problem




(b) scaled-MED

Figure 4: The mean squared error of the generalized location problem described in Eq. (11) and scaled-MED.

Finally, to investigate the generalization performance, we consider the generalized location problem and the scaled-MED. The **generalized location problem** is a multi-objective optimization problem with $L$-dimensional decision variables and $M = 3$ number of objective functions, which is defined as

$$\begin{aligned} &\underset{\boldsymbol{x}\in\mathbb{R}^L}{\text{minimize}}\ (f_1(\boldsymbol{x}), f_2(\boldsymbol{x}), f_3(\boldsymbol{x}))^\top \\ &\text{where}\quad f_i(\boldsymbol{x}) \coloneqq \|\boldsymbol{x} - \boldsymbol{e}_i\|_2^2 \quad (i = 1, 2, 3), \end{aligned}$$

and $\boldsymbol{e}_i \in \mathbb{R}^L$ for $i = 1, \ldots, 3$ are the unit vector whose $i$-th component is 1; otherwise 0.

The **scaled-MED** consists of a 3-dimensional decision variables and $M = 3$ number of objective functions, which is defined as

$$\begin{aligned} &\underset{\boldsymbol{x}\in\mathbb{R}^3}{\text{minimize}}\ (f_1(\boldsymbol{x}), f_2(\boldsymbol{x}), f_3(\boldsymbol{x}))^\top \\ &\text{where}\ f_1(\boldsymbol{x}) \coloneqq x_1^2 + 3(x_2 - 1)^2 + 2(x_3 - 1)^2, \\ &\qquad f_2(\boldsymbol{x}) \coloneqq 2(x_1 - 1)^2 + x_2^2 + 3(x_3 - 1)^2, \\ &\qquad f_3(\boldsymbol{x}) \coloneqq 3(x_1 - 1)^2 + 2(x_2 - 1)^2 + (x_3 + 1)^2. \end{aligned}$$

### 5.3 Performance Metrics

The mean squared error (MSE) is used to evaluate the performance, which is defined as

$$\mathrm{MSE}(\boldsymbol{P}) \coloneqq \frac{1}{N}\sum_{i=1}^{N} \|\boldsymbol{b}(\boldsymbol{t}_i \mid \boldsymbol{P}) - \boldsymbol{x}^*(\boldsymbol{t}_i)\|_2^2,$$

where $\{\boldsymbol{t}_i\}_{i=1}^N \subseteq \Delta^{M-1}$ are samples independently drawn from the uniform distribution $U(\Delta^{M-1})$, and $\boldsymbol{x}^* \colon \Delta^{M-1} \to \mathscr{X}$ is the optimal map from the simplex to the Pareto set. For most multi-objective optimization problems, one cannot access the map $\boldsymbol{x}^*$; however, we have a closed-form expression of $\boldsymbol{x}^*$ for the generalized location problem and scaled-MED.

Specifically, the Pareto set of the generalized location problem is known to be a convex hull of $\boldsymbol{e}_1$, $\boldsymbol{e}_2$, and $\boldsymbol{e}_3$, and every solution on the Pareto set can be given as $\boldsymbol{x}^*(\boldsymbol{t}) = t_1\boldsymbol{e}_1 + t_2\boldsymbol{e}_2 + t_3\boldsymbol{e}_3$. Regarding the scaled-MED, the optimal solution on the Pareto set can be derived as the following expression:

$$\left(\frac{2t_2 + 3t_3}{t_1 + 2t_2 + 3t_3}, \frac{3t_1 + 2t_3}{3t_1 + t_2 + 2t_3}, \frac{2t_1 + 3t_2 - t_3}{2t_1 + 3t_2 + t_3}\right)^\top.$$

### 5.4 Results

Figure 1 shows the history of the empirical and generalization losses for the group LASSO and skew-MMD. The left-hand side of both Figures 1a, 1b shows the history of the empirical loss, and the right-hand side shows the history of the generalization loss, respectively. The error bands represent 95% confidence interval for 5 trials. We can see that the proposed method (PSGD) converges much faster than the naive method (SGD) in both problems. For the proposed method, the loss decreases at almost the same rate regardless of the degree $D$, whereas the naive method shows a slower convergence as $D$ increases.

Figures 2 and 3 show the approximated Pareto set/front for the multi-objective optimization problem described in Eq. (10). We also show the true Pareto set/front by the dotted gray line. We present the approximated Pareto set/front with the control points of the Bézier simplex with degree $D \in \{2, 3, 4\}$, which is obtained by the proposed method after $K = 100$ iterations. We can see that the results of the degree $D = 3, 4$ are closer to the true Pareto set/front than the degree $D = 2$. Indeed, the result of $D = 2$ does not adequately represent the edge of the Pareto set and front.

Finally, Figure 4 shows the history of MSE for the generalized location problem and scaled-MED. Both results show that the proposed method (PSGD) outperforms the naive method (SGD) in terms of the mean squared error. These results support that the proposed method can achieve better generalization performance than the naive method.

## 6 CONCLUDING REMARKS

This paper proposes a stochastic gradient descent (SGD) based method for multi-objective optimization problems with a Bézier simplex. The proposed method updates the control points of the Bézier simplex by minimizing the empirical loss function. Unlike the standard SGD, the proposed method introduces a regularization term that prevents the current Bézier simplex from being too far apart from the previous one. We have provided a convergence analysis and demonstrated that our method achieves a faster convergence rate than the standard SGD.

As a limitation of this work, we focus on the strongly convex unconstrained multi-objective optimization problems. It would be interesting to extend the algorithm to the non-convex and constrained multi-objective optimization problems.

### Acknowledgements

This work was supported by JSPS KAKENHI Grant Numbers JP24K17465, JP20H02385 and JP22H05116.

### References

Bergou, E.-H., Diouane, Y., and Kungurtsev, V. (2020). Complexity iteration analysis for strongly convex multi-objective optimization using a newton path-following procedure. *Optimization Letters*, 15(4):1215–1227.

Chen, W. and Kwok, J. T. (2024). Efficient pareto manifold learning with low-rank structure. *arXiv preprint arXiv:2407.20734*.

Daulton, S., Balandat, M., and Bakshy, E. (2024). Parallel bayesian optimization of multiple noisy objectives with expected hypervolume improvement. In *Proceedings of the 35th International Conference on Neural Information Processing Systems*.

Daulton, S., Eriksson, D., Balandat, M., and Bakshy, E. (2022). Multi-objective Bayesian optimization over high-dimensional search spaces. In *Proceedings of the 38th Conference on Uncertainty in Artificial Intelligence*, volume 180, pages 507–517. PMLR.

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

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.

Ehrgott, M. (2005). *Multicriteria Optimization*, volume 491. Springer Science & Business Media.

Fliege, J., Vaz, A. I. F., and Vicente, L. N. (2018). Complexity of gradient descent for multiobjective optimization. *Optimization Methods and Software*, 34(5):949–959.

Garrigos, G. and Gower, R. M. (2023). Handbook of convergence theorems for (stochastic) gradient methods. *arXiv preprint arXiv:2301.11235*.

Gorbunov, E., Berard, H., Gidel, G., and Loizou, N. (2022). Stochastic extragradient: General analysis and improved rates. In *Proceedings of the 25th International Conference on Artificial Intelligence and Statistics*, volume 151, pages 7865–7901. PMLR.

Gower, R. M., Loizou, N., Qian, X., Sailanbayev, A., Shulgin, E., and Richtárik, P. (2019). SGD: General analysis and improved rates. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97, pages 5200–5209. PMLR.

Gower, R. M., Sebbouh, O., and Loizou, N. (2021). SGD for structured nonconvex functions: Learning rates, minibatching and interpolation. In *Proceedings of the 24th International Conference on Artificial Intelligence and Statistics*, volume 130, pages 1315–1323. PMLR.

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*, pages 787–794.

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 2006*, pages 659–666.

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 2007*, pages 813–820.

Hernandez-Lobato, D., Hernandez-Lobato, J., Shah, A., and Adams, R. (2016). Predictive entropy search for multi-objective bayesian optimization. In *Proceedings of The 33rd International Conference on Machine Learning*, volume 48, pages 1492–1501. 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 Jr, D. W., Lemeshow, S., and Sturdivant, R. X. (2013). *Applied Logistic Regression*. John Wiley & Sons.

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, pages 2304–2313.

Lin, X., Yang, Z., Zhang, Q., and Kwong, S. (2021). Controllable pareto multi-task learning. *arXiv preprint arXiv:2010.06313*.

Lin, X., Yang, Z., Zhang, X., and Zhang, Q. (2022). Pareto set learning for expensive multi-objective optimization. In *Advances in Neural Information Processing Systems*.

Lin, X., Zhang, X., Yang, Z., Liu, F., Wang, Z., and Zhang, Q. (2024). Smooth tchebycheff scalarization for multi-objective optimization. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 30479–30509. PMLR.

Miettinen, K. (1999). *Nonlinear Multiobjective Optimization*. Springer US.

Mizota, Y., Hamada, N., and Ichiki, S. (2021). All unconstrained strongly convex problems are weakly simplicial. *arXiv,* arXiv:2106.12704.

Navon, A., Shamsian, A., Chechik, G., and Fetaya, E. (2020). Learning the pareto front with hypernetworks. *arXiv preprint arXiv:2010.04104*.

Nesterov, Y. (2018). *Lectures on Convex Optimization*. Springer International Publishing.

Sannai, A., Hikima, Y., Kobayashi, K., Tanaka, A., and Hamada, N. (2022). Bézier flow: a surface-wise gradient descent method for multi-objective optimization. *arXiv preprint arXiv:2205.11099*.

Shamir, O. and Zhang, T. (2013). Stochastic gradient descent for non-smooth optimization: Convergence results and optimal averaging schemes. In *Proceedings of International Conference on Machine Learning*, volume 28, pages 71–79. PMLR.

Tanabe, H., Fukuda, E. H., and Yamashita, N. (2018). Proximal gradient methods for multiobjective optimization and their applications. *Computational Optimization and Applications*, 72(2):339–361.

Tanaka, A., Sannai, A., Kobayashi, K., and Hamada, N. (2020). Asymptotic risk of bézier simplex fitting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 2416–2424.

Tanaka, A., Sannai, A., Kobayashi, K., and Hamada, N. (2021). Approximate Bayesian computation of Bézier simplices. *arXiv preprint arXiv:arXiv:2104.04679*.

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

Wan, Y.-H. (1977). On the algebraic criteria for local Pareto optima–I. *Topology*, 16:113–117.

Wan, Y.-H. (1978). On the algebraic criteria for local Pareto optima. II. *Transactions of the American Mathematical Society*, 245:385–397.

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.

## Checklist

The checklist follows the references. For each question, choose your answer from the three possible options: Yes, No, Not Applicable. You are encouraged to include a justification to your answer, either by referencing the appropriate section of your paper or providing a brief inline description (1-2 sentences). Please do not modify the questions. Note that the Checklist section does not count towards the page limit. Not including the checklist in the first submission won't result in desk rejection, although in such case we will ask you to upload it during the author response period and include it in camera ready (if accepted).

**In your paper, please delete this instructions block and only keep the Checklist section heading above along with the questions/answers below.**

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes]
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Yes]
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [Yes]

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. [Yes]
   (b) Complete proofs of all theoretical results. [Yes]
   (c) Clear explanations of any assumptions. [Yes]

3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [Yes]
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Yes]
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Yes]
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [Yes]

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets. [Not Applicable]
   (b) The license information of the assets, if applicable. [Not Applicable]
   (c) New assets either in the supplemental material or as a URL, if applicable. [Yes]
   (d) Information about consent from data providers/curators. [Not Applicable]
   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Not Applicable]

5. If you used crowdsourcing or conducted research with human subjects, check if you include:
   (a) The full text of instructions given to participants and screenshots. [Not Applicable]
   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Not Applicable]
   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Not Applicable]

# Supplementary Materials

## A PROOF OF THE MAIN THEOREM

This supplementary material provides the proof of Theorem 4.5 and Theorem 4.7. We first present the known results in the field of convex analysis that are used in the proof of the main theorem. For a minibatch case, we discuss in Appendix B.

**Lemma A.1** (Nesterov (2018, Theorem 2.1.5))**.** *If $\ell\colon \mathbb{R}^d \to \mathbb{R}$ is convex and $\rho$-smooth, then for any $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^d$ we have*

$$\frac{1}{2\rho}\|\nabla\ell(\boldsymbol{y}) - \nabla\ell(\boldsymbol{x})\|^2 \leq \ell(\boldsymbol{y}) - \ell(\boldsymbol{x}) - \langle\nabla\ell(\boldsymbol{x}), \boldsymbol{y} - \boldsymbol{x}\rangle. \tag{13}$$

*Proof.* See the proof of Nesterov (2018, Theorem 2.1.5). □

**Corollary A.2.** *Let $\ell\colon \mathbb{R}^d \to \mathbb{R}$ be a function defined by $\ell(x) = \frac{1}{n}\sum_{i=1}^n \ell_i(x)$, where $\ell_i\colon \mathbb{R}^d \to \mathbb{R}$ for $i \in [n]$ is assumed to be convex and $\rho$-smooth. Suppose that $\operatorname{argmin}_{\boldsymbol{x}} \ell(\boldsymbol{x}) \neq \emptyset$. Then, for any $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^d$ we have*

$$\frac{1}{2\rho}\mathbb{E}\left[\|\nabla\ell_i(\boldsymbol{y}) - \nabla\ell_i(\boldsymbol{x})\|^2\right] \leq \ell(\boldsymbol{y}) - \ell(\boldsymbol{x}) - \langle\nabla\ell(\boldsymbol{x}), \boldsymbol{y} - \boldsymbol{x}\rangle \quad \textit{and} \quad \frac{1}{2\rho}\mathbb{E}\left[\|\nabla\ell_i(\boldsymbol{x}) - \nabla\ell_i(\boldsymbol{x}^*)\|^2\right] \leq \ell(\boldsymbol{x}) - \ell(\boldsymbol{x}^*),$$

*where $\boldsymbol{x}^* \in \operatorname{argmin}_{\boldsymbol{x}} \ell(\boldsymbol{x})$.*

*Proof.* For any $i \in [n]$, applying Lemma A.1 to $\ell_i$ and taking expectation on both side of Eq. (13), we obtain the first assertion. The second inequality is obtained by setting $\boldsymbol{y} = \boldsymbol{x}$ and $\boldsymbol{x} = \boldsymbol{x}^*$ in the first result. □

Hereinafter, we consider minimizing a finite sum which forms

$$\mathscr{L}(\boldsymbol{P}) \coloneqq \frac{1}{n}\sum_{i=1}^n \underbrace{\boldsymbol{t}_i^\top \boldsymbol{f}(\boldsymbol{b}(\boldsymbol{t}_i \,|\, \boldsymbol{P}))}_{=:\mathscr{L}_i(\boldsymbol{P})},$$

where the objective is to find the optimal control points $\boldsymbol{P}^* \in \arg\min_{\boldsymbol{P}} \mathscr{L}(\boldsymbol{P})$ of the Bézier simplex of degree $D$. In the following, we write $\mathscr{L}_i(\boldsymbol{P}) \coloneqq \boldsymbol{t}_i^\top \boldsymbol{f}(\boldsymbol{b}(\boldsymbol{t}_i \,|\, \boldsymbol{P}))$ for all $i \in [n]$. We first show that, for any $i \in [n]$, the function $\mathscr{L}_i$ is convex with respect to $\boldsymbol{P}$ under the assumption that all objective functions are convex.

**Lemma A.3.** *For all $i \in [n]$, the function $\mathscr{L}_i\colon \mathbb{R}^{|\mathbb{N}_D^M|} \to \mathbb{R}$ is convex with respect to $\boldsymbol{P}$, i.e.,*

$$\mathscr{L}_i(\lambda\boldsymbol{P}_1 + (1-\lambda)\boldsymbol{P}_2) \leq \lambda\mathscr{L}_i(\boldsymbol{P}_1) + (1-\lambda)\mathscr{L}_i(\boldsymbol{P}_2)$$

*holds for any $\boldsymbol{P}_1$, $\boldsymbol{P}_2$ and $\lambda \in [0,1]$.*

*Proof.* For all $\boldsymbol{P}_1$, $\boldsymbol{P}_2$ and $\lambda \in [0,1]$, we have

$$\begin{aligned}
\mathscr{L}_i(\lambda\boldsymbol{P}_1 + (1-\lambda)\boldsymbol{P}_2) &= \boldsymbol{t}_i^\top \boldsymbol{f}(\boldsymbol{b}(\boldsymbol{t}_i \,|\, \lambda\boldsymbol{P}_1 + (1-\lambda)\boldsymbol{P}_2)) \\
&= \boldsymbol{t}_i^\top \boldsymbol{f}(\lambda\boldsymbol{b}(\boldsymbol{t}_i \,|\, \boldsymbol{P}_1) + (1-\lambda)\boldsymbol{b}(\boldsymbol{t}_i \,|\, \boldsymbol{P}_2)) \\
&\leq \lambda\boldsymbol{t}_i^\top \boldsymbol{f}(\boldsymbol{b}(\boldsymbol{t}_i \,|\, \boldsymbol{P}_1)) + (1-\lambda)\boldsymbol{t}_i^\top \boldsymbol{f}(\boldsymbol{b}(\boldsymbol{t}_i \,|\, \boldsymbol{P}_2)) \\
&= \lambda\mathscr{L}_i(\boldsymbol{P}_1) + (1-\lambda)\mathscr{L}_i(\boldsymbol{P}_2),
\end{aligned}$$

where the second equality holds from the linearity of $\boldsymbol{b}$, and the first inequality holds from the convexity of $\boldsymbol{t}^\top \boldsymbol{f}$ since we assume that all objective functions $f_1, \ldots, f_M$ are convex. This completes the proof. □

**Corollary A.4.** *The function $\mathscr{L}$ is convex with respect to $\boldsymbol{P}$.*

*Proof.* Since $\mathscr{L}$ is a finite sum of $\mathscr{L}_i$ for $i = 1, \ldots, n$, and the sum of convex functions is also convex, we have the claim. □

In addition to the fact that the empirical loss function $\mathscr{L}$ is convex with respect to $\boldsymbol{P}$, we can show that the function $\mathscr{L}$ is strongly convex under the assumption that the matrix $\boldsymbol{Z} = (\boldsymbol{z}(\boldsymbol{t}_1), \ldots, \boldsymbol{z}(\boldsymbol{t}_n))$ is full row rank (Assumption 4.3).

**Lemma A.5** (Lemma 4.4)**.** *Under Assumption 4.3, let $\overline{\sigma} > 0$ be the smallest singular value of $\boldsymbol{Z}$. Then, the empirical loss function $\mathscr{L}$ defined by Eq.* (4) *is strongly convex with constant $\overline{\mu} \coloneqq \mu\overline{\sigma}^2/n$.*

*Proof.* For simplicity, we use $\boldsymbol{z}_i$ to denote $\boldsymbol{z}(\boldsymbol{t}_i)$. Since $\nabla_{\boldsymbol{P}} f_m(\boldsymbol{P}^\top \boldsymbol{z}) = \boldsymbol{z}\nabla f_m(\boldsymbol{P}^\top \boldsymbol{z})^\top$ for $m = 1, 2, \ldots, M$, we have the following inequalities for any $\boldsymbol{P}_1$ and $\boldsymbol{P}_2$:

$$
\begin{aligned}
\langle \nabla_{\boldsymbol{P}} \mathscr{L}(\boldsymbol{P}_1) - \nabla_{\boldsymbol{P}} \mathscr{L}(\boldsymbol{P}_1), \boldsymbol{P}_1 - \boldsymbol{P}_2 \rangle &= \frac{1}{n}\sum_{i=1}^{n}\sum_{m=1}^{M} \big\langle \boldsymbol{z}_i \big( t_{im} \big( \nabla f_m(\boldsymbol{P}_1^\top \boldsymbol{z}_i) - \nabla f_m(\boldsymbol{P}_2^\top \boldsymbol{z}_i) \big) \big), \boldsymbol{P}_1 - \boldsymbol{P}_2 \big\rangle \\
&= \frac{1}{n}\sum_{i=1}^{n}\sum_{m=1}^{M} t_{im} \big\langle \nabla f_m(\boldsymbol{P}_1^\top \boldsymbol{z}_i) - \nabla f_m(\boldsymbol{P}_2^\top \boldsymbol{z}_i), \boldsymbol{P}_1^\top \boldsymbol{z}_i - \boldsymbol{P}_2^\top \boldsymbol{z}_i \big\rangle \\
&\geq \frac{1}{n}\sum_{i=1}^{n}\sum_{m=1}^{M} t_{im}\mu \big\| \boldsymbol{P}_1^\top \boldsymbol{z}_i - \boldsymbol{P}_2^\top \boldsymbol{z}_i \big\|_2^2 \\
&= \frac{\mu}{n}\sum_{i=1}^{n} \big\| \boldsymbol{P}_1^\top \boldsymbol{z}_i - \boldsymbol{P}_2^\top \boldsymbol{z}_i \big\|_2^2 \\
&= \frac{\mu}{n} \Big\| (\boldsymbol{P}_1 - \boldsymbol{P}_2)^\top \boldsymbol{Z} \Big\|_{\mathsf{F}}^2 \\
&\geq \frac{\mu\overline{\sigma}^2}{n} \|\boldsymbol{P}_1 - \boldsymbol{P}_2\|_{\mathsf{F}}^2 =: \overline{\mu} \|\boldsymbol{P}_1 - \boldsymbol{P}_2\|_{\mathsf{F}}^2,
\end{aligned}
$$

where the third equality holds from the fact that $f_m$ is $\mu$-strongly convex and $\sum_{m=1}^{M} t_{im} = 1$ holds for $i = 1, 2, \ldots, n$. Thus, the function $\mathscr{L}$ is strongly convex with constant $\overline{\mu} \coloneqq \mu\overline{\sigma}^2/n$. □

From Lemma 4.4, the optimal solution to the problem $\min_{\boldsymbol{P}} \mathscr{L}(\boldsymbol{P})$ is unique, which we denote by $\boldsymbol{P}^*$. Next, we show the smoothness of $\mathscr{L}_i$ for $i = 1, \ldots, n$, and $\mathscr{L}$.

**Lemma A.6.** *Suppose that Assumption 4.1 holds. Let $\widetilde{\rho} \coloneqq \rho \max_{\boldsymbol{t} \in \Delta^{M-1}} \|\boldsymbol{z}(\boldsymbol{t})\|$. Then, for any $i \in [n]$, $\mathscr{L}_i$ is $\widetilde{\rho}$-smooth, that is, $\|\nabla \mathscr{L}_i(\boldsymbol{P}_1) - \nabla \mathscr{L}_i(\boldsymbol{P}_2)\| \leq \widetilde{\rho} \|\boldsymbol{P}_1 - \boldsymbol{P}_2\|$ holds for any $\boldsymbol{P}_1$ and $\boldsymbol{P}_2$. In addition, $\mathscr{L}$ is $\widetilde{\rho}$-smooth, that is, $\|\nabla \mathscr{L}(\boldsymbol{P}_1) - \nabla \mathscr{L}(\boldsymbol{P}_2)\| \leq \widetilde{\rho} \|\boldsymbol{P}_1 - \boldsymbol{P}_2\|$ holds for any $\boldsymbol{P}_1$ and $\boldsymbol{P}_2$.*

*Proof.* Firstly, we show that $\mathscr{L}_i$ is $\widetilde{\rho}$-smooth for all $i \in [n]$. For any $\boldsymbol{P}_1$ and $\boldsymbol{P}_2$, we have

$$
\begin{aligned}
\|\nabla \mathscr{L}_i(\boldsymbol{P}_1) - \nabla \mathscr{L}_i(\boldsymbol{P}_2)\| &= \left\| \sum_{m=1}^{M} t_{im} \nabla f_m(\boldsymbol{b}(\boldsymbol{t}_i \mid \boldsymbol{P}_1)) - \sum_{m=1}^{M} t_{im} \nabla f_m(\boldsymbol{b}(\boldsymbol{t}_i \mid \boldsymbol{P}_2)) \right\| \\
&\leq \sum_{m=1}^{M} t_{im} \|\nabla f_m(\boldsymbol{b}(\boldsymbol{t}_i \mid \boldsymbol{P}_1)) - \nabla f_j(\boldsymbol{b}(\boldsymbol{t}_i \mid \boldsymbol{P}_2))\| \\
&\leq \sum_{m=1}^{M} t_{im} \rho \|\boldsymbol{b}(\boldsymbol{t}_i \mid \boldsymbol{P}_1) - \boldsymbol{b}(\boldsymbol{t}_i \mid \boldsymbol{P}_2)\| \\
&= \sum_{j=1}^{M} t_{im} \rho \Big\| (\boldsymbol{P}_1 - \boldsymbol{P}_2)^\top \boldsymbol{z}(\boldsymbol{t}_i) \Big\| \\
&\leq \widetilde{\rho} \|\boldsymbol{P}_1 - \boldsymbol{P}_2\|.
\end{aligned}
$$

In the first inequality, we use the triangle inequality, and in the second inequality, we use the smoothness of $f_m$ for all $m \in [M]$. In the last inequality, we used the fact that $\sum_{m=1}^{M} t_{im} = 1$ for all $i \in [n]$ and the assumption of $\widetilde{\rho} \coloneqq \rho \max_{\boldsymbol{t} \in \Delta^{M-1}} \|\boldsymbol{z}(\boldsymbol{t})\|$.

Secondly, we show that $\mathscr{L}$ is $\widetilde{\rho}$-smooth. For any $\boldsymbol{P}_1$ and $\boldsymbol{P}_2$, we have

$$\begin{aligned}
\|\nabla \mathscr{L}(\boldsymbol{P}_1) - \nabla \mathscr{L}(\boldsymbol{P}_2)\| &= \left\| \frac{1}{n} \left( \sum_{i=1}^{n} \nabla \mathscr{L}_i(\boldsymbol{P}_1) - \nabla \mathscr{L}_i(\boldsymbol{P}_2) \right) \right\| \\
&\leq \frac{1}{n} \sum_{i=1}^{n} \|\nabla \mathscr{L}_i(\boldsymbol{P}_1) - \nabla \mathscr{L}_i(\boldsymbol{P}_2)\| \\
&\leq \frac{1}{n} \sum_{i=1}^{n} \widetilde{\rho} \|\boldsymbol{P}_1 - \boldsymbol{P}_2\| \\
&= \widetilde{\rho} \|\boldsymbol{P}_1 - \boldsymbol{P}_2\|,
\end{aligned}$$

where the first inequality holds by triangle inequality and the second inequality holds from the smoothness of $\mathscr{L}_i$ for all $i \in [n]$ by the first claim. □

Then, we have the same result as Corollary A.2 for $\mathscr{L}_i$ for any $i \in [N]$.

**Lemma A.7.** *Suppose that Assumptions 4.1 and 4.3 hold. Let $\boldsymbol{P}^* \in \operatorname{argmin}_{\boldsymbol{P}} \mathscr{L}(\boldsymbol{P})$. Then, for any $\boldsymbol{P}_1$ and $\boldsymbol{P}_2$, we have*

$$\begin{aligned}
\frac{1}{2\rho} \mathbb{E}\left[ \|\nabla \mathscr{L}_i(\boldsymbol{P}_1) - \nabla \mathscr{L}_i(\boldsymbol{P}_2)\|^2 \right] &\leq \mathscr{L}(\boldsymbol{P}_1) - \mathscr{L}(\boldsymbol{P}_2) - \langle \nabla \mathscr{L}(\boldsymbol{P}_2), \boldsymbol{P}_1 - \boldsymbol{P}_2 \rangle, \\
\frac{1}{2\rho} \mathbb{E}\left[ \|\nabla \mathscr{L}_i(\boldsymbol{P}) - \nabla \mathscr{L}_i(\boldsymbol{P}^*)\|^2 \right] &\leq \mathscr{L}(\boldsymbol{P}) - \mathscr{L}(\boldsymbol{P}^*),
\end{aligned}$$

*where $\widetilde{\rho} \coloneqq \rho \max_{\boldsymbol{t} \in \Delta^{M-1}} \|\boldsymbol{z}(\boldsymbol{t})\|$.*

*Proof.* This is a direct consequence of Corollary A.2. □

**Lemma A.8.** *Suppose that Assumptions 4.1, 4.2 and 4.3 hold. Then, for any $\boldsymbol{P}$, we have*

$$\mathbb{E}\left[ \|\nabla \mathscr{L}_i(\boldsymbol{P})\|^2 \right] \leq 4\rho(\mathscr{L}(\boldsymbol{P}) - \mathscr{L}(\boldsymbol{P}^*)) + 2\sigma_{\mathscr{L}}.$$

*Proof.* For any $\boldsymbol{P}$, we have

$$\|\nabla \mathscr{L}_i(\boldsymbol{P})\|^2 \leq 2\|\nabla \mathscr{L}_i(\boldsymbol{P}) - \nabla \mathscr{L}_i(\boldsymbol{P}^*)\|^2 + 2\|\nabla \mathscr{L}_i(\boldsymbol{P}^*)\|^2.$$

Taking expectation over the above inequality, we have

$$\begin{aligned}
\mathbb{E}\left[ \|\nabla \mathscr{L}_i(\boldsymbol{P})\|^2 \right] &\leq 2\mathbb{E}\left[ \|\nabla \mathscr{L}_i(\boldsymbol{P}) - \nabla \mathscr{L}_i(\boldsymbol{P}^*)\|^2 \right] + 2\mathbb{E}\left[ \|\nabla \mathscr{L}_i(\boldsymbol{P}^*)\|^2 \right] \\
&\leq 2\mathbb{E}\left[ 2\widetilde{\rho}(\mathscr{L}_i(\boldsymbol{P}) - \mathscr{L}_i(\boldsymbol{P}^*)) \right] + 2\mathbb{E}\left[ \|\nabla \mathscr{L}_i(\boldsymbol{P}^*)\|^2 \right] \\
&= 4\widetilde{\rho}(\mathscr{L}(\boldsymbol{P}) - \mathscr{L}(\boldsymbol{P}^*)) + 2\sigma_{\mathscr{L}},
\end{aligned}$$

where we used Lemma A.7 in the second inequality. This completes the proof. □

Now, we are ready to prove the main claims. We first show the claims for a single batch case, and then show for the minibatch case in Section B.

**Theorem A.9.** *Suppose that Assumptions 4.1, 4.2 and 4.3 hold. Let $\widetilde{\rho}$ be a positive constant defined as $\widetilde{\rho} \coloneqq \rho \max_{\boldsymbol{t} \in \Delta^{M-1}} \|\boldsymbol{z}(\boldsymbol{t})\|$. Consider a sequence of control points $\{\boldsymbol{P}^{(k)}\}_{k \in \mathbb{N}}$ generated by Algorithm 1 with stepsizes satisfying $\alpha_k \leq \min\left\{ \left(4\widetilde{\rho}\lambda_{\max}(\boldsymbol{\Sigma}^{-1})\right)^{-1}, \left(\overline{\mu}\lambda_{\min}(\boldsymbol{\Sigma}^{-1})\right)^{-1} \right\}$ for $k = 0, 1, \ldots, K-1$ and an initial control point $\boldsymbol{P}_0$. Then, for any $K > 0$, we have*

$$\mathbb{E}\left[ \mathscr{L}(\widetilde{\boldsymbol{P}}^{(K)}) - \mathscr{L}(\boldsymbol{P}^*) \right] \leq \frac{(1 - \alpha_0 \overline{\mu} \lambda_{\min}(\boldsymbol{\Sigma}^{-1}))}{\sum_{k=0}^{K-1} \alpha_k} \|\boldsymbol{P}_0 - \boldsymbol{P}^*\|_{\boldsymbol{\Sigma}}^2 + \frac{2\lambda_{\max}(\boldsymbol{\Sigma}^{-1}) \sum_{k=0}^{K-1} \alpha_k^2}{\sum_{k=0}^{K-1} \alpha_k} \sigma_{\mathscr{L}},$$

*where* $\widetilde{\boldsymbol{P}}$ *is defined as follows:*

$$\widetilde{\boldsymbol{P}}^{(K)} \coloneqq \sum_{k=0}^{K-1} q_{K,k} \boldsymbol{P}^{(k)} \quad \textit{with} \quad q_{K,k} \coloneqq \frac{\alpha_k(1-2\alpha_k\widetilde{\rho})}{2\sum_{\kappa=0}^{K-1}\tau_\kappa(1-2\alpha_\kappa\widetilde{\rho})}.$$

*Proof.* For any $k > 0$, from the update rule in Eq. (6) with single batch, we have

$$\begin{aligned}\left\|\boldsymbol{P}^{(k+1)} - \boldsymbol{P}^*\right\|_{\boldsymbol{\Sigma}}^2 &= \left\|\boldsymbol{P}^{(k)} - \alpha_k\boldsymbol{\Sigma}^{-1}\nabla\mathscr{L}_i(\boldsymbol{P}^{(k)}) - \boldsymbol{P}^*\right\|_{\boldsymbol{\Sigma}}^2 \\ &= \left\|\boldsymbol{P}^{(k)} - \boldsymbol{P}^*\right\|_{\boldsymbol{\Sigma}}^2 - 2\alpha_k\left\langle \boldsymbol{P}^{(k)} - \boldsymbol{P}^*, \boldsymbol{\Sigma}^{-1}\nabla\mathscr{L}_i(\boldsymbol{P}^{(k)})\right\rangle_{\boldsymbol{\Sigma}} + \alpha_k^2\left\|\boldsymbol{\Sigma}^{-1}\nabla\mathscr{L}_i(\boldsymbol{P}^{(k)})\right\|_{\boldsymbol{\Sigma}}^2.\end{aligned}$$

By taking expectation conditioned on $\boldsymbol{P}^{(k)}$, we have

$$\begin{aligned}\mathbb{E}_k\left[\left\|\boldsymbol{P}^{(k+1)} - \boldsymbol{P}^*\right\|_{\boldsymbol{\Sigma}}^2\right] &= \left\|\boldsymbol{P}^{(k)} - \boldsymbol{P}^*\right\|_{\boldsymbol{\Sigma}}^2 - 2\alpha_k\mathbb{E}_k\left[\left\langle \boldsymbol{P}^{(k)} - \boldsymbol{P}^*, \boldsymbol{\Sigma}^{-1}\nabla\mathscr{L}_i(\boldsymbol{P}^{(k)})\right\rangle_{\boldsymbol{\Sigma}}\right] + \alpha_k^2\mathbb{E}_k\left[\left\|\boldsymbol{\Sigma}^{-1}\nabla\mathscr{L}_i(\boldsymbol{P}^{(k)})\right\|_{\boldsymbol{\Sigma}}^2\right] \\ &\leq \left\|\boldsymbol{P}^{(k)} - \boldsymbol{P}^*\right\|_{\boldsymbol{\Sigma}}^2 - 2\alpha_k\left\langle \boldsymbol{P}^{(k)} - \boldsymbol{P}^*, \nabla\mathscr{L}(\boldsymbol{P}^{(k)})\right\rangle + \alpha_k^2\mathbb{E}_k\left[\left\|\boldsymbol{\Sigma}^{-1}\nabla\mathscr{L}_i(\boldsymbol{P}^{(k)})\right\|_{\boldsymbol{\Sigma}}^2\right].\end{aligned}$$

Here, by the strongly convexity of $\mathscr{L}$ and Lemma A.8, we have

$$\begin{aligned}-2\alpha_k\left\langle \boldsymbol{P}^{(k)} - \boldsymbol{P}^*, \nabla\mathscr{L}(\boldsymbol{P}^{(k)})\right\rangle &\leq 2\alpha_k\left(\mathscr{L}(\boldsymbol{P}^*) - \mathscr{L}(\boldsymbol{P}^{(k)}) - \frac{\overline{\mu}}{2}\left\|\boldsymbol{P}^* - \boldsymbol{P}^{(k)}\right\|^2\right) \\ &\leq 2\alpha_k\left(\mathscr{L}(\boldsymbol{P}^*) - \mathscr{L}(\boldsymbol{P}^{(k)})\right) - \alpha_k\overline{\mu}\lambda_{\min}(\boldsymbol{\Sigma}^{-1})\left\|\boldsymbol{P}^* - \boldsymbol{P}^{(k)}\right\|_{\boldsymbol{\Sigma}}^2,\end{aligned}$$

and

$$\begin{aligned}\alpha_k^2\mathbb{E}_k\left[\left\|\boldsymbol{\Sigma}^{-1}\nabla\mathscr{L}_i(\boldsymbol{P}^{(k)})\right\|_{\boldsymbol{\Sigma}}^2\right] &\leq \alpha_k^2\lambda_{\max}(\boldsymbol{\Sigma}^{-1})\mathbb{E}_k\left[\left\|\nabla\mathscr{L}_i(\boldsymbol{P}^{(k)})\right\|^2\right] \\ &\leq \alpha_k^2\lambda_{\max}(\boldsymbol{\Sigma}^{-1})\left(4\widetilde{\rho}\left(\mathscr{L}(\boldsymbol{P}^{(k)}) - \mathscr{L}(\boldsymbol{P}^*)\right) + 2\sigma_{\mathscr{L}}\right).\end{aligned}$$

Hence, we have

$$\begin{aligned}&(1-\alpha_{k+1}\overline{\mu}\lambda_{\min}(\boldsymbol{\Sigma}^{-1}))\mathbb{E}_k\left[\left\|\boldsymbol{P}^{(k+1)} - \boldsymbol{P}^*\right\|_{\boldsymbol{\Sigma}}^2\right] \\ &\leq \mathbb{E}_k\left[\left\|\boldsymbol{P}^{(k+1)} - \boldsymbol{P}^*\right\|_{\boldsymbol{\Sigma}}^2\right] \\ &\leq \left(1-\alpha_k\overline{\mu}\lambda_{\min}(\boldsymbol{\Sigma}^{-1})\right)\left\|\boldsymbol{P}^{(k)} - \boldsymbol{P}^*\right\|_{\boldsymbol{\Sigma}}^2 + 2\alpha_k\left(2\alpha_k\widetilde{\rho}\lambda_{\max}(\boldsymbol{\Sigma}^{-1}) - 1\right)\left(\mathscr{L}(\boldsymbol{P}^{(k)}) - \mathscr{L}(\boldsymbol{P}^*)\right) + 2\alpha_k^2\lambda_{\max}(\boldsymbol{\Sigma}^{-1})\sigma_{\mathscr{L}} \\ &\leq \left(1-\alpha_k\overline{\mu}\lambda_{\min}(\boldsymbol{\Sigma}^{-1})\right)\left\|\boldsymbol{P}^{(k)} - \boldsymbol{P}^*\right\|_{\boldsymbol{\Sigma}}^2 - \alpha_k\left(\mathscr{L}(\boldsymbol{P}^{(k)}) - \mathscr{L}(\boldsymbol{P}^*)\right) + 2\alpha_k^2\lambda_{\max}(\boldsymbol{\Sigma}^{-1})\sigma_{\mathscr{L}},\end{aligned}$$

where the last inequality holds by the assumption on stepsize $\alpha_k$ so that $4\alpha_k\widetilde{\rho}\lambda_{\max}(\boldsymbol{\Sigma}^{-1}) < 1$ holds. Rearranging the terms and taking expectations on both sides, we have

$$\begin{aligned}&\alpha_k\mathbb{E}\left[\mathscr{L}(\boldsymbol{P}^{(k)}) - \mathscr{L}(\boldsymbol{P}^*)\right] \\ &\leq \left(1-\alpha_k\overline{\mu}\lambda_{\min}(\boldsymbol{\Sigma}^{-1})\right)\mathbb{E}\left[\left\|\boldsymbol{P}^{(k)} - \boldsymbol{P}^*\right\|_{\boldsymbol{\Sigma}}^2\right] - \left(1-\alpha_{k+1}\overline{\mu}\lambda_{\min}(\boldsymbol{\Sigma}^{-1})\right)\mathbb{E}\left[\left\|\boldsymbol{P}^{(k+1)} - \boldsymbol{P}^*\right\|_{\boldsymbol{\Sigma}}^2\right] + 2\alpha_k^2\lambda_{\max}(\boldsymbol{\Sigma}^{-1})\sigma_{\mathscr{L}}.\end{aligned}$$

By summing over $k = 0, 1, \ldots, K-1$, we have

$$\begin{aligned}&\sum_{k=0}^{K-1}\alpha_k\mathbb{E}\left[\mathscr{L}(\boldsymbol{P}^{(k)}) - \mathscr{L}(\boldsymbol{P}^*)\right] \\ &\leq \left(1-\alpha_0\overline{\mu}\lambda_{\min}(\boldsymbol{\Sigma}^{-1})\right)\|\boldsymbol{P}_0 - \boldsymbol{P}^*\|_{\boldsymbol{\Sigma}}^2 - \left(1-\alpha_K\overline{\mu}\lambda_{\min}(\boldsymbol{\Sigma}^{-1})\right)\mathbb{E}\left[\left\|\boldsymbol{P}^{(K)} - \boldsymbol{P}^*\right\|_{\boldsymbol{\Sigma}}^2\right] + 2\lambda_{\max}(\boldsymbol{\Sigma}^{-1})\sigma_{\mathscr{L}}\sum_{k=0}^{K-1}\alpha_k^2\end{aligned}$$

$$\leq \big(1 - \alpha_0 \overline{\mu} \lambda_{\min}(\boldsymbol{\Sigma}^{-1})\big) \|\boldsymbol{P}_0 - \boldsymbol{P}^*\|_{\boldsymbol{\Sigma}}^2 + 2\lambda_{\max}(\boldsymbol{\Sigma}^{-1}) \sigma_{\mathscr{L}} \sum_{k=0}^{K-1} \alpha_k^2,$$

where the last inequality holds since $\mathbb{E}\left[\left\|\boldsymbol{P}^{(k)} - \boldsymbol{P}^*\right\|_{\boldsymbol{\Sigma}}^2\right] \geq 0$ holds and $\alpha_K \overline{\mu} \lambda_{\min}(\boldsymbol{\Sigma}^{-1}) < 1$ holds by the assumption. Dividing both sides by $\sum_{\kappa=0}^{K-1} \alpha_\kappa > 0$ gives

$$\sum_{k=0}^{K-1} \mathbb{E}\left[\frac{\alpha_k}{\sum_{\kappa=0}^{K-1} \alpha_\kappa} \Big(\mathscr{L}(\boldsymbol{P}^{(k)}) - \mathscr{L}(\boldsymbol{P}^*)\Big)\right] \leq \frac{(1 - \alpha_0 \overline{\mu} \lambda_{\min}(\boldsymbol{\Sigma}^{-1})) \|\boldsymbol{P}_0 - \boldsymbol{P}^*\|_{\boldsymbol{\Sigma}}^2}{\sum_{\kappa=0}^{K-1} \alpha_\kappa} + \frac{2\lambda_{\max}(\boldsymbol{\Sigma}^{-1}) \sigma_{\mathscr{L}} \sum_{k=0}^{K-1} \alpha_k^2}{\sum_{\kappa=0}^{K-1} \alpha_\kappa}.$$

Define $\widetilde{\boldsymbol{P}}^{(K)}$ as the weighted average of the generated control points $\boldsymbol{P}^{(k)}$ for $k = 0, 1, \ldots, K-1$, i.e.,

$$\widetilde{\boldsymbol{P}}^{(K)} \coloneqq \sum_{k=0}^{K-1} q_{K,k} \boldsymbol{P}^{(k)} \quad \text{with} \quad q_{K,k} \coloneqq \frac{\alpha_k}{\sum_{\kappa=0}^{K-1} \alpha_\kappa}.$$

Then, we have $q_{K,k} \geq 0$ for $k = 0, 1, \ldots, K-1$ and $\sum_{k=0}^{K-1} q_{K,k} = 1$. Thus, by Jensen's inequality, we have

$$\begin{aligned}
\mathbb{E}\left[\mathscr{L}(\widetilde{\boldsymbol{P}}^{(K)}) - \mathscr{L}(\boldsymbol{P}^*)\right] &\leq \sum_{k=0}^{K-1} \mathbb{E}\left[\frac{\alpha_k}{\sum_{\kappa=0}^{K-1} \alpha_\kappa} \Big(\mathscr{L}(\boldsymbol{P}^{(k)}) - \mathscr{L}(\boldsymbol{P}^*)\Big)\right] \\
&\leq \frac{(1 - \alpha_0 \overline{\mu} \lambda_{\min}(\boldsymbol{\Sigma}^{-1}))}{\sum_{\kappa=0}^{K-1} \alpha_\kappa} \|\boldsymbol{P}_0 - \boldsymbol{P}^*\|_{\boldsymbol{\Sigma}}^2 + \frac{2\lambda_{\max}(\boldsymbol{\Sigma}^{-1}) \sum_{k=0}^{K-1} \alpha_k^2}{\sum_{\kappa=0}^{K-1} \alpha_\kappa} \sigma_{\mathscr{L}}.
\end{aligned}$$

This completes the proof. □

**Corollary A.10.** *Consider the same situation as in Theorem A.9 and suppose that we choose a constant stepsize* $\alpha_k = \alpha \leq \min\Big\{\big(4\widetilde{\rho}\lambda_{\max}(\boldsymbol{\Sigma}^{-1})\big)^{-1}, \big(\overline{\mu}\lambda_{\min}(\boldsymbol{\Sigma}^{-1})\big)^{-1}\Big\}$ *for* $k = 0, 1, \ldots, K-1$. *Then, for any* $K > 0$, *we have*

$$\mathbb{E}\left[\mathscr{L}(\overline{\boldsymbol{P}}^{(K)}) - \mathscr{L}(\boldsymbol{P}^*)\right] \leq \frac{(1 - \alpha\overline{\mu}\lambda_{\min}(\boldsymbol{\Sigma}^{-1}))}{\alpha K} \|\boldsymbol{P}_0 - \boldsymbol{P}^*\|_{\boldsymbol{\Sigma}}^2 + 2\alpha\lambda_{\max}(\boldsymbol{\Sigma}^{-1})\sigma_{\mathscr{L}},$$

*where* $\overline{\boldsymbol{P}}^{(K)} \coloneqq \frac{1}{K}\sum_{t=1}^{K} \boldsymbol{P}^{(k)}$. *Additionally, if we choose a stepsize as* $\alpha_k = \frac{\alpha_0}{\sqrt{k+1}}$ *for* $\alpha_0 \leq \min\{4(\widetilde{\rho}\lambda_{\max}(\boldsymbol{\Sigma}^{-1}))^{-1}, (\overline{\mu}\lambda_{\min}(\boldsymbol{\Sigma}^{-1}))^{-1}\}$, *then we have*

$$\mathbb{E}\left[\mathscr{L}(\overline{\boldsymbol{P}}^{(K)}) - \mathscr{L}(\boldsymbol{P}^*)\right] \leq \frac{5(1 - \alpha_0\overline{\mu}\lambda_{\min}(\boldsymbol{\Sigma}^{-1}))}{4\alpha_0\sqrt{K}} \|\boldsymbol{P}_0 - \boldsymbol{P}^*\|_{\boldsymbol{\Sigma}}^2 + \frac{5\alpha_0 \ln(K+1)}{\sqrt{K}} \sigma_{\mathscr{L}} = \mathrm{O}\left(\frac{\ln K}{\sqrt{K}}\right).$$

*Proof.* It is clear that the first claim holds by setting $\alpha_k = \alpha$ for $k = 0, 1, \ldots, K-1$ in Theorem A.9. For the second claim, following the proof of Garrigos and Gower (2023, Theorem 5.7), by using the two inequalities $\sum_{k=1}^{K} \frac{1}{\sqrt{k}} \geq \frac{4}{5}\sqrt{K}$ and $\sum_{k=1}^{K} \frac{1}{k} \leq 2\ln(K+1)$, we obtain the desired result. □

We have the following results as the convergence result for the control points of the Bézier simplex generated by the proposed algorithm.

**Theorem A.11.** *Suppose that Assumptions 4.1, 4.2 and 4.3 hold. Let* $\widetilde{\rho}$ *be a positive constant defined as* $\widetilde{\rho} \coloneqq \rho \max_{\boldsymbol{t} \in \Delta^{M-1}} \|\boldsymbol{z}(\boldsymbol{t})\|$. *Consider a sequence of control points* $\{\boldsymbol{P}^{(k)}\}_{k \in \mathbb{N}}$ *generated by Algorithm 1 with a constant stepsize satisfying* $\alpha_k = \alpha \leq \big(2\widetilde{\rho}\lambda_{\max}(\boldsymbol{\Sigma}^{-1})\big)^{-1}$ *and an initial control point* $\boldsymbol{P}_0$. *Then, for any* $k > 0$, *we have*

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

*where* $\boldsymbol{P}^* \in \operatorname{argmin}_{\boldsymbol{P}} \mathscr{L}(\boldsymbol{P})$ *and* $\operatorname{cond}(\boldsymbol{\Sigma}^{-1})$ *is the condition number of* $\boldsymbol{\Sigma}^{-1}$.

*Proof.* For any $k > 0$, from the update rule in Eq. (6) with single batch, we have

$$\begin{aligned}\left\|\boldsymbol{P}^{(k+1)} - \boldsymbol{P}^*\right\|_{\boldsymbol{\Sigma}}^2 &= \left\|\boldsymbol{P}^{(k)} - \alpha\boldsymbol{\Sigma}^{-1}\nabla\mathscr{L}_i(\boldsymbol{P}^{(k)}) - \boldsymbol{P}^*\right\|_{\boldsymbol{\Sigma}}^2 \\ &= \left\|\boldsymbol{P}^{(k)} - \boldsymbol{P}^*\right\|_{\boldsymbol{\Sigma}}^2 - 2\alpha\left\langle \boldsymbol{P}^{(k)} - \boldsymbol{P}^*, \boldsymbol{\Sigma}^{-1}\nabla\mathscr{L}_i(\boldsymbol{P}^{(k)})\right\rangle_{\boldsymbol{\Sigma}} + \alpha^2\left\|\boldsymbol{\Sigma}^{-1}\nabla\mathscr{L}_i(\boldsymbol{P}^{(k)})\right\|_{\boldsymbol{\Sigma}}.\end{aligned}$$

Taking expectation conditioned on $\boldsymbol{P}^{(k)}$ on both sides, we have

$$\begin{aligned}\mathbb{E}_k\left[\left\|\boldsymbol{P}^{(k+1)} - \boldsymbol{P}^*\right\|_{\boldsymbol{\Sigma}}^2\right] &= \left\|\boldsymbol{P}^{(k)} - \boldsymbol{P}^*\right\|_{\boldsymbol{\Sigma}}^2 - 2\alpha\mathbb{E}_t\left[\left\langle \boldsymbol{P}^{(k)} - \boldsymbol{P}^*, \boldsymbol{\Sigma}^{-1}\nabla\mathscr{L}_i(\boldsymbol{P}^{(k)})\right\rangle_{\boldsymbol{\Sigma}}\right] + \alpha^2\mathbb{E}_t\left[\left\|\boldsymbol{\Sigma}^{-1}\nabla\mathscr{L}_i(\boldsymbol{P}^{(k)})\right\|_{\boldsymbol{\Sigma}}\right] \\ &= \left\|\boldsymbol{P}^{(k)} - \boldsymbol{P}^*\right\|_{\boldsymbol{\Sigma}}^2 - 2\alpha\left\langle \boldsymbol{P}^{(k)} - \boldsymbol{P}^*, \boldsymbol{\Sigma}^{-1}\nabla\mathscr{L}(\boldsymbol{P}^{(k)})\right\rangle_{\boldsymbol{\Sigma}} + \alpha^2\mathbb{E}_t\left[\left\|\boldsymbol{\Sigma}^{-1}\nabla\mathscr{L}_i(\boldsymbol{P}^{(k)})\right\|_{\boldsymbol{\Sigma}}\right] \\ &= \left\|\boldsymbol{P}^{(k)} - \boldsymbol{P}^*\right\|_{\boldsymbol{\Sigma}}^2 - 2\alpha\left\langle \boldsymbol{P}^{(k)} - \boldsymbol{P}^*, \nabla\mathscr{L}(\boldsymbol{P}^{(k)})\right\rangle + \alpha^2\mathbb{E}_t\left[\left\|\boldsymbol{\Sigma}^{-1}\nabla\mathscr{L}_i(\boldsymbol{P}^{(k)})\right\|_{\boldsymbol{\Sigma}}\right],\end{aligned}$$

where the second equality holds by the fact that $\mathbb{E}_t\left[\nabla\mathscr{L}_i(\boldsymbol{P}^{(k)})\right] = \nabla\mathscr{L}(\boldsymbol{P}^{(k)})$. By the strong convexity, the second term of the right-hand side is bounded above as follows:

$$\begin{aligned}-2\alpha\left\langle \boldsymbol{P}^{(k)} - \boldsymbol{P}^*, \nabla\mathscr{L}(\boldsymbol{P}^{(k)})\right\rangle &\leq 2\alpha\left(\mathscr{L}(\boldsymbol{P}^*) - \mathscr{L}(\boldsymbol{P}^{(k)}) - \frac{\overline{\mu}}{2}\left\|\boldsymbol{P}^* - \boldsymbol{P}^{(k)}\right\|^2\right) \\ &\leq 2\alpha\left(\mathscr{L}(\boldsymbol{P}^*) - \mathscr{L}(\boldsymbol{P}^{(k)})\right) - \alpha\overline{\mu}\lambda_{\min}(\boldsymbol{\Sigma}^{-1})\left\|\boldsymbol{P}^* - \boldsymbol{P}^{(k)}\right\|_{\boldsymbol{\Sigma}}^2.\end{aligned}$$

Thus, we have

$$\begin{aligned}&\mathbb{E}_k\left[\left\|\boldsymbol{P}^{(k+1)} - \boldsymbol{P}^*\right\|_{\boldsymbol{\Sigma}}^2\right] \\ &= \left(1 - \alpha\overline{\mu}\lambda_{\min}(\boldsymbol{\Sigma}^{-1})\right)\left\|\boldsymbol{P}^{(k)} - \boldsymbol{P}^*\right\|_{\boldsymbol{\Sigma}}^2 - 2\alpha\left(\mathscr{L}(\boldsymbol{P}^{(k)}) - \mathscr{L}(\boldsymbol{P}^*)\right) + \alpha^2\mathbb{E}_k\left[\left\|\boldsymbol{\Sigma}^{-1}\nabla\mathscr{L}_i(\boldsymbol{P}^{(k)})\right\|_{\boldsymbol{\Sigma}}^2\right].\end{aligned}$$

By taking expectations on both sides, we have

$$\begin{aligned}&\mathbb{E}\left[\left\|\boldsymbol{P}^{(k+1)} - \boldsymbol{P}^*\right\|_{\boldsymbol{\Sigma}}^2\right] \\ &\leq \left(1 - \alpha\overline{\mu}\lambda_{\min}(\boldsymbol{\Sigma}^{-1})\right)\mathbb{E}\left[\left\|\boldsymbol{P}^{(k)} - \boldsymbol{P}^*\right\|_{\boldsymbol{\Sigma}}^2\right] - 2\alpha\mathbb{E}\left[\mathscr{L}(\boldsymbol{P}^{(k)}) - \mathscr{L}(\boldsymbol{P}^*)\right] + \alpha^2\mathbb{E}\left[\left\|\boldsymbol{\Sigma}^{-1}\nabla\mathscr{L}_i(\boldsymbol{P}^{(k)})\right\|_{\boldsymbol{\Sigma}}^2\right] \\ &\leq \left(1 - \alpha\overline{\mu}\lambda_{\min}(\boldsymbol{\Sigma}^{-1})\right)\mathbb{E}\left[\left\|\boldsymbol{P}^{(k)} - \boldsymbol{P}^*\right\|_{\boldsymbol{\Sigma}}^2\right] - 2\alpha\mathbb{E}\left[\mathscr{L}(\boldsymbol{P}^{(k)}) - \mathscr{L}(\boldsymbol{P}^*)\right] + \alpha^2\lambda_{\max}(\boldsymbol{\Sigma}^{-1})\mathbb{E}\left[\left\|\nabla\mathscr{L}_i(\boldsymbol{P}^{(k)})\right\|^2\right] \\ &\leq \left(1 - \alpha\overline{\mu}\lambda_{\min}(\boldsymbol{\Sigma}^{-1})\right)\mathbb{E}\left[\left\|\boldsymbol{P}^{(k)} - \boldsymbol{P}^*\right\|_{\boldsymbol{\Sigma}}^2\right] - 2\alpha\mathbb{E}\left[\mathscr{L}(\boldsymbol{P}^{(k)}) - \mathscr{L}(\boldsymbol{P}^*)\right] + \alpha^2\lambda_{\max}(\boldsymbol{\Sigma}^{-1})\left(4\widetilde{\rho}\left(\mathscr{L}(\boldsymbol{P}^{(k)}) - \mathscr{L}(\boldsymbol{P}^*)\right) + 2\sigma_{\mathscr{L}}\right) \\ &= \left(1 - \alpha\overline{\mu}\lambda_{\min}(\boldsymbol{\Sigma}^{-1})\right)\mathbb{E}\left[\left\|\boldsymbol{P}^{(k)} - \boldsymbol{P}^*\right\|_{\boldsymbol{\Sigma}}^2\right] + 2\alpha^2\lambda_{\max}(\boldsymbol{\Sigma}^{-1})\sigma_{\mathscr{L}} + 2\alpha\left(2\alpha\widetilde{\rho}\lambda_{\max}(\boldsymbol{\Sigma}^{-1}) - 1\right)\mathbb{E}\left[\mathscr{L}(\boldsymbol{P}^{(k)}) - \mathscr{L}(\boldsymbol{P}^*)\right] \\ &\leq \left(1 - \alpha\overline{\mu}\lambda_{\min}(\boldsymbol{\Sigma}^{-1})\right)\mathbb{E}\left[\left\|\boldsymbol{P}^{(k)} - \boldsymbol{P}^*\right\|_{\boldsymbol{\Sigma}}^2\right] + 2\alpha^2\lambda_{\max}(\boldsymbol{\Sigma}^{-1})\sigma_{\mathscr{L}}.\end{aligned}$$

In the above inequalities, the third inequality holds from Lemma A.8, and the last inequality holds from the assumption on the stepsize $\alpha$ which is chosen such that $2\alpha\widetilde{\rho}\lambda_{\max}(\boldsymbol{\Sigma}^{-1}) \leq 1$ satisfies. Applying the above inequality recursively, we have

$$\begin{aligned}\mathbb{E}\left[\left\|\boldsymbol{P}^{(k)} - \boldsymbol{P}^*\right\|_{\boldsymbol{\Sigma}}^2\right] &\leq \left(1 - \alpha\overline{\mu}\lambda_{\min}(\boldsymbol{\Sigma}^{-1})\right)^k\mathbb{E}\left[\|\boldsymbol{P}_0 - \boldsymbol{P}^*\|_{\boldsymbol{\Sigma}}^2\right] + 2\alpha^2\lambda_{\max}(\boldsymbol{\Sigma}^{-1})\sigma_{\mathscr{L}} \times \sum_{j=0}^{k-1}\left(1 - \alpha\overline{\mu}\lambda_{\min}(\boldsymbol{\Sigma}^{-1})\right)^{j-1} \\ &\leq \left(1 - \alpha\overline{\mu}\lambda_{\min}(\boldsymbol{\Sigma}^{-1})\right)^k\|\boldsymbol{P}_0 - \boldsymbol{P}^*\|_{\boldsymbol{\Sigma}}^2 + 2\alpha^2\lambda_{\max}(\boldsymbol{\Sigma}^{-1})\sigma_{\mathscr{L}} \times \sum_{j=0}^{\infty}\left(1 - \alpha\overline{\mu}\lambda_{\min}(\boldsymbol{\Sigma}^{-1})\right)^{j-1} \\ &\leq \left(1 - \alpha\overline{\mu}\lambda_{\min}(\boldsymbol{\Sigma}^{-1})\right)^k\|\boldsymbol{P}_0 - \boldsymbol{P}^*\|_{\boldsymbol{\Sigma}}^2 + 2\alpha^2\lambda_{\max}(\boldsymbol{\Sigma}^{-1})\sigma_{\mathscr{L}} \times \frac{1}{1 - (1 - \alpha\overline{\mu}\lambda_{\min}(\boldsymbol{\Sigma}^{-1}))}\end{aligned}$$

$$= \left(1 - \alpha\overline{\mu}\lambda_{\min}(\boldsymbol{\Sigma}^{-1})\right)^k \|\boldsymbol{P}_0 - \boldsymbol{P}^*\|_{\boldsymbol{\Sigma}}^2 + \frac{2\alpha\lambda_{\max}(\boldsymbol{\Sigma}^{-1})\sigma_{\mathcal{L}}}{\overline{\mu}\lambda_{\min}(\boldsymbol{\Sigma}^{-1})}$$
$$= \left(1 - \alpha\overline{\mu}\lambda_{\min}(\boldsymbol{\Sigma}^{-1})\right)^k \|\boldsymbol{P}_0 - \boldsymbol{P}^*\|_{\boldsymbol{\Sigma}}^2 + \frac{2\alpha\,\mathrm{cond}\left(\boldsymbol{\Sigma}^{-1}\right)}{\overline{\mu}}\sigma_{\mathscr{L}}.$$

This concludes the proof. □

**Corollary A.12.** *Consider the same situation as in Theorem A.9. Let* $\alpha = \min\left\{\frac{\varepsilon\overline{\mu}}{4\,\mathrm{cond}(\boldsymbol{\Sigma}^{-1})\sigma_{\mathscr{L}}}, \frac{1}{2\widetilde{\rho}\lambda_{\max}(\boldsymbol{\Sigma}^{-1})}\right\}$. *Then, for any* $\varepsilon > 0$, *we can guarantee that* $\mathbb{E}\left[\left\|\boldsymbol{P}^{(k)} - \boldsymbol{P}^*\right\|_{\boldsymbol{\Sigma}}^2\right] \leq \varepsilon$ *provided that*

$$k \geq \max\left\{\frac{1}{\varepsilon}\frac{4\,\mathrm{cond}\left(\boldsymbol{\Sigma}^{-1}\right)\sigma_{\mathscr{L}}}{\overline{\mu}^2}, \frac{2\widetilde{\rho}\lambda_{\max}(\boldsymbol{\Sigma}^{-1})}{\overline{\mu}}\right\}\ln\left(\frac{2\|\boldsymbol{P}_0 - \boldsymbol{P}^*\|_{\boldsymbol{\Sigma}}^2}{\varepsilon}\right).$$

*Proof.* See Garrigos and Gower (2023, Lemma A.3). □

## B Minibatch Case

This section provides the proof of Theorem 4.5 and Theorem 4.7. We make the additional assumption on the expected smoothness (see Definition 2.4) and gradient noise for the gradient estimator with minibatch.

**Assumption B.1.** For any $B \subseteq [n]$, we assume $\nabla\mathscr{L}_B \in \mathrm{ES}(\widetilde{\rho})$, that is, there exists a constant $\widetilde{\rho}$ such that

$$\frac{1}{2\widetilde{\rho}}\mathbb{E}\left[\|\mathscr{L}_B(\boldsymbol{P}) - \mathscr{L}_B(\boldsymbol{P}^*)\|^2\right] \leq \mathscr{L}(\boldsymbol{P}) - \mathscr{L}(\boldsymbol{P}^*),$$

where $\boldsymbol{P}^*$ is a minimizer of $\mathscr{L}$.

**Assumption B.2.** The gradient noise

$$\sigma_{\mathscr{L}} = \sup_{\boldsymbol{P}^* \in \mathscr{P}} \mathbb{E}\left[\|\nabla\mathscr{L}_B(\boldsymbol{P}^*)\|^2\right]$$

is finite. Here, $\mathscr{P}$ is a set of minimizers of $\mathscr{L}$, i.e., $\mathscr{P} \coloneqq \mathrm{argmin}_{\boldsymbol{P}}\,\mathscr{L}(\boldsymbol{P})$.

As a direct consequence of Assumption B.1 and B.2, we have the following result that bounds on the norm of $\mathscr{L}(\boldsymbol{P})$.

**Lemma B.3.** *Under Assumptions B.1, and B.2, we have*

$$\mathbb{E}\left[\|\nabla\mathscr{L}_B(\boldsymbol{P})\|^2\right] \leq 4\widetilde{\rho}(\mathscr{L}(\boldsymbol{P}) - \mathscr{L}(\boldsymbol{P}^*)) + 2\sigma_{\mathscr{L}}.$$

*Proof.* For any $B \subseteq [n]$, we have

$$\begin{aligned}\mathbb{E}\left[\|\nabla\mathscr{L}_B(\boldsymbol{P})\|^2\right] &\leq 2\mathbb{E}\left[\|\mathscr{L}_B(\boldsymbol{P}) - \mathscr{L}_B(\boldsymbol{P}^*)\|^2\right] + 2\mathbb{E}\left[\|\mathscr{L}_B(\boldsymbol{P}^*)\|^2\right] \\ &\leq 4\widetilde{\rho}(\mathscr{L}(\boldsymbol{P}) - \mathscr{L}(\boldsymbol{P}^*)) + 2\sigma_{\mathscr{L}},\end{aligned}$$

where the first inequality holds for any $\boldsymbol{P}$, and the second inequality holds by the Assumptions B.1, B.2. □

**Theorem B.4** (Theorem 4.5)**.** *Assume that Assumptions 4.1, B.1, and B.2 hold. Consider a sequence of control points* $\{\boldsymbol{P}^{(k)}\}_{k\in\mathbb{N}}$ *generated by Algorithm 1 with stepsizes satisfying* $\alpha_k \leq \min\{(4\widetilde{\rho}\lambda_{\max}(\boldsymbol{\Sigma}^{-1}))^{-1}, (\overline{\mu}\lambda_{\min}(\boldsymbol{\Sigma}^{-1}))^{-1}\}$ *for* $k = 0, 1, \ldots, K-1$ *and an initial control point* $\boldsymbol{P}_0$. *Then, for any* $K > 0$, *we have*

$$\mathbb{E}\left[\mathscr{L}(\widetilde{\boldsymbol{P}}^{(K)}) - \mathscr{L}(\boldsymbol{P}^*)\right] \leq \frac{(1 - \alpha_0\overline{\mu}\lambda_{\min}(\boldsymbol{\Sigma}^{-1}))}{\sum_{k=0}^{K-1}\alpha_k}\|\boldsymbol{P}_0 - \boldsymbol{P}^*\|_{\boldsymbol{\Sigma}}^2 + \frac{2\lambda_{\max}(\boldsymbol{\Sigma}^{-1})\sum_{k=0}^{K-1}\alpha_k^2}{\sum_{k=0}^{K-1}\alpha_k}\sigma_{\mathscr{L}},$$

*where* $\widetilde{\boldsymbol{P}}^{(K)}$ *is the weighted average of control points* $\boldsymbol{P}^{(k)}$ *for* $k = 0, 1, \ldots, K-1$ *which is defined as*

$$\widetilde{\boldsymbol{P}}^{(K)} \coloneqq \sum_{k=0}^{K-1} q_{K,k}\boldsymbol{P}^{(k)}, \quad \textit{with} \quad q_{K,k} \coloneqq \frac{\alpha_k(1 - 2\alpha_k\widetilde{\rho})}{2\sum_{\kappa=0}^{K-1}\alpha_\kappa(1 - 2\alpha_\kappa\widetilde{\rho})}.$$

*Proof.* Using Lemma B.3, we can prove the same way as Theorem A.9. □

**Theorem B.5** (Theorem 4.7)**.** *Assume that Assumptions 4.1, B.1, and B.2 hold. Let $\boldsymbol{P}_0$ be an initial control points of the Bézier simplex. Consider a sequence of control points $\{\boldsymbol{P}^{(k)}\}_{t\in\mathbb{N}}$ generated by Algorithm 1 with a constant stepsize $\alpha_k = \alpha < \left(2\widetilde{\rho}\lambda_{\max}(\boldsymbol{\Sigma}^{-1})\right)^{-1}$. Then, we have*

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

*Proof.* Using Lemma B.3, we can prove the same way as Theorem A.11. □