Stochastic Tangential Pareto Dynamics Provably Samples the Whole Pareto Set

Zachary Jones, Pietro Marco Congedo, Olivier Le Maitre

STOCHASTIC TANGENTIAL PARETO DYNAMICS PROVABLY SAMPLES THE WHOLE PARETO SET *

ZACHARY JONES, PIETRO MARCO CONGEDO, OLIVIER LE MAÎTRE

Abstract. The framework of stochastic multi-objective programming allows for the inclusion of uncertainties in multi-objective optimization problems at the cost of transforming the set of objectives into a set of expectations of random quantities. The stochastic multigradient descent algorithm (SMGDA) gives a solution to these types of problems using only noisy gradient information. However, a bias in the algorithm causes it to converge to only a subset of the whole Pareto front, limiting its use. We analyze the source of this bias and prove the convergence of SMGDA to a stationary point in the nonconvex L-lipschitz smooth case. First, based on this analysis, we propose to reduce the bias of the stochastic multi-gradient calculation using an exponential smoothing technique. We then propose a novel approach to exploring the whole Pareto set by combining the debiased stochastic multigradient with an additive non-vanishing noise that guides the dynamics of the iterates tangential to the Pareto set. We finish by proving that our algorithm, Stochastic Tangential Pareto Dynamics (STPD), generates samples concentrated on the whole Pareto set.

Key words. Stochastic Programming, Multi-Objective Optimization, Stochastic Multi-Objective Optimization, Gradient, Sampling, Pareto Front, Pareto Set

MSC codes. 68Q25, 68R10, 68U05

1. Introduction.

Finding all possible solutions to an optimization problem with competing objectives and uncertainty requires specialized methods. In general there is no unique solution to an optimization problem with competing objectives but rather a collection of possible tradeoffs. A Pareto optimal point is then a design point which cannot be improved along any objective without degrading the performance of others. *the Pareto set* is the set of all Pareto optimal points. Introducing stochasticity into the objective functions complicates the process of finding the Pareto set. The objective functions then become random variables making both optimization and the determination of Pareto optimality into probabilistic tasks.

To handle the challenges of stochastic multi-objective optimization we work within

*Submitted to the editors DATE.

Funding: This work was funded by Horizon NEXTAIR.

the framework of stochastic programming. We formulate our stochastic multi-objective optimization problem as the minimization of a set of expected values of random quantities of interest. This changes the stochastic problem into a deterministic one. Unofrtunately, we rarely have access to analytic expressions for the expected values of the quantities of interest. Furthermore, estimating them can prove to be computationally costly. Therefore, we turn to stochastic approximation.

The stochastic multi-gradient descent algorithm (SMGDA) is the extension of stochastic approximation methods to the multi-objective context. It has attracted attention in engineering, operations research, and machine learning fields [5, 11, 9, 4]. At each iteration SMGDA uses stochastic gradient information from each objective to compute a direction of descent, the stochastic multigradient. This approach has two main benefits. SMGDA iterations can be performed with as few as a single gradient sample from each objective, making estimation unnecessary. It also has the advantage of provably converging to the Pareto set [9]. However, the stochastic multigradient has a known bias [5, 11]. Because of this bias, previous analysis of the convergence of SMGDA beyond the strongly convex setting have depended on strong assumptions or even concluded that the algorithm does not converge [11, 5, 4].

Our first contribution is to analyze the source and effect of this bias. We prove that, despite bias in the stochastic multigradient, SMGDA converges to a *subset* of the Pareto set in the convex L-Lipschitz smooth setting. We additionally show convergence of SMGDA in the strongly convex and nonconvex L-Lipschitz smooth settings.

Our second contribution is a sampling approach to generate candidate Pareto optimal points, stochastic tangential Pareto dynamics (STPD). It has two main ingredients, a debiased stochastic mutligradient step and an additive noise term. To debias the stochastic multigradient, based on our analysis of SMGDA, we propose a method using exponential smoothing. We then combine this debiased stochastic multigradient with a nonvanishing noise term which adds noise perpendicular to the expected stochastic multigradient. This noise term guides the sampling procedure tangentially along the Pareto set. We go on to prove that STPD generates samples

which lie arbitrarily close to the whole Pareto set in finite time.

Our companion paper then tackles the complementary problem of probabilistic inference of the Pareto set using a noisy fixed-size sample set [CITATION NEEDED].

We begin the manuscript by recalling some basic concepts in both deterministic and stochastic multi-objective optimization in Section 2. We then introduce the scalar potential for multi-objective optimization in Section 3.1, and use it to prove the convergence of MGDA in Section 3.2. We use the foundation built in Section 3 to make new, more general, proof of convergence of SMGDA in Section 4.1. We also see that SMGDA converges only to a subset of the Pareto front. We examine the bias of SMGD in Section 4.2. We introduce our sampling approach, Stochastic Tangential Pareto Dynamics, in Section 5. We continue to show that it samples the whole of the Pareto front under mild assumptions and offer discussion about the potential applications, pitfalls, improvements, and variations to our proposed approach. Finally, we give our conclusions and future outlook in Section 6

2. Background on Gradient Based (Stochastic) Multi-Objective Optimization. We start by introducing stochastic programming before reviewing the multigradient descent algorithm and Pareto concepts.

Let us look at the stochastic multi-objective programming problem. Given a probability space $(\Theta, \mathcal{F}, \mu)$ and a set of k quantities of interest collected as a vector valued function $F := [f_i(\mathbf{x}, \mathbf{W}(\theta)), ..., f_k(\mathbf{x}, \mathbf{W}(\theta))] : \mathcal{X} \subseteq \mathbb{R}^d \times \mathbf{W}(\Theta) \mapsto \mathcal{Y} \subseteq \mathbb{R}^k$, our goal is to find *all* $\mathbf{x}^*$ such that

$$\mathbf{x}^* \in \arg\min_{\mathbf{x} \in \mathcal{X}} \{\mathbb{E}[f_i(\mathbf{x}, \mathbf{W}(\theta))], ..., \mathbb{E}[f_k(\mathbf{x}, \mathbf{W}(\theta))]\}. \tag{2.1}$$

This formulation is sufficiently general to incorporate mean/variance minimization, conditional value at risk, and quantile functions as objectives. Notably, the problem is deterministic. If we have analytical expressions for $G(\mathbf{x}) = [g_1(\mathbf{x}), ..., g_k(\mathbf{x})], \mathcal{X} \mapsto \mathcal{Y} \subseteq \mathbb{R}^k$ with $g_i(\mathbf{x}) = \mathbb{E}[f_i(\mathbf{x}, \mathbf{W}(\theta))]$, we can treat this as a deterministic problem

and solve it as

$$\mathbf{x}^* \in \underset{\mathbf{x}\in\mathcal{X}}{\operatorname{argmin}}\{g_1(\mathbf{x}), ..., g_k(\mathbf{x})\}. \tag{2.2}$$

Clearly we will rarely have a unique solution to problem 2.1. Instead, given two solutions, $\mathbf{y}', \mathbf{y} \in \mathcal{Y}$, $\mathbf{y}'$ *dominates* $\mathbf{y}$ ($\mathbf{y}' \prec \mathbf{y}$) if it is lower componentwise in all dimensions, e.g. $\forall j \in \{1, ..., k\}$ $\mathbf{y}'_i < \mathbf{y}_i$. If no $\mathbf{y}'$ exists which dominates a point $\mathbf{y}$, we say it is *undominated*. The Pareto front is defined as the set of undominated points.

DEFINITION 2.1 (Pareto Front).

$$P(\mathcal{Y}) := \{\mathbf{y} \in \mathcal{Y} \mid \nexists \mathbf{y}' \in \mathcal{Y}\backslash\mathbf{y} \; s.t. \; \mathbf{y}'_\mathbf{j} \leq \mathbf{y}_\mathbf{j} \; \forall j \in \{1, ..., k\}\}. \tag{2.3}$$

This set gives a mathematical description of Pareto optimality. If $\mathbf{y} \in P$ it means that, in at least one pair of criteria, there is no global performance improvement. We can define the set of solutions in design space, the Pareto set, as the pre-image of the Pareto front of G.

DEFINITION 2.2 (Pareto Set).

$$P^{-1}(\mathcal{X}) := (P \circ G)(\mathcal{X}) = \{\mathbf{x} \in \mathcal{X} \mid G(\mathbf{x}) \in P(\mathcal{Y})\}. \tag{2.4}$$

Points $\mathbf{x}^*$ in 2.2 can be found using the multigradient descent algorithm.

It is an extension of gradient descent to the multi-objective context using the multigradient.

At each iteration we calculate

$$\nabla^{\mathcal{C}}_x\{G\}(\mathbf{x}_t) = \underset{\mathbf{d}}{\operatorname{argmax}} \max_i \langle \nabla g_i(\mathbf{x}_t), \mathbf{d}\rangle + \frac{1}{2}||\mathbf{d}||^2_2. \tag{2.5}$$

Under sufficient conditions on G, the sequence of points $\mathbf{x}_{t+1} = \mathbf{x}_t - \varepsilon\nabla^{\mathcal{C}}_x\{G\}(\mathbf{x}_t)$ will then converge to a Pareto stationary point

106 DEFINITION 2.3 (Pareto Stationary Point). $\mathbf{x}^*$ *such that*

107 (2.6) $$\nabla_x^{\mathcal{C}}\{G\}(\mathbf{x}^*) = \mathbf{0}$$

108 In the case where all objective functions are convex this forms a sufficient condition
109 for Pareto optimality [6].

110 In later sections, we will use Pareto stationarity as a tool to search for Pareto
111 optimal points.

112 In practice, solving Eq. 2.5 in its primal form is cumbersome, and we will use its
113 dual formulation, which reduces to a quadratic subproblem in the number of objec-
114 tives,

115 (2.7) $$\nabla_x^{\mathcal{C}}\{G\}(\mathbf{x}) = \sum_i \alpha_i^* \nabla g_i(\mathbf{x})$$

116 $$\textbf{s.t. } \alpha^* = \underset{\alpha \in \Delta^{k-1}}{\operatorname{argmin}} || \sum_i \alpha_i \nabla g_i(\mathbf{x})||_2^2,$$
117

118 where Δ^{k-1} is the $k-1$ dimensional simplex. This subproblem has a closed form
119 solution in the bi-objective case and can be efficiently solved for more objectives
120 using the Franke-Wolfe algorithm [3]. The multigradient, $\nabla_x^{\mathcal{C}}\{G\}(\mathbf{x})$, can be concisely
121 understood as the minimum-norm convex combination of gradients, as seen in Fig. 1.

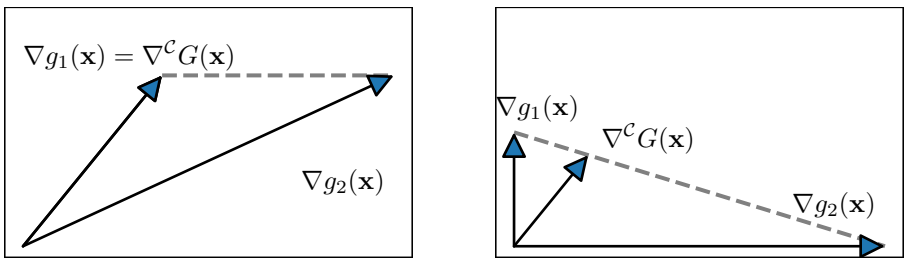


Fig. 1: Examples of Calculation of Direction of Descent $\nabla_x^{\mathcal{C}}\{G\}(\mathbf{x})$ for two objectives.

122

123 **3. Solving Deterministic Multi-Objective Problems with MGDA.** In
124 this section we reformulate and solve the deterministic multi-objective problem. To

reformulate the multi-objective problem we make an equivalent problem with only one objective, a scalar potential. This provides a new way prove convergence of MGDA and to view multi-objective optimization. In Section 4 we will extend these techniques to the stochastic case and in Section 5 we will exploit this framework to prove that our proposed approach samples points along the whole Pareto set.

3.1. Potential function for Multi-Objective Optimization. To reformulate the multi-objective problem we make an equivalent problem with only one objective, a scalar potential. Previous attempts have used the coefficients, α^*, defined in Eq. 2.7 to define a pseudo-objective $\Psi(\mathbf{x}) = \sum_i \alpha_i^* g_i(\mathbf{x})$ [5]. However, the parameters α^* are not constant with respect to $\mathbf{x}$ and so the gradient, $\nabla_x^{\mathcal{C}}\{G\}(\mathbf{x})$, is *not* $\nabla\Psi(\mathbf{x})$. Consequently $\Psi(\mathbf{x})$ is minimized only at the minima of each individual objective and not over the whole front.

We propose a potential function that has minima over the whole of the Pareto set. We define the potential, $\Phi_G(\mathbf{x}) : \mathbb{R}^d \mapsto \mathbb{R}$, as follows:

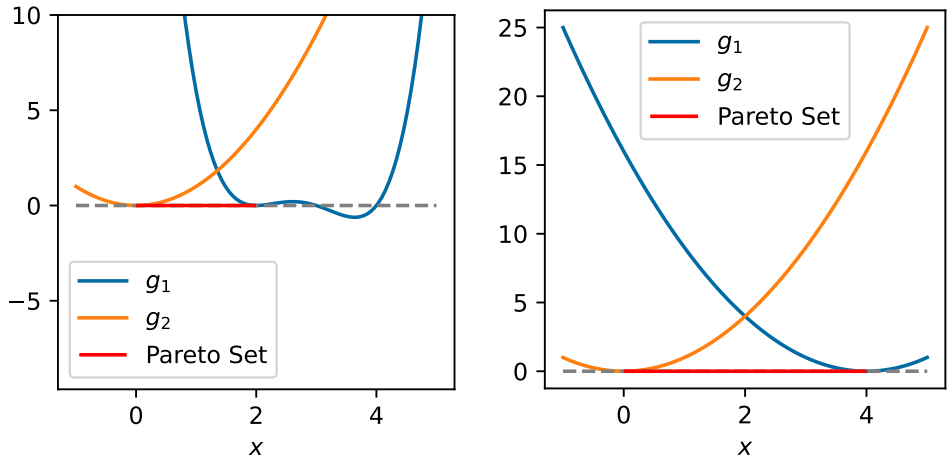
DEFINITION 3.1 (Equivalent Scalar Potential Function).

$$\Phi_G(\mathbf{x}) = \int_0^1 \langle \nabla_x^{\mathcal{C}}\{G\}(s\mathbf{x}), \mathbf{x}\rangle ds = \int_0^1 \sum_i \alpha_i^*(s\mathbf{x})\langle \nabla g_i(s\mathbf{x}), \mathbf{x}\rangle ds. \tag{3.1}$$

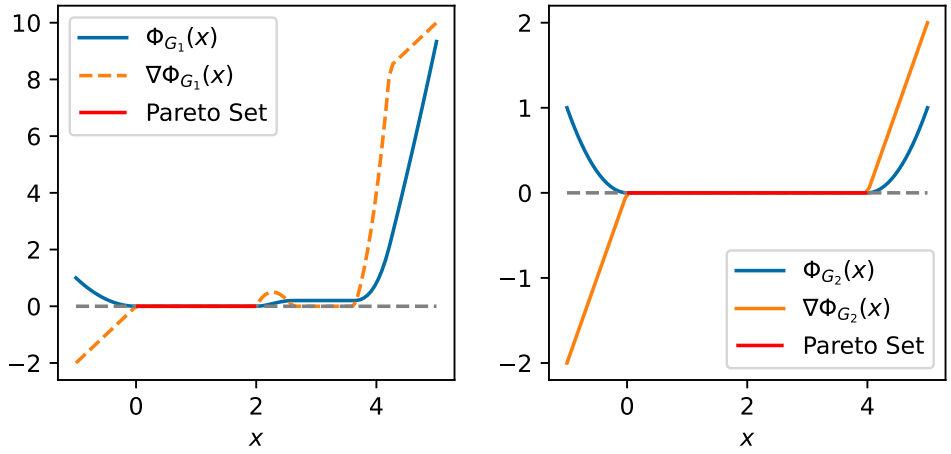
The potential $\Phi_G(\mathbf{x})$ is $\mathcal{C}^1$ smooth, locally Lipschitz continuous, and it can be seen that $\nabla\Phi(\mathbf{x}) = \nabla_x^{\mathcal{C}}\{G\}(\mathbf{x})$ by the gradient theorem. Clearly this function has minima at Pareto stationary points.

It can be readily seen in Fig. 2d that the minima of Φ_G lies on the Pareto set. In the case where there are nonconvex objectives, the set of Pareto stationary points can form a saddle point in $\Phi_G(\mathbf{x})$. The global minima remains on the Pareto set. We remark here that we have used the curve from 0 to $\mathbf{x}$ parameterized by t to perform the line integral that defines the potential function. This is done for convenience only. Any continuous curve in $\mathcal{X}$ with endpoint $\mathbf{x}$ can be used so long as the initial point is consistent and $\mathcal{X}$ is pathwise connected.

Using 3.1, we can make claims about the set of objective functions using a single scalar value.



(a) $g_1(\mathbf{x}) = (x-2)^2(x-3)(x-4)$, $g_2(\mathbf{x}) = x^2$. (b) $g_1(\mathbf{x}) = (x-4)^2$, $g_2(\mathbf{x}) = x^2$.



(c) $g_1(\mathbf{x}) = (x-2)^2(x-3)(x-4)$, $g_2(\mathbf{x}) = x^2$. (d) $g_1(\mathbf{x}) = (x-4)^2$, $g_2(\mathbf{x}) = x^2$.

Fig. 2: Top row: Competing objective functions in the cases with local minima (left) and bi-convex objectives(right). Bottom row: Minima of the equivalent scalar potential function lie on the Pareto front, with Pareto stationary nonoptimal points revealing themselves as saddle points (left).

3.2. Convergence of MGDA. We use the scalar potential to make a new proof of the convergence of MGDA. In the next section, we will use a similar strategy to show the conditions required for SMGDA to converge.

The minima of $\Phi(\mathbf{x})$ are the set of Pareto stationary points, and performing gradient descent on $\Phi(\mathbf{x})$ yields Pareto stationary points.

THEOREM 3.2. *Let $\Phi_G(\mathbf{x})$ be a potential function defined as in 3.1. With $\Phi_G : \mathcal{X} \mapsto \mathcal{Y}$, $\mathcal{X}$ compact. Then the sequence*

$$\mathbf{x}_{t+1} = \mathbf{x}_t - \varepsilon \nabla \Phi_G(\mathbf{x}). \tag{3.2}$$

converges to a Pareto stationary point.

Proof. From the assumption that Φ_G is L-Liptshitz smooth

$$\Phi_G(\mathbf{x}_{t+1}) \leq \Phi_G(\mathbf{x}_t) + \langle \nabla \Phi_G(\mathbf{x}_t), \mathbf{x}_{t+1} - \mathbf{x}_t \rangle + \frac{L}{2} ||\mathbf{x}_{t+1} - \mathbf{x}_t||^2 \tag{3.3}$$

$$\leq \Phi_G(\mathbf{x}_t) - \varepsilon(1 - \frac{L\varepsilon}{2}) ||\nabla_x^{\mathcal{C}}\{G\}(\mathbf{x}_t)||^2. \tag{3.4}$$

Picking $\varepsilon \leq \frac{2}{L}$, summing, using the fact that $\nabla \Phi(\mathbf{x}) = \nabla_x^{\mathcal{C}}\{G\}(\mathbf{x})$, and dividing by T gives us the result

$$\frac{1}{T} \sum_{t=1}^{T} ||\nabla_x^{\mathcal{C}}\{G\}(\mathbf{x}_t)||^2 \leq \frac{\Phi_G(\mathbf{x}_0) - \Phi_G(\mathbf{x}_{T+1})}{T\varepsilon}. \tag{3.5}$$

∎

4. Solving Stochastic Multi-Objective Problems with SMGDA. We now turn our attention to the stochastic formulation of the multigradient descent algorithm, SMGDA. It has already been proven that SMGDA converges to the Pareto front almost surely under strong assumptions. Previous approaches either require the individual objective functions are almost surely convex with at least one being almost surely strongly convex or they assume that the bias in the stochastic multigradient vanishes in the long time limit.[9, 4, 5]. These cases are not realistic. It has also been shown that there exist cases where there is a nonvanishing bias in the stochastic multigradient along the whole Pareto set [11]. The authors went on to conclude that SMGDA does not converge in these cases. However, this is not necessarily true.

Here, we extend the tools we have developed in the previous section to the stochastic case. This allows us to give proof of convergence of SMGDA in the nonconvex and biased case. We then go on to show that, when Pareto stationary points for which the stochastic gradient estimator is unbiased exist, that SMGDA converges to one of

those points.

In the stochastic programming setting the set of objective functions $\{g_i\}_{i=1,\dots,k}$ now have the functional form of an expectation $g_i(\mathbf{x}) = \mathbb{E}[f_i(\mathbf{x}, \mathbf{W}(\theta))]$. Unfortunately, we do not often have analytic expressions for our objectives and must work with realizations of the quantities of interest, $\{f_i(\mathbf{x}, \mathbf{W}(\theta))\}_{1,\dots,k}$ and their gradients. It is possible to estimate the value of our objectives and plug them in to a deterministic optimization algorithm. However, this strategy introduces residual error and can require costly resampling at each iteration. Ignoring the random nature of the problem by naïvely substituting samples of the quantities of interest, $f_i(\mathbf{x}_t, \mathbf{W}(\theta)_t)$, for their true value in a deterministic multi-objective method could lead to convergence to a suboptimal point.

Instead, we estimate the gradient of each objective, $\nabla g_i(\mathbf{x})$, using a single realization of the gradient of the corresponding quantity of interest $\nabla f_i(\mathbf{x}, \mathbf{W}(\theta))$. These estimates are then plugged into Eq. 2.7 to calculate the stochastic multigradient $\nabla_x^{\mathcal{C}}\{F\}(\mathbf{x}, \mathbf{W}(\theta))$. It is assumed that $\nabla f_i(\mathbf{x}, \mathbf{W}(\theta))$ is an unbiased estimator of $\nabla g(\mathbf{x})$ with finite second moment, e.g. $\mathbb{E}[\nabla f(\mathbf{x}, \mathbf{W}(\theta))] = \nabla g(\mathbf{x})$ and $\mathbb{V}(\nabla f(\mathbf{x}, \mathbf{W}(\theta))) < \infty$ with $\mathbb{V}(\cdot)$ as the variance functional. The heart of the stochastic multigradient descent algorithm is then the sequence of points generated by the relation

$$\mathbf{x}_{t+1} = \mathbf{x}_t - \varepsilon_t \nabla_x^{\mathcal{C}}\{F\}(\mathbf{x}, W). \tag{4.1}$$

We require that ε_t meet the requirement that

$$\frac{\sum_t^T \varepsilon_t^2}{\sum_t^T \varepsilon_t} \xrightarrow[T\to\infty]{} 0. \tag{4.2}$$

This is a standard requirement in stochastic optimization [1, 9].

It is straightforward then to extend the potential function from Section 3 to the stochastic multi-objective case.

DEFINITION 4.1 (Stochastic Equivalent Scalar Potential Function).

$$\Phi_F(\mathbf{x}) = \int_0^1 \langle \mathbb{E}[\nabla_x^{\mathcal{C}}\{F\}(s\mathbf{x})], \mathbf{x}\rangle ds. \tag{4.3}$$

The gradient of this potential function can be approximated using individual realizations of $\mathbf{W}(\theta)$ as $\nabla_x^{\mathcal{C}}\{F\}(\mathbf{x}, \mathbf{W}(\theta))$.

These quantities are summarized in Tab. 1.

	Objectives	Gradients	Potential	Direction of Ascent	step size
MGDA	$g(\mathbf{x})$	$\nabla g(\mathbf{x})$	Φ_G	$\nabla_x^{\mathcal{C}}\{G\}(\mathbf{x})$	ε
SMGDA	$\mathbb{E}[f(\mathbf{x}, \mathbf{W}(\theta))]$	$\nabla f(\mathbf{x}, \mathbf{W}(\theta)_i)$	Φ_F	$\nabla_x^{\mathcal{C}}\{F\}(\mathbf{x}, \mathbf{W}(\theta))$	ε_t

Table 1: Quantities used in MGDA and their stochastic programming equivalents in SMGDA.

4.1. Convergence of SMGDA. We can now show new proof of the convergence of the SMGD algorithm to a Pareto stationary point in the nonconvex case with biased gradients. Proofs of convergence for convex and strongly convex objective functions in the biased setting with bounded gradients can be found in the appendix.

We first make an assumption on the bias of the gradient estimator,

ASSUMPTION 4.2. $\exists a > 0, b \geq 0$ *such that*

$$\langle \mathbb{E}[\nabla_x^{\mathcal{C}}\{F\}(\mathbf{x}, \mathbf{W}(\theta))], \nabla_x^{\mathcal{C}}\{G\}(\mathbf{x})\rangle \geq a||\nabla_x^{\mathcal{C}}\{G\}(\mathbf{x})||_2^2 - b. \tag{4.4}$$

almost everywhere.

This assumption intuitively states that the stochastic gradient is "in-line" with the true gradient up to a constant bound, b. Many general proofs of the convergence of SGD in the single objective case make a similar assumption, setting $b = 0$ directly [1]. This assumption is more general. We can always find $b \geq 0$ such that this relationship holds with arbitrarily high probability. As we will see later, if we can find a pair a, b such that $b = 0$ and Assumption 4.2 holds almost surely over all $\mathbf{x} \in \mathcal{X}$, then SMGD converges to a Pareto stationary point. The interested reader can pursue more details

in the appendix.

We also assume that

ASSUMPTION 4.3. *$\exists a_G > 2a - 1$ and $b_G \geq 0$ such that*

$$\mathbb{E}[||\nabla_x^{\mathcal{C}}\{F\}(\mathbf{x}, \mathbf{W}(\theta))||_2^2] \leq a_G||\nabla_x^{\mathcal{C}}\{G\}(\mathbf{x})||_2^2 + b_G \tag{4.5}$$

almost everywhere.

Assumptions 4.2 and 4.3 are not strong and imply two things. The first, that the effect of the bias is lower bounded:

$$\mathbb{E}[\langle \nabla_x^{\mathcal{C}}\{F\}(\mathbf{x}, \mathbf{W}(\theta)), \nabla_x^{\mathcal{C}}\{G\}(\mathbf{x})\rangle - ||\nabla_x^{\mathcal{C}}\{G\}(\mathbf{x})||_2^2] \geq (a-1)||\nabla_x^{\mathcal{C}}\{G\}(\mathbf{x})||_2^2 - b. \tag{4.6}$$

Second, setting $M_V \geq (a_g + 1 - 2a)$ and $M_0 \geq 2b + b_G$, we see that assumptions 4.2 and 4.3 imply 4.7.

$$\mathbb{E}[||\nabla_x^{\mathcal{C}}\{F\}(\mathbf{x}, \mathbf{W}(\theta)) - \nabla_x^{\mathcal{C}}\{G\}(\mathbf{x})||_2^2] \leq M_V||\nabla_x^{\mathcal{C}}\{G\}(\mathbf{x})||_2^2 + M_0 \tag{4.7}$$

Readers may recognize Eq. 4.7 as a common assumption in stochastic approximation [1]. We now prove the convergence of SMGDA to a Pareto stationary point in the nonconvex setting.

THEOREM 4.4. *Let $F = \{f_i(\mathbf{x}, \mathbf{W}(\theta))\}_{i=1,\ldots,k}$ be a vector valued function $F : \mathcal{X} \mapsto \mathcal{Y} \subseteq \mathbb{R}^k$ with each $g_i(\mathbf{x}) = \mathbb{E}[f_i(\mathbf{x}, \mathbf{W}(\theta))]$ L-Lipschitz continuous. If there is pair of constants $a > 0$, $b = 0$ such that $\langle \nabla\Phi_F, \nabla\Phi_G\rangle \geq a||\nabla\Phi_G||_2^2 - b$ and ε_t meeting requirement 4.2, then the sequence of iterates $\{\mathbf{x}_i\}_{i\in\mathbb{N}}$ generated through the relation*

$$\mathbf{x}_{t+1} = \mathbf{x}_t - \varepsilon_t \nabla_x^{\mathcal{C}}\{F\}(\mathbf{x}_t, \mathbf{W}(\theta)_t) \tag{4.8}$$

converges almost surely to a Pareto stationary point.

Proof. Let $B_t = \mathbb{E}[\nabla\Phi_F(\mathbf{x}_t) - \nabla\Phi_G(\mathbf{x}_t)]$. We start from L-Lipschitz smoothness.

$$\Phi_G(\mathbf{x}_{t+1}) \leq \Phi_G(\mathbf{x}_t) + \langle \nabla\Phi_G(\mathbf{x}_t), \mathbf{x}_{t+1} - \mathbf{x}_t \rangle + \frac{L}{2}||\mathbf{x}_{t+1} - \mathbf{x}_t||_2^2.$$

Taking the conditional expectation, and using the definition of SMGD iterations, we have the expression:

$$\Phi_G(\mathbf{x}_{t+1}) \leq \Phi_G(\mathbf{x}_t) - \mathbb{E}[\varepsilon_t \langle \nabla\Phi_G(\mathbf{x}_t), \nabla_x^{\mathcal{C}}\{F\}(\mathbf{x}_t, \mathbf{W}(\theta)_t)\rangle] + \mathbb{E}[\frac{L\varepsilon_t^2}{2}||\nabla_x^{\mathcal{C}}\{F\}(\mathbf{x}_t, \mathbf{W}(\theta)_t||_2^2].$$

Adding and subtracting $\varepsilon_t||\Phi_G(\mathbf{x}_t)||_2^2$, substituting our expression for the bias, and using assumptions 4.2 and 4.3, we see that

$$\begin{aligned}\Phi_G(\mathbf{x}_{t+1}) &\leq \Phi_G(\mathbf{x}_t) - \varepsilon_t||\nabla\Phi_G(\mathbf{x}_t)||_2^2 - \varepsilon_t\langle B_t, \nabla\Phi_G(\mathbf{x}_t)\rangle + \frac{L\varepsilon_t^2}{2}\big(b + (a_g + 1 - 2a)||\nabla\Phi_G(\mathbf{x}_t)||_2^2\big) \\ &\leq \Phi_G(\mathbf{x}_t) - \varepsilon_t(a - \varepsilon_t\frac{L}{2}M_V)||\nabla\Phi_G(\mathbf{x}_t)||_2^2 + \varepsilon_t b + \frac{L\varepsilon_t^2}{2}M_0.\end{aligned}$$

Requiring that $\varepsilon_t < \frac{2a}{LM_V}\ \forall t$, taking the full expectation, summing, and re-arranging, we get

$$\sum_t^T \varepsilon_t||\nabla\Phi_G(\mathbf{x}_t)||_2^2 \leq \Phi_G(\mathbf{x}_0) - \mathbb{E}[\Phi_G(\mathbf{x}_T)] + b\sum_t \varepsilon_t + \frac{LM_0}{2}\sum_t \varepsilon_t^2.$$

Finally, defining $E_T = \sum_t \varepsilon_t$, we can treat ε_t/E_T as a probability measure. Choosing index t in $1, ..., T$ and applying Markov's inequality yields

$$\mathbb{P}(||\nabla\Phi_G(\mathbf{x})||_2^2 \geq \epsilon) \leq \frac{\mathbb{E}_{E_t}[||\nabla\Phi_G(\mathbf{x})||_2^2]}{\epsilon} \leq \frac{\Phi_G(\mathbf{x}_0) - \mathbb{E}[\Phi_G(\mathbf{x}_T)]}{\epsilon E_T} + \frac{b}{\epsilon} + \frac{LM_0}{2\epsilon E_T}\sum_t \varepsilon_t^2.$$

Taking the limit as $T \to \infty$ gives the result

$$\mathbb{P}(||\nabla\Phi_G(\mathbf{x})||_2^2 \geq \epsilon) \to \frac{b}{\epsilon}.$$

Therefore, if we can find a pair a, b such that $a > 0$ and $b = 0$ in assumption 4.2 then SMGD converges almost surely to a Pareto stationary point. ∎

Then, even in the case where the bias does not disappear, the SMGD algorithm can be shown to converge to a point on the Pareto set. However, the bias still has an effect. It is straightforward to show that if there is a point, $\mathbf{x}$ where $\mathbb{E}[\nabla_x^{\mathcal{C}}\{F\}(\mathbf{x}, \mathbf{W}(\theta))] = \Phi_G(()\mathbf{x}) = \mathbf{0}$ and $b = 0$ then SMGD converges to that point.

THEOREM 4.5. *Let F, G, ε_t, a, and b be defined as in Thm. 4.4. If there is an $a > 0$ and $b = 0$ meeting the requirements of assumption 4.2 and at least one point, $\mathbf{x}_* \in \mathcal{X}^* := \{\mathbf{x} | \nabla\Phi_F(\mathbf{x}) = \nabla\Phi_G(\mathbf{x}) = \mathbf{0}\}$, then SMGD converges to a point in $\mathcal{X}^*$.*

Let $d_t = \mathbf{x}_t - \mathbf{x}^*$ for some $\mathbf{x}^* \in \mathcal{X}^*$ and $B_t = \nabla\Phi_F(\mathbf{x}_t) - \nabla\Phi_G(\mathbf{x}_t)$. To prove Thm. 4.5 it is enough to show that $\langle B_t, d_t\rangle$ goes to zero. Since $B_t \perp d_t$ on the Pareto set implies that $\mathcal{X}^* = \emptyset$ there are two remaining cases. Either $B_t = 0$, indicating that $\nabla\Phi_F(\mathbf{x}_t) = \nabla\Phi_G(\mathbf{x}_t) \to 0$ and that $\mathbf{x}_t \in \mathcal{X}^*$, or $d_t = 0$ in which case $\mathbf{x}_t = \mathbf{x}_*$.

Proof. We start from the inequality

$$d_{t+1}^2 \leq d_t^2 - 2\varepsilon_t\langle\nabla_x^{\mathcal{C}}\{F\}(\mathbf{x}_t, \mathbf{W}(\theta)_t), d_t\rangle + \varepsilon_t^2||\nabla_x^{\mathcal{C}}\{F\}(\mathbf{x}_t, \mathbf{W}(\theta)_t||_2^2.$$

Then, using our definition of B_t, we have the expression

$$d_{t+1}^2 \leq d_t^2 - 2\varepsilon_t\langle B_t, d_t\rangle + \varepsilon_t^2(M_v + 1)||\nabla\Phi_G(\mathbf{x}_t)||_2^2 + \varepsilon_t^2 M_0.$$

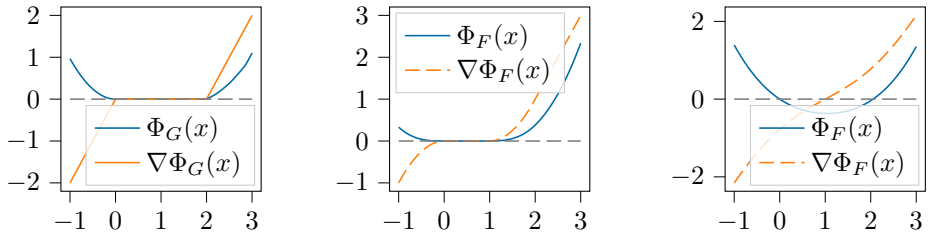
Rearranging and summing we see that

$$\sum_t^T 2\varepsilon_t\langle B_t, d_t\rangle \leq d_0^2 - d_T^2 + (M_v + 1)\sum_t^T \varepsilon_t^2||\nabla\Phi_G(\mathbf{x}_t)||_2^2 + M_0\sum_t^T \varepsilon_t^2. \tag{4.9}$$

As before, defining $E_T = \sum_t^T \varepsilon_t$, we can use Markov's inequality to show

$$\mathbb{P}_T(\langle B_t, d_t\rangle \geq \epsilon) \leq \frac{d_0^2 - d_T^2}{2\epsilon E_T} + (M_v + 1)\frac{\sum_t^T \varepsilon_t^2||\nabla\Phi_G(\mathbf{x}_t)||_2^2}{2\epsilon E_t} + M_0\frac{\sum_t^T \varepsilon_t^2}{2\epsilon E_t}. \tag{4.10}$$

Since $||\nabla\Phi[G](\mathbf{x})||_2^2 \to 0$ almost surely, we can conclude that the left hand side of Eq. 4.10 goes to zero. ∎



(a) Deterministic gradient (b) $\mathbf{W}(\theta) \sim \mathcal{U}[-1, 1]$ (c) $\mathbf{W}(\theta) \sim \mathcal{N}(0, 1)$

Fig. 3: Comparison of Potential Functions with $f_1(\mathbf{x}, \mathbf{W}(\theta)) = (x - 2 + \mathbf{W}(\theta))^2$ and $f_2(\mathbf{x}, \mathbf{W}(\theta)) = (x + \mathbf{W}(\theta))^2$. SMGDA converges to a point where $\mathbb{E}[\nabla\Phi_F] - \nabla\Phi_G = 0$. Deterministic gradient calculated as $\nabla g_i(\mathbf{x}) = \mathbb{E}[f_i(\mathbf{x}, \mathbf{W}(\theta))]$.

4.2. The Bias of the Stochastic Multigradient. Why is the stochastic multigradient biased when the stochastic gradients used to compute it are not? By assumption we have access to samples of the jacobian of F, $J_F(\mathbf{x}, \mathbf{W}(\theta))$ which is an unbiased estimator of the Jacobian of G, $J_G(\mathbf{x})$, e.g. $\mathbb{E}[J_F(\mathbf{x}, \mathbf{W}(\theta))] = J_G(\mathbf{x})$.

We look to the minimizer $\alpha^*(\mathbf{x}, \mathbf{W}(\theta)) = \operatorname{argmin}_{\alpha \in \Delta^{k-1}} \alpha^\top J_F(\mathbf{x}, \mathbf{W}(\theta)) J_F(\mathbf{x}, \mathbf{W}(\theta))^\top \alpha$ and notice that it is a quadratic function of the Jacobian. By Jensen's inequality $\mathbb{E}[J_F(\mathbf{x}, \mathbf{W}(\theta)) J_F(\mathbf{x}, \mathbf{W}(\theta))^\top] \succeq \mathbb{E}[J_F(\mathbf{x}, \mathbf{W}(\theta))]\, \mathbb{E}[J_F(\mathbf{x}, \mathbf{W}(\theta))]^\top$, and so $\mathbb{E}[\alpha^*(\mathbf{x}, \mathbf{W}(\theta))] \neq \alpha^*(\mathbf{x})$.

α^* could be debiased using Monte-Carlo estimation, resampling and averaging the Jacobian and using $\widehat{J}_G(\mathbf{x}) = \frac{1}{N} \sum_{i=1}^N J_F(\mathbf{x}, \mathbf{W}(\theta)_i)$ in place of $J_G(\mathbf{x})$ in Eq. 2.7. However, not only would it be computationally taxing to compute N gradients at each iteration, but there would still be residual variance in our estimation rendering our computational efforts somewhat phyrric. Other approaches attempt to smooth the minimizers α^* using exponential smoothing [11]. This approach effectively reduces the variance of α^* and enforces that the estimator α^* be smooth across iterations. In contrast, our approach to the problem is to mitigate the bias at its source, in the calculation of α^* itself.

5. Proposed Approach: Stochastic Tangential Pareto Dynamics. The SMGD algorithm, even if run from distinct starting points, converges to a subset of the Pareto set. Then, even if completely debiased, in order to estimate the whole Pareto

set one has to restart the algorithm from several points in design space. Not only is this approach not guaranteed to uncover the whole Pareto set but it is computationally taxing. To ameliorate this, we propose to modify the stochastic multigradient descent iterations in two ways. First, we introduce a novel debiasing strategy for the stochastic multigradient. This allows the algorithm to converge to the whole of the Pareto front. Then, we add a noise term which guides the dynamics generated by the algorithm tangentially along the Pareto set, promoting exploration as well as overcoming any residual bias. Because of the added noise term, this approach efficiently explores the area around the Pareto set and is less likely to stay stuck in a saddle point. Furthermore, iterations can be performed with single samples of the quantities of interest and their gradients at each round.

Dubbed stochastic tangential Pareto dynamics (STPD), the points generated by this approach create a dense sample set concentrated on the Pareto set.

5.1. Debiasing the Stochastic Multigradient. We would like to use the deterministic Jacobian matrix, J_G, to calculate the multigradient. This would be equivalent to having direct access to $\mathbb{E}[J_F(\mathbf{x}, \mathbf{W}(\theta))]\,\mathbb{E}[J_F(\mathbf{x}, \mathbf{W}(\theta))]^\top$. However, we only have access to samples of the matrix $J_F(\mathbf{x}, \mathbf{W}(\theta))J_F(\mathbf{x}, \mathbf{W}(\theta))^\top$. We notice that the two quantities are related,

(5.1)
$$\mathbb{E}[J_F(\mathbf{x}, \mathbf{W}(\theta))J_F(\mathbf{x}, \mathbf{W}(\theta))^\top] = \Sigma_{J_F(\mathbf{x}, \mathbf{W}(\theta))} + \mathbb{E}[J_F(\mathbf{x}, \mathbf{W}(\theta))]\,\mathbb{E}[J_F(\mathbf{x}, \mathbf{W}(\theta))]^\top,$$

where $\Sigma_{J_F(\mathbf{x}, \mathbf{W}(\theta))}$ denotes the covariance matrix of the gradients of F.

We propose to debias the stochastic multigradient using exponential smoothing. We first compute an *online* estimate of $\Sigma_{J_F(\mathbf{x}, \mathbf{W}(\theta))}$ using the recursive estimate

$$\widehat{\mu}_{t+1} = \gamma_t \widehat{\mu}_t + (1-\gamma_t) J_f(\mathbf{x}_t, \mathbf{W}(\theta)_t) \tag{5.2}$$

$$\widehat{\Sigma}_{J_F, t+1} = \gamma_t \widehat{\Sigma}_{J_F, t} + (1-\gamma_t)\big(J_F(\mathbf{x}_t, \mathbf{W}(\theta)_t) J_F(\mathbf{x}_t, \mathbf{W}(\theta)_t)^\top - \widehat{\mu}_t \widehat{\mu}_t^\top\big) \tag{5.3}$$

using individual samples of $J_F(\mathbf{x}, \mathbf{W}(\theta))$ and $\gamma_t \in (0, 1)$. We then calculate α^* using

the following modified formulation.

$$\widehat{\alpha}^*(\mathbf{x}) = \underset{\alpha \in \Delta^{k-1}}{\operatorname{argmin}} \, \alpha^\top \big(J_F(\mathbf{x}, \mathbf{W}(\theta)) J_F(\mathbf{x}, \mathbf{W}(\theta))^\top - \widehat{\Sigma}_{J_F(\mathbf{x}, \mathbf{W}(\theta))} \big) \alpha. \tag{5.4}$$

We can then calculate a debiased form of the stochastic gradient

$$\widehat{\nabla_x^{\mathcal{C}}\{F\}}(\mathbf{x}, \mathbf{W}(\theta)) := \sum_i \widehat{\alpha}_i^* \nabla f_i(\mathbf{x}, \mathbf{W}(\theta)) \tag{5.5}$$

This approach has the advantage of not requiring excess computation in individual rounds and places less assumptions on the required behavior of α^* while also effectively reducing the bias in the calculation of $\nabla_x^{\mathcal{C}}\{F\}(\mathbf{x}, \mathbf{W}(\theta))$.

5.2. Stochastic Tangential Pareto Dynamics. To find and explore the Pareto set we generate a sequence of points using the recurrence relation,

$$\mathbf{x}_{t+1} = \mathbf{x}_t - \varepsilon_t \widehat{\nabla_x^{\mathcal{C}}\{F\}}(\mathbf{x}_t) + \sqrt{2\varepsilon_t \beta_t} \Pi_t Z, \tag{5.6}$$

with $\widehat{\nabla_x^{\mathcal{C}}\{F\}}(\mathbf{x}_t)$ a direction of descent, $Z \sim \mathcal{N}(0, \mathbb{1}_{d \times d})$, Π_t a projection operator that projects the noise perpendicular to the direction of descent, ε_t a stepsize parameter, and β_t the scale of the noise. The subscript t in this case should be understood as being adapted to the filtration generated by $\mathbf{x}_t$. This algorithm can be intuitively understood as gradient descent with an additive exploratory noise term. To keep the samples from converging to a limiting distribution centered on a single point we set the stepsize parameter, ε_t, and the moving average parameter, γ_t as a pair of nonincreasing but *strictly positive* numbers:

$$\begin{aligned} \{\varepsilon_t\}_{t=0,\ldots,\infty} &:= \varepsilon_0 \geq \varepsilon_1 \geq \ldots \geq \varepsilon_\infty > 0, \\ \{\gamma_t\}_{t=0,\ldots,\infty} &:= \gamma_0 \geq \gamma_1 \geq \ldots \geq \gamma_\infty > 0. \end{aligned} \tag{5.7}$$

The moving average parameter, as we see below, is used in the estimation of the debiased stochastic multigradient and its expected value.

Ideally, we would have access to true values of the gradients for each objective.

It would then be possible to use the deterministic multigradient, $\nabla_x^{\mathcal{C}}\{G\}(\mathbf{x})$, as an unbiased direction of descent. However, having only samples of the gradients of each objective, we approximate the direction of descent. Here, the expected value of the debaised stochastic multigradient as given in Eq. 5.5 defines the gradient of the potential, $\nabla\Phi_F(\mathbf{x})$. Since we do not have access to its expected value, we approximate it using single samples.

The performance of the estimation of the debiased stochastic multigradient depends on the noise in the stochastic gradients and the strength of the additive noise term. If the noise in the individual stochastic gradients is low then exponential smoothing with a large value of γ_∞ will yield a good approximation of the gradient. As either the strength of the additive noise term or the noise in the gradients increases the autocorrelation between iterates falls and the exponential smoothing procedure breaks down. This is viewed with more detial in our companion paper.

Π_t removes the component of the additive noise that lies parallel to the multigradient. On the Pareto set its value is unity and the algorithm produces iterates that move randomly. Near the Pareto set, the direction orthogonal to the multigradient lies tangential to the Pareto set. The tangential motion of the iterates generated by STPD can then be understood as a tangential drift along the Pareto set. If Π_t was set permanently to unity then this approach would degrade into a type of stochastic Langevin dynamics. This would be unsatisfactory, however, as the variance of samples orthogonal to the Pareto set would be much higher.

To calculate Π_t we need access to the expected value of the direction of descent. However, we do not in general have access to an analytic expression for the expected value of the debiased multigradient. Therefore it must be estimated during the course of optimization. This estimation is also biased due to the nonzero stepsize.

At each iteration we update the expected value of the stochastic multigradient

$$\overline{\nabla_x^{\mathcal{C}}\{F\}}_{t+1} = \gamma_t\overline{\nabla_x^{\mathcal{C}}\{F\}}_t + (1-\gamma_t)\widehat{\nabla_x^{\mathcal{C}}\{F\}}(\mathbf{x}_t). \tag{5.8}$$

To calculate the projection operator we use the exponentially smoothed debiased

stochastic multigradient,

$$\Pi_{t,i,j} = \delta_{i,j} - \frac{1}{||\overline{\nabla_x^{\mathcal{C}}\{F\}}_t||^2}\overline{\nabla_x^{\mathcal{C}}\{F\}}_{t,i}\overline{\nabla_x^{\mathcal{C}}\{F\}}_{t,j}, \tag{5.9}$$

where $\delta_{i,j}$ denotes the dirac delta function.

The value of β_t determines the relative strength of the additive noise compared to the direction of descent and can be viewed as a type of temperature parameter. We only require that β_t be a positive upper bounded function adapted to the filtration generated by $\mathbf{x}_t$. It must be positive since a value of zero would remove the noise term, and it must be upper bounded in order to allow for convergence. We discuss possible function choices in Section 5.4. The effect of the relative strength of the noise term is studied numerically in our companion work.

This entire process is summarized in algorithm 5.1.

Algorithm 5.1 Summary of Stochastic Tangential Pareto Dynamics

Input $\{\varepsilon_t\}, \{\gamma_t\}, \mathbf{x}_0$.
Initialize $\widehat{\mu}_0 = \mathbf{0}$, $\widehat{\Sigma}_{J_F,0} = \mathbb{1}_{d\times d}$, $t = 1$, $\overline{\nabla_x^{\mathcal{C}}\{F\}}_0 = 0$.
while Running **do**
  Query Stochastic Oracle for $J_F(\mathbf{x}_t, \mathbf{W}(\theta)_t)$ ▷ Referred to as J_{F_t} below.
  $\widehat{\mu}_t \leftarrow \gamma_t\widehat{\mu}_{t-1} + (1-\gamma_t)J_{F,t}$
  $\widehat{\Sigma}_{J_F,t} \leftarrow \gamma_t\widehat{\Sigma}_{J_F,t-1} + (1-\gamma_t)(J_{F,t} - \widehat{\mu}_t)(J_{F,t} - \widehat{\mu}_t)^\top$
  $\widehat{\alpha}^* \leftarrow \text{argmin}_{\alpha\in\Delta^{k-1}}\, \alpha^\top\big(J_{F_t}J_{F_t}^\top - \widehat{\Sigma}_{J_F,t}\big)\alpha$ ▷ Eq. **??**
  $\widehat{\overline{\nabla_x^{\mathcal{C}}\{F\}}}_t \leftarrow J_{F_t}^\top\widehat{\alpha}^*$ ▷ Eq. **??**
  $\overline{\nabla_x^{\mathcal{C}}\{F\}}_t \leftarrow \gamma_t\overline{\nabla_x^{\mathcal{C}}\{F\}}_{t-1} + (1-\gamma_t)\widehat{\nabla_x^{\mathcal{C}}\{F\}}_t$ ▷ Eq. 5.8
  $\Pi_t \leftarrow \mathbb{1}_{d\times d} - \frac{\overline{\nabla_x^{\mathcal{C}}\{F\}}_t\overline{\nabla_x^{\mathcal{C}}\{F\}}_t^\top}{||\overline{\nabla_x^{\mathcal{C}}\{F\}}_t||_2^2}\overline{\nabla_x^{\mathcal{C}}\{F\}}_t\overline{\nabla_x^{\mathcal{C}}\{F\}}_t^\top$ ▷ Eq. 5.9
  $Z \sim \mathcal{N}(\mathbf{0}, \mathbb{1}_{d\times d})$
  $\mathbf{x}_{t+1} \leftarrow \mathbf{x}_t - \varepsilon_t\widehat{\nabla_x^{\mathcal{C}}\{F\}}_t + \sqrt{2\varepsilon_t\beta_t}\Pi_t Z$ ▷ Eq. 5.6
  $t \leftarrow t+1$
end while

5.3. Stochastic Tangential Pareto Dynamics Samples the Whole Pareto Front. Here we show that, given enough time, stochastic tangential Pareto dynamics samples arbitrarily close to all points on the Pareto set.

In addition to assumption 4.2 and L-Lipschitz smoothness of Φ_G, we assume coerciveness of Φ_G.

ASSUMPTION 5.1. $\exists c_a, c_b \in \mathbb{R}_+$, $\forall \mathbf{x}$, *such that*

$$||\nabla \Phi_G(\mathbf{x})||^2 \geq c_a \Phi_G(\mathbf{x}) - c_b, \qquad ||\mathbf{x}||^2 \leq c_a \Phi_G(\mathbf{x}) + c_b \quad . \tag{5.10}$$

This assumption is coercive in the sense that the strength the gradient increases the further away $\mathbf{x}$ is from the minima, "pushing" iterates back towards the minima as the value of the scalar potential gets larger. Assumption 5.1 is necessary to show the ergodicity of the SDE $dX = \nabla F(X) + dW$, and so it is appropriate in this context as well [10, 8, 2].

We can now show the main theorem.

THEOREM 5.2. *Given a set of constants* $\{\varepsilon_t\}_{t\in\mathbb{N}}$, *assuming* $\Phi_G : \mathbb{R}^d \mapsto \mathbb{R}$ *is L-Lipschitz smooth and obeys conditions 5.1 and 4.2 almost everywhere, and assuming lower bound* $B := L\beta^{-1}d + b + \frac{\varepsilon_0 L M_0}{2} + d + c_b \leq c_1 \inf \Phi_G(\mathbf{x})$, *then for any* $\mathbf{y}^* \in P^{-1}$ *and* $\epsilon, p > 0$

$$\mathbb{P}(||\mathbf{x}_t - \mathbf{y}^*|| \leq \epsilon \text{ for some } t < \infty) \geq 1 - p. \tag{5.11}$$

Our framework for proving Theorem 5.2 for STPD follows closely from [8, 2, 10]. We first show recurrence, that there is a compact sublevel set, $\Phi_G(\mathbf{x}) \leq M$, that is reached infinitely many times by STPD. We then show reachability, on a compact sublevel set, there is a nonzero probability to reach any target point $\mathbf{y}^* \in \mathcal{P}$. Once we have established both recurrence and reachability, it is straightforward to see that STPD reaches an arbitrary point $\mathbf{y}^*$ in finite time with probability arbitrarily close to 1. Since $\mathbf{y}^*$ is arbitrary, it follows that it reaches the whole Pareto set.

First we show recurrence.

LEMMA 5.3. *Recurrent visits to the sublevel set* M.
Assume that Φ_G *is L-Lipschitz smooth, obeys assumptions 5.1 and 4.2, and let* $\{\varepsilon_t\}_{t\in\mathbb{N}}$ *be a nonincreasing sequence such that* $\varepsilon_0 \geq \varepsilon_1, ..., \geq \varepsilon_\infty > 0$. *Let* $M > 0$ *and define the constant* $B := L\beta^{-1}d + b + \frac{\varepsilon_0 L M_0}{2} + c_b \leq c_a \inf \Phi_G(\mathbf{x})$. *Given a sequence of stopping*

times $\tau_{k+1} = \inf\{t : t > \tau_k, \Phi_G(\mathbf{x}_t) \leq M\}$ *then*

$$a) \quad \tau_0 = \frac{\log\left[\left[\frac{\Phi_G(\mathbf{x}_0)}{M - \frac{\varepsilon_0 B}{\varepsilon_\infty a}}\right]^2\right]}{2c_a\varepsilon_\infty} \tag{5.12}$$

and

$$b) \quad \mathbb{E}[\tau_j] = \tau_0 + (j+1)M \tag{5.13}$$

For proof see the appendix.

We then have reachability.

LEMMA 5.4 (Reachability). *Given that* $\Phi_G(\mathbf{x}_{\tau_k}) \leq M$, *let* $\mathbb{E}[\nabla\Phi_F(\mathbf{x})] \leq \mathbf{D}$ *and* $c_t > 0$. *We have, for* $\mathbf{y}^* \in \{\mathbf{y} : \Phi_G(\mathbf{y}) \leq M\} \cap \{\mathbf{y} : \mathbf{y} \in P^{-1}\}$

$$\mathbb{P}(||\mathbf{x}_{\tau_j+t} - \mathbf{y}^*||^2 \leq \epsilon) = c_t > 0 \tag{5.14}$$

For proof see the appendix.

Having shown both recurrance and reachability established in lemmas 5.3 and 5.4, we now prove Theorem 5.2

Proof. First, define a sequence of stopping times and a $\tau_{j+1} = \inf\{t : t > \tau_j, \Phi_G(\mathbf{x}_{\tau_j+1}) \leq M\}$. Then, also define a stopping time $\tau^* = \inf t : ||\mathbf{x}_t - \mathbf{y}^*|| \leq \epsilon$ for $\epsilon > 0$. We have that

$$\begin{aligned}
\mathbb{P}(\tau^* \geq T) =& \mathbb{P}(\tau^* \geq T, \tau_j \geq T) + \mathbb{P}(\tau^* \geq T, \tau_j \leq T) \\
&= \mathbb{P}(\tau_j \geq T) + \mathbb{P}(\tau^* \geq T, ||\mathbf{x}_{\tau_k} - \mathbf{y}^*|| \geq \epsilon \; \forall k = \{1, ..., j\}) \\
&\leq \mathbb{E}[\tau_j] + \prod_{k=1}^{j} \mathbb{P}(||\mathbf{x}_{\tau_k} - \mathbf{y}^*|| \geq \epsilon) && \text{(Lem. 5.3)} \\
&\leq \frac{(j+1)M + \tau_0}{T} + \prod_{k=1}^{j}(1 - c_j) && \text{(Lem. 5.4)} \\
&\leq \frac{(j+1)M + \tau_0}{T} + (1 - c^*)^j.
\end{aligned}$$

Where we have used the boundedness of c_j in the final line. Setting $\frac{(j+1)M+\tau_0}{T} = \alpha p$ and $(1-c^*)^j = (1-\alpha)p$ for $\alpha \in (0,1)$ and solving for j and T we see that setting

$$j \geq \frac{\ln((1-\alpha)p)}{1-c^*} \qquad \text{(and)} \qquad T \geq \frac{\left(\frac{\ln((1-\alpha)p)}{1-c^*}+1\right)M+\tau_0}{\alpha p}$$

gives us $\mathbb{P}(\tau^* \geq T) \leq p$. Since this is true for any p,

$$\mathbb{P}(\tau^* \leq T) \geq 1-p,$$

which can be made arbitrarily close to 1. ∎

5.4. Discussion. Here we consider implementation choices, the effect of non-stationarity of the sample distribution, parameter settings, and dealing with saddle points and nonconvexity.

We have presented here a basic version of our approach with the aim of imparting understanding about the general working of the algorithm. There are other techniques which may be included in the algorithm. They do not affect our theoretical results and may improve the sampling properties of the algorithm in practice.

We first consider variance reduction. Variance reduction techniques can be used on both $\nabla f_i(\mathbf{x}, \mathbf{W}(\theta))$ and the debiased stochastic multigradient $\widehat{\nabla_x^{\mathcal{C}}\{F\}}(\mathbf{x})$. Reducing the variance of the objective gradients would further decrease the bias of the calculation of $\widehat{\nabla_x^{\mathcal{C}}\{F\}}(\mathbf{x})$. It would also cause the points generated by the algorithm to be more concentrated on the Pareto set. Common approaches include the popular SAGA, SVRG, and minibatching [1]. The increased precision comes at a cost, for a given stepsize lower gradient variance will also decrease the speed at which the algorithm diffuses along the Pareto set. The use of variance reduction is ultimately problem dependent and will affect parameter choices for stepsize and additive noise strength.

It is also possible to include preconditioners in our approach, i.e. take steps

$$\mathbf{x}_{t+1} = \mathbf{x}_t - \varepsilon_t H_t \Phi_F(\mathbf{x}) + \sqrt{2\beta_t \varepsilon_t H_t}\Pi_t Z \tag{5.15}$$

where H_t is a positive definite matrix. There are many options for the preconditioner, as discussed in the context of single objective stochastic Langevin dynamics in [7].

The samples generated by STPD do not reach a stationary distribution. This is because of the nonvanishing stepsize. As a result the estimates for the jacobian, its covariance, and the debiased stochastic multigradient do not converge. This appears in the algorithm as a bias in both the direction of descent and the projection operator. This bias is ultimately overcome by the variation introduced by the additive noise.

The bias in the direction of descent lies tangential to the Pareto set. Clearly, immediately projecting normal to this direction would be counterproductive. Averaging over the history of the direction of descent allows us to estimate a true tangential direction. This average, however, still contains a trace of the bias, causing the projected noise to have a component normal to the Pareto set. This manifests as an increased residual variance of the samples around the Preto set.

Projecting normal to the average history of the direction of descent is not the only way to calculate the projection operator. It is also possible to average the history of the directions perpendicular to the direction of descent, e.g. calculate Π_t as:

$$\Pi_{t+1} = \gamma_t \Pi_t + \big(\mathbb{1} - \frac{1}{||\widehat{\nabla_x^{\mathcal{C}}\{F\}}(\mathbf{x}_t)||_2^2} \widehat{\nabla_x^{\mathcal{C}}\{F\}}(\mathbf{x}_t) \widehat{\nabla_x^{\mathcal{C}}\{F\}}(\mathbf{x}_t)^\top\big)$$

We have chosen exponential smoothing to model several quantities. It has the advantage of requiring little storage and being easy to calculate. However, its effectiveness depends strongly on the correlation time of the sample gradients. The correlation time, in turn, is influenced by the noise of the stochastic gradients along the Pareto set and the relative strength of the additive noise term. These interactions are not straightforward. More complex models can be used in the place of simple exponential smoothing. However, they will similarly show a dependence on the correlation time of the sample gradients and must be calibrated. This can be done simply by minimizing the predictive error of the estimation of a quantity of interest, e.g. the Jacobian.

It has been previously mentioned that the additive noise term β_t need only be a

positive upper bounded function adapted to the filtration generated by $\mathbf{x}_t$. This leaves flexibility in its choice beyond a constant function. Since the additive noise term may interfere with initial convergence of the iterates to the Pareto set it might be beneficial to damp it when iterates are far from the Pareto set. The strength of the additive noise can be damped far from the Pareto set by selecting $\beta_t = Ce^{-\overline{\nabla_x^c\{F\}}^\top A\overline{\nabla_x^c\{F\}}+D}$ with $C, D \in \mathbb{R}^{d\times d}$, $A \in \mathbb{R}_+^{d\times d}$. The dynamics generated by STPD are also strongly affected by the relative strength of the additive noise term. This suggests a scaling to augment the strength of the additive noise in proportion to the variation of the objective functions by selecting $\beta_t = \sum_i \mathrm{VAR}(\nabla f_i(\mathbf{x}_t, \mathbf{W}(\theta)_t))$. Vector valued functions can also be considered.

We take a moment to discuss saddle points in the scalar potential. We have proven that the iterates generated by STPD cover the whole Pareto set under mild assumptions. This does not preclude cases with non-convex objectives. It has been shown in the single objective case that adding Gaussian noise to stochastic gradients can allow gradient descent algorithms to escape shallow second order minima [2]. In the multi-objective case, local minima in the objective functions correspond to saddle points in the scalar potential which, if mild, may be similarly escaped. However, a potential with a split Pareto set (a 'W' shaped potential) may lead to complications. Either multiple restarts or an alternative approach more suited to nonconvex optimization may be more effective for such situations.

Diffusing across the Pareto set is not the only way to exploit the stochastic multi-gradient. An alternative approach would be to select an ensemble of random points in design space and optimize them individually using debiased SMGDA. This approach may be more effective on non-convex problems and problems with discontinuous Pareto sets. However, it has two corresponding drawbacks. One is the loss of information. Two points from distinct regions in design space may converge to similar locations on the Pareto set, making one redundant. Another drawback is resolution. Picking an ensemble of points pre-determines the resolution at which one can determine the Pareto set. This has a similar flavor to picking the number of iterations of STPD. The computational cost, however, is not the same. STPD scales linearly with

the number of samples taken, whereas adding another point in an ensemble algorithm scales multiplicatively with the number of iterations used.

6. Conclusion. We have introduced a new approach to solving multi-objective optimization problems using noise added debiased stochastic multigradients, in which the whole of the Pareto front is of interest, stochastic tangential Pareto dynamics. Along the way, we have reformulated our multi-objective optimization problem as the minimization of an equivalent scalar potential function. Using this reformulation, we have presented alternative proofs of convergence for the MGD algorithm along with new proofs of convergence for the SMGD algorithm; showing that, despite bias, the SMGD algorithm converges to a Pareto stationary point in the convex, strongly convex, and nonconvex L-Lipshitz smooth case. We have also shown that, when there are points on the Pareto set where the stochastic multigradient is unbiased, that SMGDA converges to one of those points. Finally, we have shown that STPD provably generates samples arbitrarily close to the whole of the Pareto set. Our approach produces a noisy snapshot of the Pareto front and reducing this variance through alternative gradient estimators and/or second order algorithms remains a promising avenue of future research.

A complementary line of inquiry involves characterizing the Pareto front, and set, from a fixed set of samples. In our companion paper we give a probabilistic characterization of the Pareto set across the whole of design space, giving a straightforward recipe to not only infer the Pareto set from STPD but to postprocess the results of any stochastic sampling algorithm that yields samples concentrated along the Pareto set [CITATION NEEDED]. Taken in conjunction with STPD this yields a straightforward way to estimate all likely members of the Pareto set of a stochastic program.

Acknowledgments. We would like to acknowledge ourselves for being spectacular.

REFERENCES

[1] L. Bottou, F. E. Curtis, and J. Nocedal, *Optimization methods for large-scale machine learning*, 2018, https://arxiv.org/abs/1606.04838.

[2] X. Chen, S. S. Du, and X. T. Tong, *On stationary-point hitting time and ergodicity of stochastic gradient langevin dynamics*, 2020, https://arxiv.org/abs/1904.13016.

[3] M. Frank and P. Wolfe, *An algorithm for quadratic programming*, Naval Research Logistics Quarterly, 3 (1956), pp. 95–110, https://api.semanticscholar.org/CorpusID:122654717.

[4] Z. Hu, K. Shaloudegi, G. Zhang, and Y. Yu, *Federated learning meets multi-objective optimization*, 2023, https://arxiv.org/abs/2006.11489, https://arxiv.org/abs/2006.11489.

[5] S. Liu and L. N. Vicente, *The stochastic multi-gradient algorithm for multi-objective optimization and its application to supervised machine learning*, 2021, https://arxiv.org/abs/1907.04472.

[6] M. M. Mäkelä, V.-P. Eronen, and N. Karmitsa, *On Nonsmooth Multiobjective Optimality Conditions with Generalized Convexities*, Springer New York, New York, NY, 2014, pp. 333–357, https://doi.org/10.1007/978-1-4939-0808-0_17, https://doi.org/10.1007/978-1-4939-0808-0_17.

[7] G. Marceau-Caron and Y. Ollivier, *Natural langevin dynamics for neural networks*, in Geometric Science of Information, F. Nielsen and F. Barbaresco, eds., Cham, 2017, Springer International Publishing, pp. 451–459.

[8] J. Mattingly, A. Stuart, and D. Higham, *Ergodicity for sdes and approximations: locally lipschitz vector fields and degenerate noise*, Stochastic Processes and their Applications, 101 (2002), pp. 185–232, https://doi.org/https://doi.org/10.1016/S0304-4149(02)00150-3, https://www.sciencedirect.com/science/article/pii/S0304414902001503.

[9] Q. Mercier, *Optimisation multicritère sous incertitudes : un algorithme de descente stochastique*, theses, COMUE Université Côte d'Azur (2015 - 2019), Oct. 2018, https://theses.hal.science/tel-02063322.

[10] S. Meyn and R. Tweedie, *Markov Chains and Stochastic Stability*, vol. 92, 01 1993, https://doi.org/10.2307/2965732.

[11] S. Zhou, W. Zhang, J. Jiang, W. Zhong, J. GU, and W. Zhu, *On the convergence of stochastic multi-objective gradient manipulation and beyond*, in Advances in Neural Information Processing Systems, S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, eds., vol. 35, Curran Associates, Inc., 2022, pp. 38103–38115, https://proceedings.neurips.cc/paper_files/paper/2022/file/f91bd64a3620aad8e70a27ad9cb3ca57-Paper-Conference.pdf.

Appendix A. Proofs of Recurrence and Reachability of STPD. In order to show that STPD samples the whole of the Pareto set we must show both recurrence and reachability. Recurrence shows that the algorithm iterates reach a sublevel set

upper bounded by M an infinite number of times. Reachability shows that, having reached the sublevel set upper bounded M, there is a nonzero probability of reaching any point on the Pareto set inside of that sublevel set. We begin with the proof of recurrence.

LEMMA. *Recurrent visits to the sublevel set M.*
Assume that Φ_G is L-Lipschitz smooth, obeys assumptions 5.1 and 4.2, and let $\{\varepsilon_t\}_{t\in\mathbb{N}}$ be a nonincreasing sequence such that $\varepsilon_0 \geq \varepsilon_1, ..., \geq \varepsilon_\infty > 0$. Let $M > 0$ and define the constant $B := L\max(\beta_t)d + b + \frac{\varepsilon_0 L M_0}{2} + c_b \leq c_a \inf \Phi_G(\mathbf{x})$. Given a sequence of stopping times $\tau_{k+1} = \inf\{t : t > \tau_k, \Phi_G(\mathbf{x}_t) \leq M\}$ then

$$a) \quad \tau_0 = \frac{\log\left[\left[\frac{\Phi_G(\mathbf{x}_0)}{M - \frac{\varepsilon_0 B}{\varepsilon_\infty a}}\right]^2\right]}{2c_a\varepsilon_\infty}$$

and

$$b) \quad \mathbb{E}[\tau_j] = \tau_0 + (j+1)M$$

Proof of lemma 5.12(a) (Recurrence). Let $\zeta_t = \langle \nabla\Phi_G(\mathbf{x}_t), \nabla\Phi_F(\mathbf{x}_t) - \nabla\Phi_G(\mathbf{x}_t)\rangle$. We start from L-Liptshitz smoothness of Φ_G.

$$\Phi_G(\mathbf{x}_{t+1}) \leq \Phi_G(\mathbf{x}_t) + \langle \nabla\Phi_G(\mathbf{x}_t), \mathbf{x}_{t+1} - \mathbf{x}_t\rangle + \frac{L}{2}||\mathbf{x}_{t+1} - \mathbf{x}_t||^2.$$

Substituting in $\nabla_x^{\mathcal{C}}\{F\}(\mathbf{x})$, adding and subtracting $\varepsilon_t||\Phi_G(\mathbf{x}_t)||^2$ and using 4.7,

$$\begin{aligned} &\leq \Phi_G(\mathbf{x}_t) - \varepsilon_t||\nabla\Phi_G(\mathbf{x}_t)||^2 - \varepsilon_t\zeta_t + \frac{L}{2}\left(M_{V_0} + M_V||\nabla\Phi_G(\mathbf{x}_t)||^2\right) \\ &\quad + \sqrt{2\varepsilon_t\beta_t}\langle\nabla\Phi_G(\mathbf{x}_t), \Pi_t Z\rangle + \frac{L}{2}\left(2\sqrt{2\varepsilon_t\beta_t}\langle\nabla\Phi_F(\mathbf{x}_t), \Pi_t Z\rangle + 2\varepsilon_t\beta_t\Pi_t\Pi_t^\top\langle Z, Z\rangle\right). \end{aligned}$$

Taking the conditional expectation and simplifying,

$$\begin{aligned}\mathbb{E}_t[\Phi_G(\mathbf{x}_{t+1})|\mathbf{x}_t] &\leq \Phi_G(\mathbf{x}_t) - \varepsilon_t(a - \frac{\varepsilon_t L(M_v+1)}{2})||\nabla\Phi_G(\mathbf{x}_t)||^2 + \varepsilon_t\big(b + \frac{L\varepsilon_t}{2}M_0 + L\beta_t d\big)\\ &\leq \Phi_G(\mathbf{x}_t) - \varepsilon_t||\nabla\Phi_G(\mathbf{x}_t)||^2 + \varepsilon_t\big(b + \frac{L\varepsilon_t}{2}M_0 + L\beta_t d\big)\\ &\leq (1 - c_a\varepsilon_t)\Phi_G(\mathbf{x}_t) + \varepsilon_t B\\ &\leq e^{-c_a\varepsilon_t}\Phi_G(\mathbf{x}_t) + \varepsilon_t B.\end{aligned}$$

Where we have used the inequality $1 - x \leq e^{-x}$ and enforced the requirement that $\varepsilon_0 \leq \frac{2a}{L(M_V+1)}$. Taking the full expectation, iterating, and setting the above less than equal to M gives the relation,

$$\mathbb{E}[\Phi_G(\mathbf{x}_{\tau_0})] \leq e^{-c_a\tau_0\varepsilon_\infty}\Phi_G(\mathbf{x}_0) + \frac{\varepsilon_0 B}{\varepsilon_\infty a} \leq M.$$

Solving for τ_0 gives the result. ∎

For the Proof of part b of the lemma, we will first prove that the quantity

$$\Phi_G(\mathbf{x}_{t\wedge\tau_{j+1}}) + t \wedge \tau_{j+1}$$

is a supermartingale with respect to τ_j.

LEMMA A.2.

$$\Phi_G(\mathbf{x}_{t\wedge\tau_{j+1}}) + t \wedge \tau_{j+1}$$

is a supermartingale with respect to τ_j.

Proof. Since $t \wedge \tau_{j+1}$ is a stopping time and a martingale, it suffices to show that

$$\mathbb{E}_{\tau_j}[\Phi_G(\mathbf{x}_{\tau_{j+1}})] \leq \Phi_G(\mathbf{x}_{\tau_j})$$

Note that

$$\mathbb{E}_{\tau_j}[\Phi_G(\mathbf{x}_{\tau_{j+1}})] \leq (1 - a\varepsilon_{\tau_j})\Phi_G(\mathbf{x}_{\tau_j}) + \varepsilon_{\tau_j} B \leq M$$

Is true as we can always pick $a \geq 1$. Since $\Phi_G(\mathbf{x}_{\tau_j}) \geq \inf \Phi_G(()\mathbf{x}) \geq \frac{B}{a}$,

$$\mathbb{E}_{\tau_j}[\Phi_G(\mathbf{x}_{\tau_{j+1}})] \leq \Phi_G(\tau_j)$$

And we have the result. ∎

Having proved lemma A.2 we can go on to prove part b of lemma 5.3

Proof of 5.4(b) (Recurrence). Since

$$\mathbb{E}[\Phi_G(\mathbf{x}_{t\wedge\tau_{j+1}})|\tau_j] + t \wedge \tau_{j+1} \leq \Phi_G(t \wedge \tau_j) + t \wedge \tau_j$$

we can allow $t \to \infty$ to see that

$$\mathbb{E}[\tau_{j+1}|\tau_j] \leq \mathbb{E}[\Phi_G(\mathbf{x}_{\tau_{j+1}})|\tau_j] + \tau_{j+1} \leq \Phi_G(\mathbf{x}_{\tau_j}) + \tau_j$$

Iterating, using the relation $\Phi_G(\mathbf{x}_{\tau_j}) \leq M$, and taking the full expectation we see that

∎

$$\mathbb{E}[\tau_{j+1}] \leq (j+1)M + \tau_0$$

We now prove reachability.

REACHABILITY. *Given that $\Phi_G(\mathbf{x}_{\tau_k}) \leq M$, let $\mathbb{E}[\nabla\Phi_F(\mathbf{x})] \leq \mathbf{D}$ be the componentwise maximum value of the gradient in the sublevel set, and $c_t > 0$. We have, for* $\mathbf{y}^* \in \{\mathbf{y} : \Phi_G(\mathbf{y}) \leq M\} \cap \{\mathbf{y} : \mathbf{y} \in P^{-1}\}$

$$\mathbb{P}(||\mathbf{x}_{\tau_j+t} - \mathbf{y}^*||^2 \leq \epsilon) = c_t > 0$$

Proof of Lemma 5.4 (Reachability). for a finite t we have the events

$$\mathcal{A} = ||\mathbf{x}_{\tau_j+t} - \mathbf{y}^*||^2 \leq \epsilon \quad \mathcal{B} = ||\mathbf{x}_{\tau_j} - \mathbf{y}^* + \sum_{s=\tau_j}^{\tau_j+t-1} \varepsilon_s \mathbf{D} + \sum_{s=\tau_j}^{\tau_j+t-1} \sqrt{2\varepsilon_s\beta_s}\Pi_s Z_s||^2 \leq \epsilon$$

with $Z \sim \mathcal{N}(0, \mathbb{1}_{d\times d})$. Since $t < \infty$ we conclude that $\mathbb{P}(\mathcal{B}) > 0$. From the definition of STPD, we have

$$\begin{aligned}\mathbf{x}_{\tau_j+t} =& \mathbf{x}_{\tau_j} - \sum_{s=\tau_j}^{\tau_j+t-1} \nabla\Phi_F(\mathbf{x}_s) + \sum_{s=\tau_j}^{\tau_j+t-1} \sqrt{2\varepsilon_s\beta_s}\Pi_s Z_s \\ \leq& \mathbf{x}_{\tau_j} + \sum_{s=\tau_j}^{\tau_j+t-1} D + \sum_{s=\tau_j}^{\tau_j+t-1} \sqrt{2\varepsilon_s\beta_s}\Pi_s Z_s\end{aligned}$$

Subtracting $\mathbf{y}^*$ from both sides and taking the norm, we see that

$$||\mathbf{x}_{\tau_j+t} - \mathbf{y}^*||^2 \leq ||\mathbf{x}_{\tau_j} - \mathbf{y}^* + \sum_{s=\tau_j}^{\tau_j+t-1} D + \sum_{s=\tau_j}^{\tau_j+t-1} \sqrt{2\varepsilon_s\beta_t}\Pi_s Z_s||^2$$

Where we can see that on the left hand side we have event $\mathcal{A}$. Since event $\mathcal{A}$ occurs almost surely if event $\mathcal{B}$ does, and since event $\mathcal{B}$ occurs with strictly positive probability, we have that $\mathbb{P}(\mathcal{A}) \geq \mathbb{P}(\mathcal{B}) = c_t > 0$. Where we have labeled $\mathbb{P}(\mathcal{B})$ as c_t for later convenience. ∎

Appendix B. A Note on Assumption 4.2. Here we note briefly how assumption 4.2 can be proven for arbitrary probability. Let $\mathbb{E}[||\nabla_x^{\mathcal{C}}\{F\}(\mathbf{x}) - \nabla_x^{\mathcal{C}}\{G\}(\mathbf{x})||_2^2] := \sigma^2$, $\mathbb{E}[\nabla_x^{\mathcal{C}}\{F\}(x)] := \mu_F$ and $\nabla_x^{\mathcal{C}}\{G\}(\mathbf{x}) := \mu_G$. For $b \geq 0$ we have $\mathbb{P}(a||\mu_G||_2^2 - \mu_F^\top\mu_G \geq b) \leq \frac{\sigma^2}{b^2+\sigma^2} \leq 1$ by Cantelli's concentration inequality. Therefore, $\mathbb{P}(\mu_F^\top\mu_G \geq a||\mu_G||_2^2 - b) \geq 1 - \frac{\sigma^2}{b^2+\sigma^2}$. Since σ^2 is finite, by picking $b^2 \geq \sigma^2\left(\frac{p}{1-p}\right)$ we can always set $\mathbb{P}(\mu_F^\top\mu_G \geq a||\mu_G||_2^2 - b) \geq p$ for any p.

Appendix C. Convergence of SMGDA in the Convex and Strongly Convex Cases. Here show omitted proofs of the convergence of SMGDA in the convex case and in the case where there is at least one strongly convex objective.

Let $\mathbf{x}^* = \text{argmin}_{\mathbf{s}\in P} ||\mathbf{s}-\mathbf{x}||_2^2$, e.g. the projection of $\mathbf{x}$ to the Pareto set, then we have the following three lemmas.

LEMMA C.1. *Given a set of functions G with at least one g_i m_i-strongly convex for $i \in \{1, ..., k\}$. We then have the relation*

$$\langle \mathbb{E}[\nabla_x^{\mathcal{C}}\{F\}(\mathbf{x}, \mathbf{W}(\theta))] - \nabla_x^{\mathcal{C}}\{G\}, \mathbf{x} - \mathbf{x}^* \rangle \geq 0 \tag{C.1}$$

Proof. From the definition of convexity

$$\langle \mathbb{E}[\nabla_x^{\mathcal{C}}\{F\}(\mathbf{x}, \mathbf{W}(\theta)) - \nabla_x^{\mathcal{C}}\{G\}(\mathbf{x})], \mathbf{x} - \mathbf{x}^* \rangle.$$

Adding and subtracting $\sum_i \alpha_i^*(\mathbf{x}, \mathbf{W}(\theta))\nabla g_i(\mathbf{x})$ we see that

$$= \langle \mathbb{E}[\sum_i \alpha_i^*(\mathbf{x}, \mathbf{W}(\theta))\nabla f(\mathbf{x}, \mathbf{W}(\theta)) - \alpha_i^*(\mathbf{x}, \mathbf{W}(\theta))\nabla g_i(\mathbf{x}) + (\alpha_i^*(\mathbf{x}, \mathbf{W}(\theta)) - \alpha_i^*(\mathbf{x}))\nabla g_i(\mathbf{x})], \mathbf{x} - \mathbf{x}^* \rangle,$$

$$= \langle \mathbb{E}[\sum_i (\alpha_i^*(\mathbf{x}, \mathbf{W}(\theta)) - \alpha_i^*(\mathbf{x}))\nabla g_i(\mathbf{x})], \mathbf{x} - \mathbf{x}^* \rangle.$$

Using the (strong) convexity of g_i and optimality of $\alpha^*(\mathbf{x})$ we have the result,

$$\langle \mathbb{E}[\sum_i (\alpha_i^*(\mathbf{x}, \mathbf{W}(\theta)) - \alpha_i^*(\mathbf{x}))\nabla g_i(\mathbf{x})], \mathbf{x} - \mathbf{x}^* \rangle \geq 0. \qquad \blacksquare$$

LEMMA C.2. *Given a set of functions G with at least one g_i strongly convex for $i \in \{1, ..., k\}$. We have the relation*

$$\langle \nabla_x^{\mathcal{C}}\{G\}(\mathbf{x}), \mathbf{x} - \mathbf{x}^* \rangle \geq \sum_i \alpha_i^*(\mathbf{x}) \frac{m_i}{2} ||\mathbf{x} - \mathbf{x}^*||^2 \tag{C.2}$$

Proof.

$$\sum_i \alpha_i^*(\mathbf{x}) \langle \nabla g_i(\mathbf{x}), \mathbf{x} - \mathbf{x}^* \rangle$$

Using the definition of strong convexity,

$$\geq \sum_i \alpha_i^*(\mathbf{x})\big(g_i(\mathbf{x}) - g_i(\mathbf{x}^*) + \frac{m_i}{2}||\mathbf{x} - \mathbf{x}^*||_2^2\big)$$

Using the optimality of $\sum_i \alpha_i^*(\mathbf{x}^*)g_i(\mathbf{x}^*)$ we get the result,

$$\geq \sum_i \frac{\alpha_i m_i}{2}||\mathbf{x} - \mathbf{x}^*||_2^2. \qquad \square$$

LEMMA C.3. *If all functions in the set G are convex, $\Phi_G(\mathbf{x})$ is also convex.*

Proof. If $\Phi_G(\mathbf{x})$ is convex, one has the identity $\langle \nabla\Phi_G(\mathbf{x}) - \nabla\Phi_G(\mathbf{y}), \mathbf{x} - \mathbf{y}\rangle \geq 0$. Starting from the definition of convexity and exploiting two reformulations of the inequality,

$$\langle \nabla\Phi_G(\mathbf{x}) - \nabla\Phi_G(\mathbf{y}), \mathbf{x} - \mathbf{y}\rangle =$$
$$\geq \max\big(\sum_i \langle(\alpha_i^*(\mathbf{x}) - \alpha_i^*(\mathbf{y}))\nabla g_i(\mathbf{y}), \mathbf{x} - \mathbf{y}\rangle, \sum_i \langle(\alpha_i^*(\mathbf{x}) - \alpha_i^*(\mathbf{y}))\nabla g_i(\mathbf{x}), \mathbf{x} - \mathbf{y}\rangle\big). \blacksquare$$

By the convexity of g_i and the optimality of α^* we have

$$\langle \nabla\Phi_G(\mathbf{x}) - \nabla\Phi_G(\mathbf{y}), \mathbf{x} - \mathbf{y}\rangle \geq 0 \qquad \square$$

We can now show that if all functions in G are convex, with at least one strongly convex, that the SMGD algorithm converges to the Pareto front despite the bias.

THEOREM C.4. *Let $g_i(\mathbf{x}) = \mathbb{E}[f_i(\mathbf{x}, \mathbf{W}(\theta))]$ be a collection of convex functions with bounded variance $\mathbb{V}(f(\mathbf{x}, \mathbf{W}(\theta))) \leq M$ and at least one g_i strongly convex. Let $\mathbf{x}_t^* = \text{argmin}_{\mathbf{s}\in P^{-1}} ||s - \mathbf{x}_t||^2$ be the projection of $\mathbf{x}$ to the Pareto set, define $d_t^2 = ||\mathbf{x}_t - \mathbf{x}_t^*||_2^2$ and a sequence ε_t such that $\varepsilon_t \to 0$ as $t \to \infty$ and $\frac{\sum_t \varepsilon_t^2}{\sum_t \varepsilon_t} \to 0$, then,*

$$d_t^2 \xrightarrow[t]{\infty} 0.$$

Proof. Using $\langle \nabla_x^{\mathcal{C}}\{F\}(\mathbf{x}, \mathbf{W}(\theta)) - \nabla_x^{\mathcal{C}}\{G\}(\mathbf{x}), d_t\rangle$ as $\langle B_t, d_t\rangle$. Starting from the

definition of d_t,

$$d_{t+1}^2 \leq ||\mathbf{x}_t - \varepsilon_t \nabla_x^{\mathcal{C}}\{F\}(\mathbf{x}_t) - \mathbf{x}_t^*||_2^2.$$

Expanding, and adding and subtracting $\nabla_x^{\mathcal{C}}\{G\}(\mathbf{x}_t)$,

$$\begin{aligned} &\leq d_t^2 - 2\varepsilon_t \langle \nabla_x^{\mathcal{C}}\{F\}(\mathbf{x}_t, \mathbf{W}(\theta)_t), d_t\rangle + \varepsilon_t^2 ||\nabla_x^{\mathcal{C}}\{F\}(\mathbf{x}_t, \mathbf{W}(\theta)_t)||_2^2 \\ &\leq d_t^2 - 2\varepsilon_t \langle \nabla_x^{\mathcal{C}}\{G\}(\mathbf{x}_t), d_t\rangle + -2\varepsilon_t \langle B_t, d_t\rangle + \varepsilon_t^2 ||\nabla_x^{\mathcal{C}}\{F\}(\mathbf{x}_t, \mathbf{W}(\theta)_t)||_2^2 \end{aligned}$$

Using lemmas C.2 and C.1,

$$\mathbb{E}[d_{t+1}^2] \leq \mathbb{E}[d_t^2] - 2\varepsilon_t \frac{\overline{m}}{2} \mathbb{E}[d_t^2] + \varepsilon_t^2 M.$$

Setting $\pi_t := \prod_{i=0}^{t}(1 - \varepsilon_t \overline{m})$,

$$\mathbb{E}[d_{t+1}^2] \leq \pi_t d_0^2 + \sum_s \frac{\pi_t}{\pi_s} \varepsilon_s^2 M.$$

Requiring $\varepsilon_t \leq \overline{m}\ \forall t$ and taking the limit as $t \to \infty$ gives

$$\mathbb{E}[d_{t+1}^2] \to 0. \qquad \square$$

We can also prove convergence in the convex case (with no objective function being strongly convex).

THEOREM C.5. *Let $g_i(\mathbf{x}) = \mathbb{E}[f_i(\mathbf{x}, \mathbf{W}(\theta))]$ be a collection of convex functions with bounded variance $\mathbb{V}(f(\mathbf{x}, \mathbf{W}(\theta))) \leq M$. With $\mathbf{x}_t^* = \text{argmin}_{\mathbf{s}\in P} ||s - \mathbf{x}_t||^2$ be the projection of $\mathbf{x}$ to the Pareto set, define $d_t^2 = ||\mathbf{x}_t - \mathbf{x}_t^*||_2^2$ and a sequence ε_t such that $\varepsilon_t \to 0$ as $t \to \infty$ and $\frac{\sum_t \varepsilon_t^2}{\sum_t \varepsilon_t} \to 0$, then,*

(C.3) $$\mathbb{E}[\Phi_G(\mathbf{x}_t) - \Phi_G(\mathbf{x}_t^*)] \to 0 \text{ almost surely.}$$

Proof. Starting from the definition of d_t,

$$d_{t+1}^2 \leq d_t^2 - 2\varepsilon_t \langle \nabla \Phi_F(\mathbf{x}_t), \mathbf{x}_{t+1} - \mathbf{x}_t \rangle + \varepsilon_t^2 ||\nabla \Phi_F(\mathbf{x}_t)||_2^2.$$

Using $\langle \nabla_x^{\mathcal{C}}\{F\}(\mathbf{x}, \mathbf{W}(\theta)) - \nabla_x^{\mathcal{C}}\{G\}(\mathbf{x}), d_t \rangle$ as $\langle B_t, d_t \rangle$, taking the expectation, and simplifying yields

$$\mathbb{E}[d_{t+1}^2] \leq \mathbb{E}[d_t^2] - 2\varepsilon_t \mathbb{E}\langle \nabla_x^{\mathcal{C}}\{G\}(\mathbf{x}_t), d_t \rangle + -2\varepsilon_t \langle B_t, d_t \rangle + \varepsilon_t^2 M^2.$$

Using the convexity of $\Phi(_)g$,

$$\leq \mathbb{E}[d_t^2] - 2\varepsilon_t \big(\Phi_G(\mathbf{x}_t) - \Phi_G(\mathbf{x}_t^*)\big) + \varepsilon_t^2 M^2.$$

Rearranging and summing we have,

$$\sum_{t=1}^{T} \varepsilon_t \, \mathbb{E}[\Phi_G(\mathbf{x}_t) - \Phi_G(\mathbf{x}_t^*)] \leq \sum_{t=1}^{T} \frac{\mathbb{E}[d_t^2 - d_{t+1}^2]}{2} + \frac{M^2}{2} \sum_{t=1}^{T} \varepsilon_t^2.$$

Dividing by $E_T = \sum_{t=1}^{T} \varepsilon_t$, we have the relation,

$$\sum_{t=1}^{T} \frac{\varepsilon_t}{E_t} \, \mathbb{E}[\Phi_G(\mathbf{x}_t) - \Phi_G(\mathbf{x}_t^*)] \leq \frac{d_0^2 - d_T^2}{2E_T} + \frac{M^2}{2E_T} \sum_{t=1}^{T} \varepsilon_t^2.$$

Treating $\frac{\varepsilon_t}{E_T}$ as a probability measure, we have by markov's inequality

$$\mathbb{P}_{E_T}(\mathbb{E}[\Phi_G(\mathbf{x}_t) - \Phi_G(\mathbf{x}_t^*)] \geq \epsilon) \leq \frac{d_0^2 - \mathbb{E}[d_T^2]}{2\epsilon E_T} + \frac{M^2}{2\epsilon E_T} \sum_{t=1}^{T} \varepsilon_t^2$$

The left hand side decreases to zero as $T \to \infty$, gives the result,

$$\mathbb{P}_{E_T}(\mathbb{E}[\Phi_G(\mathbf{x}_t) - \Phi_G(\mathbf{x}_t^*)] \geq \epsilon) \to 0.$$

□