Transverse Brownian Motion for Pareto Front Identification

Zachary Jones, Pietro Marco Congedo, Olivier Le Maitre

Transverse Brownian Motion for Pareto Front Identification

Zachary Jones[1], Pietro Marco Congedo[1], and Olivier Le Maître[2]

[1] *Inria, Centre de Mathématiques Appliquées, Ecole polytechnique, IPP, Route de Saclay, 91128 Palaiseau Cedex, France*

[2] *CNRS, Inria, Centre de Mathématiques Appliquées, Ecole polytechnique, IPP, Route de Saclay, 91128 Palaiseau Cedex, France*

January 11, 2024

Abstract

Including uncertainty sources in multi-objective optimization allows more robust design decisions at the cost of transforming the objective into an expectation. The stochastic multi-gradient algorithm (SMGDA)[11] extends the Robbins-Monro approach to the multi-objective case, allowing for the minimization of the expected objectives without having to directly calculate them. However, a bias in the algorithm and the inherent noise in stochastic gradients cause the algorithm to converge to only a subset of the whole Pareto front, limiting its use.

We reduce the bias of the stochastic multi-gradient calculation using an exponential smoothing technique and promote the exploration of the Pareto front by adding non-vanishing noise tangential to the front. We prove that this algorithm, Transverse Brownian Motion, generates samples in a concentrated set containing the whole Pareto front. Finally, we estimate the set of Pareto optimal design points using only the sequence generated during optimization while also providing bootstrapped confidence intervals using a nearest-neighbor model calibrated with a novel procedure based on the hypervolume metric.

Our proposed method allows for the estimation of the whole of the Pareto front using significantly fewer evaluations of the random quantities of interest when compared to a direct sample-based estimation, which is valuable in the context of costly model evaluations. We illustrate the efficacy of our approach with numerical examples in increasing dimension and discuss how to apply the method to more complex problems.

1 Notation

Symbol	Description
$(\Theta, \mathcal{A}, \mu)$	Probability space
θ	Event on a probability space
F	Set of functions of a random variable $F = \{f_1(x, \theta), ..., f_k(x, \theta)\}$
G	Set of deterministic functions $G = \{\mathbb{E}\, f_1, ..., \mathbb{E}\, f_k\}$
$\nabla_x^C\{F\}$	minimal norm vector in convex union of Clarke sub differentials acting on set F
bold	Bold font indicates a vector quantity
$d_{\mathcal{H}}$	Dimension of space $\mathcal{H}$.
$\mathcal{M}, \mathcal{N}, \mathcal{H}$	Manifold.
$\mathbf{a} \succ \mathbf{b}$	$a_i > b_i$ for a_i, b_i entries of $\mathbf{a}, \mathbf{b}$
$\mathbf{a} \succeq \mathbf{b}$	$a_i \geq b_i$ for a_i, b_i entries of $\mathbf{a}, \mathbf{b}$
$\mathbb{P}_{\not\prec \mathcal{P}}$	Probability of being undominated

2 Introduction

The true solution to a multi-objective optimization problem is not a single point, but rather a set of points corresponding to the tradeoffs inherent in minimizing a collection of objective functions. Solving multi-objective optimization problems therefore requires two steps, first, finding and collecting candidate solutions, and then assessing their optimality, or *dominance*, when compared to one another. The set of solutions with optimal trade-offs is referred to as the *Pareto front*. Determining the Pareto optimality of a solution in a deterministic problem requires only a pairwise comparison between each found solution, and so determining the Pareto front is an operation that requires cubic time.

Many problems of interest, however, are stochastic. Robust optimization, for example, incorporates noise on the design variables, constraints, and objectives to model uncertainty about their properties. The improved robusteness comes at a cost: the quantities of interest gain the functional form of an expectation.

In multi-objective problems the difficulties encountered in single-objective stochastic problems are compounded, both the process of determining candidate points and the assessment of dominance are more complex. Determining a minima in a stochastic multi-objective problem is challenging due to the difficulty imposed by working with expectations. More formally, given a probability space $(\Theta, \mathcal{F}, \mu)$ and a set of k quantities of interest $\{f_i(\mathbf{x}, \mathbf{W}(\theta))\}_{i=1,...,k}$, $f_i : \mathcal{X} \subseteq \mathbb{R}^d \times \mathbf{W}(\theta)(\Theta) \mapsto \mathbb{R}$, our goal is to find *all* $\mathbf{x}^*$ such that

$$\mathbf{x}^* \in \underset{\mathbf{x}\in\mathcal{X}}{\operatorname{argmin}}\{\mathbb{E}[f_i(\mathbf{x}, \mathbf{W}(\theta))], ..., \mathbb{E}[f_k(\mathbf{x}, \mathbf{W}(\theta))]\}. \tag{1}$$

Notice here that each objective function is a *deterministic* quantity by virtue of being the expectation of a stochastic quantity of interest. We can treat the set $\{f_i(\mathbf{x}, \mathbf{W}(\theta))\}_{i=1,...,k}$ as a vector valued function $F(\mathbf{x}, \mathbf{W}(\theta)) : \mathcal{X} \times \mathbf{W}(\Theta) \subseteq \mathbb{R}^d \mapsto \mathcal{Y} \subseteq \mathbb{R}^k$ and define the vector valued function $G := \{g_i\}_{i=1,...,k}, \mathcal{X} \subseteq \mathbb{R}^d \mapsto \mathcal{Y} \subseteq \mathbb{R}^k$ with $g_i(\mathbf{x}) = \mathbb{E}[f_i(\mathbf{x}, \mathbf{W}(\theta))]$. One can see that solving problem 1 requires finding all minima in all dimensions of G. In the setting where we do not have direct access to G, but only F and its Jacobian requires the calculation of the expectation of F. Calculating the expectation can be done in a variety of ways: sampling and averaging, integration, or a reasonable surrogate are all commonly used approaches. However, they are all computationally costly, and in the event where one has a difficult to evaluate or expensive objective function, that cost can become prohibitive.

Assessing dominance in stochastic multi-objective optimization problems becomes increasingly complex for similar reasons. A point, $\mathbf{y}$, in a set is considered dominant if it is less than or equal to all other points in the set, and strictly equal in at least one dimension. Succinctly, if the image of the deterministic vector valued function $G(\mathcal{X}) = \mathcal{Y}$, the Pareto front in objective space is the set which meets the following critera:

Definition 1 (Pareto Front)**.**

$$P(\mathcal{Y}) = \{\mathbf{y} \in \mathcal{Y} \mid \mathbf{y}' \succeq \mathbf{y}\ \exists j\ y'_j > y_j\ \forall \mathbf{y}' \in \mathcal{Y}\}. \tag{2}$$

In design space, the set of Pareto optimal design points can be defined as:

Definition 2 (Pareto Optimal Points)**.**

$$P(G\mathcal{X}) = \{\mathbf{x} \in \mathcal{X} \mid G(\mathbf{x}) \in P(\mathcal{Y})\}. \tag{3}$$

Together, these define a set of tuples $\mathcal{P} := \{(x, y)\} : x \in P(G\mathcal{X})\ y \in P(\mathcal{Y})$. A general procedure for determining the set of Pareto optimal points in design space is to first perform a series of minimizations of the objective functions G, determine the Preto front in objective space using definition 1, and then examine the preimage of the Pareto front to determine the set of Pareto optimal points. However, as in the case of finding minima, one rarely has access to exact expectations of of the objectives functions, $G(\mathbf{x})$, but rather samples $\{F(\mathbf{x}_1, \mathbf{W}(\theta)_1), ..., F(\mathbf{x}_n, \mathbf{W}(\theta)_n)\}$. The task of comparison then becomes a probabilistic one.

Both heuristic and rigorous approaches have been proposed to solve multi-objective problems. The family of heruistic approaches to multi-objective optimization, such as evolutionary strategies [5] and particle swarm optimization [1], involve maintaining a set of undominated solutions and iteratively improving on them at each round. While heuristic approaches are effective for finding the global minima of highly non-convex problems and are naturally adapted to finding a set of solutions, they are computationally extremely costly and lack theoretical guarantees. To make matters worse, applying population based approaches to stochastic problems requires the estimation the quantities of interest for each individual in the population, potentially increasing the computational complexity several fold. The family of rigorous approaches, to which our method belongs, have the advantage of strong theoretical guarantees when finding the minima of well posed problems. In addition, gradient based approaches especially, are amenable to the optimization of stochastic functions. The class of Robbins-Monro type algorithms allows for the minimization of a quantity of interest using only individual samples of the gradients, $J_{F(x)}$, avoiding calculation of the expectation entirely. Notably, the (S)tochastic (M)ulti (G)radient (A)lgorithm (SMGDA) [11] is of the class of Robbins-Monro type algorithms and builds on the work done in [6] to provably produces a sequence which converges to a point on the Pareto front. However, as we shall show later, a bias in the gradient calculation causes this algorithm to converge to a *subset* of the Pareto front, independent of the starting point chosen, making previous approaches to multi-objective optimization using SMGDA incomplete.

The task at hand then has two complementary aims:

1. One, extend the use of the SMGD algorithm to sample points from a noise ball which covers the *whole* of the Pareto front.

2. Two, to use those samples generated during optimization to identify the location of the Pareto front, and the set of Pareto optimal points, with uncertainty estimates.

For the first task, first we show that the SMGD algorithm converges only to a subset of the true Pareto front because of a bias in the direction of descent, which is the solution to a quadratic sub-problem. We then build on the SMGD algorithm in two complementary directions. One, we propose to debias the calculation of the direction of descent using an online estimate of the covariance of the Jacobian. We traverse the Pareto front by adding a novel noise term, gaussian noise perpendicular to the Pareto front, which both encourages exploration of the Pareto front and helps to prevent the algorithm from being stuck in a local minima.

For the second task, we estimate the Pareto front using a nearest neighbor model built only with points generated during the course of optimization and calibrated using the Hypervolume indicator. We estimate the joint probability of each generated point being undominated and joint confidence intervals around each point using a bootstraping approach.

3 Background: Mathematical Preliminaries, Multiple Gradient Descent Algorithm, and The Shift From Determinist to Stochastic Settings

First, a brief introduction to the Multi-gradient descent algorithm (MGDA), the Stochstic multi-gradient descent algorithm (SMGDA), and warm up with new proofs of their convergence. We note here that we have used the notation ∇g_i for the gradient of an individual objective function, g_i, but all of the following results hold if one uses a subdifferential or the Clarke subdifferential [11].

As in single-objective optimization, the minima of a multi-objective problem is the limiting point of the sequence

$$\mathbf{x}_{t+1} = \mathbf{x}_t - \varepsilon \mathbf{v}_t \tag{4}$$

where $-\mathbf{v}_t$ is a direction of descent. Assuming that all objectives are convex, L-locally lipschitz, and given some $\varepsilon \leq \frac{2}{L}$ one can prove that this sequence will converge to a point on the Pareto front.

To construct $\mathbf{v}_t$ a common strategy is to create a *scalarization* of the objective function. Given k objective functions $\{g_i(\mathbf{x})\}_{i=1,\ldots,k}$, $g_i : \mathbb{R}^d \mapsto \mathbb{R}$ generate a set of weights on the $k-1$ dimensional simplex

$w_i : \sum_{i=1}^{k} w_i = 1, w_i \geq 0$ to create the pseudo-objective:

$$\tilde{g}_w(\mathbf{x}) = \sum_{i=1}^{k} w_i g_i(\mathbf{x}). \tag{5}$$

Using the formulation of eq. 5, using a single objective optimization algorithm of choice will yield a sequence of points that converges to the Pareto front. The easily imagined strategy to find the whole pareto front using a pseudo-objective of the form 5, iteratively solving and reweighting, does not necessarily give good resolution of the whole of the Pareto front. It is often observed that even for convex problems that an evenly distributed set of weights will not lead to an even estimation of the Pareto front.

To avoid the problems associated with choosing weights for a pseudo-objective, the approach taken in [6] to find a gradient direction is to maximize the minimum improvement. To find the direction of ascent, $\mathbf{v}$, one then has to solve the subproblem

$$v^* = \arg\max_{v} \min_{i} \langle \nabla g_i(\mathbf{x}), \mathbf{v} \rangle - \frac{1}{2}||\mathbf{v}||^2. \tag{6}$$

In the dual formulation, one finds the minimium norm convex combination of gradients (see figure 1) by first solving for the weights $\alpha^*(\mathbf{x})$ in the $k-1$ dimensional simplex Δ^{k-1}

$$\alpha^* = \arg\min_{\alpha \in \Delta^{k-1}} \alpha^\top J_G(\mathbf{x}) J_G(\mathbf{x})^\top \alpha, \tag{7}$$

which is a quadratic problem in the number of objectives, k, and calculating the the direction of ascent:

$$\nabla_x^{\mathcal{C}}\{G\}(\mathbf{x}) := \sum_i \alpha_i^*(\mathbf{x}) \nabla g_i(\mathbf{x}). \tag{8}$$

One can then construct a sequence

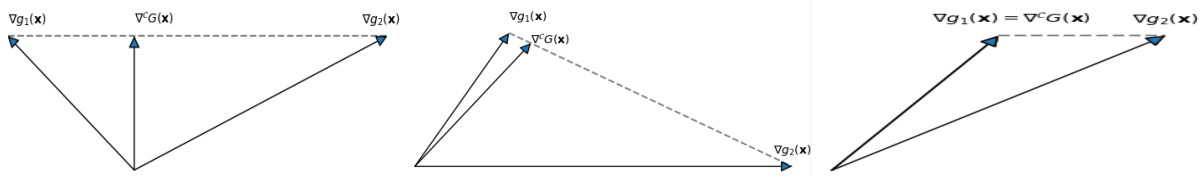


Figure 1: Examples of Calculation of Direction of Descent $\nabla_x^{\mathcal{C}}\{G\}(\mathbf{x})$ for two objectives.

One immediately notices the similarities between the pseudo-objective 5 and the formulation of the direction of descent 8. However there are two important differences. One being that the direction of descent

is recalculated at each round of optimization, meaning that there is no commitment to a particular weighting. Additionally, it is important to remember that the terms α are functions of $\mathbf{x}$ and so the peseudo-objective being minimized is *not* $\sum_i \alpha_i g_i(\mathbf{x})$, as is commonly misconstrued. The stationary points of the direction of descent as defined in eq. 8 then can be used to characterize Pareto stationarity.

Definition 3 (Pareto Stationary Point)**.** $\mathbf{x}^*$ such that

$$\nabla_x^{\mathcal{C}}\{G\}(\mathbf{x}^*) = \mathbf{0} \tag{9}$$

In the case where all objective functions are convex this forms a sufficient condition for Pareto optimality [9]. Intuitively, the pareto stationary points are the ones in which no combination of objectives can be improved without degrading the performance of at least one. In later sections, we will use Pareto stationarity as a tool to search for Pareto optimal points.

3.1 Warm Up: Potential Function Characterization of MOO problem and Convergence of MGDA

To warm up, introduce the concept of a potential function in multi-objective optimization, and establish useful results for later we will prove the convergence of the MGDA approach. Since multiple objectives are difficult to work with, in contrast to previously taken approaches [11][6][7][8], we take the conceptually simpler approach to define a pseudo-objective corresponding to the direction of descent $\nabla_x^{\mathcal{C}}\{G\}(\mathbf{x})$. We define the potential, $\Phi_G(\mathbf{x}) : \mathbb{R}^d \mapsto \mathbb{R}$

$$\Phi_G(\mathbf{x}) = \int_0^1 \nabla_x^{\mathcal{C}}\{G\}(s\mathbf{x})ds = \int_0^1 \sum_i \alpha_i^*(s\mathbf{x})\nabla g_i(s\mathbf{x})ds. \tag{10}$$

We can now replace claims made about the vector valued direction of descent $\nabla_x^{\mathcal{C}}\{G\}(\mathbf{x})$ with claims about the scalar valued $\Phi_G(\mathbf{x})$. $\Phi_G(\mathbf{x})$ is $\mathcal{C}^1$ smooth, therefore locally lipschitz continuous, and it can be readily seen that $\nabla\Phi(\mathbf{x}) = \nabla_x^{\mathcal{C}}\{G\}(\mathbf{x})$ by the gradient theorem. In addition, if all of the functions $\{g_i\}_{i=1,\ldots,k}$ are convex, $\Phi_G(\mathbf{x})$ is also convex.

Lemma 1. *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 $\langle \nabla\Phi_G(\mathbf{x}) - \nabla\Phi_G(\mathbf{y}), \mathbf{x} - \mathbf{y}\rangle \geq 0$.

$$\langle \nabla\Phi_G(\mathbf{x}) - \nabla\Phi_G(\mathbf{y}),\mathbf{x} - \mathbf{y}\rangle = \tag{11}$$

$$= \sum_i \langle \alpha_i^*(\mathbf{x})\nabla g_i(\mathbf{x}) - \alpha_i^*(\mathbf{x})\nabla g_i(\mathbf{y}) + \alpha_i^*(\mathbf{x})\nabla g_i(\mathbf{y}) - \alpha_i^*(y)\nabla g_i(\mathbf{y}), \mathbf{x} - \mathbf{y}\rangle \tag{12}$$

$$\geq \sum_i \langle (\alpha_i^*(\mathbf{x}) - \alpha_i^*(\mathbf{y}))\nabla g_i(\mathbf{y}), \mathbf{x} - \mathbf{y}\rangle \tag{13}$$

$$\geq 0 \tag{14}$$

Where in the second line we have used the convexity of the g_i and in the third we have used both the convexity of the g_i and the optimality of $\alpha^*(\mathbf{y})$ to yield the result. □

$\Phi_G(\mathbf{x})$ is, however, almost never strongly convex.

Lemma 2. *let $G := \{g_i(\mathbf{x})\}_{i=1,\ldots,k}$ be a set of convex functions, at least one of which is strongly convex, and $\Phi_G(\mathbf{x})$ be the potential function as defined in 10. Either the Pareto front is a singular point or $\Phi_G(\mathbf{x})$ is not strongly convex.*

Proof. If the pareto front is nonsingular there are two points $\mathbf{x}_1^*$ and $\mathbf{x}_2^*$ in the set of Pareto optimal points. We have then

$$\langle \nabla\Phi_G(\mathbf{x}_1^*) - \Phi_G(\mathbf{x}_2^*), \mathbf{x}_1^* - \mathbf{x}_2^*\rangle = 0 \tag{15}$$

by the pareto stationarity of points along the Pareto front. If, however, the pareto front is singular, the set of Pareto optimal points consists of only one point, $\mathbf{x}^*$ and one has

$$\langle \nabla\Phi_G(\mathbf{x}) - \nabla\Phi_G(\mathbf{x}^*), \mathbf{x} - \mathbf{x}^*\rangle \geq m||\mathbf{x} - \mathbf{x}^*||^2, \tag{16}$$

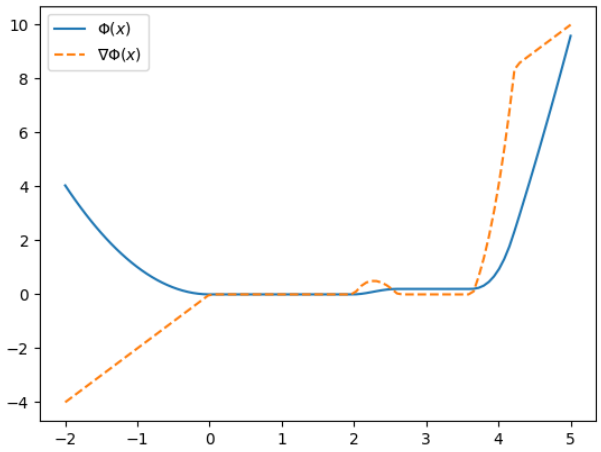
for some $m > 0$ by the pareto stationarity of $\mathbf{x}^*$ and the strong convexity of at least one member of G. □

As can be seen in figure 2b, if all of the objective functions $\{g_i(\mathbf{x})\}_{i=1,\ldots,k}$ are convex, then the full set of pareto stationary points defines the Pareto front. In the case that they are not all convex, as in figure 2a, then the Pareto front is a subset of the set of Pareto stationary points.

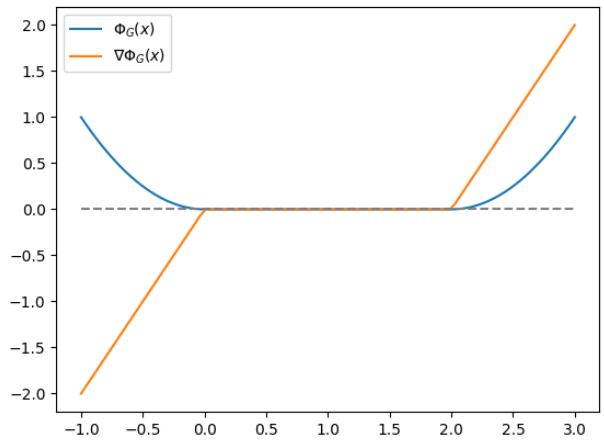
We can now show that the minima of $\Phi(\mathbf{x})$ are the set of pareto stationary points, and that performing gradient descent on $\Phi(\mathbf{x})$ yields pareto stationary points.

Theorem 3. *Let $\Phi_G(\mathbf{x})$ be a potential function defined as in 10. Then the sequence*

$$\mathbf{x}_{t+1} = \mathbf{x}_t - \varepsilon\nabla\Phi_G(\mathbf{x}). \tag{17}$$



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



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

converges to a Pareto stationary point.

Proof. From the assumption that Φ_G is locally L-liptshitz

$$\Phi_G(\mathbf{x}_{t+1}) \leq \Phi_G(\mathbf{x}_t) + \langle \nabla\Phi_G(\mathbf{x}_t), \mathbf{x}_{t+1} - \mathbf{x}_t \rangle + \frac{L}{2}||\mathbf{x}_{t+1} - \mathbf{x}_t||^2 \tag{18}$$

$$\leq \Phi_G(\mathbf{x}_t) - \varepsilon(1 - \frac{L\varepsilon}{2})||\nabla_x^{\mathcal{C}}\{G\}(\mathbf{x}_t)||^2 \tag{19}$$

Picking $\varepsilon \leq \frac{2}{L}$, summing, and dividing by T gives us the result

$$\frac{1}{T}\sum_{t=1}^{T}||\nabla_x^{\mathcal{C}}\{G\}(\mathbf{x}_t)||^2 \leq \frac{\Phi_G(\mathbf{x}_0) - \Phi_G(\mathbf{x}_{T+1})}{T\varepsilon} \tag{20}$$

Where in the second line we have used the fact that $\nabla\Phi(\mathbf{x}) = \nabla_x^{\mathcal{C}}\{G\}(\mathbf{x})$. □

3.2 Switching Gears, Deterministic to Stochastic

We now turn our attention to the stochastic formulation of the multigradient descent algorithm, SMGDA[11]. As defined in equation 1, the set of objective functions $\{g_i\}_{i=1,\dots,k}$ now have the functional form of an expectation $g_i(\mathbf{x}) = \mathbb{E}[f_i(\mathbf{x}, \mathbf{W}(\theta))]$. Instead of finding Pareto stationary points by calculating the expected value of the gradients of F, $\nabla\,\mathbb{E}[f_i(\mathbf{x}, \mathbf{W}(\theta))]$ and using the MGDA approach as shown above, we replace the gradient calculation with an estimate from a *single sample* of $\mathbf{W}(\theta)$. Where we assume that

	Objectives	Gradients	Potential	Multi-gradient
Deterministic	$g(\mathbf{x})$	$\nabla g(\mathbf{x})$	Φ_G	$\nabla_x^{\mathcal{C}}\{G\}(\mathbf{x})$
Stochastic	$\mathbb{E}[f(\mathbf{x}, \mathbf{W}(\theta))]$	$\nabla f(\mathbf{x}, \mathbf{W}(\theta)_i)$	Φ_F	$\nabla_x^{\mathcal{C}}\{F\}(\mathbf{x}, \mathbf{W}(\theta)_i)$

$\mathbb{E}[\nabla f(\mathbf{x}, \mathbf{W}(\theta))] = \nabla\,\mathbb{E}[f(\mathbf{x}, \mathbf{W}(\theta))]$ and so it is an unbiased estimator of the true gradient. It was shown

in [11] that one can make a stochastic version of MGDA, which converges to a point on the Pareto front. However, the proof given in [11] required several strongly limiting assumptions on the behavior of the individual objectives, which may not be met in practice. Instead, we can use an approach as defined above to show that SMGDA converges to a pareto stationary point defined as

$$\mathbb{E}[\nabla_x^{\mathcal{C}}\{F\}] = \mathbf{0}. \tag{21}$$

We will first define a potential function, as before, and then show that it can be minimized using stochastic gradient descent under limited assumptions about the behavior of $\nabla_x^{\mathcal{C}}\{F\}$.

Assumption 3.1. $\mathbb{V}[\nabla_x^{\mathcal{C}}\{F\}(\mathbf{x})] \leq M_{V_0} + M_V ||\, \mathbb{E}[\nabla_x^{\mathcal{C}}\{F\}(\mathbf{x})]||^2$

Theorem 4. *Let there be a function $\Phi_F(\mathbf{x})$ defined as*

$$\Phi_F(\mathbf{x}) = \int_0^1 \mathbb{E}[\nabla_x^{\mathcal{C}}\{F\}(s\mathbf{x})]ds = \int_0^1 \sum_i \mathbb{E}[\alpha(s\mathbf{x}, \mathbf{W}(\theta))\nabla f_i(s\mathbf{x}, \mathbf{W}(\theta))]ds \tag{22}$$

such that $\nabla\nabla^\top \Phi[F]\mathbf{x} \preceq L$ for some $L \geq 0$. And let there be a set of real numbers ε_t such that

$$\frac{\sum_t \varepsilon_t^2}{\sum_t \varepsilon_t} \to 0 \tag{23}$$

Then the sequence defined by

$$\mathbf{x}_{t+1} = \mathbf{x}_t - \varepsilon_t \nabla_x^{\mathcal{C}}\{F\}(\mathbf{x}_t) \tag{24}$$

converges to a Pareto stationary point as defined by 21 with probability 1.

Proof. Starting from local lipschitz continuity and the mean value theorem,

$$\begin{aligned}
\Phi_F(\mathbf{x}_{t+1}) \leq& \Phi_F(\mathbf{x}_t) + \langle \nabla\Phi_F(\mathbf{x}_t), \mathbf{x}_{t+1} - \mathbf{x}_t\rangle + \frac{L}{2}||\mathbf{x}_{t+1} - \mathbf{x}_t||^2 && (25)\\
\leq& \Phi_F(\mathbf{x}_t) - \varepsilon_t \langle \nabla\Phi_F(\mathbf{x}_t), \nabla_x^{\mathcal{C}}\{F\}(\mathbf{x}_t)\rangle + \frac{L\varepsilon_t^2}{2}||\nabla_x^{\mathcal{C}}\{F\}(\mathbf{x}_t)||^2 && (26)\\
\mathbb{E}_t[\inf \Phi_F] \leq \mathbb{E}_t[\Phi_F(\mathbf{x}_{t+1})] \leq& \Phi_F(\mathbf{x}_t) - \varepsilon_t(1 - \frac{L\varepsilon_t}{2})||\, \mathbb{E}_t[\nabla_x^{\mathcal{C}}\{F\}(\mathbf{x}_t)]||^2 + \frac{L\varepsilon_t^2}{2} M_{V_0} && (27)\\
& && (28)
\end{aligned}$$

Taking the full expectation, setting $\varepsilon_t \leq \varepsilon_0 \leq \frac{2}{L}$, summing, and rearranging yields

$$\sum_{t=1}^{T} \varepsilon_t ||\, \mathbb{E}[\nabla_x^{\mathcal{C}}\{F\}]||^2 \leq \mathbb{E}[\Phi_F(\mathbf{x}_0) - \inf \Phi_F] + \frac{LM_{V_0}}{2} \sum_{t=1}^{T} \varepsilon_t^2 \tag{29}$$

We can now define $E_T = \sum_{t=1}^{T} \varepsilon_t$ and treat the ratio $\frac{\varepsilon_t}{E_T}$ as the probability of selecting iteration t from the set $0, ..., T$, then by Markov's inequality

$$\mathbb{P}_{E_T}\left(|| \mathbb{E}[\nabla_x^{\mathcal{C}}\{F\}(\mathbf{x}_t)]||^2 \geq \epsilon \right) \leq \frac{\mathbb{E}_{E_T}[|| \mathbb{E}[\nabla_x^{\mathcal{C}}\{F\}]||^2]}{\epsilon} \leq \frac{\mathbb{E}[\Phi_F(\mathbf{x}_0) - \inf \Phi_F]}{\epsilon E_T} + \frac{LM_{V_0} \sum_{t=1}^{T} \varepsilon_t^2}{2\epsilon E_T} \tag{30}$$

Taking the limit as $T \to \infty$ and noting that $\frac{\sum_t \varepsilon_t^2}{\sum_t \varepsilon_t} \to 0$ gives the result. □

The assumptions used in this proof are commonly used in proofs of stochastic gradient descent [2] and are much lighter and applicable than used previously in [11]. It is also worth noting that they are automatically fulfilled if $\mathcal{X}$ is a subset of $\mathbb{R}^d$. However, of critical importance is the difference between the set $\mathbf{x}_1^*$ and $\mathbf{x}_2^*$ determined by

$$\mathbb{E}[\nabla_x^{\mathcal{C}}\{F\}(\mathbf{x}_2^*)] = \mathbf{0} \quad \text{and} \quad \nabla_x^{\mathcal{C}}\{G\}(\mathbf{x}_1^*) = \mathbf{0}. \tag{31}$$

As we will see in the next section, they are not the same. In fact, $\{\mathbf{x}_2\} \subset \{\mathbf{x}_1\}$. This is due to a *bias* in the calculation of $\mathbb{E}[\nabla_x^{\mathcal{C}}\{F\}]$, and is the focus of the next section.

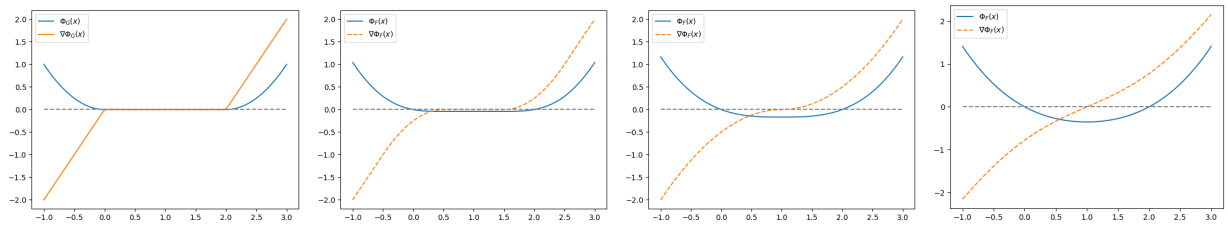
4 How the Bias Affects Convergence: Drawback of the SMGD Algorithm

The bias in the calculation of $\mathbb{E}[\nabla_x^{\mathcal{C}}\{F\}(\mathbf{x})]$ prevents the SMGD algorithm from converging to the whole of the Pareto front. Take as a motivating example the multiobjective problem

$$\mathbf{x}^* \in \underset{\mathbf{x} \in \mathcal{X}}{\operatorname{argmin}} \{\mathbb{E}[(\mathbf{x} - 2 + \mathbf{W}(\theta))^2], \mathbb{E}[(\mathbf{x} + \mathbf{W}(\theta))^2]\} \tag{32}$$

with $\mathbf{x} \in \mathbb{R}$ and $\mathbf{W}(\theta)$ distributed one of three ways, $\mathbf{W}(\theta) \sim \mathcal{U}[-.5, .5]$, $\mathbf{W}(\theta) \sim \mathcal{U}[-1, 1]$, and $\mathbf{W}(\theta) \sim \mathcal{N}(0, 1)$. In each case, the minima and Pareto front of the objectives $\mathbb{E}[(\mathbf{x}-2+\mathbf{W}(\theta))^2]$, and $\mathbb{E}[(\mathbf{x}+\mathbf{W}(\theta))^2]$ remain unchanged. As can be seen in figure 3a, where the direction of descent has been calculated using the exact expectations of the gradients, $\nabla g_i(\mathbf{x}) := \mathbb{E}[\nabla f_i(\mathbf{x}, \mathbf{W}(\theta))]$ which yields $\nabla_x^{\mathcal{C}}\{G\}(\mathbf{x})$, the Pareto front should be a straight line between 0 and 2 representing the trade-offs between the simple quadratic objectives. However, limiting ourselves to the case where one must calculate $\mathbb{E}[\nabla_x^{\mathcal{C}}\{F\}]$ as in figures 3b, 3c, 3d, we see that the noise has smoothed the calculation of the direction of descent, giving the algorithm a bias. Clearly $\mathbb{E}[\nabla_x^{\mathcal{C}}\{F\}(\mathbf{x})] \neq \nabla_x^{\mathcal{C}}\{G\}(\mathbf{x})$ in general. This effect can cause the SMGD algorithm to converge only to a single point on the Pareto front as in the case of $\mathcal{U}[-1, 1]$ and $\mathcal{N}(0, 1)$ noise seen in figures 3c and 3d respectively.

As was shown previously, the SMGD algorithm converges to the points in which $\mathbb{E}[\nabla_x^{\mathcal{C}}\{F\}] = \mathbf{0}$. We will



(a) Direct access to the expectations $\mathbb{E}[\nabla f_i]$. (b) $\mathbf{W}(\theta) \sim \mathcal{U}[-.5, .5]$ (c) $\mathbf{W}(\theta) \sim \mathcal{U}[-1, 1]$ (d) $\mathbf{W}(\theta) \sim \mathcal{N}(0, 1)$

now show that the subset determined by the level set $F^* := \{x \mid \nabla_x^{\mathcal{C}}\{F\}(\mathbf{x}) = \mathbf{0}\}$ is included in the set of points $G^*\{x \mid \nabla_x^{\mathcal{C}}\{G\}(\mathbf{x}) = \mathbf{0}\}$. We first change our assumptions on the variance, namely we assume that

Assumption 4.1.

$$\mathbb{V}[\nabla_x^{\mathcal{C}}\{F\}(\mathbf{x})] \leq M_{V_0} + M_V ||\nabla_x^{\mathcal{C}}\{G\}(\mathbf{x})||^2, \tag{33}$$

Theorem 5. *Let the sets $F^* = \{x \mid \nabla_x^{\mathcal{C}}\{F\}(\mathbf{x}) = \mathbf{0}\}$ and $G^*\{x \mid \nabla_x^{\mathcal{C}}\{G\}(\mathbf{x}) = \mathbf{0}\}$ denote the sets of minima found using SMGD and MGD on the same problem, where we use MGD on the expected gradients of the objective functions $\{f_i(\mathbf{x})\}_{i=1,\dots,k}$. $F^* \subseteq G^*$.*

Proof. Under assumption 33, define a sequence $\{\varepsilon_t\}_{t\in\mathbb{N}}$ such that $\frac{\sum_{t=1}^T \varepsilon_t^2}{\sum_{t=1}^T \varepsilon_t} \to 0$. Define $B_t = \nabla_x^{\mathcal{C}}\{F\} - \nabla_x^{\mathcal{C}}\{G\}$

$$\begin{aligned}
\Phi_G(\mathbf{x}_{t+1}) \leq& \Phi_G(\mathbf{x}_t) + \langle \nabla\Phi_G(\mathbf{x}_t), \mathbf{x}_{t+1} - \mathbf{x}_t\rangle + \frac{L}{2}||\mathbf{x}_{t+1} - \mathbf{x}_t||^2 && (34)\\
=& \Phi_G(\mathbf{x}_t) + \langle \nabla\Phi_G(\mathbf{x}_t), \mathbf{x}_{t+1} - \mathbf{x}_t\rangle + \frac{L}{2}||\mathbf{x}_{t+1} - \mathbf{x}_t||^2 && (35)\\
\mathbb{E}[\inf \Phi_G] \leq& \mathbb{E}[\Phi_G(\mathbf{x}_{t+1})] \leq \Phi_G(\mathbf{x}_t) - \varepsilon_t||\nabla_x^{\mathcal{C}}\{G\}(\mathbf{x}_t)||^2 - \varepsilon_t\langle \nabla_x^{\mathcal{C}}\{G\}(\mathbf{x}_t), \mathbb{E}[B_t]\rangle + \frac{L\varepsilon_t^2 M_{V_0}}{2} && (36)
\end{aligned}$$

Taking the full expectation, rearranging, and summing shows us the relation

$$\sum_{t=0}^{T} \mathbb{E}[||\nabla_x^{\mathcal{C}}\{G\}||^2] \leq \mathbb{E}[\Phi[G]\mathbf{x}_0 - \inf \Phi_G] - \sum_{t=1}^{T} \varepsilon_t\langle \nabla_x^{\mathcal{C}}\{G\}(\mathbf{x}_t), \mathbb{E}[B_t]\rangle + \frac{LM_{V_0}}{2}\sum_{t=1}^{T}\varepsilon_t^2 \tag{37}$$

Where as before, defining $E_T = \sum_t^T \varepsilon_t$, we have

$$\mathbb{P}_{E_T}\left(\mathbb{E}[||\nabla_x^{\mathcal{C}}\{G\}||^2] \geq \epsilon\right) \leq \frac{\mathbb{E}[\Phi[G]\mathbf{x}_0 - \inf\Phi_G]}{\epsilon E_T} - \underbrace{\frac{\sum_{t=1}^T \varepsilon_t\langle \nabla_x^{\mathcal{C}}\{G\}(\mathbf{x}_t), \mathbb{E}[B_t]\rangle}{\epsilon E_T}}_{\text{bias term}} + \frac{LM_{V_0}}{2}\frac{\sum_{t=1}^T \varepsilon_t^2}{\epsilon E_T}. \tag{38}$$

for some t' such that $B_{t'} = \mathbf{0}\ \forall t \geq t'$ we have $\nabla_x^{\mathcal{C}}\{F\}(\mathbf{x}_\infty) = \nabla_x^{\mathcal{C}}\{G\}(\mathbf{x}_\infty) = \mathbf{0}$ and so $\mathbf{x}_\infty \in F^*, G^*$. However, the converse is not true. Take $\mathbf{x}'_\infty = \text{argmin}_x\, \mathbb{E}[f_j(\mathbf{x}, \mathbf{W}(\theta))]$ for some $j \in \{1, \dots, k\}$. Clearly,

$\nabla_x^{\mathcal{C}}\{G\}(x[\infty]') = \mathbf{0}$, however $\mathbb{E}[\nabla_x^{\mathcal{C}}\{F\}(\mathbf{x}'_\infty)] \neq 0$ since $\nabla f_j(\mathbf{x}'_\infty, \mathbf{W}(\theta))$ is dominated entirely by noise. So $\mathbf{x}'_\infty \in G^*$ but $\mathbf{x}'_\infty \notin F^*$ and $F^* \subseteq G^*$. □

The natural question to ask is, what is the source of this bias? 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. $\mathbb{E}[J_F(\mathbf{x}, \mathbf{W}(\theta))] = J_G(\mathbf{x})$. In such a situation it would be straightforward to assume then that the gradient estimation is unbiased. However, the minimizer $\alpha^*(\mathbf{x}, \mathbf{W}(\theta)) = \operatorname{argmin}_{\alpha \in \Delta^{k-1}} \alpha^\top J_F(\mathbf{x}, \mathbf{W}(\theta)) J_F(\mathbf{x}, \mathbf{W}(\theta))^\top \alpha$ is a quadratic function of the Jacobian, by Jensen's inequality

$$\mathbb{E}[J_F(\mathbf{x}, \mathbf{W}(\theta)) J_F(\mathbf{x}, \mathbf{W}(\theta))^\top] \geq \mathbb{E}[J_F(\mathbf{x}, \mathbf{W}(\theta))]\, \mathbb{E}[J_F(\mathbf{x}, \mathbf{W}(\theta))]^\top, \tag{39}$$

and so $\mathbb{E}[\alpha^*(\mathbf{x}, \mathbf{W}(\theta))] \neq \alpha^*(\mathbf{x})$.

Debiasing the computation of α^* could be as straightforward as resampling and averaging the Jacobian, using $\widehat{J}_F(\mathbf{x}) = \frac{1}{N} \sum_{i=1}^N J_F(\mathbf{x}, \mathbf{W}(\theta)_i)$ in place of $J_F(\mathbf{x})$ and calculating α^* as

$$\widehat{\alpha}^* = \operatorname*{argmin}_{\alpha \in \Delta^{k-1}} \alpha^\top \widehat{J}_F(\mathbf{x}) \widehat{J}_F(\mathbf{x})^\top \alpha. \tag{40}$$

However, not only would it be computationally taxing to compute N extra gradients at each time step, but there would also be residual variance in our estimation rendering our computational efforts somewhat phyrric. Instead, noting that

$$\mathbb{E}[J_F(\mathbf{x}, \mathbf{W}(\theta)) J_F(\mathbf{x}, \mathbf{W}(\theta))^\top] = \Sigma_{J_F(\mathbf{x}, \mathbf{W}(\theta))} + \mathbb{E}[J_F(\mathbf{x}, \mathbf{W}(\theta))]\, \mathbb{E}[J_F(\mathbf{x}, \mathbf{W}(\theta))]^\top \tag{41}$$

where $\Sigma_{J_F(\mathbf{x}, \mathbf{W}(\theta))}$ denotes the covariance matrix of the gradients of F, we estimate $\widehat{\Sigma}_{J_F(\mathbf{x}, \mathbf{W}(\theta))}$ *on-line* using individual samples of $J_F(\mathbf{x}, \mathbf{W}(\theta))$ and calculate α^* using the following modified formulation.

$$\widetilde{\alpha}^*(\mathbf{x}) = \operatorname*{argmin}_{\alpha \in \Delta^{k-1}} \alpha^\top \big(J_F(\mathbf{x}, \mathbf{W}(\theta)) J_F(\mathbf{x}, \mathbf{W}(\theta))^\top - \widetilde{\Sigma}_{J_F(\mathbf{x}, \mathbf{W}(\theta))} \big) \alpha. \tag{42}$$

We can then calculate a debiased form of the stochastic gradient

$$\widetilde{\nabla_x^{\mathcal{C}}\{F\}}(\mathbf{x}, \mathbf{W}(\theta)) := \sum_i \widetilde{\alpha}_i^* \nabla f_i(\mathbf{x}, \mathbf{W}(\theta)) \tag{43}$$

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

5 Proposed Approach: Transverse Brownian Motion

Our work takes two distinct lines. Firstly, the SMGD algorithm, even if run from distinct starting points, converges to a subset of the Pareto front, which is undesirable. To ameliorate this we propose to augment the standard gradient descent iterates from the SMGD algorithm in two ways. The debiasing strategy discussed in section 4 to allows the algorithm to converge to the whole of the Pareto front. To explore the Pareto front as well as overcome any remaining bias, we add a noise term which will allow the algorithm to explore directions tangential to the Pareto front. Because of the added noise term, this approach efficiently explores the area around the minima and will not stay stuck in a saddle point, leading to a dense characterization of the Pareto front with less wasteful computation when compared to multiple restarts. In the second phase, oonce we have a collection of samples from a noise ball around the Pareto front, we must determine the set of Pareto optimal design points and estimate the location of the true Pareto front. We estimate the Pareto front using local averages in order to give a preliminary estimation the Pareto front. Then, we create both bootstrap confidence intervals and estimate the fitness of each potential design point to give a characterization of the whole of the Pareto front with uncertainty. Our strategy does not require the full objective functions or gradients to ever be evaluated, making the extra gradient descent steps worthwhile and computationally tractable.

5.1 Transverse Brownian Motion

To find and explore the Pareto front we generate a sequence using the recurrence relation,

$$\mathbf{x}_{t+1} = \mathbf{x}_t - \varepsilon_t \widetilde{\nabla_x^{\mathcal{C}}\{F\}}(\mathbf{x}_t) + \sqrt{2\varepsilon_t \beta^{-1}} \mathbb{1}_{\mathcal{P}} T_t, \tag{44}$$

with $T_t \sim \mathcal{N}(0, PP^\top)$ with P such that T_t is perpendicular to the Pareto front, $\widetilde{\nabla_x^{\mathcal{C}}\{F\}}(\mathbf{x}_t)$ is the direction of descent calculated with debiased parameters $\widetilde{\alpha}^*$ explained in section 4,β an inverse temperature parameter set by the practitioner, and $\mathbb{1}_{\mathcal{P}}$ is an indicator that $x[t]$ is near the Pareto front.

Since in this setting we do not have a priori information, particularly about the Pareto front, we have to estimate the direction tangent to the Pareto front. To do this, we can use the *stochastic multi-gradient* information since the stochastic multi-gradient always has a component normal to the tangential direction

of the Pareto front. At each iteration, we calculate a set of moving averages

$$\widetilde{\mu}_{t+1} = \widetilde{\mu}_t + \gamma_t\big(J_F(\mathbf{x}_t, \mathbf{W}(\theta)_t) - \widetilde{\mu}_t\big) \tag{45}$$

$$\widetilde{\Sigma}_{J_F,t+1} = \widetilde{\Sigma}_{J_F,t} + \gamma_t\big(J_F(\mathbf{x}_t, \mathbf{W}(\theta)_t) - \widetilde{\mu}_t\big)\big(J_F(\mathbf{x}_t, \mathbf{W}(\theta)_t) - \widetilde{\mu}_t\big)^\top \tag{46}$$

$$\widetilde{\Pi}_{t+1} = \widetilde{\Pi}_t + \gamma_t\big(\widetilde{\nabla_x^{\mathcal{C}}\{F\}}(\mathbf{x}_t, \mathbf{W}(\theta)_t) - \widetilde{\Pi}_t\big) \tag{47}$$

With $\gamma_t \in (0,1)$ and using the value of $\widetilde{\Sigma}_{J_F,t}$ to estimate the covariance of the gradients for use in the debiasing procedure outlined in equations 42 and 43. We then calculate the projection matrix, P

$$P_{t,i,j} = \delta_{i,j} - \frac{1}{||\widetilde{\Pi}_t||^2}\widetilde{\Pi}_{t,i}\widetilde{\Pi}_{t,j} \tag{48}$$

where $\delta_{i,j}$ denotes the dirac delta function. Setting $T_t = P_t Z_t$ with $Z \sim \mathcal{N}(0, \mathbb{1}_{d\times d})$ allows us to take a step forward using 44. In addition, since the gradient has been debiased, the quantity $\widetilde{\nabla_x^{\mathcal{C}}\{F\}}(\mathbf{x}_t)$ to give us more information about the pareto stationarity of the point $\mathbf{x}_t$, allowing us to perform a relaxation of the indicator function, approximating it with the stand-in

$$\mathbb{1}_{\mathcal{P}} \approx e^{-||\widetilde{\Pi}_t||}. \tag{49}$$

This damps the random dynamics far away from stationary points while not hindering either the exploration of the Pareto front or the ability of the algorithm to escape saddle points. The goal of this approach is to explore the whole of the Pareto front, and allowing ε_t or γ_t to go to zero would create a unique limiting distribution centered at $\mathbf{x}_\infty$. In order to insure that all points are reached we are willing to allow for a slightly higher amount of noise in the sampled points and so we define a pair of sequences

$$\{\varepsilon_t\}_{t=0,\dots,\infty} := \varepsilon_0 \geq \varepsilon_1 \geq \dots \geq \varepsilon_\infty > 0 \tag{50}$$

$$\{\gamma_t\}_{t=0,\dots,\infty} := \gamma_0 \geq \gamma_1 \geq \dots \geq \gamma_\infty > 0 \tag{51}$$

Which prevents the algorithm from having a limiting distribution centered on a single point.

5.2 Pareto Front Estimation

Having a set of samples $\{(\mathbf{x}_1, \mathbf{F}(\mathbf{x}_1, \mathbf{W}(\theta)_1)), \dots, (\mathbf{x}_n, \mathbf{F}(\mathbf{x}_n, \mathbf{W}(\theta)_n))\}$, our task now decomposes into three problems

1. To map samples $\mathbf{x}_i$ to objective space in order to compare them we must find a mapping $\widehat{G}(\mathbf{x}) : \mathbb{R}^d \mapsto$

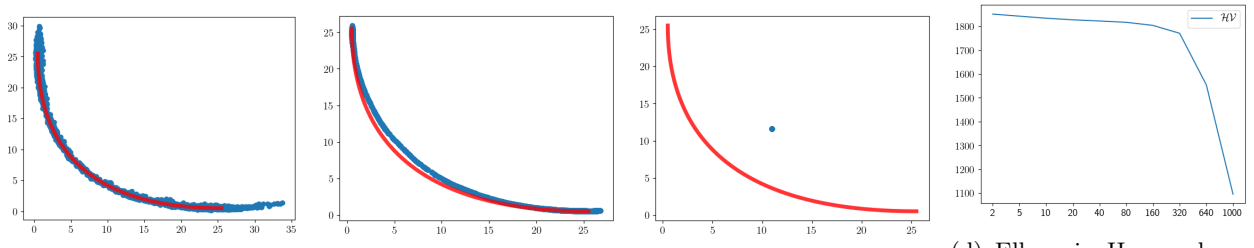
$\mathbb{R}^k, \mathbf{x} \mapsto [\mathbb{E}[f_1(\mathbf{x}, \mathbf{W}(\theta))|X], ..., \mathbb{E}[f_k(\mathbf{x}, \mathbf{W}(\theta))|X]]$ where $[\mathbb{E}[f_1(\mathbf{x}, \mathbf{W}(\theta))|X], ..., \mathbb{E}[f_k(\mathbf{x}, \mathbf{W}(\theta))|X]]$ which is now meant to be understood as the k dimensional vector of objectives.

2. To assess the Pareto efficiency of each point $\mathbf{x}_i$ and determine the Pareto front in objective space $\mathcal{P}(\widehat{G})$.

3. To determine the set of pareto optimal designs through the preimage of the mapping to the Pareto front $P(\widehat{G}\mathcal{X})$

The set of samples $\{(\mathbf{x}_1, \mathbf{F}(\mathbf{x}_1, \mathbf{W}(\theta)_1)), ..., (\mathbf{x}_n, \mathbf{F}(\mathbf{x}_n, \mathbf{W}(\theta)_n))\}$ are not directly on the Pareto front almost surely. The conditional expectation of the sample function evaluations $\{\mathbb{E}[f_1(\mathbf{x}, \mathbf{W}(\theta))|X], ..., \mathbb{E}[f_k(\mathbf{x}, \mathbf{W}(\theta))|X]\}$, however, *are* on the Pareto front. We calculate the conditional expectation using k-nearest-neighbor averages, a flexible fully nonparametric approach with few hyperparameters which converges under light assumptions [4].

Calculating the number of neighbors is equivalent to finding the best smoothing of our data. Using too few neighbors to estimate the vector of conditional expectations results in a granular and overly optimistic Pareto front. Too many, though, and we overwrite the local nature of the estimate and instead compute a global average. To track the quality of our estimate and calibrate the number of neighbors we sequentially add neighbors and search for an elbow in the *hypervolume indicator*, a commonly used performance metric indicating the quality of a set of solutions in multi-objective optimization. As can be seen in figure 4d, the rapid change in the hypervolume indicator corresponds to an oversmoothing of the local average. To prevent oversmoothing, we set the number of neighbors to be equal to half of the 'elbow' value.



(a) 5 neighbors (b) 160 neighbors (c) 1000 neighbors (d) Elbow in Hypervolume at $\sim$ 160 neighbors

Figure 4: Convex Pareto front estimated through KNN average.

Having a mapping to the space of objectives allows us to assess the relative dominance between points. Denoting the result of the local average at sampled point $\mathbf{x}_i$ as $\widehat{G}(\mathbf{x}_i)$, we would like to know if an individual $\mathbf{I}_k$ in the set of n tuples $\mathcal{I} = \{(\mathbf{x}_1, \widehat{G}(\mathbf{x}_1)), ..., (\mathbf{x}_n, \widehat{G}(\mathbf{x}_n))\}$ is undominated. e.g.

$$\nexists \mathbf{I}_j \in \mathcal{I} \text{ s.t. } \mathbf{I}_j \prec \mathbf{I}_k. \tag{52}$$

Because the values of $\widehat{G}(\mathbf{x})$ are estimates of the the true vector valued function $G(\mathbf{x})$, the expression above in Eq. 52 is now a probabilistic notion, and we seek to find

$$\nexists \mathbf{I}_j \in \mathcal{I} \text{ s.t. } \mathbb{P}_{\mathcal{I}}(\mathbf{I}_j \prec \mathbf{I}_k) \geq \epsilon \tag{53}$$

This condition is well generalized by the Pareto optimal Probability [13]

$$\mathcal{POP}(\mathbf{I}_k) = \mathbb{P}_{\neg \mathcal{P}} = 1 - P_{\mathcal{I}}(\cap_j \mathbf{I}_j \prec \mathbf{I}_k) \tag{54}$$

The probability $P_{\mathcal{I}}$ must be the joint pdf of all of the elements in $\mathcal{I}$. Elements in $\mathcal{I}$ are not independent. Since the set of samples are, in expectation, on the Pareto front we can *resample* and perform a bootstrap estimate of the Pareto optimal probability for each individual in the sample $\mathbf{I}_k$. The procedure is as follows

1. for each $\mathbf{x}_i$ resample individual $\mathbf{I}_i$ m times, generating $\{\widehat{\mathbf{I}}_i^{(1)}, ..., \widehat{\mathbf{I}}_i^{(m)}\}_{i=1}^n$
2. for each $\mathbf{x}_i$ calculate a bootstrap estimate of the Pareto optimal probability $\mathcal{P}_i = 1 - \frac{1}{m} \sum_{p=1}^{m} \mathbb{1}(\exists j \text{ s.t. } \widehat{\mathbf{I}}_j^{(p)} \prec \widehat{\mathbf{I}}_i^{(p)})$

This yields a set of quality assessments for each sampled point $\mathbf{x}_i$ Since we resample across the whole set of points $\mathcal{I}$, the bootstrap samples are drawn from the joint PDF for $\mathcal{I}$. To estimate the set of optimal design points, we take as $P(\widehat{G}\mathcal{X})$ the set

$$P(\widehat{G}\mathcal{X}) = \{\mathbf{x}_i \mid \mathcal{P}_i > c\} \tag{55}$$

With this estimation of the set of Pareto optimal points and corresponding bootstrap estimates of $\widehat{G}(\mathbf{x})$ we can also generate bootstrapped confidence intervals in objective space.

1. first discretize the space into a set of boxes of dimension k, labeled $B_{1,...,k}$.
2. A confidence interval for row d can be made by ordering the elements in the union of all the boxes in the d^{th} row $\{\cup_{i,...,l}^{k,...,k} B_{i,...,d,...,l}\}$, and taking appropriate quantiles.

See figure 5 for a visual representation of this procedure for one point in a bi-objective problem.

6 Numerical Experiments

In this section, we show the numerical effectiveness of our approach in both converging to the whole Pareto front in objective space and in inferring the set of Pareto optimal points. First, we point out that constructing example problems for stochastic multi-objective optimization requires some care. One may naïvely transform

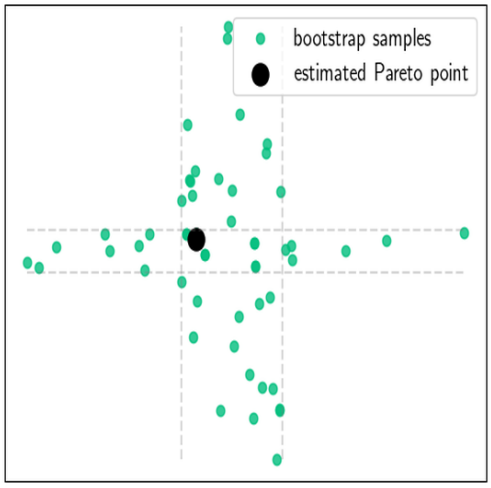


Figure 5: Example estimated Pareto efficient point and corresponding box $B_{i,j}$

a set of deterministic objectives by simply adding noise to the design variables $f(\mathbf{x}) \to f(\mathbf{x}+\mathbf{W}(\theta))$. However, adding noise to the design variables has a smoothing effect on the objective in expectation and can change the resulting Pareto front in unpredictable ways, even changing it from concave to convex. In addition, one one has added noise to the design variables, changes in dimensionality can also change the shape of the Pareto front, and the length of the Pareto front increases in proportion to the change in volume as one goes to higher dimension. To make a fair comparison across dimension we generate an example problem, the quadratic gaussian, a higher dimensional example of that seen in section 4. Our goal is to find all $\mathbf{x}^*$

$$\mathbf{x}^* \in \underset{\mathbf{x}}{\text{argmin}}\{\mathbb{E}[f_1(\mathbf{x}, \mathbf{W}(\theta))], \mathbb{E}[f_2(\mathbf{x}, \mathbf{W}(\theta))]\} \tag{56}$$

with

$$f_1(\mathbf{x}, \mathbf{W}(\theta)) = (\mathbf{x} - \frac{\mathbf{s}}{\sqrt{d}} + \mathbf{W}(\theta))||^2 \qquad \mathbb{E}[f_1(\mathbf{x}, \mathbf{W}(\theta))] = ||\mathbf{x} - \frac{\mathbf{s}}{\sqrt{d}}||^2 + \sigma^2 \tag{57}$$

$$f_2(\mathbf{x}, \mathbf{W}(\theta)) = ||(\mathbf{x} + \mathbf{W}(\theta))||^2 \qquad \mathbb{E}[f_2(\mathbf{x}, \mathbf{W}(\theta))] = ||\mathbf{x}||^2 + \sigma^2 \tag{58}$$

$$\mathbf{W}(\theta) \sim \mathcal{N}(0, \frac{\sigma^2}{d}) \tag{59}$$

For our examples, we set $s = 5\mathbb{1}_d$ with $\mathbb{1}_d$ the vector of all ones. The set of Pareto optimal points is straight line between $x_1^* = \frac{\mathbf{5}}{\sqrt{d}}$ and $\mathbf{0}$ and has a length of 5 in all dimensions.

As seen in section 4, this type of problem is difficult to solve with vanilla SMGD, the noise has infinite support and, without debiasing the gradient, it would posess a limiting point at the center of the Pareto front. To conduct our tests, we allow only 1000 *total* evaluations of the stochastic functions $\{f_1(\mathbf{x}, \mathbf{W}(\theta)), f_2(\mathbf{x}, \mathbf{W}(\theta))\}$,

and we give ε_t and γ_t the functional form

$$\varepsilon_t = \gamma_t = \max(\frac{\varepsilon_0}{\sqrt{1+\lambda_0 t}}, \varepsilon_\infty), \tag{60}$$

with $\varepsilon_0 > \varepsilon_\infty > 0$, as our decreasing set of parameters.

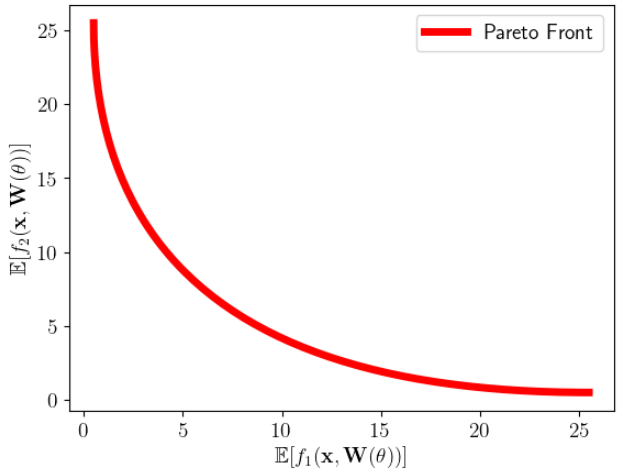


Figure 6: Pareto front for example problem in Objective space

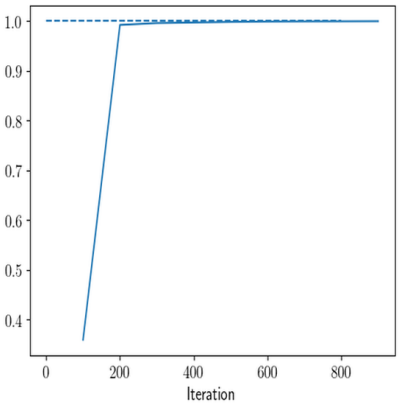
6.1 2D Gaussian

We now turn our attention fully to the quadratic gaussian example in two dimensions. Particularly, we aim to assess our approach at estimating the Pareto front in both objective space and design space (the set of Pareto optimal points).

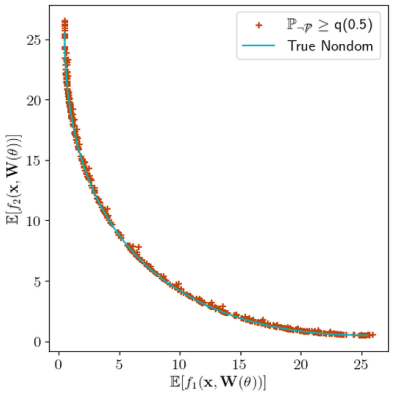
Since we have access to the true expected value of our objectives for this toy problem, we directly compute the true Hypervolume dominated by one run of optimization. Examining the ratio $\frac{\mathcal{HV}_{true}}{\mathcal{HV}_{algo}}$ shown in figure 7a, we can see that the transverse brownian motion algorithm captures 100% of the possible hypervolume.

We also have a set of points estimated to be undominated. Using the true expectations, we can compare the set of probabilistic undomianted points to the true Pareto front as in figure 7b. We can see that points which we predict to be undominated are near the front, meaning that if their true expectations were to be evaluated they would likely be Pareto optimal points. One also sees that the preimage of the set of probabilistically undominated points is centered around the set of Pareto optimal points with low variance.

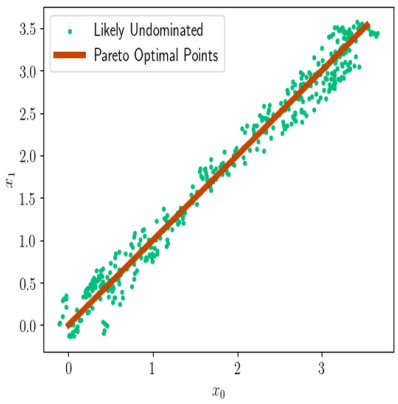
Finally, we turn our attention to the 90% confidence intervals shown in figures 7d and 7e. We can see that they cover both the Pareto front and the set of Pareto optimal points, indicating that we have effectively captured the Pareto front.



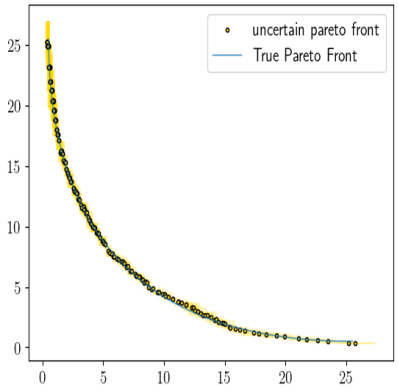
(a) Proportion of the hypervolume captured by the algorithm to the true hypervolume.



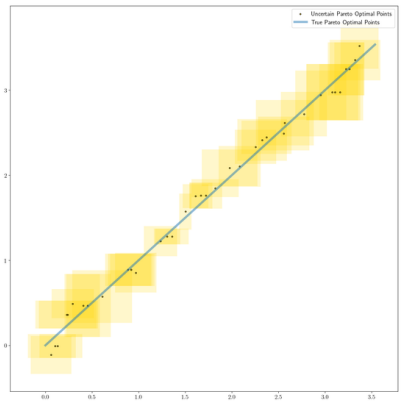
(b) Probability of Being Undominated Captures Pareto front.



(c) Local Mean is Concentrated around True Undominated Points.



(d) 90% Bootstrapped CI for Pareto Front in Objective Space.



(e) 90% Bootstrapped CI for Pareto Optimal Design Points.

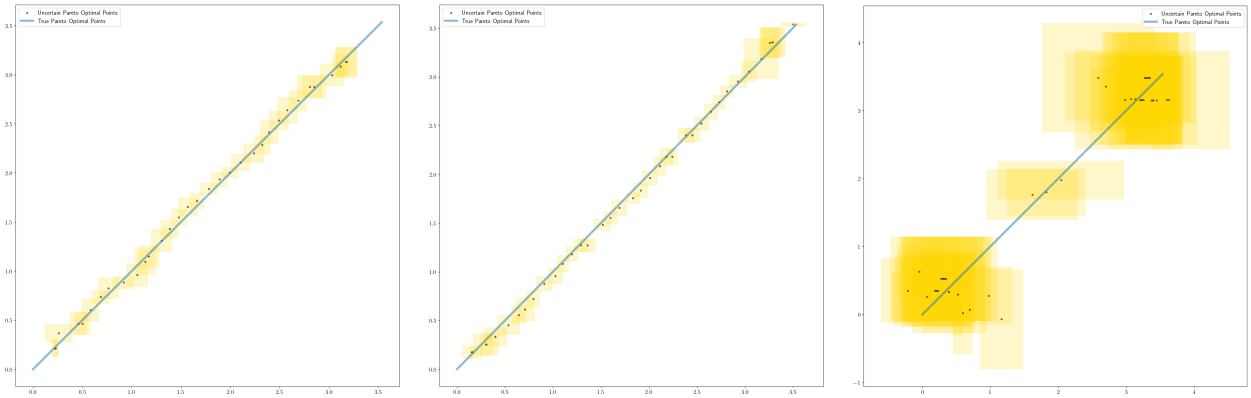
Figure 7: 1000 samples, $\varepsilon_0 = .9$, $\lambda_0 = 1$, and $\varepsilon_\infty = .15$.

6.2 Tradeoff: Resolution and Efficiency

The sampling approach contains four hyperparameters, ε_0, λ_0, γ_t, and ε_∞. The first three control the optimization, particularly how transverse brownian motion iterates approach a minima; however, ε_∞ has a direct effect on the sampling efficiency of our algorithm. It is therefore important to see how the value of ε_∞ changes our ability to infer design points. Specifically, given some value of ε_∞, we would like to know the probability that our design point, $\mathbf{x}$ will be further than $\epsilon > 0$ from the closest point in the set of Pareto optimal points, $\mathbf{x}^* = \text{argmin}_{\mathbf{x}^* \in \mathcal{P}} ||\mathbf{x} - \mathbf{x}^*||^2$, e.g. $\mathbb{P}_\epsilon(||\mathbf{x} - \mathbf{x}^*|| \geq \epsilon | \varepsilon_\infty)$. This can be viewed as a type of concentration about the set of Pareto optimal points. Since we want to know the value for all epsilon, we define the error as the measurement

$$E_{\varepsilon_\infty} = 1 - \int_0^\infty \mathbb{P}_\epsilon(||\mathbf{x} - \mathbf{x}^*|| \geq \epsilon) d\epsilon \tag{61}$$

As we can see in the set of figures 8 showing estimates of the set of design variables given 1000 samples, as the value of the parameter ε_∞ increases, we increase the coverage of design space. However, this coverage comes at a cost to concentration about the set of true design points, as the error measure also increases. There are still no silver-bullets.

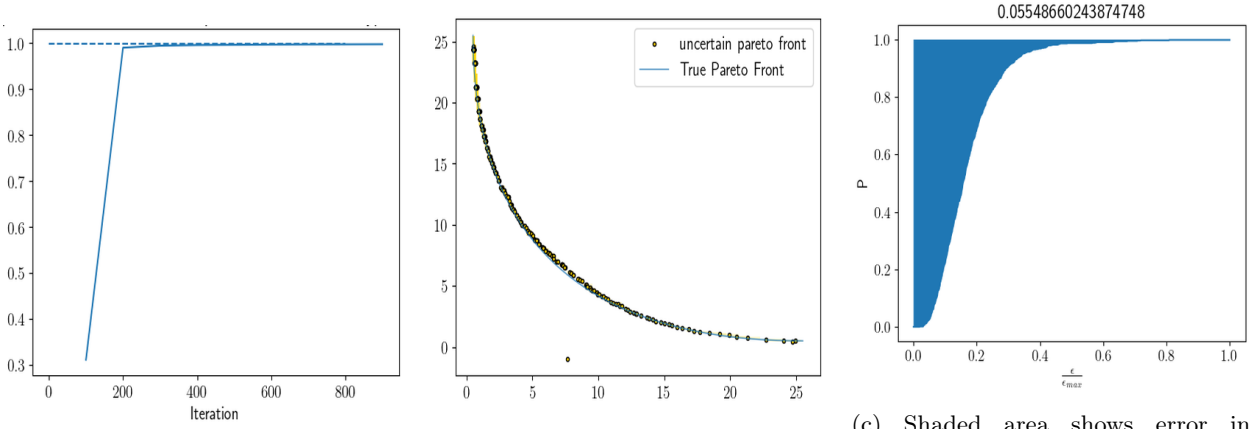


(a) $\varepsilon_\infty = 0.01$, $E_{\varepsilon_\infty} = .013$ (b) $\varepsilon_\infty = 0.1$, $E_{\varepsilon_\infty} = .032$ (c) $\varepsilon_\infty = 0.5$, $E_{\varepsilon_\infty} = .376$

Figure 8

6.3 Revisiting the Quadratic Gaussian Example in HD

We can now examine the quadratic gaussian problem in higher dimensions first tackled in section 6.1. Keeping all hyperparameters the same but increasing the dimension to 10 yields the convergence statistics shown in figure 9. Since the dimension is high, it becomes prohibitive to show the full set of 55 comparisons between all design dimensions, so the confidence intervals for the set of Pareto optimal points are relegated to the appendix. We can see, however, that not only does the true hypervolume of the samples generated by transverse brownian motion converge, but the Pareto front is estimated effectively with only 1000 evaluations of the stochastic objectives.



(a) Hypervolume Converges

(b) 90% Confidence Bands around estimated Pareto front cover the front.

(c) Shaded area shows error in concentration measurement, $E_{\lambda_1} = 0.055$.

Figure 9

7 Transverse Brownian Motion Reaches All Points on the Pareto Front

We are particularly interested in the ability of our approach to reach any point on the Pareto front in finite time. Similar to the quantity introduced in section 6.2, we want to show that, for *any* point on the Pareto front $\mathbf{y}^* \in \mathcal{P}$ and $\epsilon > 0$, we have

Theorem 6. *for any* $\mathbf{y}^* \in \mathcal{P}$ *and some* $t < \infty$

$$\mathbb{P}(||\mathbf{x}_t - \mathbf{y}^*|| \leq \epsilon) > 0. \tag{62}$$

As long as this quantity is strictly greater than zero then transverse brownian motion will eventually reach every point on the Pareto front. Our framework for proving theorem 6 for transverse brownian motion

follows from that used by [10], [3] and [12], we will first show recurrence, that there is a compact sublevel set, $\Phi_G(\mathbf{x}) \leq M$, that is reached infinitely many times by transverse brownian motion. We will 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 transverse brownian motion reaches an arbitrary point $\mathbf{y}^*$ in finite time with nonzero probability. Since $\mathbf{y}^*$ is arbitrary, the result follows.

We will make two assumptions.

Assumption 7.1. there are constants a and b such that

$$||\nabla\Phi_G(\mathbf{x})||^2 \geq a\Phi_G(\mathbf{x}) - b, \qquad ||\mathbf{x}||^2 \leq a\Phi_G(\mathbf{x}) + b \quad \forall\mathbf{x}. \tag{63}$$

Assumption 7.2. There are constants $c \leq 1$ and d such that

$$|\,\mathbb{E}[\langle\nabla\Phi_G(()\mathbf{x}), \nabla\Phi_F(()\mathbf{x}) - \nabla\Phi_G(\mathbf{x})\rangle] \leq c||\nabla\Phi_G(()\mathbf{x})||^2 + d. \tag{64}$$

Assumption 7.1 is necessary to show the ergodicity of the SDE $dX = \nabla F(X) + dW$ and is mild if intimidating. It corresponds to a type of coerciveness, the further away $\mathbf{x}$ is from the minima, the stronger the gradient. The second assumption is a mild one, we only require that the bias is a linear function of the gradient, note also that $\langle\nabla\Phi_F(\mathbf{x}), \Phi_G(\mathbf{x})\rangle = cos(\theta)|\nabla\Phi_F(\mathbf{x})||\nabla\Phi_G(\mathbf{x})| \leq |\nabla\Phi_F(\mathbf{x})||\nabla\Phi_G(\mathbf{x})|$ Note that assumptions 7.2 and 7.1 are automatically met if $\mathbf{x}$ is confined to a bounded subset ($\mathbf{x} \in \mathcal{X} \subset \mathbb{R}^d$).

We are now able to prove our first result, recurrence.

Lemma 7. *Recurrent visits to the sublevel set M.*
Under assumptions **??**, *7.1, and 7.2 and defining a decreasing sequence $\{\varepsilon_t\}_{t\in\mathbb{N}}$ such that $\varepsilon_0 \geq \varepsilon_1, ..., \geq \varepsilon_\infty > 0$. Defining a sublevel set $M > 0$, a constant $B := L\beta^{-1}d + b + \frac{\varepsilon_0 L M_{V_0}}{2} + d \leq a \inf\Phi_G(\mathbf{x})$, and a sequence of stopping times $\tau_{k+1} = \inf\{t : t > \tau_k, \Phi_G(\mathbf{x}_t\} \leq M)$ then*

$$a) \quad \tau_0 = \frac{\log\left[\left[\frac{\Phi_G(\mathbf{x}_0)}{M - \frac{\varepsilon_0 B}{\varepsilon_\infty a}}\right]^2\right]}{2a\varepsilon_\infty} \tag{65}$$

and

$$b) \quad \mathbb{E}[\tau_j] = \tau_0 + (j+1)M \tag{66}$$

Proof of a). Using lipschitz continuity of $\Phi_G(\mathbf{x})$ and the definition of transverse brownian motion, and

defining $\zeta_t = \langle \nabla \Phi_G((\mathbf{x}_t), \nabla \Phi_F((\mathbf{x}_t) - \nabla \Phi_G((\mathbf{x}_t) \rangle$

$$\Phi_G(\mathbf{x}_{t+1}) \leq \Phi_G(\mathbf{x}_t) + \langle \nabla \Phi_G(\mathbf{x}_t), \mathbf{x}_{t+1} - \mathbf{x}_t \rangle + \frac{L}{2} ||\mathbf{x}_{t+1} - \mathbf{x}_t||^2 \tag{67}$$

$$\leq \Phi_G(\mathbf{x}_t) - \varepsilon_t ||\nabla \Phi_G(\mathbf{x}_t)||^2 - \varepsilon_t \zeta_t + \frac{L}{2} \big(M_{V_0} + M_V ||\nabla \Phi_G(\mathbf{x}_t)||^2 \big) \tag{68}$$

$$+ \sqrt{2\varepsilon_t \beta^{-1}} \langle \nabla \Phi_G(\mathbf{x}_t), T_t \rangle + \frac{L}{2} \big(2\sqrt{2\varepsilon_t \beta^{-1}} \langle \nabla \Phi_F(\mathbf{x}_t), T_t \rangle + 2\varepsilon_t \beta^{-1} \langle T_t, T_t \rangle \big) \tag{69}$$

$$\mathbb{E}_t[\Phi_G(\mathbf{x}_{t+1})] \leq \Phi_G(\mathbf{x}_t) - \varepsilon_t (1 - c - \frac{\varepsilon_t M_v}{2}) ||\nabla \Phi_G(\mathbf{x}_t)||^2 + \varepsilon_t \big(d + \frac{L\varepsilon_t}{2} M_{V_0} + L\beta^{-1} d \big) \tag{70}$$

$$\leq \Phi_G(\mathbf{x}_t) - \varepsilon_t ||\nabla \Phi_G(\mathbf{x}_t)||^2 + \varepsilon_t \big(d + \frac{L\varepsilon_t}{2} M_{V_0} + L\beta^{-1} d \big) \tag{71}$$

$$\leq (1 - a\varepsilon_t) \Phi_G(\mathbf{x}_t) + \varepsilon_t B \tag{72}$$

$$\leq e^{-a\varepsilon_t} \Phi_G(\mathbf{x}_t) + \varepsilon_t B \tag{73}$$

Where we have used the inequality $1 - x \leq e^{-x}$. Taking the full expectation, iterating, and setting the above less than equal to M gives the relation

$$\mathbb{E}[\Phi_G(\mathbf{x}_{\tau_0})] \leq e^{-a\tau_0 \varepsilon_\infty} \Phi_G(\mathbf{x}_0) + \frac{\varepsilon_0 B}{\varepsilon_\infty a} \leq M. \tag{74}$$

Solving for τ_0 gives the result. □

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

$$\Phi_G(\mathbf{x}_{t \wedge \tau_{j+1}}) + t \wedge \tau_{j+1} \tag{75}$$

is a supermartingale with respect to τ_j.

Lemma 8.

$$\Phi_G(\mathbf{x}_{t \wedge \tau_{j+1}}) + t \wedge \tau_{j+1} \tag{76}$$

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

$$\mathbb{E}_{\tau_j}[\Phi_G(\mathbf{x}_{\tau_{j+1}})] \leq \Phi_G(\mathbf{x}_{\tau_j}) \tag{77}$$

Note that

$$\mathbb{E}_{\tau_j}[\Phi_G(\mathbf{x}_{\tau_{j+1}})] \leq (1 - a\varepsilon_{\tau_j}) \Phi_G(\mathbf{x}_{\tau_j}) + \varepsilon_{\tau_j} B \leq M \tag{78}$$

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

$$\mathbb{E}_{\tau_j}[\Phi_G(\mathbf{x}_{\tau_{j+1}})] \leq \Phi_G(\tau_j) \tag{79}$$

And we have the result. □

Having proved lemma 8 we can go on to prove part b of theorem 7

part b). Since

$$\mathbb{E}_{\tau_j}[\Phi_G(\mathbf{x}_{t\wedge\tau_{j+1}})] + t \wedge \tau_{j+1} \leq \Phi_G(t \wedge \tau_j) + t \wedge \tau_j \tag{80}$$

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

$$\mathbb{E}_{\tau_j}[\tau_{j+1}] \leq \mathbb{E}_{\tau_j}[\Phi_G(\mathbf{x}_{\tau_{j+1}})] + \tau_{j+1} \leq \Phi_G(\mathbf{x}_{\tau_j}) + \tau_j \tag{81}$$

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

$$\mathbb{E}[\tau_{j+1}] \leq (j+1)M + \tau_0 \tag{82}$$

□

Having shown recurrence, we can now prove reachability.

Lemma 9. *Given that $\Phi_G(\mathbf{x}_{\tau_k}) \leq M$, let $\mathbb{E}[\nabla\Phi_F(\mathbf{x})] \leq D$. We have, for $\mathbf{y}^* \in \{\mathbf{y} : \Phi_G(\mathbf{y}) \leq M\} \cap \{\mathbf{y} : \mathbf{y} \in \mathcal{P}(G\mathcal{X})\}$*

$$P(||\mathbf{x}_{\tau_j+t} - \mathbf{y}^*||^2 \leq \epsilon) > 0 \tag{83}$$

Proof. for a finite t we have the events

$$\mathcal{A} = ||\mathbf{x}_{\tau_j+t} - \mathbf{y}^*||^2 \leq \epsilon \qquad \mathcal{B} = ||\mathbf{x}_{\tau_j} - \mathbf{y}^* + \sum_{s=\tau_j}^{\tau_j+t-1} \varepsilon_s D + \sum_{s=\tau_j}^{\tau_j+t-1} \sqrt{2\varepsilon_s\beta^{-1}} Z_s||^2 \leq \epsilon \tag{84}$$

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 transverse brownian

motion, we have

$$\mathbf{x}_{\tau_j+t} = \mathbf{x}_{\tau_j} - \sum_{s=\tau_j}^{\tau_j+t-1} \nabla_x^{\mathcal{C}}\{F\}(\mathbf{x}_s) + \sum_{s=\tau_j}^{\tau_j+t-1} \sqrt{2\varepsilon_s\beta^{-1}}T_s \tag{85}$$

$$= \mathbf{x}_{\tau_j} - \sum_{s=\tau_j}^{\tau_j+t-1} \nabla_x^{\mathcal{C}}\{F\}(\mathbf{x}_s) + \sum_{s=\tau_j}^{\tau_j+t-1} \sqrt{2\varepsilon_s\beta^{-1}}P_s Z_s \tag{86}$$

$$= \mathbf{x}_{\tau_j} + \sum_{s=\tau_j}^{\tau_j+t-1} D + \sum_{s=\tau_j}^{\tau_j+t-1} \sqrt{2\varepsilon_s\beta^{-1}}Z_s \tag{87}$$

$$\tag{88}$$

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

$$||\mathbf{x}_{\tau_j+t} - \mathbf{y}||^2 \leq ||\mathbf{x}_{\tau_j} - \mathbf{y}^* + \sum_{s=\tau_j}^{\tau_j+t-1} D + \sum_{s=\tau_j}^{\tau_j+t-1} \sqrt{2\varepsilon_s\beta^{-1}}Z_s||^2 \tag{89}$$

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, one can say that $\mathbb{P}(\mathcal{A}) \geq \mathbb{P}(\mathcal{B}) = c_t > 0$. □

We are now in position to prove theorem 6

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

$$\mathbb{P}(\tau^* \geq T) = \mathbb{P}(\tau^* \geq T\tau_j \geq T) + \mathbb{P}(\tau^* \geq T\tau_j \leq T) \tag{90}$$

$$= \mathbb{P}(\tau_j \geq T) + \mathbb{P}(\tau^* \geq T, ||\mathbf{x}_{\tau_k} - \mathbf{y}^*|| \geq \epsilon \; \forall k = \{1, ..., j\}) \tag{91}$$

$$\leq \mathbb{E}[\tau_j] + \prod_{k=1}^{j} \mathbb{P}(||\mathbf{x}_{\tau_k} - \mathbf{y}^*|| \geq \epsilon) \tag{92}$$

$$\leq \frac{(j+1)M + \tau_0}{T} + \prod_{k=1}^{j}(1 - c_j) \tag{93}$$

Where we have used our proof of recurrence in the third line, and reachability in the fourth. We can always choose j and T large enough such that this probability vanishes. □

8 Conclusion

We have introduced a new approach to solving multi-objective optimization problems in which the whole of the pareto front is of interest, transverse brownian motion. We have seen its effectiveness in estimating

both the Pareto front and Pareto optimal points on a problem which could not be effectively treated with its predecessor algorithm, SMGDA, which has a bias that leads it to converge to a subset of the Pareto front. We have also proven, under mild assumptions, that transverse brownian motion samples the whole of the Pareto front in finite time, making it reliably usable in multi-objective optimization problems.

References

[1] Sanghamitra Bandyopadhyay et al. "A Simulated Annealing-Based Multiobjective Optimization Algorithm: AMOSA". In: *IEEE Transactions on Evolutionary Computation* 12.3 (2008), pp. 269–283. DOI: 10.1109/TEVC.2007.900837.

[2] Léon Bottou, Frank E. Curtis, and Jorge Nocedal. *Optimization Methods for Large-Scale Machine Learning*. 2018. arXiv: 1606.04838 [stat.ML].

[3] Xi Chen, Simon S. Du, and Xin T. Tong. *On Stationary-Point Hitting Time and Ergodicity of Stochastic Gradient Langevin Dynamics*. 2020. arXiv: 1904.13016 [stat.ML].

[4] Thomas M. Cover and Peter E. Hart. "Nearest neighbor pattern classification". In: *IEEE Trans. Inf. Theory* 13 (1967), pp. 21–27. URL: https://api.semanticscholar.org/CorpusID:5246200.

[5] K. Deb et al. "A fast and elitist multiobjective genetic algorithm: NSGA-II". In: *IEEE Transactions on Evolutionary Computation* 6.2 (2002), pp. 182–197. DOI: 10.1109/4235.996017.

[6] Jean-Antoine Désidéri. "Multiple-gradient descent algorithm (MGDA) for multiobjective optimization". In: *Comptes Rendus Mathematique* 350.5 (2012), pp. 313–318. ISSN: 1631-073X. DOI: https://doi.org/10.1016/j.crma.2012.03.014. URL: https://www.sciencedirect.com/science/article/pii/S1631073X12000738.

[7] Zeou Hu et al. "Federated Learning Meets Multi-Objective Optimization". In: *IEEE Transactions on Network Science and Engineering* 9.4 (2022), pp. 2039–2051. DOI: 10.1109/TNSE.2022.3169117.

[8] Suyun Liu and Luis Nunes Vicente. *The stochastic multi-gradient algorithm for multi-objective optimization and its application to supervised machine learning*. 2021. arXiv: 1907.04472 [math.NA].

[9] Marko M. Mäkelä, Ville-Pekka Eronen, and Napsu Karmitsa. "On Nonsmooth Multiobjective Optimality Conditions with Generalized Convexities". In: *Optimization in Science and Engineering: In Honor of the 60th Birthday of Panos M. Pardalos*. Ed. by Themistocles M. Rassias, Christodoulos A. Floudas, and Sergiy Butenko. New York, NY: Springer New York, 2014, pp. 333–357. ISBN: 978-1-4939-0808-0. DOI: 10.1007/978-1-4939-0808-0_17. URL: https://doi.org/10.1007/978-1-4939-0808-0_17.

[10] J.C. Mattingly, A.M. Stuart, and D.J. Higham. "Ergodicity for SDEs and approximations: locally Lipschitz vector fields and degenerate noise". In: *Stochastic Processes and their Applications* 101.2 (2002), pp. 185–232. ISSN: 0304-4149. DOI: https://doi.org/10.1016/S0304-4149(02)00150-3. URL: https://www.sciencedirect.com/science/article/pii/S0304414902001503.

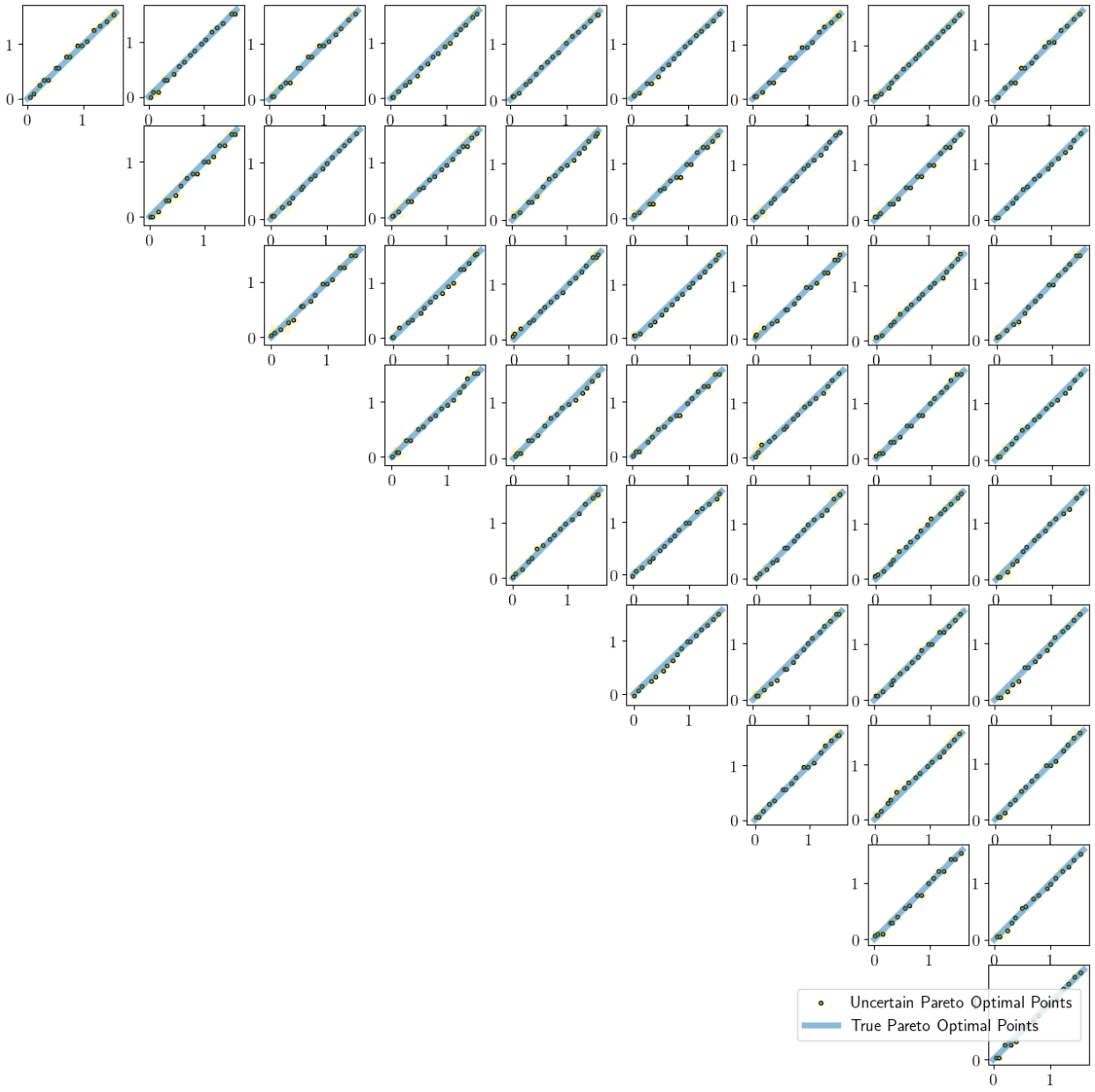
[11] Quentin Mercier. "Optimisation multicritère sous incertitudes : un algorithme de descente stochastique". Theses. COMUE Université Côte d'Azur (2015 - 2019), Oct. 2018. URL: https://theses.hal.science/tel-02063322.

[12] Sean Meyn and R. Tweedie. *Markov Chains and Stochastic Stability*. Vol. 92. Jan. 1993. DOI: 10.2307/2965732.

[13] Mickael Rivier, Nassim Razaaly, and Pietro Marco Congedo. "Non-Parametric Measure Approximations for Constrained Multi-Objective Optimisation under Uncertainty". working paper or preprint. Sept. 2022. URL: https://inria.hal.science/hal-03781832.

A Appendex A: Pareto Optimal Points for 10 dimensional Quadratic Gaussian Example

Confidence intervals for Pareto optimal points for the high dimensional Gaussian example.


Uncertain Pareto Optimal Points
True Pareto Optimal Points

B Appendix B: Convergence Results for SMGDA

We collect here a series of convergence results for SMGDA which serve to round out the literature. Particularly, exploiting the potential function developed in the previous secitons we can show convergence results using weaker assumptions than in [11] and previously explored in prior works.

B.1 Strongly Convex Case

For a set of k objective functions $\{f_i(\mathbf{x}, \mathbf{W}(\theta))\}_{i=1,...,k}$ such that $\mathbb{E}[f_i(\mathbf{x}, \mathbf{W}(\theta))] = g_i(\mathbf{x})$ with $\{g_i(\mathbf{x})\}$ – $i = 1, ..., k$ convex and at least one g_i strongly convex and lipsthitz continuous We will show that, given a set of positive real numbers, $\{varepsilon_t\}_{t\in\mathbb{N}}$ such that

$$\sum_t \varepsilon_t = \infty \qquad\qquad \sum_t \varepsilon_t^2 < \infty, \tag{94}$$

a sequence of the form:

$$\mathbf{x}_{t+1} = \mathbf{x}_t - \varepsilon_t \nabla_x^{\mathcal{C}}\{F\}(\mathbf{x}_t) \tag{95}$$

converges to a point on the pareto front.

First we will prove a pair of lemmas which will be useful for the result. Let $\mathbf{x}_t^*$ be the projection of a point $\mathbf{x}_t$ to the pareto front, that is

$$\mathbf{x}_t^* = \underset{y\in\mathcal{P}}{\operatorname{argmin}} ||\mathbf{x}_t - y||^2. \tag{96}$$

We have the relation that

Lemma 10.

$$\langle \nabla\Phi_G(\mathbf{x}), \mathbf{x} - \mathbf{x}^*\rangle \geq \sum_i \alpha_i^*(\mathbf{x}) m_i ||\mathbf{x} - \mathbf{x}^*||^2 \tag{97}$$

where $m_i \geq 0$

and

Lemma 11.

$$\mathbb{E}[\nabla\Phi_F(\mathbf{x}) - \nabla\Phi_G(\mathbf{x})] = \sum_i \big(\mathbb{E}[\alpha_i^*(\mathbf{x}, \mathbf{W}(\theta))] - \alpha_i^*(\mathbf{x})\big)\nabla g_i(\mathbf{x}) \tag{98}$$

Proof of 10. From (strong) convexity we have that

$$\langle \nabla g_i(\mathbf{x}), \mathbf{x} - \mathbf{x}^*\rangle \geq m_i ||\mathbf{x} - \mathbf{x}^*||^2 \tag{99}$$

with $m_i = 0$ for g_i convex and $m_i > 0$ for g_i strongly convex. Multiply both sides by $\alpha_i^*(\mathbf{x})$ and sum to see

that

$$\langle \alpha^*(\mathbf{x}) \nabla g_i(\mathbf{x}), \mathbf{x} - \mathbf{x}^* \rangle \geq \sum_i m_i \alpha_i^*(\mathbf{x}) ||\mathbf{x} - \mathbf{x}^*||^2. \tag{100}$$

Using the definition of $\nabla \Phi_G(\mathbf{x})$ gives the result. □

Proof of 11.

$$\begin{aligned}
\nabla \Phi_F(\mathbf{x}) - \nabla \Phi_G(\mathbf{x}) &= \mathbb{E}[\nabla_x^{\mathcal{C}}\{F\}(\mathbf{x})] - \nabla_x^{\mathcal{C}}\{G\}(\mathbf{x}) && (101)\\
&= \sum_i \mathbb{E}[\alpha^*(\mathbf{x}, \mathbf{W}(\theta))(\nabla f_i(\mathbf{x}) - \nabla g_i(\mathbf{x})) + (\alpha^*(\mathbf{x}, \mathbf{W}(\theta)) - \alpha^*(\mathbf{x}))\nabla g_i(\mathbf{x})] && (102)\\
&= \sum_i (\mathbb{E}[\alpha_i^*(\mathbf{x}, \mathbf{W}(\theta))] - \alpha_i^*(\mathbf{x}))\nabla g_i(\mathbf{x}) && (103)
\end{aligned}$$

□

We can now prove the theorem.

Theorem 12. *Given a set of k objectives $\{f(\mathbf{x}, \mathbf{W}(\theta))\}_{i=1,\dots,k}$ such that all member of the set $\{\mathbb{E}[f_i(\mathbf{x}, \mathbf{W}(\theta))]\}_{i=1,\dots,k}$ are convex and M-lipschitz continuous and defining a set of real numbers $\{\varepsilon_t\}_{t\in\mathbb{N}}$ meeting the criteria stated in 94, Then the SMGD algorithm converges to a point on the Pareto front.*

$$||\mathbf{x}_{T+1} - \mathbf{x}_{T+1}^*||^2 \xrightarrow[\infty]{T} 0 \tag{104}$$

Proof. As a shorthand notation, let $d_t^2 := ||\mathbf{x}_t - \mathbf{x}_t^*||^2$, $B_t := \nabla \Phi_F(\mathbf{x}_t) - \nabla \Phi_G(\mathbf{x}_t)$, and $\overline{m}_\alpha = \sum_i \alpha_i^*(\mathbf{x}, \mathbf{W}(\theta)) m_i$.

$$\begin{aligned}
||\mathbf{x}_{t+1} - \mathbf{x}_{t+1}^*||^2 \leq & ||\mathbf{x}_{t+1} - \mathbf{x}_t^*||^2 = ||\mathbf{x}_t - \varepsilon_t \nabla_x^{\mathcal{C}}\{F\}(\mathbf{x}) - \mathbf{x}_t^*||^2 && (105)\\
\mathbb{E}_t[d_{t+1}^2] &= d_t^2 - 2\varepsilon_t \langle \Phi_G(\mathbf{x}_t), \mathbf{x}_t - \mathbf{x}_t^* \rangle - 2\varepsilon_t \langle B_t, d_t \rangle + \varepsilon_t^2 \, \mathbb{E}[||\nabla_x^{\mathcal{C}}\{F\}(\mathbf{x})||^2] && (106)\\
&= d_t^2 - 2\varepsilon_t \big(\sum_i \alpha_i^*(\mathbf{x}, \mathbf{W}(\theta)) m_i \big) d_t^2 + \varepsilon_t^2 M^2 && (107)\\
&= (1 - 2\varepsilon_t \overline{m}_\alpha) d_t^2 + \varepsilon_t^2 M^2 && (108)
\end{aligned}$$

Where in the third line we have used lemmas 11 and 10 in addition to the M-lipschitz continuity of $\Phi_F(\mathbf{x})$. Picking $\varepsilon_t \leq \varepsilon_0 \leq \frac{1}{2\overline{m}_\alpha}$ and defining $\pi_t = \prod_{i=1}^t (1 - 2\varepsilon_i \overline{m}_\alpha)$, taking the full expectation, and iterating, we have

$$\mathbb{E}[d_T^2] = \pi_t d_0^2 + \sum_{s=0}^{t} \frac{\pi_t}{\pi_s} \varepsilon_s^2 M^2 \xrightarrow[\infty]{T} 0 \tag{109}$$

□

B.2 Online Convex Multi-Gradient and Online Stochastic Convex Multi-Gradient

Using the fact that $\Phi_F(\mathbf{x})$ is convex, as shown in section 4, we can show convergence in the online setting. Assume that $||\mathbf{x} - \mathbf{x}^*|| \leq D$ and $\Phi_F()$ is lipschitz continuous.

Theorem 13. *Given a set of k stochastic functions $F := \{f_i(\mathbf{x}, \mathbf{W}(\theta))\}_{i=1,\dots,k}$ such that, for each i, $\mathbb{E}[f_i(\mathbf{x}, \mathbf{W}(\theta))]$ is convex and M-lipschitz continuous over a bounded set $||\mathbf{x} - \mathbf{x}^*|| \leq D$, we have a sublinear regret in $\Phi_F(\mathbf{x})$ with respect to its minimizer on the Pareto front $\mathbf{x}^* \in \operatorname{argmin}_{\mathbf{x}} \Phi_F(\mathbf{x})$.*

$$\sum_{t=1}^{T} \big(\Phi_F(\mathbf{x}_t) - \Phi_F(\mathbf{x}^*)\big) \leq \frac{3DM\sqrt{T}}{2} \tag{110}$$

Proof.

$$||\mathbf{x}_{t+1} - \mathbf{x}^*||^2 = ||\mathbf{x}_t - \mathbf{x}^* - \varepsilon_t \nabla_x^{\mathcal{C}}\{F\}(\mathbf{x}_t)||^2 \tag{111}$$

$$= ||\mathbf{x}_t - \mathbf{x}^*||^2 - 2\varepsilon_t \langle \nabla_x^{\mathcal{C}}\{F\}(\mathbf{x}_t), \mathbf{x} - \mathbf{x}^* \rangle + \varepsilon_t^2 ||\nabla_x^{\mathcal{C}}\{F\}(\mathbf{x}_t)||^2 \tag{112}$$

$$\mathbb{E}_t[||\mathbf{x}_{t+1} - \mathbf{x}^*||^2] \leq ||\mathbf{x}_t - \mathbf{x}^*||^2 - 2\varepsilon_t \big(\Phi_F(\mathbf{x}_t) - \Phi_F(\mathbf{x}^*)\big) + \varepsilon_t^2 M^2 \tag{113}$$

$$\tag{114}$$

Taking the full expectation, rearranging, summing, and setting $\varepsilon_t = \frac{D}{M\sqrt{t}}$ with $\varepsilon_0 := 0$ we see that

$$\sum_{t=1}^{T} \big(\Phi_F(\mathbf{x}_t) - \Phi_F(\mathbf{x}^*)\big) \leq \sum_{t=1}^{T} \frac{||\mathbf{x}_t - \mathbf{x}^*||^2}{2} \big(\frac{1}{\varepsilon_{t+1}} - \frac{1}{\varepsilon_t}\big) + \sum_{t=1}^{T} \varepsilon_t M^2 \tag{115}$$

$$\leq \frac{D^2}{2} \frac{1}{\varepsilon_T} + M^2\sqrt{T} \tag{116}$$

$$\leq \frac{3DM\sqrt{T}}{2} \tag{117}$$

Where we have used the inequality $\int_0^T \frac{1}{\sqrt{t}} \leq 2\sqrt{T}$ □

An important distinction must be made to minimizers $x_1^* \in \operatorname{argmin}_{\mathbf{x}} \Phi_F(\mathbf{x})$ and $x_2^* \in \operatorname{argmin}_{\mathbf{x}} \Phi_G(\mathbf{x})$. While this proof works in both the stochastic and deterministic cases, functions F and functions G, the minimizers of the stochastic problem are a subset of the deterministic problem solved using exact gradients.