Proceedings of the 2025 Winter Simulation Conference
E. Azar, A. Djanatliev, A. Harper, C. Kogler, V. Ramamohan, A. Anagnostou, and S. J. E. Taylor, eds.

THE DERIVATIVE-FREE FULLY-CORRECTIVE FRANK-WOLFE ALGORITHM FOR OPTIMIZING FUNCTIONALS OVER PROBABILITY SPACES

Di Yu[1], Shane G. Henderson[2], and Raghu Pasupathy[1]

[1]Dept. of Statistics, Purdue University, West Lafayette, IN, USA
[2]Operations Research and Information Eng., Cornell University, Ithaca, NY, USA

ABSTRACT

The challenge of optimizing a smooth convex functional over probability spaces is highly relevant in experimental design, emergency response, variations of the problem of moments, etc. A viable and provably efficient solver is the fully-corrective Frank-Wolfe (FCFW) algorithm. We propose an FCFW recursion that rigorously handles the zero-order setting, where the derivative of the objective is known to exist, but only the objective is observable. Central to our proposal is an estimator for the objective's *influence function*, which gives, roughly speaking, the directional derivative of the objective function in the direction of point mass probability distributions, constructed via a combination of Monte Carlo, and a projection onto the orthonormal expansion of an L_2 function on a compact set. A bias-variance analysis of the influence function estimator guides step size and Monte Carlo sample size choice, and helps characterize the recursive rate behavior on smooth non-convex problems.

1 INTRODUCTION

The challenge of optimizing a smooth functional on the space of compactly supported probability measures is stated as follows:

$$\begin{aligned} \text{min.} \quad & J(\mu) \\ \text{s.t.} \quad & \mu \in \mathscr{P}(\mathscr{X}), \end{aligned} \tag{P}$$

where $\mathscr{X} \subset \mathbb{R}^d$ is compact, $\mathscr{P}(\mathscr{X})$ is the space of probability measures supported on $\mathscr{X}$, and $J : \mathscr{P}(\mathscr{X}) \to \mathbb{R}$ is a smooth functional. The problem (P) has recently received a lot of attention on account of its direct applicability in various contexts. Of particular interest in this paper is the frequently encountered *zero-order* setting, also known as the *derivative-free* setting, where J is smooth in a sense to be defined in Section 1.2, $J(\mu) \in \mathbb{R}$ is observable at any $\mu \in \mathscr{P}(\mathscr{X})$, but its derivative $J'_\mu(\nu - \mu)$ (to be defined precisely in Section 1.2) is not directly observable. The derivative operator $J'_\mu(\cdot - \mu)$, a crucial aid in the search for a critical point associated with J, will therefore have to be appropriately estimated as the search evolves. The analogue of this setting in the Euclidean context is widely understood to be useful — see Shashaani et al. (2018), Conn et al. (2009). Correctly converging iterative algorithms to solve (P) generate a sequence $\mu_k \subseteq \mathscr{P}(\mathscr{X})$ of probability measures that satisfy a reasonable criterion such as $J(\mu_k) \to \inf\{J(\mu), \mu \in \mathscr{P}(\mathscr{X})\}$ as $k \to \infty$ in some appropriate sense, assuming the infimum is finite.

1.1 Illustrative Examples

The question in (P) arises in numerous and varied contexts. In what follows, we outline three examples as illustration.

Example 1 (Experimental Design) Consider the question of how best to sample points from a set $\mathscr{X} \subseteq \mathbb{R}^d$ in the service of estimating the parameter vector $\beta^* \in \mathbb{R}^d$ of a *parametric response surface*

$$Y(x) = f(x, \beta^*) + \varepsilon(x),$$

979-8-3315-8726-0/25/$31.00 ©2025 IEEE

where $Y(x)$ is the *response* at x, $f(\cdot,\beta^*):\mathbb{R}^d\to\mathbb{R}$ is some response surface, e.g., a Scheffé polynomial (Cornell 2011; Draper and Pukelsheim 1999), a power-mean-mixture model (Coetzer and Focke 2010), or a model using Padé approximants (Focke et al. 2021), and $\varepsilon(x)$ satisfies $\mathbb{E}[\varepsilon(x)]=0, \mathrm{var}[\varepsilon(x)]=\sigma^2<\infty$. To pose the problem precisely, suppose we obtain observations $Y_1,Y_2,\ldots,Y_n$ at the locations $X_1,X_2,\ldots,X_n\overset{\text{iid}}{\sim}\mu\in\mathscr{P}(\mathscr{X})$, and construct the "least-squares" estimator $\hat{\beta}_n$ of β^*, that is,

$$\hat{\beta}_n := \arg\min\left\{\sum_{i=1}^{n}(Y_i-f(X_i,\beta))^2 : \beta\in\mathbb{R}^d\right\}.$$

The experimental design question is then that of identifying a measure $\mu\in\mathscr{P}$ that optimizes a performance measure associated with $\hat{\beta}_n$. For example, A-optimal design in classical statistics selects the trace of the covariance of matrix as the performance measure, $J(\mu):=\mathrm{tr}\left(\mathrm{cov}(\hat{\beta}_n)\right)$, whereas D-optimal design chooses the determinant of the inverse covariance matrix, $J(\mu):=-\det\left(\mathrm{cov}^{-1}(\hat{\beta}_n)\right)$. Various other performance measures may be of interest — see, for instance, Atkinson et al. (2007).

If J and its derivative J'_μ (defined in Section 1.2) can be observed without error, then a first-order method such as the fully corrective Frank Wolfe (FCFW) method (detailed in Yu et al. (2024) and Section 2) can be used to minimize J. However, it is sometimes the case that this is not possible because the response surface f has a complicated form and J'_μ, while guaranteed to exist with linear structure (see Definition 2 for details), might be difficult or at least inconvenient to compute or approximate. One might then reasonably seek a Kiefer-Wolfowitz type (Kiefer and Wolfowitz 1952) derivative-free analogue of first-order FCFW in which the derivative J'_μ is estimated upon demand using observations of J at appropriate locations.

Example 2 (Problem of Moments) Let $\mathscr{P}_C(\mathscr{X})$ denote the set of probability measures supported on $\mathscr{X}\subseteq\mathbb{R}^d$ and having a density function, that is, corresponding to each $\mu\in\mathscr{P}_C(\mathscr{X})$ there exists a Lebesgue integrable function $f_\mu:\mathbb{R}^d\to[0,\infty)$ such that $\mu(A)=\int_A f_\mu(x)\,dx$ for all (Lebesgue) measurable sets $A\in\mathscr{L}(\mathbb{R}^d)$. For $\mathscr{X}=[a,b]$ being some finite interval, it is well-known that the uniform probability measure maximizes entropy, that is, $\mu^*(A)=\int_{A\cap[a,b]}dx/(b-a)$ solves the optimization problem

$$\begin{aligned}\min.\quad & H(\mu):=\int_{\mathbb{R}^d}f_\mu(x)\log f_\mu(x)dx\\ \text{s.t.}\quad & \mu\in\mathscr{P}_C(\mathscr{X}),\quad \mathscr{X}=[a,b].\end{aligned}\tag{1}$$

Similarly, the exponential distribution with parameter λ has the largest entropy among all continuous distributions supported on $\mathbb{R}$ and constrained to have mean λ^{-1}, that is, the measure $\mu^*(A)=\int_{A\cap[0,\infty)}\lambda\exp\{-\lambda x\}\,dx$, for all $A\in\mathscr{L}(\mathbb{R})$ solves the optimization problem

$$\begin{aligned}\min.\quad & H(\mu):=\int_{\mathbb{R}}f_\mu(x)\log f_\mu(x)dx\\ \text{s.t.}\quad & \int_{\mathbb{R}}x\,d\mu(x)=\lambda^{-1};\quad \mu\in\mathscr{P}_C(\mathscr{X}),\quad \mathscr{X}=[0,\infty).\end{aligned}\tag{2}$$

A third classical result asserts that the normal distribution with mean zero and variance σ^2 has the maximum entropy among all continuous distributions supported on $\mathbb{R}$ and having specified variance $\sigma^2<\infty$.

As a generalization of (1) and (2), we might consider functionals other than entropy and ask which among the probability measures in $\mathscr{P}(\mathscr{X})$ minimizes a specified statistical functional $J:\mathscr{P}(\mathscr{X})\to\mathbb{R}$, while subject to a finite set of linear (in μ) constraints, that is,

$$\begin{aligned}\min.\quad & J(\mu)\\ \text{s.t.}\quad & A\mu=b;\quad \mu\in\mathscr{P}(\mathscr{X}),\end{aligned}\tag{3}$$

where $A : \mathscr{P}(\mathscr{X}) \to \mathbb{R}^c$ is a linear operator, and $b \in \mathbb{R}^c$. Depending on the nature of the statistical functional J, only an algorithmic solution to (3) may be possible in the sense that no "named distribution" might solve (3). Furthermore, while J might have a derivative J'_μ with linear structure, its calculation may be too challenging or inconvenient, calling for a derivative-free method such as the one we outline in this paper. Admittedly, however, the derivative-free method we propose here does not treat constraints, and needs to be augmented appropriately to solve (3).

Example 3 (P-means Problem) The P-means problem (Molchanov and Zuyev 2002; Okabe, Boots, Sugihara, and Chiu 2000), often described as the randomized version of the k-means clustering problem, is stated as follows. Suppose *demand sources* located at $\ell_1, \ell_2, \ldots, \ell_{n_0} \in \mathscr{X} \subset \mathbb{R}^d$ are to be serviced by *responders* located in $\mathscr{X}$, where $\mathscr{X}$ is some known compact set. For randomization, suppose that the responders are located in $\mathscr{X}$ according to a spatial Poisson process X_μ having mean measure μ, with $\mu(\mathscr{X}) = 1$, so that $\mu \in \mathscr{P}(\mathscr{X})$. Also, assume that the i-th demand source is serviced by one of the responders $(X_{\mu,1}, X_{\mu,2}, \ldots, X_{\mu,N})$ of X_μ with incurred cost $c_i(X_\mu)$. In the simplest setting,

$$c_i(X_\mu) := \begin{cases} \min_{j=1,2,\ldots,N} \|X_{\mu,j} - \ell_i\|_2 & N > 0; \\ \operatorname{diam}(\mathscr{X}) = \sup_{x,y\in\mathscr{X}} \|x-y\|_2 & N = 0, \end{cases}$$

but can take a much more complicated form depending on the specific application. The P-means problem, seeking $\mu \in \mathscr{P}(\mathscr{X})$ that minimizes the expected total cost, is then stated as:

$$\begin{aligned} \min. \quad & J(\mu) = \sum_{i=1}^{n_0} \int_0^\infty P_X\left(c_i(X_\mu) > t\right) \mathrm{d}t \\ \text{s.t.} \quad & \mu \in \mathscr{P}(\mathscr{X}). \end{aligned}$$

As in the previous examples, while the derivative J'_μ endowed with linear structure might exist under mild assumptions, it may not be easy to compute, especially if the cost function c_i has a complicated form. Such settings call for derivative-free methods such as the one we outline in this paper.

1.2 Preliminaries

In what follows, we provide definitions of some key mathematical machinery used in the paper.

Definition 1 (Measure, Signed Measure, Probability Measure) Let $(\mathscr{X}, \Sigma)$ be a measurable space. A set function $\mu : \Sigma \to \mathbb{R}^+ \cup \{\infty\}$ is called a *measure* if (a) $\mu(A) \geq 0 \; \forall A \in \Sigma$, (b) $\mu(\emptyset) = 0$, and (c) $\mu\left(\bigcup_{j=1}^\infty \mu(A_i)\right) = \sum_{j=1}^\infty \mu(A_i)$ for a countable collection $\{A_j, j \geq 1\}$ of pairwise disjoint sets in Σ. The set function μ is called a *signed measure* if the non-negativity condition in (a) is dropped and the infinite sum in (c) converges absolutely. It is called a σ*-finite measure* if there exists a countable collection $\{A_j, j \geq 1\}$ such that $\mu(A_j) < \infty, j \geq 1$ and $\bigcup_{j=1}^\infty A_j = \mathscr{X}$, and a *probability measure* if $\mu(\mathscr{X}) = 1$. In the current paper $\mathscr{X} \subseteq \mathbb{R}^d$, $\Sigma \equiv \mathscr{B}(\mathscr{X})$ is the Borel σ-algebra on $\mathscr{X}$, and $\mathscr{P}(\mathscr{X})$ refers to the set of probability measures on $(\mathscr{X}, \mathscr{B}(\mathscr{X}))$.

Definition 2 (Influence function and von Mises Derivative) Suppose $J : \mathscr{P}(\mathscr{X}) \to \mathbb{R}$ is a real-valued function, the *influence function* $h_\mu : \mathscr{X} \to \mathbb{R}$ of J at $\mu \in \mathscr{P}(\mathscr{X})$ is defined as

$$h_\mu(x) = \lim_{t\to 0^+} \frac{1}{t}\left\{ J(\mu + t(\delta_x - \mu)) - J(\mu) \right\}, \tag{4}$$

where $\delta_x : \mathscr{B}(\mathscr{X}) \to \{0,1\}$, defined by $\delta_x(A) := \mathbb{I}_A(x)$ for $A \in \mathscr{B}(\mathscr{X})$, is the Dirac measure (or atomic mass) concentrated at $x \in \mathscr{X}$ (Fernholz 1983). The influence function should be loosely understood as the rate of change in the objective J at μ, due to a perturbation of μ by a point mass δ_x. The *von Mises derivative* is defined as

$$J'_\mu(\nu - \mu) := \lim_{t\to 0^+} \frac{1}{t}\left\{ J(\mu + t(\nu - \mu)) - J(\mu) \right\}, \quad \mu, \nu \in \mathscr{P}(\mathscr{X}),$$

provided $J'_\mu(\cdot)$ is *linear* in its argument, that is, there exists a function $\phi_\mu : \mathscr{X} \to \mathbb{R}$, integrable with respect to both μ and ν, such that

$$J'_\mu(\nu - \mu) = \int \phi_\mu(x)\, \mathrm{d}(\nu - \mu)(x). \qquad (5)$$
$$=: \mathbb{E}_{X \sim \nu}[\phi_\mu(X)] - \mathbb{E}_{X \sim \mu}[\phi_\mu(X)].$$

When (5) holds, we can see that ϕ_μ in (5) and h_μ in (4) coincide to within a constant since $\nu - \mu$ has total measure zero.

2 FRANK-WOLFE METHODS FOR OPTIMIZATION OVER PROBABILITY MEASURES

We now introduce Frank-Wolfe methods for optimizing measures over $\mathscr{P}(\mathscr{X})$. To motivate our approach, recall the Frank-Wolfe recursion (Dunn and Harshbarger 1978), also known as the *conditional gradient* method (Bubeck 2015), in finite-dimensional Euclidean spaces. When minimizing a smooth function $f : \mathbb{R}^d \to \mathbb{R}$ over a compact convex set $Z \subset \mathbb{R}^d$, the Frank-Wolfe recursion is given by

$$y_{k+1} = (1 - \eta_k) y_k + \eta_k s_k, \quad s_k := \operatorname*{arg\,min}_{s \in Z} \nabla f(y_k)^\top s,$$

where $\eta_k \in (0,1]$ is a step size. This method ensures that the iterates $\{y_k\}$ remain feasible, and the linear subproblem can be solved efficiently.

To extend the Frank-Wolfe idea to $\mathscr{P}(\mathscr{X})$, consider a smooth functional $J : \mathscr{P}(\mathscr{X}) \to \mathbb{R}$. Instead of gradients in the finite-dimensional setting, the von Mises derivative J'_μ serves as the first-order approximation:

$$J(u) \approx J(\mu) + J'_\mu(u - \mu), \quad \text{for } u \in \mathscr{P}(\mathscr{X}).$$

This suggests the following recursion for measures:

$$\mu_{k+1} = (1 - \eta_k)\mu_k + \eta_k u_k, \quad u_k := \operatorname*{arg\,min}_{u \in \mathscr{P}(\mathscr{X})} J'_{\mu_k}(u - \mu_k).$$

As shown in Lemma 5 of Yu et al. (2024), the minimizer of $J'_\mu(u - \mu)$ is a Dirac measure $\delta_{x^*(\mu_k)}$, where $x^*(\mu)$ minimizes the influence function $h_\mu(x)$ over $\mathscr{X}$. Thus, the recursion simplifies to

$$\mu_{k+1} = (1 - \eta_k)\mu_k + \eta_k \delta_{x^*(\mu_k)}, \quad x^*(\mu_k) \in \operatorname*{arg\,min}_{x \in \mathscr{X}} h_{\mu_k}(x). \qquad \text{(dFW)}$$

The approach encapsulated by (dFW) solves an infinite-dimensional optimization problem by iteratively accumulating point masses at optimally chosen locations in $\mathscr{X}$. Unlike traditional finite-dimensionalization techniques such as gridding, the (dFW) recursion dynamically updates its support set without requiring a predefined discretization of the search space. Under standard assumptions of convexity and L-smoothness of J, the (dFW) method guarantees an $O(k^{-1})$ convergence rate in objective value, as established in Theorem 1. This result follows from the complexity analysis in Yu et al. (2024).

Lemma 1 (Complexity; Yu et al. 2024) Suppose J is convex and L-smooth. Then the iterates (dFW) satisfy

$$J(\mu_k) - J^* \le \frac{2LR^2}{k+2}, \quad k \ge 1$$

where $J^* := \inf\{J(\mu) : \mu \in \mathscr{P}(\mathscr{X})\}$ and $R = \sup\{\|\mu_1 - \mu_2\| : \mu_1, \mu_2 \in \mathscr{P}(\mathscr{X})\}$.

A natural extension of Frank-Wolfe is the fully corrective version, which often yields improved practical performance (Bredies and Pikkarainen 2013; Boyd et al. 2017). A fully corrective modification of (dFW) is presented in Algorithm 1. Unlike (dFW), where the update is a convex combination of the previous

Algorithm 1 Fully-corrective Frank Wolfe (FCFW) on $\mathscr{P}(\mathscr{X})$

```
**Input:** Initial measure $\mu_0 \in \mathscr{P}(\mathscr{X})$
**Output:** Iterates $\mu_1, \dots, \mu_K \in \mathscr{P}(\mathscr{X})$
1 $A_0 \leftarrow \emptyset$ or $\{\mu_0\}$
2 **for** $k = 0, 1, \dots, K$ **do**
3     $x^*(\mu_k) \leftarrow \arg\min_{x \in \mathscr{X}} h_{\mu_k}(x)$
4     $A_{k+1} \leftarrow A_k \cup \{\delta_{x^*(\mu_k)}\}$
5     $\mu_{k+1} \leftarrow \arg\min_{\mu \in \mathrm{conv}(A_{k+1})} J(\mu)$
6 **end for**
```

iterate μ_k and the Dirac measure $\delta_{x^*(\mu_k)}$, fully corrective Frank Wolfe (FCFW) maintains and optimizes over an expanding set of Dirac measures. Specifically, FCFW starts with an initial measure μ_0 and an empty set A_0. At each iteration k, a new Dirac measure $\delta_{x^*(\mu_k)}$ is added to A_k, forming the updated set:

$$A_{k+1} = \{\delta_{x^*(\mu_0)}, \dots, \delta_{x^*(\mu_k)}\}.$$

The FCFW algorithm then minimizes $J(\mu)$ over the convex hull of A_{k+1}, incorporating all selected atoms, as shown in Step 5. This fully corrective step is equivalent to solving:

$$\min_{p_0,\dots,p_k \in \mathbb{R}} J\left(\sum_{i=0}^{k} p_i \delta_{x^*(\mu_i)}\right) \quad \text{s.t.} \quad \sum_{i=0}^{k} p_i = 1, \quad p_i \geq 0.$$

Since J is convex, this results in a finite-dimensional convex optimization problem, which remains computationally feasible in practice.

As an extension of (dFW), FCFW also achieves an $O(k^{-1})$ convergence rate under the same assumptions as Theorem 1, with a similar proof. However, FCFW introduces additional theoretical properties that enhance its practical performance. In particular, Yu et al. (2024) establish a sufficient decrease property under a weaker assumption, requiring only $h_{\mu_k}(x^*(\mu_k)) \leq 0$ rather than the global optimality of $x^*(\mu_k)$.

3 ESTIMATING THE DERIVATIVE

The von Mises derivative J'_μ (or the influence function h_μ) of the functional J is a key mathematical object that determines the "direction" along which iterates are updated within FCFW. As we have described through illustrative examples in Section 1.1, it is sometimes the case that while J'_μ might exist, it is not directly observable. We thus turn to the question of how to *estimate* the von Mises derivative J'_μ.

3.1 Recall Estimation in $\mathbb{R}^d$

To provide intuition for the method we outline, we first consider the derivative estimation problem in Euclidean space. Suppose $f : \mathscr{D} \subseteq \mathbb{R}^d \to \mathbb{R}$ is a real-valued differentiable function with domain $\mathscr{D} \subseteq \mathbb{R}^d$, whose gradient $\nabla f(x)$ we want to estimate. Assume for simplicity that $\mathscr{D}$ has a non-empty interior. Now, suppose Z is an $\mathbb{R}^d$-valued random vector such that

$$\mathbb{E}[ZZ^\intercal] = I_d,$$

where I_d is the $d \times d$ identity matrix. Then, we see that

$$\mathbb{E}\left[ZZ^\intercal \nabla f(x)\right] = \nabla f(x). \tag{6}$$

The left-hand side of (6) cannot be used directly to construct an unbiased estimator of $\nabla f(x)$ because $Z^\intercal \nabla f(x)$ is not directly observable. However, notice that for small $t > 0$, $Z^\intercal \nabla f(x)$ can be approximated through finite-differencing, that is,

$$Z^{\intercal}\nabla f(x) \approx \frac{1}{t}\left(f(x+tZ)-f(x)\right). \tag{7}$$

The expressions in (6) and (7) thus suggest the following (biased) estimator for $\nabla f(x)$:

$$\hat{\nabla} f(x;t) := Z \times \frac{1}{t}\left(f(x+tZ)-f(x)\right), \quad t>0. \tag{8}$$

The estimator in (8) seems to have been independently discovered by various authors (Nesterov and Spokoiny 2017; Spall 1998) and has recently become the work-horse of zeroth-order optimization methods in Euclidean space (Duchi et al. 2015). The properties of $\hat{\nabla} f(x;t)$ have been investigated thoroughly (Spall 1998) but, considering our purpose, we will not go into details here.

3.2 Estimation in $\mathscr{P}(\mathscr{X})$ using L_2 Approximation

Recall that the von Mises derivative J'_{μ} is given by

$$J'_{\mu}(\nu-\mu) = \int_{\mathscr{X}} h_{\mu}(x)\, d(\nu-\mu), \tag{9}$$

where $h_{\mu}:\mathscr{X}\to\mathbb{R}$ is the influence function of J at $\mu\in\mathscr{P}(\mathscr{X})$. Due to the linear structure in (9), the problem of estimating J'_{μ} essentially amounts to the problem of estimating the influence function h_{μ}. We next mimic the development in Section 3.1 to obtain an estimator. A significant challenge is the infinite dimensionality of h_{μ}, which we attempt to overcome using a projection framework.

Suppose that $h_{\mu}\in L_2(\mathscr{X})$, the (Hilbert) space of square integrable functions on $\mathscr{X}$, and that $\mathrm{diam}(\mathscr{X})<\infty$. Suppose further that $\{u_n, n\geq 1\}$ is a complete orthonormal basis of $L_2(\mathscr{X})$, so that we can write

$$h_{\mu}(x) = \sum_{j=1}^{\infty} a_j u_j(x) = \sum_{j=1}^{d} a_j u_j(x) + r(x); \quad r(x) := \sum_{j=d+1}^{\infty} a_j u_j(x), \tag{10}$$

where $\langle u_i, u_j\rangle = \int_{\mathscr{X}} u_i(x)u_j(x)\,dx = \delta_{i,j}$ and $\langle u_j, r\rangle = 0, j=1,2,\ldots,d$. Standard results (e.g., Luenberger (1997), Section 3.8) guarantee the existence of a complete orthonormal sequence on $L_2(\mathscr{X})$.

Since $(u_n : n\geq 1)$ forms an orthonormal set, we have

$$a_j = \langle h_{\mu}, u_j\rangle = \int_{\mathscr{X}} h_{\mu}(x)u_j(x)\,dx, \quad j=1,2,\ldots \tag{11}$$

The expression for a_j in (11), and the expression for h_{μ} in (10) suggests the Monte Carlo estimator

$$\tilde{h}_{\mu}(x) = \sum_{j=1}^{d} \tilde{a}_j(m)u_j(x) \text{ with } \tilde{a}_j(m) = \frac{v}{m}\sum_{t=1}^{m} h_{\mu}(X_t)u_j(X_t), \tag{12}$$

where $X_t\sim\mathrm{Unif}(\mathscr{X})$ and v is the volume of $\mathscr{X}$, which is finite by assumption. The estimator $\tilde{a}_j(m)$ of a_j in (12) is still not observable because $h_{\mu}(X_t)$ is not observable. So, we next construct a finite-difference estimator for $h_{\mu}(X_t)$ as

$$FD_{s,\mu}(X_t) = \frac{1}{s}\left(J((1-s)\mu+s\delta_{X_t})-J(\mu)\right), \tag{13}$$

where $s>0$ is the step-size. The estimator (13) can now be plugged into (12) to obtain an observable (but biased) Monte Carlo estimator of the influence function,

$$\hat{h}_{\mu}(p,x) = \sum_{j=1}^{d} \hat{a}_j(m,s)u_j(x), \quad x\in\mathscr{X}, \tag{14}$$

where the "parameter vector" triple $p := (m,s,d)$, and

$$\hat{a}_j(m,s) = \frac{v}{m}\sum_{t=1}^{m} FD_{s,\mu}(X_t)u_j(X_t), \quad X_t \sim \text{Unif}(\mathscr{X}).$$

Observe that the (random) function estimator $\hat{h}_\mu(p,\cdot)$ (of $h_\mu(\cdot)$) appearing in (14) has two sources of bias. The first is due to the truncation (to d terms) of the infinite sum appearing in (10), and the second is due to the finite-difference approximation appearing in (13). Both of these sources of bias, along with a term due to the variance associated with Monte Carlo sampling, appear in the following result quantifying the quality of the function estimator $\hat{h}$.

Theorem 1 (Estimator Quality) Suppose that the following assumptions hold:

A.1 the functional $J : \mathscr{P}(\mathscr{X}) \to \mathbb{R}$ is L-smooth;
A.2 $\text{diam}(\mathscr{X}) < \infty$;
A.3 the orthonormal vectors $u_j, j = 1,2,\ldots$ are such that $\|u\|_\infty := \sup_{j\geq 1} \|u_j\|_\infty < \infty$; and
A.4 the second moment $\mathbb{E}[h_\mu(X)^2] < \infty$, where $X \sim \text{Unif}(\mathscr{X})$, which then implies, in view of A3, that $\sigma_\mu^2 := \sup_{j\geq 1} \text{Var}(h_\mu(X)u_j(X)) < \infty$.

Then,

$$\mathbb{E}\left[\left\|\hat{h}_\mu(p,\cdot) - h_\mu(\cdot)\right\|_\infty\right] \leq \|u\|_\infty \left(\frac{2dLs}{\sqrt{v}} + \sum_{j=d+1}^{\infty} |a_j| + \frac{dv\sigma_\mu}{\sqrt{m}}\right), \tag{15}$$

and

$$\mathbb{E}\left[\left\|\hat{h}_\mu(p,\cdot) - h_\mu(\cdot)\right\|_\infty^2\right] \leq 2\|u\|_\infty^2 \left(\frac{8d^2L^2s^2}{v} + \left(\sum_{j=d+1}^{\infty} |a_j|\right)^2 + \frac{2d^2v^2\sigma_\mu^2}{m}\right). \tag{16}$$

Suppose further that

A.5 the coefficients a_j decay polynomially, that is, $\exists C_1 > 0, c > 1$ such that for all $j \geq 1$, $|a_j| \leq C_1 j^{-c}$.

Then the function estimator $\hat{h}_\mu((m,s),\cdot)$ satisfies

$$\mathbb{E}\left[\left\|\hat{h}_\mu(p,\cdot) - h_\mu(\cdot)\right\|_\infty\right] \leq \|u\|_\infty \left(\frac{2dLs}{\sqrt{v}} + \frac{C_1}{c-1}d^{-c+1} + \frac{dv\sigma_\mu}{\sqrt{m}}\right), \tag{17}$$

and

$$\mathbb{E}\left[\left\|\hat{h}_\mu(p,\cdot) - h_\mu(\cdot)\right\|_\infty^2\right] \leq 2\|u\|_\infty^2 \left(\frac{8d^2L^2s^2}{v} + \frac{C_1^2}{(c-1)^2}d^{-2(c-1)} + \frac{2d^2v^2\sigma_\mu^2}{m}\right). \tag{18}$$

Proof Sketch. We only provide a proof sketch for (16) and (18); the proofs for (15) and (17) follow similarly and are omitted. First, by Assumption A.3, we have

$$
\begin{aligned}
\|\hat{h}_\mu(p,\cdot) - h_\mu(\cdot)\|_\infty^2 &= \sup_{x \in \mathscr{X}} \left| \sum_{j=1}^{d} (a_j - \hat{a}_j(m,s)) u_j(x) + \sum_{j=d+1}^{\infty} a_j u_j(x) \right|^2 \\
&\le \left(\sum_{j=1}^{d} |a_j - \hat{a}_j(m,s)| + \sum_{j=d+1}^{\infty} |a_j| \right)^2 \|u\|_\infty^2 \\
&\le 2 \left(\left(\sum_{j=1}^{d} |a_j - \hat{a}_j(m,s)| \right)^2 + \left(\sum_{j=d+1}^{\infty} |a_j| \right)^2 \right) \|u\|_\infty^2 \\
&\le 2 \left(d \sum_{j=1}^{d} (a_j - \hat{a}_j(m,s))^2 + \left(\sum_{j=d+1}^{\infty} |a_j| \right)^2 \right) \|u\|_\infty^2. \qquad (19)
\end{aligned}
$$

Next,

$$
\begin{aligned}
\mathbb{E}\left[(a_j - \hat{a}_j(m,s))^2\right] &= \mathbb{E}\left[(a_j - \tilde{a}_j(m) + \tilde{a}_j(m) - \hat{a}_j(m,s))^2\right] \\
&\le 2\mathbb{E}\left[(a_j - \tilde{a}_j(m))^2\right] + 2\mathbb{E}\left[(\tilde{a}_j(m) - \hat{a}_j(m,s))^2\right] =: I + II
\end{aligned}
$$

From Assumption A.4 and properties of the unbiased Monte Carlo estimator of a_j, we have $I \le \frac{2v^2\sigma_\mu^2}{m}$.

Using the L-smoothness of J, the finite-difference error $w(X) := FD_{s,\mu}(X) - h_\mu(X)$ satisfies $|w(X)| \le 2sL$. Since $\mathbb{E}[u_j^2(X)] = v^{-1}$ and thus $\mathbb{E}[|u_j(X)|] \le v^{-1/2}$, we obtain the bound

$$
\begin{aligned}
II &= 2\mathbb{E}\left[\left(\frac{1}{m} \sum_{t=1}^{m} w(X_t) u_j(X_t) \right)^2 \right] \\
&= 2\frac{1}{m^2} \left(m\mathbb{E}\left[w^2(X) u_j^2(X)\right] + m(m-1)\,\mathbb{E}\left[w(X_1)u(X_1)w(X_2)u(X_2)\right] \right) \\
&\le 2\frac{4s^2L^2}{m^2} \left(m\mathbb{E}\left[u_j^2(X)\right] + m(m-1)\left(\mathbb{E}\,|u(X_1)|\right)^2 \right) \le \frac{8s^2L^2}{v}.
\end{aligned}
$$

Taking the expectation of (19), we obtain

$$
\begin{aligned}
\mathbb{E}\left[\|\hat{h}_\mu(p,\cdot) - h_\mu(\cdot)\|_\infty^2\right] &\le 2 \left(d \sum_{j=1}^{d} \mathbb{E}\left[(a_j - \hat{a}_j(m,s))^2\right] + \left(\sum_{j=d+1}^{\infty} |a_j| \right)^2 \right) \|u\|_\infty^2 \\
&= 2\|u\|_\infty^2 \left(\frac{8d^2L^2s^2}{v} + \left(\sum_{j=d+1}^{\infty} |a_j| \right)^2 + \frac{2d^2v^2\sigma_\mu^2}{m} \right).
\end{aligned}
$$

Finally, by Assumption A.5, we have $\sum_{j=d+1}^{\infty} |a_j| \le (C_1) d^{-c+1}/(c-1)$, which yields (18). □

Assumptions 1 and 2 of Theorem 1 are standard. Assumption 3 is not overly restrictive, e.g., it is satisfied by Fourier series. Assumption 4 is satisfied, e.g., if $\mathscr{X}$ is compact and h_μ is continuous, and might be verified in other settings through, e.g., Lyapunov methods. Some form of coefficient decay like in Assumption 5 is necessary to ensure the tails of the expansion are ultimately negligible and is common in Fourier analysis. We leave the question of how to verify that condition to future work.

4 THE DERIVATIVE-FREE FCFW ALGORITHM

The function estimator $\hat{h}_\mu(p,\cdot)$ appearing in (14) for the influence function h_μ suggests a straightforward zeroth-order extension of the first-order FCFW algorithm in Algorithm 1 by simply replacing $h_{\mu_k}(\cdot)$ with $\hat{h}_{\mu_k}(p_k,\cdot)$, $p_k := (m_k, s_k, d_k)$ with the sample size sequence $(m_k : k \geq 1)$, the step-size sequence $(s_k : k \geq 1)$, and the number of terms in the expansion $(d_k : k \geq 1)$ determined appropriately.

Algorithm 2 Derivative-Free Fully Corrective Frank Wolfe on $\mathscr{P}(\mathscr{X})$

```
Input: Initial probability measure $\mu_0 \in \mathscr{P}(\mathscr{X})$; sample-size sequence $(m_k : k \geq 1)$;
       step-size sequence $(s_k : k \geq 1)$ and truncation sequence $(d_k : k \geq 1)$
Output: Iterates $\mu_1, \ldots, \mu_K \in \mathscr{P}(\mathscr{X})$
1 $A_0 \leftarrow \emptyset$ or $\{\mu_0\}$
2 for $k = 0, 1, \ldots, K$ do
3   find $x_k^* \in \mathscr{X}$ such that $\hat{h}_{\mu_k}(p_k, x_k^*) \leq 0$
4   $A_{k+1} \leftarrow A_k \cup \{\delta_{x^*(\mu_k)}\}$
5   $\mu_{k+1} \leftarrow \arg\min_{\mu \in \mathrm{conv}(A_{k+1})} J(\mu_k)$
6 end for
```

It is plausible that as long as $m_k \to \infty$, $s_k \to 0$, and $d_k \to \infty$ (at appropriate rates), then *consistency* of some form, e.g., $|h_{\mu_k}| \to 0$, should result. Even more interesting than such consistency is the question of convergence rate, and in particular, how the decay rate c appearing in A.5, and the parameter sequence $p_k = (m_k, s_k, d_k)$ interact to determine the convergence rate.

In preparation for such a consistency and rate result, we now state an important sufficient decrease lemma quantifying the behavior of the sequence $(J(\mu_k) : k \geq 1)$ as a function of the quality of the solution obtained in Step 3 of Algorithm 2.

Lemma 2 (Almost sufficient decrease) Suppose J is L-smooth. Then the sequence $\{\mu_k\}$ generated by Algorithm 2 satisfies

$$J(\mu_{k+1}) - J(\mu_k) \leq -\min\left\{\frac{1}{2LR^2}\hat{h}^2_{\mu_k}(p_k, x_k^*), \frac{1}{2}LR^2\right\} + \gamma_k \varepsilon_k, \quad k \geq 0, \tag{20}$$

where

$$\varepsilon_k := h_{\mu_k}(x_k^*) - \hat{h}_{\mu_k}(p_k, x_k^*); \quad \gamma_k := \min\left\{-\frac{\hat{h}_{\mu_k}(p_k, x_k^*)}{LR^2}, 1\right\} \leq 1,$$

and $R = \sup\{\|\mu_1 - \mu_2\| : \mu_1, \mu_2 \in \mathscr{P}(\mathscr{X})\}$.

Proof Sketch. Define $\mu_{k+0.5} := (1-\gamma_k)\mu_k + \gamma_k \delta_{x_k^*}$, where $\gamma_k \in [0,1]$ was defined in the statement of the lemma. Using the smooth function inequality on J, and after (lots of) algebra, we have that

$$J(\mu_{k+0.5}) - J(\mu_k) \leq -\min\left\{\frac{1}{2LR^2}\hat{h}^2_{\mu_k}(p_k, x_k^*), \frac{1}{2}LR^2\right\} + \gamma_k\left(h_{\mu_k}(x_k^*) - \hat{h}_{\mu_k}(p_k, x_k^*)\right).$$

Since $\mu_{k+1} \in \arg\min_{\mu \in \mathrm{conv}(A_{k+1})} J(\mu_k)$, we have that $J(\mu_{k+1}) \leq J(\mu_{k+0.5})$, and thus

$$J(\mu_{k+1}) - J(\mu_k) \leq -\min\left\{\frac{1}{2LR^2}\hat{h}^2_{\mu_k}(p_k, x_k^*), \frac{1}{2}LR^2\right\} + \gamma_k\left(h_{\mu_k}(x_k^*) - \hat{h}_{\mu_k}(p_k, x_k^*)\right).$$

□

It is important that Lemma 2 does not guarantee descent, that is, the objective function does not necessarily decrease at each iteration because the term ε_k that appears on the right-hand side of (20) can be positive. This is the key manner in which the current context of derivative-free smooth optimization differs from the first-order context where an inequality analogous to (20) has $\varepsilon_k = 0$, and the estimator $\hat{h}_{\mu_k}(p_k,\cdot)$ is replaced by $h_{\mu_k}(\cdot)$, thereby guaranteeing descent at every step.

Given the inequality in (2) and the nature of ε_k, it stands to reason that if the parameter vector $p_k = (m_k, s_k, d_k)$ is such that the estimator error $\|h_{\mu_k}(\cdot) - \hat{h}_{\mu_k}(p_k,\cdot)\|_\infty$ is appropriately controlled across iterations, then consistency is obtained, as demonstrated in the following result.

Theorem 2 (Consistency and Complexity) Suppose the assumptions A.1–A.5 in Theorem 1 hold. Suppose further that

$$\sum_{k=1}^{\infty} \frac{d_k}{\sqrt{m_k}} < \infty; \quad \sum_{k=1}^{\infty} d_k s_k < \infty; \quad \text{and } \sum_{k=1}^{\infty} d_k^{-(c-1)} < \infty. \tag{21}$$

Let $J^* := \inf\{J(\mu) : \mu \in \mathscr{P}(\mathscr{X})\}$. Then, the random sequence $\{\mu_k\}$ generated using Algorithm 2 satisfies

$$\lim_{k\to\infty} h_{\mu_k}(x_k^*) = 0 \quad \text{a.s.}$$

Furthermore, there exists (a random) $K_0 < \infty$ such that

$$\min_{K_0 \le k \le n} \left|h_{\mu_k}(x_k^*)\right| \le \frac{1}{\sqrt{n - K_0 + 1}} \left(2LR^2(J(\mu_{K_0}) - J^*) + \sum_{i=K_0}^{n} \varepsilon_k^2\right)^{\frac{1}{2}} \tag{22}$$

where $R = \sup\{\|\mu_1 - \mu_2\| : \mu_1, \mu_2 \in \mathscr{P}(\mathscr{X})$.

Proof Sketch. First, due to (21) and (17),

$$\sum_{k=1}^{\infty} |\varepsilon_k| < \infty \quad \text{and} \quad \sum_{k=1}^{\infty} \varepsilon_k^2 < \infty \quad \text{a.s.}, \tag{23}$$

where ε_k is as in (20). Also, from (20), we have

$$J(\mu_{k+1}) - J(\mu_k) \le -\min\left\{\frac{1}{2LR^2}\hat{h}_{\mu_k}(p_k, x_k^*)^2, \frac{LR^2}{2}\right\} + |\varepsilon_k|. \tag{24}$$

Adding the inequalities (24) for $k = 1, 2, \ldots, n$, we see that

$$-\infty < J^* - J(\mu_1) \le J(\mu_{n+1}) - J(\mu_1) \le -\sum_{k=1}^{n} \min\left\{\frac{1}{2LR^2}\hat{h}_{\mu_k}(p_k, x_k^*)^2, \frac{LR^2}{2}\right\} + \sum_{k=1}^{n} |\varepsilon_k|. \tag{25}$$

From (23) and (25),

$$\sum_{k=1}^{\infty} \hat{h}_{\mu_k}(p_k, x_k^*)^2 < \infty \quad \text{a.s.} \tag{26}$$

From (26) and (23), and observing that

$$\sum_{k=1}^{n} h_{\mu_k}(x_k^*)^2 \le 2\sum_{k=1}^{n} \hat{h}_{\mu_k}(p_k, x_k^*)^2 + 2\sum_{k=1}^{n} \varepsilon_k^2 \quad \text{a.s.},$$

we conclude that $|h_{\mu_k}(x_k^*)| \to 0$ as $k \to \infty$ a.s. Then, there exists $K_0 < \infty$ such that $|\hat{h}_{\mu_k}(p_k, x_k^*)| \le LR^2$ for all $k \ge K_0$. Using (20) and after some computation, we obtain

$$2LR^2(J(\mu_{k+1}) - J(\mu_k)) \le -h_{\mu_k}(x_k^*)^2 + \varepsilon_k^2, \quad \forall k \ge K_0.$$

Summing over $k = K_0, \dots, n$ and simplifying gives

$$\min_{K_0 \leq k \leq n} |h_{\mu_k}(x_k^*)| \leq \frac{1}{\sqrt{n - K_0 + 1}} \left(2LR^2 (J(\mu_{K_0}) - J^*) + \sum_{k=K_0}^{n} \varepsilon_k^2 \right)^{1/2},$$

which proves (22). □

Theorem 2 establishes that the value of the influence function at the point x_k^* chosen at iteration k converges to 0, and also gives the rate result (22) on how quickly this happens. One might be tempted to view this result as saying that the derivative converges to 0 as the iterations progress, but that is a stronger statement than is proved in Theorem 2. The difficulty stems from the fact that we chose x_k^* arbitrarily, except that it has a negative estimated influence function value. For example, one could select a sequence of points that have negative true influence function value, but of smaller and smaller absolute value, while there exist other points with negative influence function values that are bounded away from 0. In short, our very weak selection criterion for x_k^* leads to correspondingly weak convergence results. With a stronger selection criterion, stronger conclusions are possible using the tools we have developed, here. Indeed, a close reading of Theorem 2 makes clear that the better the choice of x_k^*, the better the asymptotic result that can be achieved.

5 CONCLUDING REMARKS

The infinite dimensional problem (P) that we have tackled in this paper is extremely challenging when we do not have access to first-order information, in particular, the influence function that, roughly speaking, provides directional derivatives in the direction of point mass probability distributions. We developed an estimator of the influence function that has a finite-dimensional representation. The estimator is obtained by expressing the influence function as a linear combination of orthonormal basis functions, then neglecting all but finitely many terms in the expansion. The remaining basis function coefficients are estimated using Monte Carlo. The estimator is biased because of both truncation of the expansion and finite differencing used to estimate certain directional derivatives. We provided an error analysis of the resulting influence-function estimator.

We then provided a fully corrective Frank-Wolfe algorithm that exploits the influence function estimator, along with a result on the asymptotics of the sequence of measures that result. It is of central importance in applications that the sequence of measures that are attained are finite discrete measures, and thus readily stored, sampled and represented. This is a significant advantage of the approach explored in this paper where we directly tackle an infinite-dimensional problem, relative to an approach where one discretizes the domain $\mathscr{X}$ at the outset and then attempts to solve a finite-dimensional optimization problem. In both approaches, one obtains algorithms that work with finite dimensional representations, but the approach advocated herein does not need to impose arbitrary discretizations at the outset.

There is much more to be explored. For example, what are the implications of the conditions (21) on the various parameters of the influence function estimator, and are there versions of Theorem 2 that, under stronger conditions, provide stronger convergence guarantees? Such results might rely, for example, on some combination of stronger assumptions on the functional, J, or on the selection x_k^* at each iteration k. We assumed that the functional, J, could be exactly observed at a measure, μ, but in many problems, $J(\mu)$ can only be observed through unbiased estimates of $J(\mu)$ at any given measure μ. Extensions to that setting would be welcome. Implementing, testing and comparing algorithms is a central goal. Moreover, there are many application-specific ideas that could be explored, e.g., how should one select the orthonormal basis and the ordering of functions within the basis for a given application?

ACKNOWLEDGMENTS

This work was partially supported by National Science Foundation grant OAC 2410950.

REFERENCES

Atkinson, A. C., A. N. Donev, and R. D. Tobias. 2007. *Optimum Experimental Designs, with SAS*. Oxford: Oxford University Press https://doi.org/10.1093/oso/9780199296590.001.0001.

Boyd, N., G. Schiebinger, and B. Recht. 2017. "The Alternating Descent Conditional Gradient Method for Sparse Inverse Problems". *SIAM Journal on Optimization* 27(2):616–639 https://doi.org/10.1137/15M1035793.

Bredies, K., and H. Pikkarainen. 2013. "Inverse Problems in Spaces of Measures". *ESAIM: Control, Optimisation and Calculus of Variations* 19(1):190–218 https://doi.org/10.1051/cocv/2011205.

Bubeck, S. 2015. "Convex Optimization: Algorithms and Complexity". *Foundations and Trends in Machine Learning* 8(3–4):231–358 https://doi.org/10.1561/2200000050.

Coetzer, R. L. J., and W. W. Focke. 2010. "Optimal Designs for Estimating the Parameters in Weighted Power-Mean-Mixture Models". *Journal of Chemometrics* 24(1):34–42 https://doi.org/10.1002/cem.1271.

Conn, A. R., K. Scheinberg, and L. N. Vicente. 2009. *Introduction to Derivative-Free Optimization*. Philadelphia: Society for Industrial and Applied Mathematics https://doi.org/10.1137/1.9780898718768.

Cornell, J. A. 2011. *A Primer on Experiments with Mixtures*. 4th ed. Hoboken: John Wiley & Sons https://doi.org/10.1002/9780470907443.

Draper, N. R., and F. Pukelsheim. 1999. "Kiefer Ordering of Simplex Designs for First- and Second-Degree Mixture Models". *Journal of Statistical Planning and Inference* 79(2):325–348 https://doi.org/10.1016/S0378-3758(98)00263-8.

Duchi, J. C., M. I. Jordan, M. J. Wainwright, and A. Wibisono. 2015. "Optimal Rates for Zero-Order Convex Optimization: The Power of Two Function Evaluations". *IEEE Transactions on Information Theory* 61(5):2788–2806 https://doi.org/10.1109/TIT.2015.2409256.

Dunn, J. C., and S. Harshbarger. 1978. "Conditional Gradient Algorithms with Open Loop Step Size Rules". *Journal of Mathematical Analysis and Applications* 62(2):432–444 https://doi.org/10.1016/0022-247X(78)90137-3.

Fernholz, L. T. 1983. *Von Mises Calculus for Statistical Functionals*. New York: Springer https://doi.org/10.1007/978-1-4612-5604-5.

Focke, W. W., S. Endres, E. L. du Toit, M. T. Loots, and R. L. J. Coetzer. 2021. "Revisiting the Classic Activity Coefficient Models". *Industrial & Engineering Chemistry Research* 60(15):5639–5650 https://doi.org/10.1021/acs.iecr.0c06330.

Kiefer, J., and J. Wolfowitz. 1952. "Stochastic Estimation of the Maximum of a Regression Function". *Annals of Mathematical Statistics* 23(3):462–466 https://doi.org/10.1214/aoms/1177729392.

Luenberger, D. G. 1997. *Optimization by Vector Space Methods*. New York: John Wiley & Sons.

Molchanov, I., and S. Zuyev. 2002. "Steepest Descent Algorithms in a Space of Measures". *Statistics and Computing* 12(2):115–123 https://doi.org/10.1023/A:1014878317736.

Nesterov, Y., and V. Spokoiny. 2017. "Random Gradient-Free Minimization of Convex Functions". *Foundations of Computational Mathematics* 17(2):527–566 https://doi.org/10.1007/s10208-015-9296-2.

Okabe, A., B. Boots, K. Sugihara, and S. N. Chiu. 2000. *Spatial Tessellations: Concepts and Applications of Voronoi Diagrams*. 2nd ed. Chichester: Wiley https://doi.org/10.1002/9780470317013.

Shashaani, S., F. S. Hashemi, and R. Pasupathy. 2018. "ASTRO-DF: A Class of Adaptive Sampling Trust-Region Algorithms for Derivative-Free Simulation Optimization". *SIAM Journal on Optimization* 28(4):3145–3176 https://doi.org/10.1137/15M1042425.

Spall, J. C. 1998. "Implementation of the Simultaneous Perturbation Algorithm for Stochastic Optimization". *IEEE Transactions on Aerospace and Electronic Systems* 34(3):817–823 https://doi.org/10.1109/7.705889.

Yu, D., S. G. Henderson, and R. Pasupathy. 2024. "Deterministic and Stochastic Frank-Wolfe Recursion on Probability Spaces". *arXiv preprint arXiv:2407.00307*.

AUTHOR BIOGRAPHIES

DI YU is a Ph.D. candidate in the Department of Statistics at Purdue University. His research focuses on stochastic optimization, infinite-dimensional optimization, and simulation-based methods. His email address is yu1128@purdue.edu.

SHANE G. HENDERSON holds the Charles W. Lake, Jr. Chair in Productivity in the School of Operations Research and Information Engineering at Cornell University. His research interests include simulation theory and a range of applications including emergency services. He is an INFORMS Fellow. He is a co-creator of SimOpt, a testbed of simulation optimization problems and solvers. His email address is sgh9@cornell.edu and his homepage is http://people.orie.cornell.edu/shane.

RAGHU PASUPATHY is Professor of Statistics at Purdue University. His current research interests lie broadly in stochastic optimization, uncertainty quantification, and simulation methodology. He has been actively involved with the Winter Simulation Conference for the past 20 years. Raghu Pasupathy's email address is pasupath@purdue.edu, and his web page https://web.ics.purdue.edu/~pasupath contains links to papers, software codes, and other material.