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.

ASTROMORF: ADAPTIVE SAMPLING TRUST-REGION OPTIMIZATION WITH DIMENSIONALITY REDUCTION

Benjamin Rees[1], Christine S.M. Currie[1], Phan Tu Vuong[1]

[1]School of Mathematical Sciences, University of Southampton, Southampton, UK

ABSTRACT

High dimensional simulation optimization problems have become prevalent in recent years. In practice, the objective function is typically influenced by a lower dimensional combination of the original decision variables, and implementing dimensionality reduction can improve the efficiency of the optimization algorithm. In this paper, we introduce a novel algorithm ASTROMoRF that combines adaptive sampling with dimensionality reduction, using an iterative trust-region approach. Within a trust-region algorithm a series of surrogates or metamodels is built to estimate the objective function. Using a lower dimensional subspace reduces the number of design points needed for building a surrogate within each trust-region and consequently the number of simulation replications. We explain the basis for the algorithm within the paper and compare its finite-time performance with other state-of-the-art solvers.

1 INTRODUCTION

Simulation optimization (SO), or optimization via simulation, describes a class of optimization algorithms used to solve problems where the objective function can only be observed through replications of a Monte Carlo simulation. Such problems are intrinsically stochastic and in recent years, high-dimensional SO problems have become prevalent, e.g., inventory management (Wang and Hong 2023), uncertainty quantification (Xie et al. 2014), and transportation optimization (Tay and Osorio 2024). A review of high-dimensional SO is presented by Fan et al. (2024), which highlights the need to develop dimensionality reduction methods to improve the efficiency of solvers on high-dimensional SO problems. The algorithm we develop here, ASTROMoRF, combines adaptive sampling with dimensionality reduction using an iterative trust-region approach.

The unconstrained continuous SO problem is formally stated as

$$\min_{\mathbf{x}\in\mathbb{R}^n} \mathrm{f}(\mathbf{x}) = \min_{\mathbf{x}\in\mathbb{R}^n} \mathbb{E}[\mathrm{F}(\mathbf{x},\zeta)], \tag{1}$$

where $\mathrm{f}:\mathbb{R}^n\to\mathbb{R}$ can only be observed through responses of the simulation model. The function $\mathrm{F}:\mathbb{R}^n\times\Omega\to\mathbb{R}$ is a stochastic function representing the responses from the simulation model. The sample average approximation (SAA) at design point $\mathbf{x}$ is defined as the sample mean of n replications,

$$\overline{\mathrm{F}}(\mathbf{x};n) = \frac{1}{n}\sum_{i=1}^{n}\mathrm{F}(\mathbf{x},\zeta_i), \tag{2}$$

which provides an unbiased estimator of $\mathrm{f}(\mathbf{x})$.

Simulation replications can be computationally expensive, and as a result, appropriate measures of efficiency for SO algorithms use the total number of calls to the simulation model rather than the number of algorithm iterations. Stochastic approximation algorithms that use gradient descent such as *Kiefer-Wolfowitz* (Kiefer and Wolfowitz 1952), and *simultaneous perturbation stochastic approximation* (SPSA) (Spall 1992) can be computationally inefficient in the derivative-free SO setting, requiring multiple replications to obtain accurate gradient estimations. Furthermore, the choice of step-size for finite differencing under a stochastic

setting requires fine-tuning, which can result in gradient estimators with poor quality (Asmussen and Glynn 2007). SO solvers that prioritize stability and accuracy over asymptotically fast convergence are consequently more promising. Deterministic model-based derivative-free trust-region (TRO-DF) algorithms provide stable trajectories in a finite-run (Conn et al. 2009). In each iteration, a TRO-DF algorithm solves the optimization problem on a surrogate model within a constrained (trust) region around the incumbent solution. The acceptance of a new solution and the trust-region's size adapt based on how well the surrogate model predicts the simulation model's behavior. The restriction of the feasible region to a neighborhood around the current solution in each iteration allows a more stable trajectory to be achieved during the run of the algorithm.

The first development of TRO-DF for SO was the deterministic *unconstrained optimization by quadratic approximation* (UOBYQA) (Powell 2002). An adaptation to stochastic functions is provided by Deng and Ferris (2006). UOBYQA ensures that the candidate solution is suitable for the maintained design set by solving an optimization problem on the interpolation functions. In its extension to the stochastic setting, UOBYQA controls the random error present from sampling by making repeated replications of the simulation, deciding on the number of samples through Bayesian techniques. A new version of quadratic approximation (NEWUOA) (Powell 2006) and its bounded complement *bound constrained optimization with quadratic approximation* (BOBYQA) (Powell et al. 2009) build on UOBYQA. They both attempt to reduce the number of design points sampled to $2n+1$. Further developments include *derivative-free adaptive sampling trust-region optimization* (ASTRO-DF) (Shashaani et al. 2016), which utilizes adaptive sampling of the simulation model to ensure that the stochastic sampling error is in lock-step with the model bias. The benefit of keeping these errors in lock-step is improved efficiency and stability of the algorithm during its run. ASTRO-DF also conducts a direct search on the design set used to construct the surrogate model at every iteration. When the estimated response at the candidate solution is less than the response of the design points used to construct the surrogate model, the direct search solution is accepted over the candidate solution. This allows for progress to be made in unsuccessful runs.

The TRO-DF algorithms avoid calculating first-order derivatives through construction of surrogate models using interpolation. UOBYQA uses Lagrange interpolation with $\frac{1}{2}(n+1)(n+2)$ design points, where n is the dimension of the decision vector in the SO problem, whereas, ASTRO-DF uses $2n+1$ points with a fixed geometry (Ha and Shashaani 2024). This results in the model construction stage of UOBYQA taking $\mathcal{O}(n^2)$ simulation replications per iteration and ASTRO-DF taking $\mathcal{O}(n)$ simulation replications per iteration. Furthermore, ASTRO-DF performs a first-order criticality check on the accuracy of the surrogate model and if it fails to meet the accuracy threshold, the interpretation set is reconstructed under a smaller trust-region. This first-order criticality check is crucial for ensuring convergence to a first-order stationary point; however it can lead to multiple simulation replications within one iteration. The simulation cost to construct accurate surrogate models through interpolation increases exponentially as the number of decision variables increases. For a high-dimensional SO problem, the model construction stage can expend a large number of simulation replications per iteration, which is prohibitive under a fixed replication budget.

In practice, the objective function in (1) is often largely affected by a relatively small subset of decision variables in the decision vector around any point. Active subspaces (Constantine et al. 2014; Russi 2010) can be employed to identify this subset of components and isolate them into a reduced dimensional subspace, mitigating the curse of dimensionality associated with model construction in TRO-DF. Active subspace methods have been applied in uncertainty quantification (Smith 2024), where simulation model inputs are sampled using Monte Carlo methods and the outputs are treated as a data set for statistical analysis. The application of active subspaces in combination with TRO-DF has been successfully applied in a deterministic setting. Two such solvers include the *Ridge-Informed Trust Region Solver* (RITR) (Gross et al. 2020) and *Optimization with Moving Ridge Functions* (OMoRF) (Gross and Parks 2022). These solvers have reduced the number of responses needed from the simulation model in a single run.

An alternative dimensionality-reduction method is applying global sensitivity analysis to determine which decision variables have the most influence on the simulation model's response before the solver is

run (Kleijnen 2005), which has proven popular in the simulation community. Sensitivity analysis ranks the coordinates of the inputs on the basis of which coordinate directions exhibit the most variability. Dimensions that do not exhibit high variability are removed from the decision space, resulting in a reduced dimension function after appropriate rotation (Saltelli et al. 2008). Within TRO-DF, the algorithm makes progress by reducing the problem to a local optimization around the incumbent solution. Reducing the dimension of the decision space through a global search may ignore local variability in the sensitivity of the objective to different variables (Gould et al. 2005).

Our contribution is the proposal of a novel TRO-DF algorithm, *Adaptive Sampling Trust Region Optimization with Moving Ridge Functions* (ASTROMoRF). ASTROMoRF combines the adaptive sampling and design point selection from ASTRO-DF with techniques from OMoRF to construct local active subspace matrices to project the surrogate model into a lower-dimensional subspace $\mathbb{R}^d$.

The subspace dimension size d, is selected a priori by the experimenter. In practice, the optimal choice for d is unknown but knowledge of the problem may suggest a good choice. Choosing a small value for d reduces the simulation budget expended per iteration. We introduce novel adaptations to the selection of design points in the projected subspace to construct a quadratic stochastic interpolation model. We also introduce novel adaptations to the ridge function recovery method presented in Hokanson and Constantine (2018) to ensure that the constructed surrogate model is fully-linear. The application of a local active subspace allows the selection of coordinate directions that exhibit the most variability on the response surface, enabling more substantial steps to be made in each iteration.

In Section 2, we discuss key implementation aspects of ASTROMoRF that justify its fast convergence to a first-order critical point with respect to the number of simulation replications, for high-dimensional problems. Section 3 reports numerical results that demonstrate the performance of ASTROMoRF against ASTRO-DF and OMoRF on three high-dimensional SO test problems. Finally, we conclude in Section 4.

2 METHODOLOGY

ASTROMoRF makes use of two main features of ASTRO-DF: (i) certification to ensure that the surrogate model meets the conditions of being fully linear; and (ii) an adaptive sampling rule used in ASTRO-DF to suggest the number of replications to be made for model construction and evaluation of the candidate step. The addition we offer is the construction of a new basis matrix in each iteration, allowing us to project design points into a subspace of the decision space. This permits us to sample a significantly reduced number of design points, detaching the number of replications needed for model construction from the dimension of the problem, and solving a known issue with TRO-DF solvers, whose main simulation expense derives from the number of design points needing to be simulated in each iteration. Algorithm 1 describes ASTROMoRF, where the presentation of the algorithm closely follows the notation of Nocedal and Wright (1999) and Ha and Shashaani (2024). In algorithm 1, we carry out a direct search on the finite set of evaluated design points in the design set, by sorting the responses of the design set at each design point and selecting the design point corresponding to the smallest response (Ha and Shashaani 2024).

In the following sections, we discuss key implementation characteristics of ASTROMoRF that ensure its finite-time performance. First, we discuss how our algorithm selects design points for constructing the surrogate model through interpolation, how we ensure that the design points selected span the trust-region within the projected subspace, and how we update the design points if they do not ensure that the interpolation matrix is poised. Second, we present a method of variable projection to recover the ridge function and active subspace matrix to construct the trust-region's surrogate model and the corresponding active subspace matrix. This is used to project the current iteration's candidate solution to the reduced subspace. Third, we discuss the adaptive sampling rule used to decide on the number of replications to make at a design point in the current iteration.

Algorithm 1 Adaptive Sampling Trust-Region Optimization with Moving Ridge Functions (ASTROMoRF)

Require: Initial solution $\mathbf{x}_0 \in \mathbb{R}^n$, initial trust-region radius $\Delta_0 > 0$, acceptance thresholds $0 < \eta_1 < \eta_2 < 1$, minimum sample size $\lambda_{\min}$, subspace dimension $d \leq n$, trust-region radius expansion $\gamma_1 > 1$, trust-region radius shrinkage $\gamma_2 \in (0,1)$, adaptive sampling constant $\kappa > 0$, sufficient reduction constant $\theta > 0$.

Set $k = 0$

while expended budget $<$ budget **do**

Build a design set $\mathcal{X}_k$ from $2d+1$ points from the trust-region $\mathcal{B}(\mathbf{x}_k;\Delta_k)$

Obtain $\mathrm{N}_k(\mathbf{x}_i)$ replications of each design point of $\mathcal{X}_k = \{\mathbf{x}_0,\ldots,\mathbf{x}_{2d}\}$ using (12) and evaluate the SAA (2) response at each design point $\overline{\mathrm{F}}(\mathbf{x}_i;\mathrm{N}_k(\mathbf{x}_i))$ for $i = 0,\ldots,2d$.

Construct a d-dimensional quadratic model m_k and active subspace matrix $\mathrm{U}_k \in \mathbb{R}^{n\times d}$ using $\mathcal{X}_k$.

Find the stepsize $\mathbf{s}_k$ by solving the trust-region subproblem:

$$\mathbf{s}_k = \min_{\mathbf{s}\in\mathcal{B}(\mathbf{0};\Delta_k)} m_k(\mathrm{U}_k^T(\mathbf{x}_k+\mathbf{s})).$$

Evaluate the candidate solution over $\mathrm{N}_k(\mathbf{x}_k+\mathbf{s}_k)$ replications using (12) and evaluate the SAA (2) of these replications: $\overline{\mathrm{F}}(\mathbf{x}_k+\mathbf{s}_k;\mathrm{N}_k(\mathbf{x}_k+\mathbf{s}_k))$.

Find the direct search reduction, through a pattern search on $\hat{\mathbf{x}}_k = \min_{\mathbf{x}\in\mathcal{X}^k\setminus\mathbf{x}_k}\overline{F}(\mathbf{x};\mathrm{N}_k(\mathbf{x}))$

Calculate the ratio r_k using the rate of change in the simulation model response against the surrogate model values between the incumbent solution and the candidate solution:

$$r_k = \frac{\overline{\mathrm{F}}(\mathbf{x}_k;\mathrm{N}_k(\mathbf{x}_k)) - \overline{\mathrm{F}}(\mathbf{x}_k+\mathbf{s}_k;\mathrm{N}_k(\mathbf{x}_k+\mathbf{s}_k))}{m_k(\mathrm{U}_k^T\mathbf{x}_k) - m_k(\mathrm{U}_k^T(\mathbf{x}_k+\mathbf{s}_k))} = \frac{\delta\overline{\mathrm{F}}(\mathbf{x}_k+\mathbf{s}_k)}{\delta m_k(\mathbf{x}_k+\mathbf{s}_k)}.$$

Accept/Reject the Candidate solution $\mathbf{x}_k+\mathbf{s}_k$ and update the trust-region radius based on the following criteria:

$$\mathbf{x}_{k+1} = \begin{cases} \hat{\mathbf{x}}_k, & \delta\overline{\mathrm{F}}(\hat{\mathbf{x}}_k) > \max\{\delta\overline{\mathrm{F}}(\mathbf{x}_k+\mathbf{s}_k), \theta\Delta_k^2\} \\ \mathbf{x}_k+\mathbf{s}_k, & r_k \geq \eta_1 \\ \mathbf{x}_k, & r_k < \eta_1 \end{cases} \qquad \Delta_{k+1} = \begin{cases} \min\{\gamma_1\Delta_k, \gamma_1\|\mathbf{s}_k\|, \Delta_{\max}\}, & r_k \geq \eta_2 \\ \min\{\gamma_1\Delta_k, \Delta_{\max}\}, & \eta_1 \leq r_k < \eta_2 \\ \gamma_2\Delta_k, & r_k < \eta_1 \end{cases}$$

Update $k = k+1$.

end while

return $\{\mathbf{x}_0, \mathbf{x}_1, \ldots\}$ and the intermediate budgets.

2.1 Ensuring Full-Linearity of the Surrogate Model

The quality of the surrogate model is highly dependent on the choice of design set. This dependency affects the numerical stability of the interpolation matrix $\mathrm{V}(\phi;\mathcal{X})$. A numerically stable $\mathrm{V}(\phi;\mathcal{X})$ implies that the design set is *poised* (Conn et al. 2008). The numerical stability of $\mathrm{V}(\phi;\mathcal{X})$ is crucial as we obtain the coefficients $\mathbf{c}$ for the surrogate model by solving the system of linear equations,

$$\mathrm{V}(\phi;\mathcal{X})\mathbf{c} = \mathbf{f}(\mathcal{X}). \tag{3}$$

Each row of the matrix $\mathrm{V}(\phi;\mathcal{X})$ represents the evaluation of a design point in the design set $\mathcal{X} = \{\mathbf{x}_1,\ldots,\mathbf{x}_p\}$ at each element of the polynomial basis $\phi(\mathbf{x}) = \{\phi_1(\mathbf{x}),\ldots,\phi_q(\mathbf{x})\}$, such that $\mathrm{V}(\phi;\mathcal{X})_{i,j} = \phi_j(\mathbf{x}_i)$. The vector $\mathbf{f}(\mathcal{X})$ consists of the responses of the model at each design point. Solving (3) for the coefficients of the model $\mathbf{c} = [c_1,\ldots,c_q]$, allows us to obtain the polynomial surrogate model:

$$m_k(\mathbf{x}) = \sum_{i=1}^{q} c_i\phi_i(\mathbf{x}). \tag{4}$$

A poised design set ensures that $\mathrm{V}(\phi;\mathcal{X})$ is non-singular, implying that a solution to the linear system (3) exists. Conn et al. (2009) show that that the design set is nonpoised if all the design points lie on a polynomial manifold of degree r or less, where r is the degree of the surrogate model. The property of the design set being poised is a necessary condition for the surrogate model to be fully-linear. A fully-linear surrogate model is one that satisfies the Taylor-like bounds,

$$\begin{aligned} \|f(\mathbf{x}) - m_k(\mathbf{x})\| &\leq \kappa_{ef}\Delta_k^2 \\ \|\nabla f(\mathbf{x}) - \nabla m_k(\mathbf{x})\| &\leq \kappa_{eg}\Delta_k \end{aligned} \tag{5}$$

and ensures that the TRO-DF converges to a first-order optimal solution. These bounds must hold for all design points $\mathbf{x}$ within the current trust-region $\mathcal{B}(\mathbf{x}_k;\Delta_k)$ at iteration k, where the trust-region is defined as the ball with center $\mathbf{x}_k$ and radius Δ_k. The constants $\kappa_{ef}, \kappa_{eg} > 0$ are defined a priori.

2.2 Choosing Design Points for the Model Construction Step

We construct the design set for the surrogate model using an approach taken from Ha et al. (2024). Their method selects $2n+1$ design points, where n is the dimension of the decision space, resulting in a quadratic model on a fixed geometry with a diagonal Hessian (Coope and Tappenden 2021). This fixed geometry has been shown to have an optimal design (Ragonneau and Zhang 2024).

The design set construction for ASTROMoRF differs by constructing a design set that spans the trust-region in the projected subspace and whose points are mutually orthogonal, such that they do not lie on a polynomial manifold of degree 2 or less. For the set, $2d$ points form the basis of the trust-region with the additional point being the projected incumbent solution at the center. The interpolation design of Ha et al. (2024), allows for reuse of previously visited design points in subsequent iterations to conserve computational resources, as well as a rotated coordinate basis to obtain more precise estimates of the gradient at $\mathbf{x}_k$. In our case, we only require $2d+1 \leq 2n+1$ design points to ensure a solution to (3), because we are working in a reduced dimension d (Ragonneau and Zhang 2024).

The $2d+1$ design points are selected by first projecting the incumbent solution $\mathbf{x}_k$ into the subspace by U_k. If there exists previously visited design points within the trust-region after projection, then the design point furthest from the projected incumbent solution is selected. We then select the remaining $2d-1$ points within the projected trust-region using the design set construction presented in Ha et al. (2024). After the design set is constructed, it is projected back into the original space.

In the case that there are no previously visited points within the projected trust-region, we apply the design set construction presented in Shashaani et al. (2016), which relies on a coordinate-stencil approach. This set construction does not always result in $\mathrm{V}(\phi;\mathrm{U}_k^T\mathcal{X}_k)$ being poised. In the case that our design set is ill-poised, we apply a pivoting algorithm adapted from Conn et al. (2008) to improve the design set by removing a point that causes the set to be ill-poised and replacing it with a better design point. This relies on using Gaussian elimination with row pivoting on $\mathrm{V}(\phi;\mathrm{U}_k^T\mathcal{X}_k)$. We continually improve the design set until we can certify that the surrogate model constructed under the design set is fully-linear. The benefit of applying a geometry improving algorithm over completely reconstructing the design set is that it reduces the number of design points resampled in each iteration.

2.3 Constructing the Stochastic Polynomial Interpolation Model and Active Subspace

We adapt the method presented in Hokanson and Constantine (2018) to obtain better approximations of the response surface on the trust-region at every iteration k. The algorithm simultaneously obtains the active subspace matrix that is near-optimal at identifying the effective dimension within the trust-region. We first introduce Hokanson's method and then go on to describe how we adapt it to our problem.

2.3.1 Hokanson's Method for Polynomial Ridge Recovery

Hokanson's method presents an alternating algorithm to recover the surrogate model coefficients and the active subspace matrix by solving the least-squares problem

$$\min_{\substack{m_k \in \mathbb{P}^r(\mathbb{R}^d) \\ \text{Range } \mathrm{U}_k \in \mathbb{G}(d,\mathbb{R}^n)}} \sum_{\mathbf{x} \in \mathcal{X}_k} \left[f(\mathbf{x}) - m_k(\mathrm{U}_k^T \mathbf{x}) \right]^2. \tag{6}$$

Assuming that U_k has orthonormal columns (Constantine et al. 2017), and that the approximated ridge function m_k has the structure equivalent to the interpolation function (4), which is characterized by the vector of coefficients $\mathbf{c}$, we can reformulate this problem and solve it with respect to $\mathbf{c}$ and U_k. We solve (3) to fix $\mathbf{c}$, and express the problem in terms of U_k

$$\min_{\text{Range } \mathrm{U}_k \in \mathbb{G}(d,\mathbb{R}^n)} \|\mathbf{f}(\mathcal{X}_k) - \mathrm{V}(\boldsymbol{\phi}; \mathrm{U}_k^T \mathcal{X}_k) \mathrm{V}(\boldsymbol{\phi}; \mathrm{U}_k^T \mathcal{X}_k)^+ \mathbf{f}(\mathcal{X}_k)\|_2^2. \tag{7}$$

This is the norm of the orthogonal projector of $\mathbf{f}(\mathcal{X}_k)$ onto the range of $\mathrm{V}(\boldsymbol{\phi}; \mathbf{U}_k^T \mathcal{X}_k)$, which we denote as $\mathrm{P}^{\perp}_{\mathrm{V}(\boldsymbol{\phi}; \mathrm{U}_k^T \mathcal{X})} \mathbf{f}(\mathcal{X}_k)$. The optimization problem (7) can be solved using Newton's method on the Grassmann manifold. Golub and Pereyra (1973) provides an explicit formulation of the Jacobian of $\mathrm{P}^{\perp}_{\mathrm{V}(\boldsymbol{\phi}; \mathrm{U}_k^T \mathcal{X})} \mathbf{f}(\mathcal{X}_k)$ with respect to U_k, denoted as $\mathcal{J}(\mathrm{U}_k)$. Furthermore, U_k is orthogonal with slices of $\mathcal{J}(\mathrm{U}_k)$. Therefore, the second-order derivative information of Newton's method can be approximated using the Gauss-Newton approximation, allowing the normal equations,

$$\mathrm{W}_1^T \Delta = \mathbf{0}, \quad \mathrm{Hess}\boldsymbol{\varphi}(\Delta, \mathbf{X}) = -\langle \mathrm{grad}\ \boldsymbol{\varphi}, \mathbf{X} \rangle, \quad \forall \mathbf{X}\ \text{ such that }\ \mathrm{U}_k^T \mathbf{X} = 0 \tag{8}$$

that selects a search direction $\Delta \in \mathbb{R}^{n \times d}$ tangent to the current estimate of U_k to be replaced with a better conditioned least squares problem

$$\begin{aligned} \min_{\Delta \in \mathbb{R}^{n \times d}} \quad & \|\mathrm{vec}\,(\mathcal{J}(\mathrm{U}_k))\,\mathrm{vec}\,(\Delta) - \mathrm{P}^{\perp}_{\mathrm{V}(\boldsymbol{\phi}; \mathrm{U}_k^T \mathcal{X}_k)} \mathbf{f}(\mathcal{X}_k)\|_2^2 \\ \text{such that} \quad & \mathrm{U}_k^T \Delta = \mathbf{0}. \end{aligned} \tag{9}$$

The function $\boldsymbol{\varphi}$ in (9) is an arbitrary function of the active subspace matrix U_k on the Grassmann manifold, and $\mathbf{X} \in \mathbb{R}^{n \times d}$ is a test matrix that, along with Δ, defines the Hessian of $\boldsymbol{\varphi}$ on the Grassmann manifold. The algorithm presented in Hokanson and Constantine (2018) undergoes a variable projection scheme, computing the ridge function coefficients $\mathbf{c}$ before solving (7) through the Gauss-Newton algorithm on Grassmann manifolds, through a step on the geodesic

$$\mathrm{U}(t) = \mathrm{U}_k \mathrm{Z} \cos(\Sigma t) \mathrm{Z}^T + \mathrm{Y} \sin(\Sigma t) \mathrm{Z}^T, \tag{10}$$

where the SVD of Δ is $\mathrm{Y}\Sigma\mathrm{Z}^T$. It iterates through this variable projection scheme until the active subspace matrix U_k converges. Hokanson's variable projection algorithm provides much faster convergence than the alternating algorithm presented in Constantine et al. (2017).

The method of Hokanson and Constantine (2018) requires an initial estimate of the active subspace matrix U_k; they compute this initial estimate by sampling design points from the normal distribution and finding U_k through the QR decomposition. Furthermore, the projected points are scaled and shifted into the d-dimensional hypercube $[-1,1]^d$. This transformation is done so that the selected polynomial basis $\boldsymbol{\phi}$ can be chosen to ensure that the matrix $\mathrm{V}(\boldsymbol{\phi}; \mathrm{U}_k^T \mathcal{X}_k)$ is well-conditioned. The polynomial basis chosen to ensure a well-conditioned $\mathrm{V}(\boldsymbol{\phi}; \mathrm{U}_k^T \mathcal{X}_k)$ is the Legendre basis.

2.3.2 Applying Hokanson's Method

We first obtain our initial estimate for U_k by sampling d n-dimensional design points around the current trust-region as an $n \times d$ design matrix, and obtain the initial estimate of U_k through a QR decomposition of the design matrix. This initial estimate requires no additional replications of the simulation model. Second, in approximating the surrogate model, we do not map the design points into the d-dimensional hypercube. Avoiding this transformation ensures the surrogate model constructed is an interpolation model. We also apply the natural monomial polynomial basis instead of the Legendre basis that is used in Hokanson and Constantine (2018). Our method provides certifications on the constructed surrogate model after evaluating $\mathbf{c}$, by checking the inequality $\Delta_k \leq \mu_k \|\nabla m_k(\mathbf{x}_k)\|$ holds. This bound ensures the model will converge to a first-order critical point. A similar "criticality check" is common in modern TRO-DF solvers, making the algorithm more numerically efficient, albeit with more complicated trust-region management and convergence analysis. If the model does not provide a large enough gradient, we update the design set using the pivoting algorithm mentioned above until the condition is met.

2.4 The Adaptive Sampling Rule

In handling errors derived from the sampling of the response of the simulation model, we apply the adaptive sampling rule described in Ha et al. (2024), which improves on the adaptive sampling rule introduced in Shashaani et al. (2016). The rule provides us with an optimal number of replications $\mathrm{N}_k(\mathbf{x}_k)$ to make at a particular design point $\mathbf{x}_k \in \mathbb{R}^n$ at iteration k of the solver,

$$\mathrm{N}_k(\mathbf{x}_k) = \min\left\{ t \geq \lambda_k : \frac{\hat{\sigma}(\mathbf{x}_k, t)}{\sqrt{t}} \leq \frac{\kappa \Delta_k^2}{\sqrt{\lambda_k}} \right\}, \tag{11}$$

where $\{\lambda_k\}$ is a deterministically increasing sequence with a logarithmic growth. The logarithmic growth allows for the simulation budget to be saved later on in the run of the algorithm. This sequence is used to determine the minimum sample size at iteration k. Here, $\hat{\sigma}^2(\mathbf{x}_k, t)$ denotes the estimated variance from t replications of the model at $\mathbf{x}_k$ and $\kappa > 0$ is a constant defined a priori by the user. The adaptive element of the sampling rule, defined by the inequality within the set construction of (11), ensures that the stochastic sampling error, defined as the standard error of the estimated response by SAA, is in lock-step with the first-order optimality gap, defined by the right hand side of the inequality. As the first-order optimality gap cannot be calculated during the run, a proxy of the true gradient norm is used. Enforcing that the estimation error is in lock-step with the optimization error allows us to identify a sample size that is neither too small nor too large, reducing the possibility of oversampling and inefficiently expending the simulation budget.

Ha et al. (2024) improved this rule by reducing the number of times the variance $\hat{\sigma}^2(\mathbf{x}_k, t)$ needs to be calculated. The initial sample size for iteration k is set to λ_k. If λ_k is sufficient to attain an accurate estimate of the response of the simulation model at $\mathbf{x}_k$ then there is no need to obtain additional responses from the simulation model at $\mathbf{x}_k$, giving the sample size as $\mathrm{N}_k(\mathbf{x}_k) = \lambda_k$. For the case that λ_k is an insufficient sample size, we can set $\mathrm{N}_k(\mathbf{x}_k) = \lambda_k \hat{\sigma}^2(\mathbf{x}_k, \lambda_k) \left(\kappa \Delta_k^4\right)^{-1}$. This means that $\mathrm{N}_k(\mathbf{x}_k)$ will always satisfy (11) under an arbitrary choice of k and $\mathbf{x}_k$. The resulting two-stage adaptive sampling rule is formalized as

$$\mathrm{N}_k(\mathbf{x}_k) = \lambda_k \max\left\{ 1, \frac{\hat{\sigma}^2(\mathbf{x}_k, \lambda_k)}{\kappa \Delta_k^4} \right\}. \tag{12}$$

This sampling process is a heuristic of the sampling process presented in (11) that reduces the number of times $\hat{\sigma}^2(\mathbf{x}_k, \lambda_k)$ needs to be evaluated. Note that if λ_k is mis-specified, and set to be too small then $\hat{\sigma}^2(\mathbf{x}_k, \lambda_k)$ may be a poor estimator of the variance. The method can also fail if $\hat{\sigma}^2(\mathbf{x}_k, \lambda_k)$ is very large, leading to a potentially large $\mathrm{N}_k(\mathbf{x}_k)$.

The implementation of adaptive sampling in ASTROMoRF has two main benefits. First, the responses of the simulation model at each design point during model construction have a lower variance, allowing for

more accurate surrogate models to be constructed. This results in candidate solutions being found during the subproblem that are more likely to be accepted. Second, we can obtain more accurate responses of the candidate solution, improving the chance of the candidate step being accepted as the new solution. These benefits are evident when comparing ASTROMoRF with deterministic TRO-DF algorithms that use active subspaces. For example, the finite-time performances of OMoRF and RITR are heavily affected by the stochastic noise in SO problems.

2.5 Effect of the Active Subspace Matrix

We conclude with a brief discussion on the effect the active subspace matrix has on the performance of ASTROMoRF. Unless the objective function is an exact ridge function, the projected points in the active subspace $\mathbf{y} = \mathrm{U}^T\mathbf{x}$ will exhibit additional noise when replicating the model at $\mathbf{y}$. This is because the mapping $\mathbf{y} \mapsto \mathbf{x}$ is not unique, due to a non-unique choice of basis matrix. The projection error can be treated as additive noise on the evaluations of the surrogate model and is dependent on the active subspace and $\mathbf{x}$. It can be interpreted as the information loss from mapping into the active subspace and then back into the original subspace. This projection error can be mitigated by an adaptive sampling rule. By considering the projection error as a form of model bias, an adaptive sampling rule can be developed that will keep the model bias bounded by controlling the stochastic sampling error as well as the projection error through SAA. The adaptive sampling rule presented above (11) can be extended to keep this model error in lock-step with the optimization error.

RITR actively handles this projection noise differently, by checking algorithmic progress and detecting if progress is halted under the same conditions as BOBYQA (Cartis and Roberts 2019). However, this method may not be beneficial to the performance of the solver under a fixed simulation budget, due to multiple warm restarts of the algorithm.

3 NUMERICAL RESULTS

In this section, we report on ASTROMoRF's overall finite-time performance against state-of-the-art SO solvers included in the SimOpt library (Eckman et al. 2023). We test primarily against ASTRO-DF as a benchmark due to ASTROMoRF being an extension of ASTRO-DF and because of ASTRO-DF's superiority against other SO solvers seen in Ha and Shashaani (2024).

We consider three problems in our analysis. '*DYNAMNEWS-1*' is provided by the SimOpt library and consists of selecting an initial inventory level for a one-period inventory model that is modeled after a newsvendor problem with dynamic consumer substitution (Mahajan and Van Ryzin 2001). This problem has a dimension of 10 and contains a single source of randomness, in the form of a multinomial logit model, that assigns a utility value for each customer-product pairing in the model. The other two problems presented are minimization problems of the Rosenbrock function (De Jong 1975) and the Zakharov function (Surjanovic and Bingham 2013), each of dimension 15. Both of these functions have feasible regions defined as the hypercube in $\mathbb{R}^{15}$, $[-5,10]^{15}$. The source of randomness comes from normally-distributed additive noise on f, $\mathrm{F}_i(\mathbf{x}) = f(\mathbf{x}) + \xi_i$, where $\xi_i \sim \mathrm{N}(0,\sigma^2)$; and in our experiments $\sigma^2 = 0.1$. We present the Rosenbrock and Zakharov functions in their deterministic form in (13) and (14) respectively,

$$f(\mathbf{x}) = \sum_{i=1}^{n-1} \left(100(x_{i+1} - x_i^2)^2 + (x_i - 1)^2\right), \tag{13}$$

$$f(\mathbf{x}) = \sum_{i=1}^{n} x_i^2 + \left(\sum_{i=1}^{d} \frac{1}{2} i x_i\right)^2 + \left(\sum_{i=1}^{d} \frac{1}{2} i x_i\right)^4. \tag{14}$$

Before running our experiments, we run an initial test for each solver that applies dimensionality-reduction techniques on each problem. We run each problem-solver pair at each possible subspace dimension for their respective problems. We then identify the subspaces that perform the best after 10 macroreplications and 100

postreplications on the SimOpt library for each problem-solver, selecting these optimal subspace dimensions in our comparison of the finite-time performance against other SO solvers on the same problem suite. This preliminary investigation into the optimal subspace for OMoRF and ASTROMoRF, helps to understand the underlying structure of each problem. In practice, it is useful for an experimenter to understand the structure of the simulation model and its source of randomness. Heuristically, choosing a smaller subspace dimension without this preliminary investigation will achieve comparable results; however these results may be present with more variability between macroreplications due to poor handling of projection error.

Each problem-solver pair is executed until a specified simulation budget of $10,000$ calls to the simulation is exhausted. We run 20 macroreplications on each problem-solver pair before running 200 postreplications on each problem-solver pair at the intermediate solutions of each macroreplication to estimate the unbiased objective function. For each of the experiments, we keep the parameters the same, set to the initial values in Table 1, which are common within the literature. We set an initial solution for the newsvendor problem with dynamic substitution to $(3,\ldots,3) \in \mathbb{R}^{10}$. The initial solutions for the Rosenbrock and the Zakharov functions are both set to $(2,\ldots,2) \in \mathbb{R}^{15}$.

Table 1: Initial Parameters for Solvers.

Solver	Subspace Dimension	η_1	η_2	γ_1	γ_2	$\lambda_{\min}$
ASTROMoRF	DYNAMNEWS-1: 1 ROSENBROCK-1: 5 ZAKHAROV-1: 8	0.1	0.8	2.5	0.5	5
OMoRF	DYNAMNEWS-1: 1 ROSENBROCK-1: 1 ZAKHAROV-1: 7	0.1	0.8	2.5	0.5	NA
ASTRO-DF	NA	0.1	0.8	2.5	0.5	5

Figure 1 and Table 2 suggest that ASTROMoRF solves the three problems extremely well with rapid progress towards stationary points of different high-dimensional problems. In fact, within less than 200 replications of the simulation model, ASTROMoRF has solved almost 85% of the problems within a 0.1-optimality gap. One of the issues highlighted with ASTRO-DF, which is present in Figure 1 is that, for higher-dimensional problems, there is a transient phase of progress early on as it makes steady progression towards optimal solutions. With ASTROMoRF, we have eliminated this transient phase for high-dimensional problems. As seen in Table 2, ASTRO-DF presents much larger relative optimality gaps with fewer iterations in comparison to ASTROMoRF. This contributes to much larger average decreases in the objective function value on successful iterations. When comparing ASTRO-DF and ASTROMoRF on DYNAMNEWS-1, where both solvers undergo a similar number of iterations, we see that ASTROMoRF has a larger average decrease on successful iterations. This implies that while both algorithms display steady trajectories ASTROMoRF seemingly exploits the geometry of the space and makes a consistent and substantial jump to a neighborhood of the optimal solution early on, before engaging in a very stable exploitation phase of converging around the optimal solution. It can be seen in Figure 1 that OMoRF has not performed well, which we assume is because of the stochastic noise present in the problems. OMoRF is not designed to handle stochastic sampling error and the poor performance of OMoRF justifies our use of adaptive sampling on the replications of the model to improve the performance in a stochastic setting.

4 CONCLUSION

Within the past few years, we have seen a rise in simulation problems with decision spaces that have become increasingly higher dimensional. We have shown that applying a popular variational dimensionality-reduction method of active subspaces is a viable method for handling these high-dimensional SO problems. In this paper, we have described a novel solver, ASTROMoRF, that combines the active subspace dimensionality-reduction seen in deterministic-solvers like OMoRF, with the adaptive sampling rule of algorithms like ASTRO-DF. We have discussed implementation aspects of ASTROMoRF that allow the algorithm to

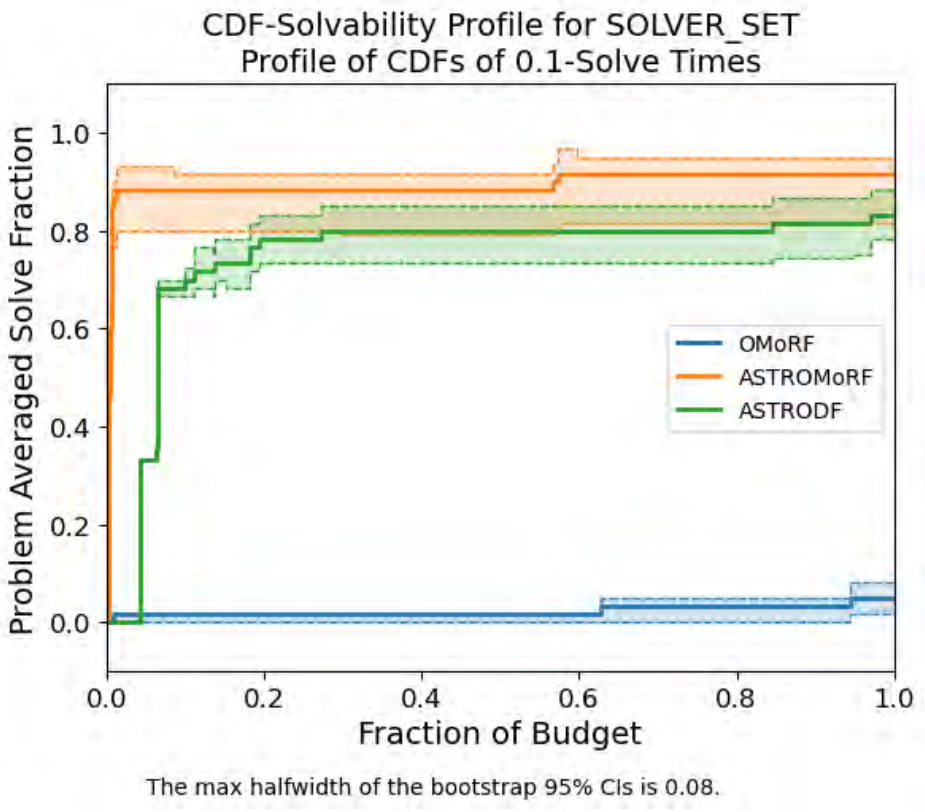

Figure 1: Fraction of problems solved within the SimOpt library solved to a 0.1-optimality with 95% confidence intervals from 20 runs of each algorithm.

Table 2: Computational times and performance benchmarks of each Problem-Solver pair.

Solver	Computational Times (s)	Iterations	Relative Optimality Gap	Average Decrease on Successful Iterations
ASTROMoRF	DYNAMNEWS-1: 1.706 ROSENBROCK-1: 0.2305 ZAKHAROV-1: 0.3545	DYNAMNEWS-1: 9.2 ROSENBROCK-1: 64.95 ZAKHAROV-1: 96.25	DYNAMNEWS-1: 9.80 ROSENBROCK-1: 5.93 ZAKHAROV-1: 0.40	DYNAMNEWS-1: 18.41 ROSENBROCK-1: 352.56 ZAKHAROV-1: 7.05×10^6
OMoRF	DYNAMNEWS-1: 2.06 ROSENBROCK-1: 0.068 ZAKHAROV-1: 2.138	DYNAMNEWS-1: 11.8 ROSENBROCK-1: 17.25 ZAKHAROV-1: 632.3	DYNAMNEWS-1: 26.28 ROSENBROCK-1: 3049.69 ZAKHAROV-1: 8.26×10^7	DYNAMNEWS-1: 29.26 ROSENBROCK-1: 1.13×10^8 ZAKHAROV-1: 3291.10
ASTRO-DF	DYNAMNEWS-1: 1.834 ROSENBROCK-1: 0.0745 ZAKHAROV-1: 0.082	DYNAMNEWS-1: 9.15 ROSENBROCK-1: 19.95 ZAKHAROV-1: 20.75	DYNAMNEWS-1: 4.35 ROSENBROCK-1: 17.17 ZAKHAROV-1: 5.36	DYNAMNEWS-1: 10.60 ROSENBROCK-1: 441.17 ZAKHAROV-1: 1.66×10^7

be convergent and reduce the number of simulation replications when dealing with high-dimensional SO problems. The numerical results reported suggest that ASTROMoRF is a reasonable framework for simulation models with high dimensionality and low to high variability.

Ongoing research will focus on an extended analysis of theoretical and practical matters surrounding ASTROMoRF. These include undergoing more testing on a larger problem suite, including more problems that are stochastic simulation models instead of deterministic functions with added stochastic noise, to observe how the solver performs against problems that are less prone to artificial solution-dependent estimators (Ha and Shashaani 2024). Thought also needs to be given to identifying the optimal subspace dimension in practice and the penalties associated with mis-identification. Future work will include a more comprehensive set of convergence results. Currently ASTROMoRF is ensured convergence by the notion

that the model is ensured to be fully linear using results from Bandeira et al. (2014). An analysis of the asymptotic convergence rate of ASTROMoRF, and its comparison to the optimal asymptotic convergence rate of gradient-based solvers of $\mathcal{O}(n^{-\frac{1}{2}})$ is needed, as well as an analysis of the iteration complexity to reach ε-optimality in expectation.

REFERENCES

Asmussen, S., and P. W. Glynn. 2007. *Stochastic Simulation: Algorithms and Analysis*, Volume 57 of *Stochastic Modelling and Applied Probability*. Springer.

Bandeira, A. S., K. Scheinberg, and L. N. Vicente. 2014. "Convergence of Trust-Region Methods based on Probabilistic Models". *SIAM Journal on Optimization* 24(3):1238–1264.

Cartis, C., and L. Roberts. 2019. "A Derivative-Free Gauss-Newton Method". *Mathematical Programming Computation* 11:631–674.

Conn, A. R., K. Scheinberg, and L. N. Vicente. 2008. "Geometry of Interpolation Sets in Derivative Free Optimization". *Mathematical programming* 111:141–172.

Conn, A. R., K. Scheinberg, and L. N. Vicente. 2009. *Introduction to Derivative-Free Optimization*. SIAM.

Constantine, P. G., E. Dow, and Q. Wang. 2014. "Active Subspace Methods in Theory and Practice: Applications to Kriging Surfaces". *SIAM Journal on Scientific Computing* 36(4):A1500–A1524.

Constantine, P. G., A. Eftekhari, J. Hokanson, and R. A. Ward. 2017. "A Near-Stationary Subspace for Ridge Approximation". *Computer Methods in Applied Mechanics and Engineering* 326:402–421.

Coope, I. D., and R. Tappenden. 2021. "Gradient and Diagonal Hessian Approximations using Quadratic Interpolation Models and Aligned Regular Bases". *Numerical Algorithms* 88:767–791.

De Jong, K. A. 1975. *An Analysis of the Behavior of a Class of Genetic Adaptive Systems*. University of Michigan.

Deng, G., and M. C. Ferris. 2006. "Adaptation of the UOBYQA Algorithm for Noisy Functions". In *2006 Winter Simulation Conference (WSC)*, 312–319 https://doi.org/10.1109/WSC.2006.323088.

Eckman, D. J., S. G. Henderson, and S. Shashaani. 2023. "SimOpt: A Testbed for Simulation-Optimization Experiments". *INFORMS Journal on Computing* 35(2):495–508.

Fan, W., L. J. Hong, G. Jiang, and J. Luo. 2024. "Review of Large-Scale Simulation Optimization". *arXiv preprint arXiv:2403.15669*.

Golub, G. H., and V. Pereyra. 1973. "The Differentiation of Pseudo-Inverses and Nonlinear Least Squares Problems whose Variables Separate". *SIAM Journal on numerical analysis* 10(2):413–432.

Gould, N. I., D. Orban, A. Sartenaer, and P. L. Toint. 2005. "Sensitivity of Trust-Region Algorithms to their Parameters". *4OR* 3:227–241.

Gross, J. C., and G. T. Parks. 2022. "Optimization by Moving Ridge Functions: Derivative-Free optimization for Computationally Intensive Functions". *Engineering Optimization* 54(4):553–575.

Gross, J. C., P. Seshadri, and G. Parks. 2020. "Optimisation with Intrinsic Dimension Reduction: A Ridge Informed Trust-Region Method". In *AIAA Scitech 2020 Forum*.

Ha, Y., and S. Shashaani. 2024. "Iteration Complexity and Finite-Time Efficiency of Adaptive Sampling Trust-Region Methods for Stochastic Derivative-Free Optimization". *IISE Transactions* (just-accepted):1–26.

Ha, Y., S. Shashaani, and M. Menickelly. 2024. "Two-Stage Estimation and Variance Modeling for Latency-Constrained Variational Quantum Algorithms". *INFORMS Journal on Computing*.

Hokanson, J. M., and P. G. Constantine. 2018. "Data-Driven Polynomial Ridge Approximation using Variable Projection". *SIAM Journal on Scientific Computing* 40(3):A1566–A1589.

Kiefer, J., and J. Wolfowitz. 1952. "Stochastic Estimation of the Maximum of a Regression Function". *The Annals of Mathematical Statistics* 23(3):462–466.

Kleijnen, J. P. 2005. "An Overview of the Design and Analysis of Simulation Experiments for Sensitivity Analysis". *European Journal of Operational Research* 164(2):287–300.

Mahajan, S., and G. Van Ryzin. 2001. "Stocking Retail Assortments under Dynamic Consumer Substitution". *Operations Research* 49(3):334–351.

Nocedal, J., and S. J. Wright. 1999. *Numerical Optimization*, Volume 2 of *Springer Series in Operations Research and Financial Engineering*. Springer.

Powell, M. J. 2002. "UOBYQA: Unconstrained Optimization by Quadratic Approximation". *Mathematical Programming* 92(3):555–582.

Powell, M. J. 2006. "The NEWUOA Software for Unconstrained Optimization without Derivatives". In *Large-scale nonlinear optimization*, 255–297. Springer.

Powell, M. J. *et al*. 2009. "The BOBYQA Algorithm for Bound Constrained Optimization without Derivatives". *Cambridge NA Report NA2009/06, University of Cambridge, Cambridge* 26:26–46.

Ragonneau, T. M., and Z. Zhang. 2024. "An Optimal Interpolation Set for Model-Based Derivative-Free Optimization Methods". *Optimization Methods and Software* 39(4):898–910.

Russi, T. M. 2010. *Uncertainty Quantification with Experimental Data and Complex System Models*. University of California, Berkeley.

Saltelli, A., M. Ratto, T. Andres, F. Campolongo, J. Cariboni, D. Gatelli, *et al*. 2008. *Global Sensitivity Analysis: the Primer*. John Wiley & Sons.

Shashaani, S., S. R. Hunter, and R. Pasupathy. 2016. "ASTRO-DF: Adaptive Sampling Trust-Region Optimization Algorithms, Heuristics, and Numerical Experience". In *2016 Winter Simulation Conference (WSC)*, 554–565 https://doi.org/10.1109/WSC.2016.7822121.

Smith, R. C. 2024. *Uncertainty Quantification: Theory, Implementation, and Applications, Second Edition*. Philadelphia, PA: Society for Industrial and Applied Mathematics.

Spall, J. C. 1992. "Multivariate Stochastic Approximation using a Simultaneous Perturbation Gradient Approximation". *IEEE transactions on automatic control* 37(3):332–341.

Sonja Surjanovic and Derek Bingham 2013. "Zakharov Function". https://www.sfu.ca/~ssurjano/zakharov.html. Accessed: 26th March 2025.

Tay, T., and C. Osorio. 2024. "A Sampling Strategy for High-Dimensional, Simulation-Based Transportation Optimization Problems". *Transportation Science* 58(5):947–972.

Wang, T., and L. J. Hong. 2023. "Large-scale Inventory Optimization: A Recurrent Neural Networks-Inspired Simulation Approach". *INFORMS Journal on Computing* 35(1):196–215.

Xie, W., B. L. Nelson, and R. R. Barton. 2014. "A Bayesian Framework for Quantifying Uncertainty in Stochastic Simulation". *Operations Research* 62(6):1439–1452.

ACKNOWLEDGMENTS

We are grateful for funding from the Engineering and Physical Sciences Research Council under grant EP/W524621/1 that provided support for this work.

AUTHOR BIOGRAPHIES

BENJAMIN REES is a PhD student in the School of Mathematical Sciences at the University of Southampton. His email address is B.Rees@soton.ac.uk.

CHRISTINE S.M. CURRIE is a Professor of Operational Research in Mathematical Sciences at the University of Southampton and a member of the Centre for Operational Research, Management Sciences and Information Systems (CORMSIS). Her email address is christine.currie@soton.ac.uk and her homepage is https://www.southampton.ac.uk/people/5wzzxf/professor-christine-currie.

VUONG PHAN is an Associate Professor of Operational Research in Mathematical Sciences at the University of Southampton and the Deputy director of CORMSIS. His email address is t.v.phan@soton.ac.uk and his homepage is https://www.southampton.ac.uk/people/5y2ds9/doctor-vuong-phan.