Proceedings of the 2021 Winter Simulation Conference
S. Kim, B. Feng, K. Smith, S. Masoud, Z. Zheng, C. Szabo, and M. Loper, eds.

FLAT CHANCE! USING STOCHASTIC GRADIENT ESTIMATORS TO ASSESS PLAUSIBLE OPTIMALITY FOR CONVEX FUNCTIONS

David J. Eckman

Industrial and Systems Engineering
Texas A&M University
Emerging Technologies Building
College Station, TX 77843, USA

Matthew Plumlee
Barry L. Nelson

Industrial Engineering and Management Sciences
Northwestern University
2145 Sheridan Road
Evanston, IL 60208, USA

ABSTRACT

This paper studies methods that identify plausibly near-optimal solutions based on simulation results obtained from only a small subset of feasible solutions. We do so by making use of both noisy estimates of performance and their gradients. Under a convexity assumption on the performance function, these inference methods involve checking only a system of inequalities. We find that these methods can yield more powerful inference at less computational expense compared to methodological predecessors that do not leverage stochastic gradient estimators.

1 INTRODUCTION

Consider a stochastic simulation model of a system parameterized by a real-valued vector. We refer to a specific setting of input parameters as a solution and assume it is one of many (possibly infinite) in a space of solutions under consideration. The performance of a solution is unknown but can be estimated by averaging noisy observations produced by running independent replications of the simulation with the corresponding inputs. In simulation optimization, our goal is to determine which of the feasible solutions result in optimal (or near-optimal) performance using these noisy observations.

Some simulation models, such as structured queueing systems (Plambeck et al. 1996) and stochastic activity networks (Fu 2015), give rise to convex performance functions over continuous solution spaces. While some simulation-optimization algorithms have theoretical guarantees (e.g., convergence rates) when the performance function is convex, many do not directly exploit this property to improve empirical performance. Plumlee and Nelson (2018) and Eckman et al. (2021) laid the groundwork for plausible screening (PS), a framework in which known or assumed functional properties can be incorporated to screen out unacceptable solutions based on sampling at only a small subset of solutions. These methods borrow ideas from constrained statistical inference (Silvapulle and Sen 2005) to trade simulation for optimization, screening solutions by solving linear or quadratic programs.

In the absence of information about the performance function, the PS approach reduces to traditional subset selection (Eckman et al. 2020). When equipped with functional information, PS offers significant advantages over other subset-selection approaches, namely, one can deliver statistical guarantees on the quality of solutions without having to simulate them. However, existing deployments of PS have their limitations. These methods are designed with large, but not continuous, solution spaces in mind, because optimization is required to determine plausibility. Moreover, existing PS methods assume that estimators of only the performance are available, not estimators of any other functional (Eckman et al. 2021). The goal of this paper is to enhance current PS methodology when stochastic gradient estimators are available.

We consider continuous solution spaces with simulation models that admit stochastic gradient estimators of the performance function. The approach of discretizing a continuous solution space and screening a subset of solutions becomes computationally onerous, especially in higher dimensions. In this setting, a simulation-optimization algorithm can instead use PS methods to check the plausible acceptability of some candidate solutions before simulating them. When available, a stochastic gradient estimator provides a first-order approximation of the performance function within a neighborhood of a simulated solution. Direct gradient estimators—those obtained by simulating only the solution in question—are especially appealing since the solution is already being simulated to estimate its performance. Also desirable are estimators with low variance, as is typically the case with those resulting from infinitesimal perturbation analysis.

Stochastic gradient estimators have been exploited in different settings to enhance simulation-optimization algorithms or output analysis. Gradients from deterministic computer experiments have been used to build Gaussian-process-based metamodels of the performance function for either prediction (Morris et al. 1993) or optimization (Forrester and Keane 2009). These approaches directly model the gradients with another Gaussian process. Chen et al. (2013) similarly exploit gradient estimators for prediction in the context of stochastic simulation experiments (i.e., stochastic kriging) and study the effects of different correlations between performance and gradient estimators and their relative variability. Qu and Fu (2014) take a different approach to incorporating gradients in stochastic kriging, effectively adding solutions to the experimental set by extrapolating their performances using the gradient from a nearby solution. Placing a Gaussian process prior on the performance function and its gradients is also common practice in Bayesian optimization. In this area, stochastic gradient estimators have been incorporated to direct sequential sampling (Wu et al. 2017). In a related vein of research, Jian and Henderson (2020) propose a sequential sampling procedure that assesses the posterior probability that there exists a convex function interpolating the unknown performance function at a finite set of simulated solutions. Their methods, however, are not designed to incorporate stochastic gradient estimators.

This paper describes simple approaches that find plausibly near-optimal solutions of convex performance functions by checking systems of inequalities featuring performance and gradient estimators. These approaches have desirable statistical properties of confidence and consistency. The remainder of the paper is organized as follows. We introduce the experimental setup in Section 2, explaining how performances and gradients at solutions are estimated and describing the statistical guarantees sought by PS methods. Section 3 adopts the PS approach of incorporating known or assumed properties and presents a mathematical formulation of near-optimal solutions of convex performance functions that forms the basis for our methods. Sections 4 and 5 propose two subsets of solutions defined by inequalities where the second features only gradients of the performance function, and in Section 6 we establish the consistency of the methods. We state a number of theorems, with proofs omitted, that formalize our discussion. In Section 7, we illustrate our new methods on an artificial two-dimensional example, comparing them to the original PS, which does not make use of gradient estimators. We lay out future research directions in Section 8.

2 SETTING AND GOALS

Much of the setup and notation follows from Eckman et al. (2021).

2.1 Stochastic Simulation with Direct Gradient Estimators

Consider a continuous space of candidate solutions $\mathscr{X} \subseteq \mathbb{R}^q$, where a solution is represented by a vector of parameters $x \in \mathscr{X}$ that fully specifies a simulated system. Each solution x has an associated scalar performance, $\mu(x)$, that is a key performance indicator of interest to a decision-maker, e.g., the expected cost or expected throughput of the system. The performance $\mu(x)$ is unknown, but can be estimated by obtaining replications from the simulation model described by x. We assume that the performance—when viewed as a function $\mu\colon \mathscr{X} \mapsto \mathbb{R}$—is differentiable at all $x \in \mathscr{X}$, and we let the column vector $\nabla\mu(x) \in \mathbb{R}^q$

denote the gradient of the performance function at x. Throughout the paper, it will be notationally convenient to concatenate the performance and gradient of a solution x, hence we define $\zeta(x) \equiv (\mu(x), \nabla\mu(x)^\top)^\top$.

The decision-maker initially selects a set of k solutions to simulate, denoted by $\mathsf{X} \equiv \{x_1, x_2, \ldots, x_k\}$, which we refer to as the experimental set. The experimental set may consist of solutions chosen to fill the space $\mathscr{X}$, for instance. Let $\mu(\mathsf{X}) \equiv (\mu(x_1), \mu(x_2), \ldots, \mu(x_k))^\top$ denote the restriction of the performance function μ to X and let $\nabla\mu(\mathsf{X}) \equiv (\nabla\mu(x_1)^\top, \nabla\mu(x_2)^\top, \ldots, \nabla\mu(x_k)^\top)^\top$ likewise denote the restriction of the gradient $\nabla\mu$, when viewed as a vector-valued function. Concatenating the performances and gradients for each solution in X, we define

$$\zeta(\mathsf{X}) \equiv \left(\zeta(x_1)^\top, \zeta(x_2)^\top, \ldots, \zeta(x_k)^\top\right)^\top = \left(\mu(x_1), \nabla\mu(x_1)^\top, \mu(x_2), \nabla\mu(x_2)^\top, \ldots, \mu(x_k), \nabla\mu(x_k)^\top\right)^\top.$$

A single generic replication at a solution x_i yields a stochastic output, $Y(x_i) \in \mathbb{R}$, which is assumed to be unbiased, i.e., $\mathbb{E}[Y(x_i)] = \mu(x_i)$ for $i = 1, 2, \ldots, k$. A replication at x_i simultaneously yields a stochastic estimator of the gradient, $G(x_i) \in \mathbb{R}^q$, which is also assumed to be unbiased, i.e., $\mathbb{E}[G(x_i)] = \nabla\mu(x_i)$. We correspondingly define $Z(x_i) \equiv (Y(x_i), G(x_i)^\top)^\top$, thus $\mathbb{E}[Z(x_i)] = \zeta(x_i)$. Examples of direct gradient estimators include the likelihood ratio (LR) or score function (SF) estimator (Glynn 1987; Rubinstein 1989) and the infinitesimal perturbation analysis (IPA) estimator (Glasserman 1991). Assuming certain technical conditions are met, the LR/SF and IPA estimators are both unbiased (Fu 2015).

For any solution $x_i \in \mathsf{X}$, the performance estimator $Y(x_i)$ and gradient estimator $G(x_i)$ are almost certainly dependent, as they come from the same replication. Let $\Psi(x_i)$ be the real-valued joint variance-covariance matrix of $Y(x_i)$ and $G(x_i)$, i.e., the variance-covariance matrix of $Z(x_i)$:

$$\Psi(x_i) \equiv \mathbb{E}\left[Z(x_i)Z(x_i)^\top\right] - \zeta(x_i)\zeta(x_i)^\top,$$

which is assumed to be positive definite.

Obtaining a single replication at each solution in the experimental set yields a noisy estimate $Z(\mathsf{X}) \equiv (Z(x_1)^\top, Z(x_2)^\top, \ldots, Z(x_k)^\top)^\top$, and we let $\Psi(\mathsf{X})$ be the variance-covariance matrix of $Z(\mathsf{X})$, i.e.,

$$\Psi(\mathsf{X}) \equiv \mathbb{E}\left[Z(\mathsf{X})Z(\mathsf{X})^\top\right] - \zeta(\mathsf{X})\zeta(\mathsf{X})^\top.$$

We assume that solutions in the experimental set are simulated independently, thus $\Psi(\mathsf{X})$ has a block-diagonal structure with $\Psi(\mathsf{X}) = \mathrm{diag}(\Psi(x_1), \Psi(x_2), \ldots, \Psi(x_k))$. The more general case when solutions are simulated dependently can be developed, but we do not do so here. For a basic treatment of plausible screening with dependent sampling, see Eckman et al. (2021). In our finite-sample results, we will additionally assume that the performance and gradient estimators are jointly normally distributed, meaning

$$Z(x_i) = \begin{pmatrix} Y(x_i) \\ G(x_i) \end{pmatrix} \sim \mathscr{N}\left(\begin{pmatrix} \mu(x_i) \\ \nabla\mu(x_i) \end{pmatrix}, \Psi(x_i)\right) = \mathscr{N}\left(\zeta(x_i), \Psi(x_i)\right) \quad \text{for } i = 1, 2, \ldots, k. \tag{1}$$

At each solution $x_i \in \mathsf{X}$, we run n_i replications and obtain independent and identically distributed (i.i.d.) observations $Z_1(x_i), Z_2(x_i), \ldots, Z_{n_i}(x_i)$, which consist of i.i.d. performance estimators $Y_1(x_i), Y_2(x_i), \ldots, Y_{n_i}(x_i)$ and i.i.d. gradient estimators $G_1(x_i), G_2(x_i), \ldots, G_{n_i}(x_i)$. The sample means

$$\widehat{\mu}_i \equiv \frac{1}{n_i}\sum_{\ell=1}^{n_i} Y_\ell(x_i) \quad \text{and} \quad \widehat{\nabla}\mu_i \equiv \frac{1}{n_i}\sum_{\ell=1}^{n_i} G_\ell(x_i)$$

are unbiased estimators of $\mu(x_i)$ and $\nabla\mu(x_i)$, respectively. Furthermore,

$$\widehat{\zeta}_i \equiv \frac{1}{n_i}\sum_{\ell=1}^{n_i} Z_\ell(x_i) \equiv \left(\widehat{\mu}_i, \widehat{\nabla}\mu_i^\top\right)^\top$$

is an unbiased estimator for $\zeta(x_i)$. Let $\widehat{\zeta} \equiv \left(\widehat{\zeta}_1^\top, \widehat{\zeta}_2^\top, \ldots, \widehat{\zeta}_k^\top\right)^\top$ denote the resulting estimator of $\zeta(\mathsf{X})$ from simulating all solutions in the experimental set with sample sizes described by $\mathsf{n} \equiv (n_1, \ldots, n_k)$. When sampling independently across solutions, the variance-covariance matrix $\Psi(\mathsf{X})$ can be estimated by $\widehat{\Psi} \equiv \text{diag}(\widehat{\Psi}_1, \widehat{\Psi}_2, \ldots \widehat{\Psi}_k)$, where $\widehat{\Psi}_i$ is the sample variance-covariance matrix from the replications taken at solution x_i:

$$\widehat{\Psi}_i \equiv \frac{1}{n_i - 1} \sum_{\ell=1}^{n_i} \left(Z_\ell(x_i) - \widehat{\zeta}_i\right)\left(Z_\ell(x_i) - \widehat{\zeta}_i\right)^\top \quad \text{for } i = 1, 2, \ldots, k.$$

2.2 Acceptable Solutions

At the most basic level, the decision-maker seeks to determine whether the performance of an arbitrary solution $x_0 \in \mathscr{X}$ is *acceptable*. Eckman et al. (2021) provide several definitions of acceptability that reflect different objectives, e.g., optimal, feasible, better than a control, and equal to a target. In all of these cases, the acceptability of a solution x_0 is expressed in terms of its performance, $\mu(x_0)$, and possibly its relationship to the performances of other solutions. To ground our presentation, we focus on the case where a solution is deemed acceptable if its performance is near optimal. Specifically, we define the set of acceptable solutions as $\mathscr{A} \equiv \{x_0 \in \mathscr{X} : \mu(x_0) \leq \inf_{x \in \mathscr{X}} \mu(x) + \delta\}$, where $\delta \geq 0$ is some user-specified tolerance and the choice of minimization is without loss of generality. Enlarging the set of solutions treated as "good enough" affords the decision-maker an opportunity to make a final selection from among high-quality solutions based on secondary performance measures.

2.3 Confidence and Consistency

We refer to the operation of determining whether a solution's performance is acceptable as *screening*. Our methods can screen unsimulated solutions by using replications obtained from solutions in the experimental set and any other available information about the performance function or its gradient.

An important quality of a screening procedure is its ability to make correct screening decisions. We desire that a screening procedure be able to—with high probability—retain a solution with acceptable performance and screen out (i.e., remove from consideration) a solution with unacceptable (δ-suboptimal) performance. We describe properties of a screening procedure in terms of how it would perform if applied to *all* solutions in $\mathscr{X}$. We denote the resulting set of retained solutions by $\mathscr{S}_\mathsf{n} \subseteq \mathscr{X}$ and note that it is random, since it depends on the estimated performances and gradients at solutions in the experimental set. For continuous solution spaces, constructing $\mathscr{S}_\mathsf{n}$ is likely impossible; a screening procedure could instead screen a large finite set of solutions or a sequence of solutions identified by a simulation-optimization algorithm.

The following definitions, which are reproduced from Eckman et al. (2021), presume that the performance function μ belongs to some function space $\mathscr{M}$, which we make more precise in Section 3.

Definition 1 (Finite-sample confidence) A subset $\mathscr{S}_\mathsf{n}$ achieves *finite-sample confidence* $1-\alpha$ for $1-\alpha \in (1/2, 1)$ if for sufficiently large $\min_{i=1,\ldots,k} n_i$ and any $\mu \in \mathscr{M}$, $\mathbb{P}(x_0 \in \mathscr{S}_\mathsf{n}) \geq 1-\alpha$ for all $x_0 \in \mathscr{A}$.

Definition 2 (Asymptotic confidence) A subset $\mathscr{S}_\mathsf{n}$ achieves *asymptotic confidence* $1-\alpha$ for $1-\alpha \in (1/2, 1)$ if for any $\mu \in \mathscr{M}$, $\mathbb{P}(x_0 \in \mathscr{S}_\mathsf{n}) \gtrsim 1-\alpha$ as $\min_{i=1,\ldots,k} n_i \to \infty$ for all $x_0 \in \mathscr{A}$.

Definition 3 (Consistency) A subset $\mathscr{S}_\mathsf{n}$ achieves *consistency* if for any $\mu \in \mathscr{M}$, $\mathbb{P}(x_0 \in \mathscr{S}_\mathsf{n}) \to 0$ as $\min_{i=1,\ldots,k} n_i \to \infty$ for all $x_0 \notin \mathscr{A}$.

In Definition 2, the notation $\mathbb{P}(x_0 \in \mathscr{S}_\mathsf{n}) \gtrsim 1-\alpha$ means that for any $\varepsilon > 0$, there exists a minimum sample size $n(\varepsilon, x_0)$ such that $\mathbb{P}(x_0 \in \mathscr{S}_\mathsf{n}) \geq 1-\alpha-\varepsilon$ for all n for which $\min_{i=1,\ldots,k} n_i \geq n(\varepsilon, x_0)$.

Confidence and consistency describe a procedure's ability to retain acceptable solutions and screen out unacceptable solutions, respectively. Delivering finite-sample confidence requires knowing the family of distributions for $\widehat{\zeta}_1, \widehat{\zeta}_2, \ldots, \widehat{\zeta}_k$, as in (1). We achieve asymptotic confidence by designing procedures for normally distributed performance and gradient outputs, since the multivariate form of the Central Limit

Theorem implies that even if $\widehat{\zeta}_i$ is not normally distributed,

$$\sqrt{n_i}\left(\widehat{\zeta}_i - \zeta(x_i)\right) \xrightarrow{d} \mathcal{N}\left(\mathbf{0}_{q+1}, \Psi(x_i)\right) \text{ as } n_i \to \infty \quad \text{for } i = 1,2,\ldots,k,$$

where $\xrightarrow{d}$ denotes convergence in distribution and $\mathbf{0}_{q+1}$ is a column vector of $q+1$ zeros; see Theorem 3.4.3 of Anderson (1984). Consistency, on the other hand, is generally unattainable, since only a subset of solutions has been simulated. In Section 6, we introduce a relaxed version of consistency that represents a reasonable goal for a procedure's screening power as sample sizes increase to infinity.

3 CONVEX SIMULATION OPTIMIZATION WITH GRADIENTS

The plausible screening framework assumes that the decision-maker possesses known or assumed properties of the unknown performance function. Examples include convexity of μ and Lipschitz continuity of μ or $\nabla\mu$. Eckman et al. (2021) provide examples of simulation models for which some of these properties are verifiable and point to sample-path arguments and stochastic orders as tools that can help to establish them. Knowledge of properties of the performance function restricts the space of functions $\mathscr{M}$ to which μ can belong and thereby facilitates sharper inferences regarding a solution's acceptability. In this paper, we focus on the case where μ is convex over $\mathscr{X}$ and everywhere differentiable.

As an illustration of the power of gradient information, consider the simple case where replications are obtained at a single solution x and produce estimators $\widehat{\mu}(x)$ and $\widehat{\nabla}\mu(x)$. If only $\widehat{\mu}(x)$ were available, it would be impossible to conclude with high confidence that any given solution x_0 is δ-suboptimal—there are simply no means of inferring the performance of any other solution *relative* to $\mu(x)$. On the other hand, if $\widehat{\nabla}\mu(x)$ were available, controlling its associated error could allow some solutions located in (estimated) directions of ascent from x to be screened out. To be precise, $\widehat{\mu}(x) + (x_0 - x)^\top \widehat{\nabla}\mu(x)$ is a noisy estimate of the hyperplane with respect to x_0 that passes through $\mu(x)$ at x and supports (i.e., lies below) μ. Were it known, the true hyperplane $\mu(x) + (x_0 - x)^\top \nabla\mu(x)$ could be manipulated to describe a halfspace of solutions that are all δ-suboptimal, namely those solutions x_0 for which $(x_0 - x)^\top \nabla\mu(x) > \delta$. We show in Section 5 that an estimate of the gradient alone at a given solution x allows a cone of solutions to be screened out. However, if the estimated hyperplane is too flat, i.e, if $\|\widehat{\nabla}\mu(x)\|$ is too small, no solutions can be screened out, since no directions of ascent can be determined with high confidence.

We next formalize our mathematical framework for screening. Let $\mathscr{M}$ be the space of functions that are convex over $\mathscr{X}$ and everywhere differentiable. When screening a given solution x_0, we focus on the subset of functions in $\mathscr{M}$ for which x_0 is acceptable, denoted by $\mathscr{M}(x_0) \equiv \{m \in \mathscr{M} : x_0 \in \mathscr{A}(m)\}$, where m is a generic function and $\mathscr{A}(m)$ indicates the set of acceptable solutions given m. Thus, $\mathscr{M}(x_0)$ is the space of differentiable convex functions for which solution x_0 is δ-optimal.

Rather than work with an abstract space like $\mathscr{M}(x_0)$, we project $\mathscr{M}(x_0)$ to a finite-dimensional space described by the performances and gradients at the solutions in the experimental set:

$$\left\{\mathsf{m} \in \mathbb{R}^k, \mathsf{g} \in \mathbb{R}^{kq} : \text{ there exists } m \in \mathscr{M}(x_0) \text{ such that } m(\mathsf{X}) = \mathsf{m} \text{ and } \nabla m(\mathsf{X}) = \mathsf{g}\right\}, \tag{2}$$

where $\mathsf{m} \equiv (\mathsf{m}_1, \mathsf{m}_2, \ldots, \mathsf{m}_k)$, $\mathsf{g} \equiv \left(\mathsf{g}_1^\top, \mathsf{g}_2^\top, \ldots, \mathsf{g}_k^\top\right)^\top$, and $\mathsf{g}_i \in \mathbb{R}^q$ for $i = 1,2,\ldots,k$. In words, (2) is the set of k performances and k gradients that admit a convex function that interpolates them at the corresponding solutions in X and for which x_0 is δ-optimal. We can derive an outer approximation to (2) described by a finite system of linear inequalities:

$$\begin{aligned}\mathsf{Z}(x_0) = \Big\{\mathsf{m} \in \mathbb{R}^k, \mathsf{g} \in \mathbb{R}^{kq} : \text{there exists } \mathsf{m}_0 \in \mathbb{R} \text{ such that} \\ \mathsf{m}_i - \mathsf{m}_0 + (x_0 - x_i)^\top \mathsf{g}_i \le 0 \text{ for all } i = 1,2,\ldots,k \\ -\mathsf{m}_i + \mathsf{m}_0 \le \delta \text{ for all } i = 1,2,\ldots,k\Big\}.\end{aligned} \tag{3}$$

The first set of inequalities asserts that the performance of solution x_0—represented by the term m_0—lies above each of the hyperplanes described by the gradients and performances at each solution in X. The second set of inequalities asserts that solution x_0 is δ-optimal relative to the other solutions in X. We could also include constraints of the form $\mathsf{m}_i - \mathsf{m}_j + (x_j - x_i)^\top \mathsf{g}_i \leq 0$ for all $i, j = 1, 2, \ldots, k$, which collectively assert that the performance function is convex over $x_1, x_2, \ldots, x_k$. In experiments we observed insignificant benefit when including these constraints as they provide little information for determining the near-optimality of solution x_0; we therefore restricted our attention to constraints pertaining to x_0. It is possible that these additional constraints might be critical for inferring other aspects of the function, such as if the performance function is convex.

Projecting out m_0 from (3) makes it evident that $\mathsf{Z}(x_0)$ is a polyhedron in terms of m and g:

$$\mathsf{Z}(x_0) = \left\{ \mathsf{m} \in \mathbb{R}^k, \mathsf{g} \in \mathbb{R}^{kq} \colon \mathsf{m}_i - \mathsf{m}_j + (x_0 - x_i)^\top \mathsf{g}_i \leq \delta \text{ for all } i, j = 1, 2, \ldots, k \right\}. \tag{4}$$

This observation will be helpful in the next section when we use the machinery of mathematical programming to design new screening methods.

4 PLAUSIBLE SCREENING WITH GRADIENTS

We develop a computationally cheap screening method that involves checking whether $\widehat{\zeta}$ belongs to a relaxation of $\mathsf{Z}(x_0)$. The amount by which $\mathsf{Z}(x_0)$, or more precisely, the right-hand-side vector of (4), is relaxed is chosen to compensate for the error associated with using $\widehat{\zeta}$ as an estimator for $\zeta(\mathsf{X})$. The following sequence of ideas incorporates stochastic gradient estimators into the *relaxed* PS approach proposed in Eckman et al. (2021). When referencing the original and adapted methods, we drop the "relaxed" prefix.

Consider a generic vector $\mathsf{z} \equiv \left(\mathsf{m}_1, \mathsf{g}_1^\top, \mathsf{m}_2, \mathsf{g}_2^\top, \ldots, \mathsf{m}_k, \mathsf{g}_k^\top\right)^\top$ representing a concatenation of performances and gradients associated with solutions in the experimental set, X. As shown in (4), $\mathsf{Z}(x_0)$ can be expressed as a polyhedron in terms of z:

$$\mathsf{Z}(x_0) = \left\{ \mathsf{z} \in \mathbb{R}^{k(q+1)} \colon A(x_0)\mathsf{z} \leq b \right\},$$

for some matrix $A(x_0) \in \mathbb{R}^{k^2 \times k(q+1)}$ and vector $b \in \mathbb{R}^{k^2}$, where $A(x_0)$ depends on x_0 (and X). We relax $\mathsf{Z}(x_0)$ by inflating its right-hand-side vector, defining

$$b'_l = b_l + \max_{\mathsf{z} \in \mathbb{R}^{k(q+1)}} \left\{ a_l^\top \left(\widehat{\zeta} - \mathsf{z} \right) : d_{\mathsf{n}}(\mathsf{z}, \widehat{\zeta}, \widehat{\Psi}) \leq \mathsf{D} \right\} \text{ for all } l = 1, \ldots, k^2,$$

where a_l is the lth row of A, expressed as a column vector, $d_{\mathsf{n}}(\mathsf{z}, \widehat{\zeta}, \widehat{\Psi})$ is a measure of the agreement between the vector z and the estimator $\widehat{\zeta}$, and D is a cutoff value suitably chosen to deliver the confidence guarantees.

The term $d_{\mathsf{n}}(\mathsf{z}, \widehat{\zeta}, \widehat{\Psi})$ is referred to as the *standardized discrepancy*, and larger values of $d_{\mathsf{n}}(\mathsf{z}, \widehat{\zeta}, \widehat{\Psi})$ indicate less agreement between z and $\widehat{\zeta}$. We consider the specific example of

$$d_{\mathsf{n}}(\mathsf{z}, \widehat{\zeta}, \widehat{\Psi}) \equiv \max_{i=1,\ldots,k} n_i \left(\widehat{\zeta}_i - \mathsf{z}_i \right)^\top \widehat{\Psi}_i^{-1} \left(\widehat{\zeta}_i - \mathsf{z}_i \right).$$

This standardized discrepancy builds upon the metric Eckman et al. (2021) proposed for dependent sampling, but in this case, the dependence comes from the performance and gradient estimators at a given solution, as opposed to performance estimators obtained at different solutions in X.

Our method, which we refer to as plausible screening with gradients (PSG), retains those solutions x_0 for which $\widehat{\zeta}$ belongs to the aforementioned relaxation of $\mathsf{Z}(x_0)$, i.e.,

$$\mathscr{S}_{\mathsf{n}}^{\text{PSG}} \equiv \left\{ x_0 \in \mathscr{X} \colon A(x_0)\widehat{\zeta} \leq b' \right\},$$

where $b' \equiv (b'_1, b'_2, \ldots, b'_{k^2})^\top$. It can be shown that

$$\mathscr{S}_{\mathsf{n}}^{\mathrm{PSG}} = \left\{ x_0 \in \mathscr{X} : \max_{i=1,\ldots,k} \left\{ \widehat{\mu}_i + (x_0 - x_i)^\top \widehat{\nabla} \mu_i - \sqrt{\frac{\mathsf{D}}{n_i} (1, (x_0 - x_i)^\top) \, \widehat{\Psi}_i \, (1, (x_0 - x_i)^\top)^\top} \right\} \right.$$
$$\leq \min_{j=1,\ldots,k} \left\{ \widehat{\mu}_j + \sqrt{\frac{\mathsf{D}}{n_j} \widehat{\Psi}_{j,11}} \right\} + \delta$$
$$\left. \max_{i=1,\ldots,k} \left\{ (x_0 - x_i)^\top \widehat{\nabla} \mu_i - \sqrt{\frac{\mathsf{D}}{n_i} (0, (x_0 - x_i)^\top) \, \widehat{\Psi}_i \, (0, (x_0 - x_i)^\top)^\top} \right\} \leq \delta \right\}, \quad (5)$$

where $\widehat{\Psi}_{j,11}$ is the upper-left element of $\widehat{\Psi}_j$, i.e., the sample variance of $Y_1(x_j), Y_2(x_j), \ldots, Y_{n_j}(x_j)$.

From (5), we see that screening a solution with PSG entails checking two inequalities. The first states that the maximum lower confidence bound for the k supporting hyperplanes, evaluated at x_0, must be less than the minimum upper confidence bound for the performances of the solutions in X, plus a tolerance of δ. The second states that the maximum lower confidence bound for the difference in performances between each simulated solution and that of x_0 is no more than δ. Given the relative ease of checking the two inequalities in (5), PSG offers a substantial computational advantage over PS, which requires solving a linear program whose size grows with both k and q.

Theorem 1 Let D satisfy

$$\mathbb{P}\left(\max_{i=1,\ldots,k} \frac{2(n_i - 1)}{n_i - 2} F_{2,\, n_i - 2} \leq \mathsf{D} \right) = 1 - \alpha, \quad (6)$$

where $F_{2,n_1-2}, F_{2,n_2-2}, \ldots, F_{2,n_k-2}$ are k independent F-distributed random variables with 2 numerator degrees of freedom and $n_i - 2$ denominator degrees of freedom. For this choice of D and $\min_{i=1,\ldots,k} n_i \geq 3$, the set $\mathscr{S}_{\mathsf{n}}^{\mathrm{PSG}}$ achieves finite-sample confidence under (1) and asymptotic confidence.

The cutoff specified in (6) is notably independent of the dimension q.

5 PLAUSIBLE SCREENING WITH ONLY GRADIENTS

For convex performance functions, performance and gradient estimators together provide global lower bounds in the form of supporting hyperplanes. For some simulation models, however, estimators of the performance can be much more variable than those for the gradient, resulting in less reliable functional constraints. This observation motivates us to explore an alternative to PSG that ignores performance estimators and therefore does not pay for them with a larger cutoff D.

With this objective in mind, we consider a different projection of $\mathscr{M}(x_0)$ which features only the gradients at solutions in the experimental set:

$$\left\{ \mathsf{g} \in \mathbb{R}^{kq} : \text{ there exists } m \in \mathscr{M}(x_0) \text{ such that } \nabla m(\mathsf{X}) = \mathsf{g} \right\}. \quad (7)$$

We obtain an outer approximation of (7) by adding together pairs of constraints describing $\mathsf{Z}(x_0)$ in (3):

$$\mathsf{m}_i - \mathsf{m}_0 + (x_0 - x_i)^\top \mathsf{g}_i \leq 0 \quad \text{and} \quad -\mathsf{m}_i + \mathsf{m}_0 \leq \delta \quad \Rightarrow \quad (x_0 - x_i)^\top \mathsf{g}_i \leq \delta$$

for all $i = 1, 2, \ldots, k$. This operation cancels the performance terms $\mathsf{m}_0, \mathsf{m}_1, \mathsf{m}_2, \ldots, \mathsf{m}_k$ and yields the set

$$\overline{\mathsf{Z}}(x_0) = \left\{ \mathsf{g} \in \mathbb{R}^{kq} : (x_0 - x_i)^\top \mathsf{g}_i \leq \delta \text{ for all } i = 1, 2, \ldots, k \right\}.$$

We then apply the same technique as before to offset the right-hand-side vector of $\overline{\mathsf{Z}}(x_0)$, in this case to compensate only for our uncertainty regarding $\nabla \mu(\mathsf{X})$. The resulting method, referred to as plausible

screening with only gradients (PSOG), retains the set of solutions

$$\mathscr{S}_{\mathsf{n}}^{\text{PSOG}} = \left\{ x_0 \in \mathscr{X} : \max_{i=1,\ldots,k} \left\{ (x_0 - x_i)^\top \widehat{\nabla}\mu_i - \sqrt{\frac{\mathsf{D}}{n_i} (0, (x_0 - x_i)^\top) \widehat{\Psi}_i (0, (x_0 - x_i)^\top)^\top} \right\} \leq \delta \right\}. \quad (8)$$

The constraint in (8) is exactly the second constraint in (5). Decomposing the maximum in (8) shows that each solution $x_i \in \mathsf{X}$ effectively screens out a second-order cone of solutions, namely, those solutions x_0 for which

$$(x_0 - x_i)^\top \widehat{\nabla}\mu_i - \sqrt{\frac{\mathsf{D}}{n_i} (0, (x_0 - x_i)^\top) \widehat{\Psi}_i (0, (x_0 - x_i)^\top)^\top} > \delta.$$

To account for the tolerance term, δ, the vertex of the cone is offset some distance from x_i, in the direction of the estimated gradient $\widehat{\nabla}\mu_i$.

To determine a suitable choice of cutoff D, we rearrange terms in (8), giving

$$\mathscr{S}_{\mathsf{n}}^{\text{PSOG}} = \left\{ x_0 \in \mathscr{X} : \max_{i=1,\ldots,k} \frac{(x_0 - x_i)^\top \widehat{\nabla}\mu_i - \delta}{\sqrt{\frac{1}{n_i} (0, (x_0 - x_i)^\top) \widehat{\Psi}_i (0, (x_0 - x_i)^\top)^\top}} \leq \sqrt{\mathsf{D}} \right\}. \quad (9)$$

For an acceptable solution x_0, $(x_0 - x_i)^\top \nabla\mu(x_i) \leq \delta$, hence the numerator terms on the left-hand side of (9) are all normally distributed with means less than zero and variances $\left(0, (x_0 - x_i)^\top\right) \Psi_i \left(0, (x_0 - x_i)^\top\right)^\top$. As for the denominators, $\widehat{\Psi}_i$ has a Wishart distribution with dimension $q+1$ and degrees of freedom $n_i - 1$, for $i = 1, 2, \ldots, k$. Therefore the quadratic term

$$\left(0, (x_0 - x_i)^\top\right) \widehat{\Psi}_i \left(0, (x_0 - x_i)^\top\right)^\top \overset{d}{=} \left(0, (x_0 - x_i)^\top\right) \Psi_i \left(0, (x_0 - x_i)^\top\right)^\top \chi^2_{n_i - 1},$$

where χ^2_ν denotes a χ^2-distributed random variables with ν degrees of freedom (Rao 2002). Putting these results together, we have that each term in the maximum in (9) is stochastically dominated by a t-distributed random variable with $n_i - 1$ degrees of freedom.

Theorem 2 Let D satisfy

$$\mathbb{P}\left(\max_{i=1,\ldots,k} T_{n_i - 1} \leq \sqrt{\mathsf{D}} \right) = 1 - \alpha, \quad (10)$$

where $T_{n_1 - 1}, T_{n_2 - 1}, \ldots, T_{n_k - 1}$ are k independent t-distributed random variables with $n_i - 1$ degrees of freedom. For this choice of D and $\min_{i=1,\ldots,k} n_i \geq 2$, the set $\mathscr{S}_{\mathsf{n}}^{\text{PSOG}}$ achieves finite-sample confidence under (1) and asymptotic confidence.

The choice of cutoff in (10) is also independent of the dimension q.

6 CONSISTENCY

Like previous PS methods, PSG and PSOG do not achieve consistency when only a strict subset of solutions, $\mathsf{X} \subset \mathscr{X}$, is simulated. We introduce a relaxed version of consistency featuring a generic subset $\mathsf{S}(\mathsf{X})$ that contains *all* acceptable solutions, where the notation reflects the dependence of this set on X.

Definition 4 ($\mathsf{S}(\mathsf{X})$ Consistency) A subset $\mathscr{S}_{\mathsf{n}}$ achieves $\mathsf{S}(\mathsf{X})$ *consistency* for a subset $\mathsf{S}(\mathsf{X}) \supseteq \mathscr{A}$ if for any $\mu \in \mathscr{M}$, $\mathbb{P}(x_0 \in \mathscr{S}_{\mathsf{n}}) \to 0$ as $\min_{i=1,\ldots,k} n_i \to \infty$ for all $x_0 \notin \mathsf{S}(\mathsf{X})$.

$\mathsf{S}(\mathsf{X})$ consistency states that for any solution $x_0 \notin \mathsf{S}(\mathsf{X})$, the probability we screen it out goes to one as the sampling effort goes to infinity. We proceed to give instances of $\mathsf{S}(\mathsf{X})$ based on the behavior of each screening procedure in the simplifying setting where solutions in X are evaluated without estimation error. That is, the decision-maker observes the true performances $\mu(\mathsf{X})$ and true gradients $\nabla\mu(\mathsf{X})$.

In this case, the set of solutions that are *possibly* acceptable (δ-optimal) is

$$\mathsf{S}^{\mathsf{G}}(\mathsf{X}) \equiv \left\{ x_0 \in \mathscr{X} : \max_{i=1,\ldots,k} \left\{ \mu(x_i) + (x_0 - x_i)^\top \nabla\mu(x_i) \right\} \leq \min_{j=1,\ldots,k} \mu(x_j) + \delta \right\}.$$

From its construction, the set $\mathsf{S}^{\mathsf{G}}(\mathsf{X})$ is a polyhedron containing all acceptable solutions, i.e., $\mathscr{A} \subseteq \mathsf{S}^{\mathsf{G}}(\mathsf{X})$. The set $\mathsf{S}^{\mathsf{G}}(\mathsf{X})$ can be interpreted as those solutions x_0 for which there is a nonnegative gap between the highest of the k supporting hyperplanes, evaluated at x_0, and the performance of the best solution in X plus δ. In other words, the condition compares a lower bound for $\mu(x_0)$ due to convexity (the maximum term) to an upper bound on $\mu(x_0)$ for x_0 to be δ-optimal (the minimum term plus δ).

Theorem 3 For the choice of D in (6), the set $\mathscr{S}_{\mathsf{n}}^{\mathrm{PSG}}$ achieves $\mathsf{S}^{\mathsf{G}}(\mathsf{X})$ consistency.

We next define the set of solutions that are possibly δ-optimal if we observed only the gradients at solutions in X without estimation error, as is the basis for PSOG:

$$\mathsf{S}^{\mathsf{OG}}(\mathsf{X}) \equiv \left\{ x_0 \in \mathscr{X} : \nabla\mu(\mathsf{X}) \in \overline{\mathsf{Z}}(x_0) \right\} = \left\{ x_0 \in \mathscr{X} : \max_{i=1,\ldots,k} \left\{ (x_0 - x_i)^\top \nabla\mu(x_i) \right\} \leq \delta \right\}.$$

It can be seen that $\mathsf{S}^{\mathsf{OG}}(\mathsf{X})$ is also a polyhedron containing all acceptable solutions.

Theorem 4 For the choice of D in (10), the set $\mathscr{S}_{\mathsf{n}}^{\mathrm{PSOG}}$ achieves $\mathsf{S}^{\mathsf{OG}}(\mathsf{X})$ consistency.

We also compare with the screening methods of Eckman et al. (2021), which do not make use of gradient estimators. Their methods were consistent with respect to a different limiting set,

$$\mathsf{S}^{\mathsf{O}}(\mathsf{X}) \equiv \left\{ x_0 \in \mathscr{X} : \text{there exists } \xi_1, \ldots, \xi_k \in \mathbb{R}^d \text{ such that } \max_{i=1,\ldots,k} \left\{ \mu(x_i) + (x_0 - x_i)^\top \xi_i \right\} \leq \min_{j=1,\ldots,k} \mu(x_j) + \delta \right\}.$$

The set $\mathsf{S}^{\mathsf{O}}(\mathsf{X})$ closely resembles $\mathsf{S}^{\mathsf{G}}(\mathsf{X})$, expect that the true gradients $\nabla\mu(x_1), \nabla\mu(x_2), \ldots, \nabla\mu(x_k)$ are replaced by free variables $\xi_1, \xi_2, \ldots, \xi_k$.

Theorem 5 establishes how the sets $\mathsf{S}^{\mathsf{G}}(\mathsf{X})$, $\mathsf{S}^{\mathsf{OG}}(\mathsf{X})$, and $\mathsf{S}^{\mathsf{O}}(\mathsf{X})$ are related

Theorem 5 For any solution space $\mathscr{X}$, any experimental set X, and any performance function $\mu \in \mathscr{M}$, $\mathsf{S}^{\mathsf{G}}(\mathsf{X}) \subseteq \mathsf{S}^{\mathsf{OG}}(\mathsf{X})$ and $\mathsf{S}^{\mathsf{G}}(\mathsf{X}) \subseteq \mathsf{S}^{\mathsf{O}}(\mathsf{X})$.

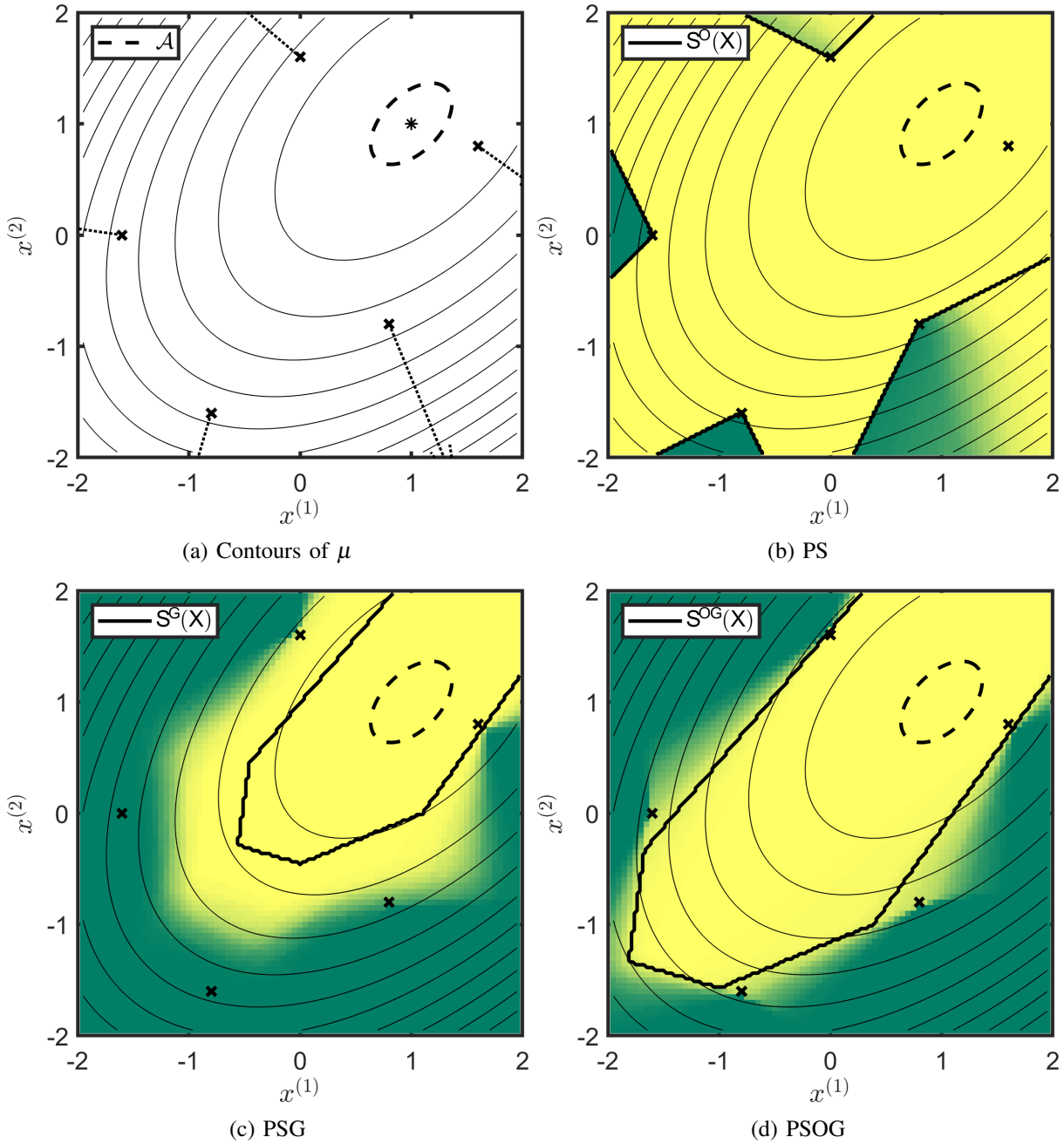
Theorem 5 implies that when taking a very large number of replications at each solution in the experimental set, using both the performance and gradient estimators screens out the most unacceptable solutions. We have no guarantee that, in the limit, using only gradients will screen out more solutions than using only performances, but in many cases we imagine this is true. However, as will be shown in the next section, we find that for small sample sizes, the situation can be more complicated. In particular, we observe that some solutions can be screened out when using only gradients (PSOG) but not when using both performance and gradients together (PSG). The reason is that PSG pays the price for both performance and gradient estimation with a larger cutoff D.

7 NUMERICAL EXPERIMENTS

We test the proposed methods on a synthetic problem with performance and gradient functions

$$\mu(x) = (x - \mathbf{1}_2)^\top \begin{bmatrix} 1 & -1/2 \\ -1/2 & 1 \end{bmatrix} (x - \mathbf{1}_2) \quad \text{and} \quad \nabla\mu(x) = \begin{bmatrix} 2 & -1 \\ -1 & 2 \end{bmatrix} (x - \mathbf{1}_2),$$

for all $x \in \mathscr{X} = [-2,2] \times [-2,2] \subset \mathbb{R}^2$, where $\mathbf{1}_2 = (1,1)^\top$. The associated Hessian matrix is positive definite and thus μ is convex with a unique global minimum of 0 at $x^* = \mathbf{1}_2$. The set of δ-optimal solutions, $\mathscr{A}$, is an ellipsoid corresponding to the δ sub-level set of μ; we set $\delta = 0.1$. Figure 1a shows the contours of the performance function, the optimal solution x^*, and the set of δ-optimal solutions. The experimental set is taken to be



(a) Contours of μ (b) PS

(c) PSG (d) PSOG

Figure 1: Results for synthetic example where $x \equiv (x^{(1)}, x^{(2)})$. (a) Contours of the performance function μ with the experimental set X, gradients in $\nabla\mu(\mathsf{X})$, acceptable solutions $\mathscr{A}$, and optimal solution x^*. Heat maps of $\mathbb{P}(x_0 \in \mathscr{S}_{\mathsf{n}})$ for (b) PS, (c) PSG, and (d) PSOG showing limiting sets $\mathsf{S}^{\mathsf{O}}(\mathsf{X})$, $\mathsf{S}^{\mathsf{G}}(\mathsf{X})$, and $\mathsf{S}^{\mathsf{OG}}(\mathsf{X})$, respectively. Lighter shading indicates values near 1, while darker shading indicates values near 0.

a Latin hypercube design of $k = 5$ solutions: $\mathsf{X} = \{(-1.6, 0), (-0.8, -1.6), (0, 1.6), (0.8, -0.8), (1.6, 0.8)\}$. Figure 1a also shows the gradients in $\nabla\mu(\mathsf{X})$.

A single replication at a solution $x_i \in \mathsf{X}$ produces an estimator

$$Z(x_i) \equiv \begin{bmatrix} Y(x_i) \\ G(x_i) \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \mu(x_i) \\ \nabla\mu(x_i) \end{bmatrix}, \begin{bmatrix} 1 & \rho & \rho \\ \rho & 1 & \rho \\ \rho & \rho & 1 \end{bmatrix} \sigma^2(x_i) \right),$$

where $\sigma^2(x_i) = 0.5 + \|x_i - \mathbf{1}_2\|_2$ is a variance term—common to the performance estimator and components of the gradient estimator—that increases away from the global minimizer, and $\rho = 0.5$ is a common correlation coefficient. We obtain $n_i = 20$ replications from each solution $x_i \in \mathsf{X}$ and the outputs at a given solution are i.i.d. and independent across solutions.

We compare three methods: (1) (relaxed) plausible screening (PS) from Eckman et al. (2021), (2) PSG, and (3) PSOG. PS screens solutions by solving linear programs, while PSG and PSOG require no optimization. We run 100 macroreplications of each method with $1 - \alpha = 0.95$, using common random numbers across methods to ensure that all three methods observe the same performance and gradient estimates on any given macroreplication. We screen a dense grid of solutions, so as to approximate the sets of solutions each procedure would retain if applied to all solutions in $\mathscr{X}$, i.e., $\mathscr{S}_{\mathsf{n}}^{\mathrm{PS}}$, $\mathscr{S}_{\mathsf{n}}^{\mathrm{PSG}}$, and $\mathscr{S}_{\mathsf{n}}^{\mathrm{PSOG}}$. The probabilities each solution is included in $\mathscr{S}_{\mathsf{n}}^{\mathrm{PS}}$, $\mathscr{S}_{\mathsf{n}}^{\mathrm{PSG}}$, and $\mathscr{S}_{\mathsf{n}}^{\mathrm{PSOG}}$ are shown in the heat maps of Figures 1b, 1c, and 1d, respectively. Figure 1b shows that PS struggles to screen out many unacceptable solutions, even those in the lower-left corner that are far from optimal. In Figures 1c and 1d, we see that many more solutions are screened out by PSG and PSOG. PSG tends to screen out more solutions than PSOG, but not in a nested way. In particular, we see that PSOG screens out more solutions around $(1.6, 0.8)$, but struggles to screen out solutions around $(-1.6, 0)$ and $(-0.8, -1.6)$ because it does not use the performance estimators that indicate the inferiority of those two solutions.

Figures 1b, 1c, and 1d also depict the limiting sets characterizing the consistency of the three procedures: $\mathsf{S}^{\mathsf{O}}(\mathsf{X})$ for PS, $\mathsf{S}^{\mathsf{G}}(\mathsf{X})$ for PSG, and $\mathsf{S}^{\mathsf{OG}}(\mathsf{X})$ for PSOG. It can be seen that $\mathsf{S}^{\mathsf{O}}(\mathsf{X})$, $\mathsf{S}^{\mathsf{G}}(\mathsf{X})$, and $\mathsf{S}^{\mathsf{OG}}(\mathsf{X})$ satisfy the nested relationships claimed in Theorems 5. Furthermore, $\mathsf{S}^{\mathsf{G}}(\mathsf{X})$ and $\mathsf{S}^{\mathsf{OG}}(\mathsf{X})$ are polyhedra while $\mathsf{S}^{\mathsf{O}}(\mathsf{X})$ is highly non-convex. The subset $\mathsf{S}^{\mathsf{O}}(\mathsf{X})$ is also much larger than the other two subsets, which illustrates the additional screening power offered by leveraging gradient estimators, at least in the limit as sample sizes increase. The gap between $\mathscr{S}_{\mathsf{n}}^{\mathrm{PSG}}$ and $\mathsf{S}^{\mathsf{G}}(\mathsf{X})$ also suggests that there is an opportunity for a tighter cutoff D for PSG.

8 CONCLUSION

We borrowed concepts from plausible screening to identify near-optimal solutions to convex simulation-optimization problems when unbiased stochastic gradient estimators are available. Aside from the time required to derive and implement them, gradient estimators are leveraged to enhance screening in a computationally cheap way. Numerical experiments demonstrate that the methods screen out more solutions than previous plausible screening methods, which use only performance estimators. We envision IPA gradient estimators being especially powerful for screening due to the fact that they tend to be less variable compared to LR/SF gradient estimators.

We presented a standardized discrepancy that handles the possible dependence between the performance and gradient estimators at a given solution. This new metric could be adapted to simulation applications with one or more dependent responses or stochastic constraints. Future research directions include extending definitions of acceptability and incorporating functional properties that include other forms of first-order information, e.g., identifying first-order stationary solutions for a performance function with Lipschitz gradients. We also intend to study how screening power is affected by the relative variability of performance and gradient estimators, as well as by the variability in the true values of these quantities over the experimental set.

ACKNOWLEDGMENTS

This work was supported by the National Science Foundation under grant nos. DMS–1854562 and DMS–1953111. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation (NSF).

REFERENCES

Anderson, T. W. 1984. *An Introduction to Multivariate Statistical Analysis*. 2nd ed. New York: John Wiley & Sons.

Chen, X., B. E. Ankenman, and B. L. Nelson. 2013. "Enhancing Stochastic Kriging Metamodels with Gradient Estimators". *Operations Research* 61(2):512–528.

Eckman, D. J., M. Plumlee, and B. L. Nelson. 2020. "Revisiting Subset Selection". In *Proceedings of the 2020 Winter Simulation Conference*, edited by K.-H. Bae, B. Feng, S. Kim, S. Lazarova-Molnar, Z. Zheng, T. Roeder, and R. Thiesing, 2972–2983. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

Eckman, D. J., M. Plumlee, and B. L. Nelson. 2021. "Plausible Screening Using Functional Properties for Simulations with Large Solution Spaces". Under Review.

Forrester, A. I., and A. J. Keane. 2009. "Recent Advances in Surrogate-Based Optimization". *Progress in Aerospace Sciences* 45(1-3):50–79.

Fu, M. 2015. "Stochastic Gradient Estimation". In *Handbook of Simulation Optimization*, edited by M. Fu, Chapter 5, 105–147. New York, New York: Springer.

Glasserman, P. 1991. *Gradient Estimation Via Perturbation Analysis*. Dordrecht, The Netherlands: Kluwer.

Glynn, P. W. 1987. "Likelihood Ratio Gradient Estimation: An Overview". In *Proceedings of the 1987 Winter Simulation Conference*, edited by A. Thesen, H. Grant, and W. D. Kelton, 366–375. New York: Association for Computing Machinery.

Jian, N., and S. G. Henderson. 2020. "Estimating the Probability That a Function Observed With Noise is Convex". *INFORMS Journal on Computing* 32(2):376–389.

Morris, M. D., T. J. Mitchell, and D. Ylvisaker. 1993. "Bayesian Design and Analysis of Computer Experiments: Use of Derivatives in Surface Prediction". *Technometrics* 35(3):243–255.

Plambeck, E. L., B.-R. Fu, S. M. Robinson, and R. Suri. 1996. "Sample-Path Optimization of Convex Stochastic Performance Functions". *Mathematical Programming* 75(2):137–176.

Plumlee, M., and B. L. Nelson. 2018. "Plausible Optima". In *Proceedings of the 2018 Winter Simulation Conference*, edited by M. Rabe, A. A. Juan, N. Mustafee, A. Skoogh, S. Jain, and B. Johansson, 1981–1992. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

Qu, H., and M. C. Fu. 2014. "Gradient Extrapolated Stochastic Kriging". *ACM Transactions on Modeling and Computer Simulation (TOMACS)* 24(4):Article 23, 1–25.

Rao, C. R. 2002. *Linear Statistical Inference and Its Applications*. 2nd ed. New York: John Wiley & Sons, Inc.

Rubinstein, R. Y. 1989. "Sensitivity Analysis and Performance Extrapolation for Computer Simulation Models". *Operations Research* 37(1):72–81.

Silvapulle, M. J., and P. K. Sen. 2005. *Constrained Statistical Inference: Inequality, Order, and Shape Restrictions*. Hoboken, New Jersey: John Wiley & Sons, Inc.

Wu, J., M. Poloczek, A. G. Wilson, and P. I. Frazier. 2017. "Bayesian Optimization with Gradients". In *Advances in Neural Information Processing Systems*, edited by I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Volume 30, 5267–5278: Curran Associates, Inc.

AUTHOR BIOGRAPHIES

DAVID J. ECKMAN is an Assistant Professor in the Wm Michael Barnes '64 Department of Industrial and Systems Engineering at Texas A&M University. His research interests deal with optimization and output analysis for stochastic simulation models. His e-mail address is eckman@tamu.edu.

MATTHEW PLUMLEE is an Assistant Professor in the Department of Industrial Engineering and Management Sciences at Northwestern University. He primarily researches uncertainty quantification methods for computational models of systems. His e-mail address is mplumlee@northwestern.edu.

BARRY L. NELSON is the Walter P. Murphy Professor in the Department of Industrial Engineering and Management Sciences at Northwestern University. He is a Fellow of INFORMS and IISE. His research centers on the design and analysis of computer simulation experiments on models of stochastic systems, and he is the author of *Foundations and Methods of Stochastic Simulation: A First Course*, from Springer. His e-mail address is nelsonb@northwestern.edu.

9 APPENDIX

9.1 Proof that $\mathsf{Z}(x_0)$ is a relaxation of (2).

For a given μ, let $\mathscr{A}(\mu) = \{x \in \mathscr{X} : \mu(x) \leq \inf_{x' \in \mathscr{X}} \mu(x') + \delta\}$ and suppose that μ is known to be convex over a continuous domain $\mathscr{X} \subseteq \mathbb{R}^q$. A differentiable function m is convex over $\mathscr{X}$ if for all $x, x' \in \mathscr{X}$, $m(x) - m(x') + (x' - x)^\top \nabla m(x) \leq 0$. In terms of our notation,

$$\mathscr{M} = \left\{ m \colon m(x) - m(x') + (x' - x)^\top \nabla m(x) \leq 0 \text{ for all } x, x' \in \mathscr{X} \right\}.$$

For a given solution $x_0 \in \mathscr{X}$, we add the constraint that $m(x_0) \leq m(x) + \delta$ for all $x \in \mathscr{X}$ to obtain the set of convex functions for which solution x_0 is δ-optimal,

$$\mathscr{M}(x_0) = \left\{ m \colon m(x_0) \leq m(x) + \delta \text{ and } m(x) - m(x') + (x' - x)^\top \nabla m(x) \leq 0 \text{ for all } x, x' \in \mathscr{X} \right\}.$$

We show that the projection of $\mathscr{M}(x_0)$ onto $\mathbb{R}^{k(q+1)}$ is contained in

$$\begin{aligned} \mathsf{Z}(x_0) = \Big\{ & \mathsf{m} \in \mathbb{R}^k, \mathsf{g} \in \mathbb{R}^{kq} \colon \text{there exists } \mathsf{m}_0 \in \mathbb{R} \text{ such that} \\ & \mathsf{m}_i - \mathsf{m}_0 + (x_0 - x_i)^\top \mathsf{g}_i \leq 0 \text{ for all } i = 1, 2, \ldots, k \\ & -\mathsf{m}_i + \mathsf{m}_0 \leq \delta \text{ for all } i = 1, 2, \ldots, k \Big\}. \end{aligned}$$

Fix an arbitrary function $m \in \mathscr{M}(x_0)$ and define $\mathsf{m}_i = m(x_i)$ for $i = 0, 1, \ldots, k$ and $\mathsf{g}_i = \nabla m(x_i)$ for $i = 1, \ldots, k$. Since $m \in \mathscr{M}(x_0)$, $\mathsf{m}_i - \mathsf{m}_0 + (x_0 - x_i)^\top \mathsf{g}_i = m(x_i) - m(x_0) + (x_0 - x_i)^\top \nabla m(x_i) \leq 0$ for all $i = 1, 2, \ldots, k$ and $-\mathsf{m}_i + \mathsf{m}_0 = -m(x_i) + m(x_0) \leq \delta$ for all $i = 1, 2, \ldots, k$. Hence, for this choice of m_0, we have that $(\mathsf{m}, \mathsf{g}) \in \mathsf{Z}(x_0)$. This proves that $\mathsf{Z}(x_0)$ is a relaxation of (2), the projection of $\mathscr{M}(x_0)$ onto $\mathbb{R}^{k(q+1)}$.

9.2 Projecting out m_0 from $\mathsf{Z}(x_0)$.

We apply Fourier-Motzkin elimination to our representation of $\mathsf{Z}(x_0)$ to project out m_0.

$$\begin{aligned} \mathsf{Z}(x_0) = & \Big\{ \mathsf{m} \in \mathbb{R}^k, \mathsf{g} \in \mathbb{R}^{kq} \colon \text{there exists } \mathsf{m}_0 \in \mathbb{R} \text{ such that} \\ & \mathsf{m}_i - \mathsf{m}_0 + (x_0 - x_i)^\top \mathsf{g}_i \leq 0 \text{ for all } i = 1, 2, \ldots, k \\ & -\mathsf{m}_i + \mathsf{m}_0 \leq \delta \text{ for all } i = 1, 2, \ldots, k \Big\} \\ = & \Big\{ \mathsf{m} \in \mathbb{R}^k, \mathsf{g} \in \mathbb{R}^{kq} \colon \text{there exists } \mathsf{m}_0 \in \mathbb{R} \text{ such that} \\ & \mathsf{m}_0 \geq \mathsf{m}_i + (x_0 - x_i)^\top \mathsf{g}_i \text{ for all } i = 1, 2, \ldots, k \\ & \mathsf{m}_0 \leq \mathsf{m}_i + \delta \text{ for all } i = 1, 2, \ldots, k \Big\} \\ = & \left\{ \mathsf{m} \in \mathbb{R}^k, \mathsf{g} \in \mathbb{R}^{kq} \colon \ \mathsf{m}_i + (x_0 - x_i)^\top \mathsf{g}_i \leq \mathsf{m}_j + \delta \text{ for all } i, j = 1, 2, \ldots, k \right\} \\ = & \left\{ \mathsf{m} \in \mathbb{R}^k, \mathsf{g} \in \mathbb{R}^{kq} \colon \ \mathsf{m}_i - \mathsf{m}_j + (x_0 - x_i)^\top \mathsf{g}_i \leq \delta \text{ for all } i, j = 1, 2, \ldots, k \right\}. \end{aligned}$$

9.3 Useful Lemmas for PSG

Lemma 1 For any $\mathsf{v} \in \mathbb{R}^{k(q+1)}$,

$$\max_{\mathsf{z} \in \mathbb{R}^{k(q+1)}} \left\{ \mathsf{v}^\top \left(\widehat{\zeta} - \mathsf{z} \right) : \sum_{i=1}^{k} n_i \left(\widehat{\zeta}_i - \mathsf{z}_i \right)^\top \widehat{\Psi}_i^{-1} \left(\widehat{\zeta}_i - \mathsf{z}_i \right) \leq \mathsf{D} \right\} = \sqrt{\sum_{i=1}^{k} \frac{\mathsf{D}}{n_i} \mathsf{v}_i^\top \widehat{\Psi}_i \mathsf{v}_i},$$

where $\mathsf{v} \equiv \left(\mathsf{v}_1^\top, \mathsf{v}_2^\top, \ldots, \mathsf{v}_k^\top \right)^\top$ with $\mathsf{v}_i \in \mathbb{R}^{q+1}$ for $i = 1, 2, \ldots, k$.

Proof of Lemma 1. Let $\mathbf{1}_{q+1,q+1}$ be a $(q+1)\times(q+1)$ matrix of ones and let $A\otimes B$ denote the Kronecker product of matrices A and B. Then

$$\begin{aligned}
&\max_{\mathsf{z}\in\mathbb{R}^{k(q+1)}}\left\{\mathsf{v}^\top\left(\widehat{\zeta}-\mathsf{z}\right):\ \sum_{i=1}^{k}n_i\left(\widehat{\zeta}_i-\mathsf{z}_i\right)^\top\widehat{\Psi}_i^{-1}\left(\widehat{\zeta}_i-\mathsf{z}_i\right)\le\mathsf{D}\right\}\\
&\quad=\max_{\mathsf{z}\in\mathbb{R}^{k(q+1)}}\left\{\mathsf{v}^\top\left(\widehat{\zeta}-\mathsf{z}\right):\ \left(\widehat{\zeta}-\mathsf{z}\right)^\top\left(\left(\operatorname{diag}(\mathsf{n})^{-1/2}\otimes\mathbf{1}_{q+1,q+1}\right)^\top\widehat{\Psi}\left(\operatorname{diag}(\mathsf{n})^{-1/2}\otimes\mathbf{1}_{q+1,q+1}\right)\right)^{-1}\left(\widehat{\zeta}-\mathsf{z}\right)\le\mathsf{D}\right\}\\
&\quad=\max_{\mathsf{z}\in\mathbb{R}^{k(q+1)}}\left\{\mathsf{v}^\top\left(\widehat{\zeta}-\mathsf{z}\right):\ \left(\widehat{\zeta}-\mathsf{z}\right)^\top\left(\left(\sqrt{\mathsf{D}}\operatorname{diag}(\mathsf{n})^{-1/2}\otimes\mathbf{1}_{q+1,q+1}\right)^\top\widehat{\Psi}\left(\sqrt{\mathsf{D}}\operatorname{diag}(\mathsf{n})^{-1/2}\otimes\mathbf{1}_{q+1,q+1}\right)\right)^{-1}\left(\widehat{\zeta}-\mathsf{z}\right)\le 1\right\}\\
&\quad=\max_{\mathsf{z}\in\mathbb{R}^{k(q+1)}}\left\{\mathsf{v}^\top\left(\widehat{\zeta}-\mathsf{z}\right):\ \left\|\widehat{\zeta}-\mathsf{z}\right\|_{\left(\left(\sqrt{\mathsf{D}}\operatorname{diag}(\mathsf{n})^{-1/2}\otimes\mathbf{1}_{q+1,q+1}\right)^\top\widehat{\Psi}\left(\sqrt{\mathsf{D}}\operatorname{diag}(\mathsf{n})^{-1/2}\otimes\mathbf{1}_{q+1,q+1}\right)\right)^{-1}}\le 1\right\}\\
&\quad=\|\mathsf{v}\|_{\left(\sqrt{\mathsf{D}}\operatorname{diag}(\mathsf{n})^{-1/2}\otimes\mathbf{1}_{q+1,q+1}\right)^\top\widehat{\Psi}\left(\sqrt{\mathsf{D}}\operatorname{diag}(\mathsf{n})^{-1/2}\otimes\mathbf{1}_{q+1,q+1}\right)}\\
&\quad=\sqrt{\mathsf{v}^\top\left(\sqrt{\mathsf{D}}\operatorname{diag}(\mathsf{n})^{-1/2}\otimes\mathbf{1}_{q+1,q+1}\right)^\top\widehat{\Psi}\left(\sqrt{\mathsf{D}}\operatorname{diag}(\mathsf{n})^{-1/2}\otimes\mathbf{1}_{q+1,q+1}\right)\mathsf{v}}\\
&\quad=\sqrt{\sum_{i=1}^{k}\frac{\mathsf{D}}{n_i}\mathsf{v}_i^\top\widehat{\Psi}_i\mathsf{v}_i}.
\end{aligned}$$

□

Lemma 2 For any $\mathsf{v}\in\mathbb{R}^{k(q+1)}$,

$$\max_{\mathsf{z}\in\mathbb{R}^{k(q+1)}}\left\{\mathsf{v}^\top\left(\widehat{\zeta}-\mathsf{z}\right):d_{\mathsf{n}}(\mathsf{z},\widehat{\zeta},\widehat{\Psi})\le\mathsf{D}\right\}=\max_{\mathsf{z}\in\mathbb{R}^{k(q+1)}}\left\{\mathsf{v}^\top\left(\widehat{\zeta}-\mathsf{z}\right):\ \max_{i=1,\ldots,k}n_i\left(\widehat{\zeta}_i-\mathsf{z}_i\right)^\top\widehat{\Psi}_i^{-1}\left(\widehat{\zeta}_i-\mathsf{z}_i\right)\le\mathsf{D}\right\}=\sum_{i=1}^{k}\sqrt{\frac{\mathsf{D}}{n_i}\mathsf{v}_i^\top\widehat{\Psi}_i\mathsf{v}_i},$$

where $\mathsf{v}\equiv\left(\mathsf{v}_1^\top,\mathsf{v}_2^\top,\ldots,\mathsf{v}_k^\top\right)^\top$ with $\mathsf{v}_i\in\mathbb{R}^{q+1}$ for $i=1,2,\ldots,k$.

Proof of Lemma 2.

$$\begin{aligned}
\max_{\mathsf{z}\in\mathbb{R}^{k(q+1)}}\left\{\mathsf{v}^\top\left(\widehat{\zeta}-\mathsf{z}\right):d_{\mathsf{n}}(\mathsf{z},\widehat{\zeta},\widehat{\Psi})\le\mathsf{D}\right\}&=\max_{\mathsf{z}\in\mathbb{R}^{k(q+1)}}\left\{\mathsf{v}^\top\left(\widehat{\zeta}-\mathsf{z}\right):\ \max_{i=1,\ldots,k}n_i\left(\widehat{\zeta}_i-\mathsf{z}_i\right)^\top\widehat{\Psi}_i^{-1}\left(\widehat{\zeta}_i-\mathsf{z}_i\right)\le\mathsf{D}\right\}\\
&=\max_{\mathsf{z}\in\mathbb{R}^{k(q+1)}}\left\{\mathsf{v}^\top\left(\widehat{\zeta}-\mathsf{z}\right):n_i\left(\widehat{\zeta}_i-\mathsf{z}_i\right)^\top\widehat{\Psi}_i^{-1}\left(\widehat{\zeta}_i-\mathsf{z}_i\right)\le\mathsf{D}\ \text{ for all }\ i=1,2,\ldots,k\right\}\\
&=\max_{\mathsf{z}\in\mathbb{R}^{k(q+1)}}\left\{\sum_{i=1}^{k}\mathsf{v}_i^\top\left(\widehat{\zeta}_i-\mathsf{z}_i\right):n_i\left(\widehat{\zeta}_i-\mathsf{z}_i\right)^\top\widehat{\Psi}_i^{-1}\left(\widehat{\zeta}_i-\mathsf{z}_i\right)\le\mathsf{D}\ \text{ for all }\ i=1,2,\ldots,k\right\}\\
&=\sum_{i=1}^{k}\max_{\mathsf{z}_i\in\mathbb{R}^{q+1}}\left\{\mathsf{v}_i^\top\left(\widehat{\zeta}_i-\mathsf{z}_i\right):n_i\left(\widehat{\zeta}_i-\mathsf{z}_i\right)^\top\widehat{\Psi}_i^{-1}\left(\widehat{\zeta}_i-\mathsf{z}_i\right)\le\mathsf{D}\right\}\\
&=\sum_{i=1}^{k}\sqrt{\frac{\mathsf{D}}{n_i}\mathsf{v}_i^\top\widehat{\Psi}_i\mathsf{v}_i}.
\end{aligned}$$

The last equality comes from applying Lemma 1. □

9.4 Derivation of $\mathscr{S}_{\mathsf{n}}^{\mathrm{PSG}}$

We defined

$$\mathsf{Z}(x_0)=\left\{\mathsf{m}\in\mathbb{R}^k,\mathsf{g}\in\mathbb{R}^{kq}:\ \mathsf{m}_i-\mathsf{m}_j+(x_0-x_i)^\top\mathsf{g}_i\le\delta\ \text{ for all }\ i,j=1,2,\ldots,k\right\},$$

which can be viewed as a polyhedron

$$\mathsf{Z}(x_0)=\left\{\mathsf{z}\in\mathbb{R}^{k(q+1)}:A(x_0)\mathsf{z}\le b\right\},$$

where $A(x_0)\in\mathbb{R}^{k^2\times k(q+1)}$ and $b=\delta\mathbf{1}_{k^2}$.

The subset of solutions returned by PSG is defined as

$$\mathscr{S}_{\mathsf{n}}^{\mathrm{PSG}}\equiv\left\{x_0\in\mathscr{X}:A(x_0)\widehat{\zeta}\le b'\right\},\tag{11}$$

where $b' \equiv (b'_1, b'_2, \ldots, b'_{k^2})^\top$ and

$$b'_l = b_l + \max_{\mathsf{z} \in \mathbb{R}^{k(q+1)}} \left\{ a_l^\top \left(\widehat{\zeta} - \mathsf{z} \right) : d_{\mathsf{n}}(\mathsf{z}, \widehat{\zeta}, \widehat{\Psi}) \leq \mathsf{D} \right\} \text{ for all } l = 1, \ldots, k^2.$$

We showed in Lemma 2 that

$$\max_{\mathsf{z} \in \mathbb{R}^{k(q+1)}} \left\{ a_l^\top \left(\widehat{\zeta} - \mathsf{z} \right) : d_{\mathsf{n}}(\mathsf{z}, \widehat{\zeta}, \widehat{\Psi}) \leq \mathsf{D} \right\} = \sum_{i=1}^{k} \sqrt{\frac{\mathsf{D}}{n_i} a_{li}^\top \widehat{\Psi}_i a_{li}}.$$

Substituting this result into (11) gives

$$\begin{aligned}
\mathscr{S}_{\mathsf{n}}^{\mathrm{PSG}} = \Bigg\{ & x_0 \in \mathscr{X} : \widehat{\mu}_i - \widehat{\mu}_j + (x_0 - x_i)^\top \widehat{\nabla} \mu_i \leq \delta + \sqrt{\frac{\mathsf{D}}{n_i} \left(1, (x_0 - x_i)^\top\right) \widehat{\Psi}_i \left(1, (x_0 - x_i)^\top\right)^\top} + \sqrt{\frac{\mathsf{D}}{n_j} \widehat{\Psi}_{j,11}} \quad \text{for all } i \neq j \\
& (x_0 - x_i)^\top \widehat{\nabla} \mu_i \leq \delta + \sqrt{\frac{\mathsf{D}}{n_i} \left(0, (x_0 - x_i)^\top\right) \widehat{\Psi}_i \left(0, (x_0 - x_i)^\top\right)^\top} \quad \text{for all } i = 1, 2, \ldots, k \Bigg\} \\
= \Bigg\{ & x_0 \in \mathscr{X} : \widehat{\mu}_i - \widehat{\mu}_j + (x_0 - x_i)^\top \widehat{\nabla} \mu_i - \sqrt{\frac{\mathsf{D}}{n_i} \left(1, (x_0 - x_i)^\top\right) \widehat{\Psi}_i \left(1, (x_0 - x_i)^\top\right)^\top} - \sqrt{\frac{\mathsf{D}}{n_j} \widehat{\Psi}_{j,11}} \leq \delta \quad \text{for all } i \neq j \\
& (x_0 - x_i)^\top \widehat{\nabla} \mu_i - \sqrt{\frac{\mathsf{D}}{n_i} \left(0, (x_0 - x_i)^\top\right) \widehat{\Psi}_i \left(0, (x_0 - x_i)^\top\right)^\top} \leq \delta \quad \text{for all } i = 1, 2, \ldots, k \Bigg\}. \qquad (12) \\
= \Bigg\{ & x_0 \in \mathscr{X} : \max_{i=1,\ldots,k} \left\{ \widehat{\mu}_i + (x_0 - x_i)^\top \widehat{\nabla} \mu_i - \sqrt{\frac{\mathsf{D}}{n_i} \left(1, (x_0 - x_i)^\top\right) \widehat{\Psi}_i \left(1, (x_0 - x_i)^\top\right)^\top} \right\} \leq \min_{j=1,\ldots,k} \left\{ \widehat{\mu}_j + \sqrt{\frac{\mathsf{D}}{n_j} \widehat{\Psi}_{j,11}} \right\} + \delta \\
& \max_{i=1,\ldots,k} \left\{ (x_0 - x_i)^\top \widehat{\nabla} \mu_i - \sqrt{\frac{\mathsf{D}}{n_i} \left(0, (x_0 - x_i)^\top\right) \widehat{\Psi}_i \left(0, (x_0 - x_i)^\top\right)^\top} \right\} \leq \delta \Bigg\}.
\end{aligned}$$

9.5 Derivation of $\mathscr{S}_{\mathsf{n}}^{\mathrm{PSOG}}$

We defined

$$\overline{\mathsf{Z}}(x_0) = \left\{ \mathsf{g} \in \mathbb{R}^{kq} : (x_0 - x_i)^\top \mathsf{g}_i \leq \delta \text{ for all } i = 1, 2, \ldots, k \right\},$$

which can be viewed as a polyhedron

$$\overline{\mathsf{Z}}(x_0) = \left\{ \mathsf{z} \in \mathbb{R}^{k(q+1)} : A(x_0) \mathsf{z} \leq b \right\},$$

where $A(x_0) \in \mathbb{R}^{k \times k(q+1)}$ and $b = \delta \mathbf{1}_k$. Note that these $A(x_0)$ and b differ from those in Appendix 9.4 and some columns of $A(x_0)$ are zeroed out.

Applying the same construction method as in PSG, the subset of solutions returned by PSOG is defined as

$$\mathscr{S}_{\mathsf{n}}^{\mathrm{PSOG}} \equiv \left\{ x_0 \in \mathscr{X} : A(x_0) \widehat{\zeta} \leq b' \right\},$$

where $b' \equiv (b'_1, b'_2, \ldots, b'_{k^2})^\top$ and

$$b'_l = b_l + \max_{\mathsf{z} \in \mathbb{R}^{k(q+1)}} \left\{ a_l^\top \left(\widehat{\zeta} - \mathsf{z} \right) : d_{\mathsf{n}}(\mathsf{z}, \widehat{\zeta}, \widehat{\Psi}) \leq \mathsf{D} \right\} \text{ for all } l = 1, \ldots, k.$$

The subset of retained solutions is given by

$$\begin{aligned}
\mathscr{S}_{\mathsf{n}}^{\mathrm{PSOG}} &= \left\{ x_0 \in \mathscr{X} : (x_0 - x_i)^\top \widehat{\nabla} \mu_i \leq \delta + \sqrt{\frac{\mathsf{D}}{n_i} \left(0, (x_0 - x_i)^\top\right) \widehat{\Psi}_i \left(0, (x_0 - x_i)^\top\right)^\top} \quad \text{for all } i = 1, 2, \ldots, k \right\} \\
&= \left\{ x_0 \in \mathscr{X} : \max_{i=1,\ldots,k} \left\{ (x_0 - x_i)^\top \widehat{\nabla} \mu_i - \sqrt{\frac{\mathsf{D}}{n_i} \left(0, (x_0 - x_i)^\top\right) \widehat{\Psi}_i \left(0, (x_0 - x_i)^\top\right)^\top} \right\} \leq \delta \right\}.
\end{aligned}$$

9.6 Confidence Results

Proof of Theorem 1. Fix arbitrary $1 - \alpha \in (1/2, 1)$ and $\mu \in \mathscr{M}$. Fix an arbitrary solution $x_0 \in \mathscr{A}$. Since $x_0 \in \mathscr{A}$, we have that $\mu \in \mathscr{M}(x_0)$ and therefore $\zeta(\mathsf{X}) \in \mathsf{Z}(x_0)$ where

$$\mathsf{Z}(x_0) = \left\{ \mathsf{m} \in \mathbb{R}^k, \mathsf{g} \in \mathbb{R}^{kq} : \mathsf{m}_i - \mathsf{m}_j + (x_0 - x_i)^\top \mathsf{g}_i \leq \delta \text{ for all } i, j = 1, 2, \ldots, k \right\}. \qquad (13)$$

We can write $\mathsf{Z}(x_0) = \left\{\mathsf{z} \in \mathbb{R}^{k(q+1)} : A(x_0)\mathsf{z} \leq b\right\}$ for $A(x_0) \in \mathbb{R}^{k^2 \times k(q+1)}$ and $b = \delta \mathbf{1}_{k^2}$. Thus the statement $\zeta(\mathsf{X}) \in \mathsf{Z}(x_0)$ implies that $a_l^\top \zeta(\mathsf{X}) \leq b_l$ for all $l = 1,2,\ldots,k$.

For the subset of solutions retained by PSG,

$$\begin{aligned}
\mathbb{P}\left(x_0 \in \mathscr{S}_{\mathsf{n}}^{\mathrm{PSG}}\right) &= \mathbb{P}\left(A(x_0)\widehat{\zeta} \leq b'\right) \\
&= \mathbb{P}\left(a_l^\top \widehat{\zeta} \leq b_l + \max_{\mathsf{z} \in \mathbb{R}^{k(q+1)}} \left\{a_l^\top\left(\widehat{\zeta} - \mathsf{z}\right) : d_{\mathsf{n}}(\mathsf{z}, \widehat{\zeta}, \widehat{\Psi}) \leq \mathsf{D}\right\} \text{ for all } l = 1,\ldots,k^2\right) \\
&= \mathbb{P}\left(a_l^\top\left(\widehat{\zeta} - \zeta(\mathsf{X})\right) + a_l^\top \zeta(\mathsf{X}) \leq b_l + \max_{\mathsf{z} \in \mathbb{R}^{k(q+1)}} \left\{a_l^\top\left(\widehat{\zeta} - \mathsf{z}\right) : d_{\mathsf{n}}(\mathsf{z}, \widehat{\zeta}, \widehat{\Psi}) \leq \mathsf{D}\right\} \text{ for all } l = 1,\ldots,k^2\right) \\
&\geq \mathbb{P}\left(a_l^\top\left(\widehat{\zeta} - \zeta(\mathsf{X})\right) \leq \max_{\mathsf{z} \in \mathbb{R}^{k(q+1)}} \left\{a_l^\top\left(\widehat{\zeta} - \mathsf{z}\right) : d_{\mathsf{n}}(\mathsf{z}, \widehat{\zeta}, \widehat{\Psi}) \leq \mathsf{D}\right\} \text{ for all } l = 1,\ldots,k^2\right). \qquad (14)
\end{aligned}$$

Closer inspection of (13) reveals that $A(x_0)\mathsf{z} = HC(x_0)^\top \mathsf{z}$ where $C(x_0) \in \mathbb{R}^{k(q+1)\times 2k}$ is a block-diagonal matrix defined as

$$C(x_0) \equiv \mathrm{diag}\left(C_1(x_0), C_2(x_0), \ldots, C_k(x_0)\right) \quad \text{with} \quad C_i(x_0) \equiv \begin{bmatrix} 1 & 0 \\ \mathbf{0}_q & (x_0 - x_i) \end{bmatrix} \quad \text{for } i = 1,2,\ldots,k,$$

and $H \in \mathbb{R}^{k^2 \times 2k}$ is chosen accordingly. For $l = 1,2,\ldots,k^2$, let $h_l \equiv \left(h_{l1}^\top, h_{l2}^\top, \ldots, h_{lk}^\top\right)^\top \in \mathbb{R}^{2k}$ denote the lth row of H, expressed as a column vector, where $h_{li} \in \mathbb{R}^2$ for $i = 1,2,\ldots,k$. Substituting $A(x_0) = HC(x_0)^\top$ into (14), we have

$$\mathbb{P}\left(x_0 \in \mathscr{S}_{\mathsf{n}}^{\mathrm{PSG}}\right) \geq \mathbb{P}\left(h_l^\top C(x_0)^\top\left(\widehat{\zeta} - \zeta(\mathsf{X})\right) \leq \max_{\mathsf{z} \in \mathbb{R}^{k(q+1)}} \left\{h_l^\top C(x_0)^\top\left(\widehat{\zeta} - \mathsf{z}\right) : d_{\mathsf{n}}(\mathsf{z}, \widehat{\zeta}, \widehat{\Psi}) \leq \mathsf{D}\right\} \text{ for all } l = 1,\ldots,k^2\right) \qquad (15)$$

For all $l = 1,2,\ldots,k^2$, Lemma 2 implies that

$$\begin{aligned}
\max_{\mathsf{z} \in \mathbb{R}^{k(q+1)}} \left\{h_l^\top C(x_0)^\top\left(\widehat{\zeta} - \mathsf{z}\right) : d_{\mathsf{n}}(\mathsf{z}, \widehat{\zeta}, \widehat{\Psi}) \leq \mathsf{D}\right\} &= \max_{\mathsf{z} \in \mathbb{R}^{k(q+1)}} \left\{h_l^\top C(x_0)^\top\left(\widehat{\zeta} - \mathsf{z}\right) : \max_{i=1,\ldots,k} n_i\left(\widehat{\zeta}_i - \mathsf{z}_i\right)^\top \widehat{\Psi}_i^{-1}\left(\widehat{\zeta}_i - \mathsf{z}_i\right) \leq \mathsf{D}\right\} \\
&= \sum_{i=1}^k \sqrt{\frac{\mathsf{D}}{n_i} h_{li}^\top C_i(x_0)^\top \widehat{\Psi}_i C_i(x_0) h_{li}} \\
&= \sum_{i=1}^k \sqrt{\frac{\mathsf{D}}{n_i} h_{li}^\top \left(C_i(x_0)^\top \widehat{\Psi}_i C_i(x_0)\right) h_{li}} \\
&= \max_{\mathsf{w} \in \mathbb{R}^{2k}} \left\{h_l^\top(\widehat{\omega} - \mathsf{w}) : \max_{i=1,\ldots,k} n_i(\widehat{\omega}_i - \mathsf{w}_i)^\top \left(C_i(x_0)^\top \widehat{\Psi}_i C_i(x_0)\right)^{-1}(\widehat{\omega}_i - \mathsf{w}_i) \leq \mathsf{D}\right\},
\end{aligned}$$

where $\widehat{\omega} \equiv C(x_0)^\top \widehat{\zeta} \in \mathbb{R}^{2k}$ and $\widehat{\omega}_i \equiv C_i(x_0)^\top \widehat{\zeta}_i$ for $i = 1,2,\ldots,k$.

Letting $\omega(\mathsf{X}) \equiv C(x_0)^\top \zeta(\mathsf{X})$, (15) can be rewritten as

$$\begin{aligned}
\mathbb{P}\left(x_0 \in \mathscr{S}_{\mathsf{n}}^{\mathrm{PSG}}\right) &\geq \mathbb{P}\Bigg(h_l^\top(\widehat{\omega} - \omega(\mathsf{X})) \\
&\qquad \leq \max_{\mathsf{w} \in \mathbb{R}^{2k}} \left\{h_l^\top(\widehat{\omega} - \mathsf{w}) : \max_{i=1,\ldots,k} n_i(\widehat{\omega}_i - \mathsf{w}_i)^\top \left(C_i(x_0)^\top \widehat{\Psi}_i C_i(x_0)\right)^{-1}(\widehat{\omega}_i - \mathsf{w}_i) \leq \mathsf{D}\right\} \text{ for all } l = 1,\ldots,k^2\Bigg) \\
&\geq \mathbb{P}\left(\max_{i=1,\ldots,k} n_i(\widehat{\omega}_i - \mathsf{w}_i)^\top \left(C_i(x_0)^\top \widehat{\Psi}_i C_i(x_0)\right)^{-1}(\widehat{\omega}_i - \mathsf{w}_i) \leq \mathsf{D}\right) \qquad (16)
\end{aligned}$$

Under the normality assumption (1), $\widehat{\omega}_i \sim \mathscr{N}\left(C_i(x_0)^\top \zeta_i, C_i(x_0)^\top \Psi_i C_i(x_0)\right)$ for $i = 1,2,\ldots,k$. Provided $n_i \geq 3$ for all $i = 1,2,\ldots,k$, Theorem 5.2.2 of Anderson (1984) implies that

$$n_i(\widehat{\omega}_i - \mathsf{w}_i)^\top \left(C_i(x_0)^\top \widehat{\Psi}_i C_i(x_0)\right)^{-1}(\widehat{\omega}_i - \mathsf{w}_i) \stackrel{d}{=} \frac{2(n_i - 1)}{n_i - 2} F_{2,n_i-2} \quad \text{for all } i = 1,2,\ldots,k.$$

Therefore the cutoff D is appropriately defined as the $1-\alpha$ quantile of

$$\max_{i=1,\ldots,k} \frac{2(n_i - 1)}{n_i - 2} F_{2,n_i-2},$$

where the random variables $F_{2,n_1-2}, F_{2,n_2-2}, \ldots, F_{2,n_k-2}$ are independent. Under the normality assumption (1),

$$\mathbb{P}\left(\max_{i=1,\ldots,k} n_i\left(\widehat{\omega}_i - \mathsf{w}_i\right)^\top \left(C_i(x_0)^\top \widehat{\Psi}_i C_i(x_0)\right)^{-1}\left(\widehat{\omega}_i - \mathsf{w}_i\right) \leq \mathsf{D}\right) = 1-\alpha.$$

From (16), this establishes that $\mathscr{S}_{\mathsf{n}}^{\mathrm{PSG}}$ achieves finite-sample confidence.

By the Central Limit Theorem and the Continuous Mapping Theorem,

$$n_i\left(\widehat{\omega}_i - \mathsf{w}_i\right)^\top \left(C_i(x_0)^\top \widehat{\Psi}_i C_i(x_0)\right)^{-1}\left(\widehat{\omega}_i - \mathsf{w}_i\right) \xrightarrow{d} \chi^2_2,$$

as $n_i \to \infty$ for $i = 1,\ldots,k$ where χ^2_ν denotes a chi-squared random variable with ν degree of freedom. Our choice of D converges to the $1-\alpha$ quantile of the maximum of k independent χ^2_2 random variables. Therefore,

$$\mathbb{P}\left(\max_{i=1,\ldots,k} n_i\left(\widehat{\omega}_i - \mathsf{w}_i\right)^\top \left(C_i(x_0)^\top \widehat{\Psi}_i C_i(x_0)\right)^{-1}\left(\widehat{\omega}_i - \mathsf{w}_i\right) \leq \mathsf{D}\right) \to 1-\alpha \quad \text{as} \quad \min_{i=1,\ldots,k} n_i \to \infty,$$

and $\mathscr{S}_{\mathsf{n}}^{\mathrm{PSG}}$ achieves asymptotic confidence. □

Proof of Theorem 2. Fix arbitrary $1-\alpha \in (1/2,1)$ and $\mu \in \mathscr{M}$. Fix an arbitrary solution $x_0 \in \mathscr{A}$. Since $x_0 \in \mathscr{A}$, we have that $\mu \in \mathscr{M}(x_0)$ and therefore $\nabla\mu(\mathsf{X}) \in \overline{\mathsf{Z}}(x_0)$, i.e., $(x_0 - x_i)^\top \nabla\mu(x_i) \leq \delta$ for all $i = 1,2,\ldots,k$. Then

$$\begin{aligned}
\mathbb{P}\left(x_0 \in \mathscr{S}_{\mathsf{n}}^{\mathrm{PSOG}}\right) &= \mathbb{P}\left((x_0-x_i)^\top \widehat{\nabla}\mu_i \leq \delta + \sqrt{\frac{\mathsf{D}}{n_i}\left(0,(x_0-x_i)^\top\right)\widehat{\Psi}_i\left(0,(x_0-x_i)^\top\right)^\top} \quad \text{for all } i=1,2,\ldots,k\right) \\
&= \mathbb{P}\left(\frac{(x_0-x_i)^\top \widehat{\nabla}\mu_i - \delta}{\sqrt{\frac{1}{n_i}\left(0,(x_0-x_i)^\top\right)\widehat{\Psi}_i\left(0,(x_0-x_i)^\top\right)^\top}} \leq \sqrt{\mathsf{D}} \quad \text{for all } i=1,2,\ldots,k\right) \\
&\geq \mathbb{P}\left(\frac{(x_0-x_i)^\top \widehat{\nabla}\mu_i - (x_0-x_i)^\top \nabla\mu(x_i)}{\sqrt{\frac{1}{n_i}\left(0,(x_0-x_i)^\top\right)\widehat{\Psi}_i\left(0,(x_0-x_i)^\top\right)^\top}} \leq \sqrt{\mathsf{D}} \quad \text{for all } i=1,2,\ldots,k\right). \qquad (17)
\end{aligned}$$

Under (1), each term on the left-hand side of (17) is a t-distributed random variable with $n_i - 1$ degrees of freedom. Therefore our choice of D as the squared $1-\alpha$ quantile of the maximum of independent random variables $T_{n_1-1}, T_{n_2-1}, \ldots, T_{n_k-1}$ ensures that $\mathbb{P}\left(x_0 \in \mathscr{S}_{\mathsf{n}}^{\mathrm{PSOG}}\right) \geq 1-\alpha$ for all n such that $n_i \geq 2$ for all $i = 1,2,\ldots,k$. Thus $\mathscr{S}_{\mathsf{n}}^{\mathrm{PSOG}}$ achieves finite-sample confidence.

For the asymptotic confidence result, the Multivariate Central Limit Theorem and Continuous Mapping Theorem together imply that each term on the left-hand side of (17) is asymptotically distributed as a standard normal random variable as $n_i \to \infty$ for all $i = 1,2,\ldots,k$. Our choice of cutoff D meanwhile converges to the squared $1-\alpha$ quantile of the maximum of k independent standard normal random variables as $\min_{i=1,\ldots,k} n_i \to \infty$. Therefore $\mathbb{P}\left(x_0 \in \mathscr{S}_{\mathsf{n}}^{\mathrm{PSOG}}\right) \gtrsim 1-\alpha$ and $\mathscr{S}_{\mathsf{n}}^{\mathrm{PSOG}}$ achieves asymptotic confidence. □

9.7 Consistency Results

Proof of Theorem 3. Fix arbitrary $\mu \in \mathscr{M}$ and an arbitrary solution $x_0 \notin \mathsf{S}^{\mathsf{G}}(\mathsf{X})$. From the definition of $\mathsf{S}^{\mathsf{G}}(\mathsf{X})$, there exists a pair of indices (i^*, j^*) such that $\mu(x_{i^*}) + (x_0 - x_{i^*})^\top \nabla\mu(x_{i^*}) > \mu(x_{j^*}) + \delta$. If $i^* = j^*$, then $(x_0 - x_{i^*})^\top \nabla\mu(x_{i^*}) > \delta$. From (12),

$$\begin{aligned}
\mathbb{P}\left(x_0 \in \mathscr{S}_{\mathsf{n}}^{\mathrm{PSG}}\right) = \mathbb{P}\Bigg(&\widehat{\mu}_i - \widehat{\mu}_j + (x_0-x_i)^\top \widehat{\nabla}\mu_i - \sqrt{\frac{\mathsf{D}}{n_i}\left(1,(x_0-x_i)^\top\right)\widehat{\Psi}_i\left(1,(x_0-x_i)^\top\right)^\top} - \sqrt{\frac{\mathsf{D}}{n_j}\widehat{\Psi}_{j,11}} \leq \delta \quad \text{for all } i \neq j \\
&(x_0-x_i)^\top \widehat{\nabla}\mu_i - \sqrt{\frac{\mathsf{D}}{n_i}\left(0,(x_0-x_i)^\top\right)\widehat{\Psi}_i\left(0,(x_0-x_i)^\top\right)^\top} \leq \delta \quad \text{for all } i=1,2,\ldots,k\Bigg).
\end{aligned}$$

If $i^* = j^*$,

$$\mathbb{P}\left(x_0 \in \mathscr{S}_{\mathsf{n}}^{\mathrm{PSG}}\right) \leq \mathbb{P}\left((x_0-x_{i^*})^\top \widehat{\nabla}\mu_{i^*} - \sqrt{\frac{\mathsf{D}}{n_{i^*}}\left(0,(x_0-x_{i^*})^\top\right)\widehat{\Psi}_{i^*}\left(0,(x_0-x_{i^*})^\top\right)^\top} \leq \delta\right). \qquad (18)$$

As $\min_{i=1,\ldots,k} n_i \to \infty$, $\widehat{\nabla}\mu_{i^*} \to \nabla\mu(x_{i^*})$ a.s., $\widehat{\Psi}_{i^*} \to \Psi_{i^*}$ a.s. and D converges to the $1-\alpha$ quantile of the maximum of k independent χ^2_2 random variables. By the Continuous Mapping Theorem,

$$(x_0-x_{i^*})^\top \widehat{\nabla}\mu_{i^*} - \sqrt{\frac{\mathsf{D}}{n_{i^*}}\left(0,(x_0-x_{i^*})^\top\right)\widehat{\Psi}_{i^*}\left(0,(x_0-x_{i^*})^\top\right)^\top} \xrightarrow{a.s.} (x_0-x_{i^*})^\top \nabla\mu(x_{i^*}) > \delta.$$

Therefore the probability on the right-hand side of (18) goes to zero as $\min_{i=1,\ldots,k} n_i \to \infty$.

If instead $i^* \neq j^*$,

$$\mathbb{P}\left(x_0 \in \mathscr{S}_{\mathsf{n}}^{\mathrm{PSG}}\right) \leq \mathbb{P}\left(\widehat{\mu}_{i^*} - \widehat{\mu}_{j^*} + (x_0 - x_{i^*})^\top \widehat{\nabla}\mu_{i^*} - \sqrt{\frac{\mathsf{D}}{n_{i^*}}\left(1, (x_0 - x_{i^*})^\top\right)\widehat{\Psi}_{i^*}\left(1, (x_0 - x_{i^*})^\top\right)^\top} - \sqrt{\frac{\mathsf{D}}{n_{j^*}}\widehat{\Psi}_{j^*,11}} \leq \delta\right). \quad (19)$$

As $\min_{i=1,\ldots,k} n_i \to \infty$, $\widehat{\mu}_{i^*} \to \mu(x_{i^*})$ a.s., $\widehat{\nabla}\mu_{i^*} \to \nabla\mu(x_{i^*})$ a.s., $\widehat{\Psi}_{i^*} \to \Psi_{i^*}$ a.s., and as $n_{j^*} \to \infty$, $\widehat{\mu}_{j^*} \to \mu(x_{j^*})$ a.s. and $\widehat{\Psi}_{j^*} \to \Psi_{j^*}$ a.s. By the Continuous Mapping Theorem,

$$\begin{aligned}
&\widehat{\mu}_{i^*} - \widehat{\mu}_{j^*} + (x_0 - x_{i^*})^\top \widehat{\nabla}\mu_{i^*} - \sqrt{\frac{\mathsf{D}}{n_{i^*}}\left(1, (x_0 - x_{i^*})^\top\right)\widehat{\Psi}_{i^*}\left(1, (x_0 - x_{i^*})^\top\right)^\top} - \sqrt{\frac{\mathsf{D}}{n_{j^*}}\widehat{\Psi}_{j^*,11}} \\
&\overset{a.s.}{\to} \mu(x_{i^*}) - \mu(x_{j^*}) + (x_0 - x_{i^*})^\top \nabla\mu(x_{i^*}) > \delta.
\end{aligned}$$

Therefore the probability on the right-hand side of (19) goes to zero as $\min_{i=1,\ldots,k} n_i \to \infty$.

Together, these two cases show that $\mathbb{P}\left(x_0 \in \mathscr{S}_{\mathsf{n}}^{\mathrm{PSG}}\right) \to 0$ as $\min_{i=1,\ldots,k} n_i \to \infty$, hence $\mathscr{S}_{\mathsf{n}}^{\mathrm{PSG}}$ achieves $\mathsf{S}^{\mathsf{G}}(\mathsf{X})$ consistency. □

Proof of Theorem 4. Fix arbitrary $\mu \in \mathscr{M}$ and an arbitrary solution $x_0 \notin \mathsf{S}^{\mathsf{OG}}(\mathsf{X})$. Since $x_0 \notin \mathsf{S}^{\mathsf{OG}}(\mathsf{X})$, $\max_{i=1,\ldots,k}\left\{(x_0 - x_i)^\top \nabla\mu(x_i)\right\} > \delta$, thus there exists an index $i^* \in \{1,2,\ldots,k\}$ for which $(x_0 - x_{i^*})^\top \nabla\mu(x_{i^*}) > \delta$.

From the analysis in the proof of Theorem 2,

$$\begin{aligned}
\mathbb{P}\left(x_0 \in \mathscr{S}_{\mathsf{n}}^{\mathrm{PSOG}}\right) &= \mathbb{P}\left((x_0 - x_i)^\top \widehat{\nabla}\mu_i \leq \delta + \sqrt{\frac{\mathsf{D}}{n_i}\left(0, (x_0 - x_i)^\top\right)\widehat{\Psi}_i\left(0, (x_0 - x_i)^\top\right)^\top} \quad \text{for all } i = 1,2,\ldots,k\right) \\
&= \mathbb{P}\left(\frac{(x_0 - x_i)^\top \widehat{\nabla}\mu_i - \delta}{\sqrt{\frac{1}{n_i}\left(0, (x_0 - x_i)^\top\right)\widehat{\Psi}_i\left(0, (x_0 - x_i)^\top\right)^\top}} \leq \sqrt{\mathsf{D}} \quad \text{for all } i = 1,2,\ldots,k\right\} \\
&\leq \mathbb{P}\left(\frac{(x_0 - x_{i^*})^\top \widehat{\nabla}\mu_{i^*} - \delta}{\sqrt{\frac{1}{n_{i^*}}\left(0, (x_0 - x_i^*)^\top\right)\widehat{\Psi}_i\left(0, (x_0 - x_i^*)^\top\right)^\top}} \leq \sqrt{\mathsf{D}}\right) \\
&= \mathbb{P}\left(\frac{(x_0 - x_{i^*})^\top \widehat{\nabla}\mu_{i^*} - (x_0 - x_{i^*})^\top \nabla\mu(x_{i^*})}{\sqrt{\frac{1}{n_{i^*}}\left(0, (x_0 - x_i^*)^\top\right)\widehat{\Psi}_i\left(0, (x_0 - x_i^*)^\top\right)^\top}} + \frac{(x_0 - x_{i^*})^\top \nabla\mu(x_{i^*}) - \delta}{\sqrt{\frac{1}{n_{i^*}}\left(0, (x_0 - x_i^*)^\top\right)\widehat{\Psi}_i\left(0, (x_0 - x_i^*)^\top\right)^\top}} \leq \sqrt{\mathsf{D}}\right). \quad (20)
\end{aligned}$$

By the Multivariate Central Limit Theorem and the Continuous Mapping Theorem, the first term on the left-hand side of (20) is asmpytotically distributed as a standard normal random variable. The second term on the left-hand side has a positive numerator (from the definition of i^*) and its denominator converges in probability to 0 as $n_{i^*} \to \infty$. The choice of cutoff D converges to a constant: the square of the $1-\alpha$ quantile of the maximum of k independent standard normal random variables. Therefore $\mathbb{P}\left(x_0 \in \mathscr{S}_{\mathsf{n}}^{\mathrm{PSOG}}\right) \to 0$ and we have that $\mathscr{S}_{\mathsf{n}}^{\mathrm{PSOG}}$ achieves $\mathsf{S}^{\mathsf{O}}(\mathsf{X})$ consistency. □

9.8 Limiting Set Results

Proof of Theorem 5. Fix arbitrary $\mathscr{X}$, X, and $\mu \in \mathscr{M}$. Fix an arbitrary solution $x_0 \in \mathsf{S}^{\mathsf{G}}(\mathsf{X})$.

For the first result, notice that

$$\begin{aligned}
\max_{i=1,\ldots,k}\left\{\mu(x_i) + (x_0 - x_i)^\top \nabla\mu(x_i)\right\} \leq \min_{j=1,\ldots,k} \mu(x_j) + \delta \quad &\Leftrightarrow \quad \max_{i=1,\ldots,k}\left\{\mu(x_i) + (x_0 - x_i)^\top \nabla\mu(x_i)\right\} - \min_{j=1,\ldots,k} \mu(x_j) \leq \delta \\
&\Rightarrow \quad \max_{i=1,\ldots,k}\left\{(x_0 - x_i)^\top \nabla\mu(x_i)\right\} \leq \delta.
\end{aligned}$$

Thus $x_0 \in \mathsf{S}^{\mathsf{OG}}(\mathsf{X})$ and $\mathsf{S}^{\mathsf{G}}(\mathsf{X}) \subseteq \mathsf{S}^{\mathsf{OG}}(\mathsf{X})$.

For the second result, set $\xi_i = \nabla\mu(x_i)$ for $i = 1,2,\ldots,k$. It can be seen from the definition of $\mathsf{S}^{\mathsf{O}}(\mathsf{X})$ that $x_0 \in \mathsf{S}^{\mathsf{O}}(\mathsf{X})$. Thus $\mathsf{S}^{\mathsf{G}}(\mathsf{X}) \subseteq \mathsf{S}^{\mathsf{O}}(\mathsf{X})$. □