Plausible Screening Using Functional Properties for Simulations with Large Solution Spaces

David J. Eckman, Matthew Plumlee, and Barry L. Nelson
Department of Industrial Engineering and Management Sciences
Northwestern University
Evanston, Illinois 60208

Abstract

When working with models that allow for many candidate solutions, simulation practitioners can benefit from screening out unacceptable solutions in a statistically controlled way. However, for large solution spaces, estimating the performance of all solutions through simulation can prove impractical. We propose a statistical framework for screening solutions even when only a relatively small subset of them are simulated. Our framework derives its superiority over exhaustive screening approaches by leveraging available properties of the function that describes the performance of solutions. The framework is designed to work with a wide variety of available functional information and provides guarantees on both the confidence and consistency of the resulting screening inference. We provide explicit formulations for the properties of convexity and Lipschitz continuity and show through numerical examples that our procedures can efficiently screen out many unacceptable solutions.

1 Introduction

Operations researchers have increasingly relied on stochastic simulations to understand complex systems. These simulation models are typically endowed with a vector of parameterized inputs we term a *solution*. Each solution has an associated performance that can be estimated by running replications of the simulation with the corresponding inputs. Motivating applications of this general approach arise in simulation optimization, feasibility determination, and model calibration.

When there are many candidate solutions, it can be difficult to thoroughly evaluate the performance of all solutions through exhaustive simulation. A more reasonable approach is to first screen out, meaning remove from consideration, solutions regarded as unacceptable based on initial experiments. However, obtaining even a single replication from all candidate solutions is sometimes impractical. Thus, our goal is to provide a method for screening solutions that can work even when simulating only a small subset of them. Screening procedures can be employed to efficiently remove unacceptable solutions before running a more intensive algorithm [Nelson et al., 2001] or for post hoc analysis [Boesel et al., 2003]. This

use of the term "screening" differs from "factor screening" which entails removing solution-defining variables having minimal impact on the performance [Bettonvil and Kleijnen, 1997, Wan et al., 2006].

Although we discuss our methodological framework in generality, we at times focus on its uses for simulation optimization, where either optimal or near-optimal solutions are deemed acceptable. Within this setting, classical subset-selection methods [Gupta, 1965, Nelson et al., 2001, Boesel et al., 2003] guarantee that the optimal solution is retained with high probability. While these methods are highly effective and have been extended to parallel computing environments [Ni et al., 2014], they do not solve our problem as posed, as they still require simulating all candidate solutions. These methods treat the performances of solutions as being unrelated to their location in the solution space and therefore fail to exploit any structural properties of the performance function.

A separate technique that directly targets the performance function is simulation metamodeling. Here, one builds an approximate model of the performance function, often based on statistical or machine-learning models. These metamodels allow one to predict performances at unsimulated solutions. Some metamodeling methods formalize functional properties as constraints and determine the metamodel that best fits the simulation outputs subject to those constraints, e.g., convex and polynomial regression [Lim and Glynn, 2012, Kleijnen, 2015]. Others impose a probabilistic structure, e.g., Gaussian process regression [Ankenman et al., 2010] or Gaussian Markov random fields [Salemi et al., 2019a]. While metamodels are central to some simulation-optimization searches, e.g., stochastic trust-region methods like STRONG [Chang et al., 2013] and ASTRO-DF [Shashaani et al., 2018], to the best of our knowledge metamodels have not been used for screening. Furthermore, metamodels do not naturally lend themselves to probabilistic guarantees without extremely strong assumptions [Wan et al., 2016]. For example, the commonly used commercial software `OptQuest` employs neural networks to remove solutions from consideration [Laguna, 2011], but the procedure lacks statistical guarantees on the screening inference. Our methods achieve the best of both: screening out unsimulated solutions while providing a statistical guarantee akin to that of subset selection.

Our framework converts general information about the performance function into a screening approach delivering statistical guarantees. This is valuable because for some simulation models it is possible to analytically or empirically establish properties of the performance function, such as Lipschitz continuity, convexity, or bounds. More specifically, we propose screening solutions by measuring the discrepancy between the observed data and the space of performance functions having certain known properties; our framework thus shares some concepts with constrained statistical inference [Silvapulle and Sen, 2005]. When further restricting the space of functions to those for which a particular solution is acceptable, this discrepancy measures the plausible acceptability of said solution. A very large discrepancy at a solution implies it is implausible that the solution is acceptable. Our methods accordingly remove from consideration solutions for which the discrepancy is sufficiently large—an act we term *plausible screening*. We prescribe reasonable discrepancies and cutoffs that achieve standard statistical properties desired in screening. With proper care, our methods can provide *confidence*—which can be thought of as the probability of correct selection guarantee from subset selection—and *consistency*—the concept that any unacceptable solution is screened out in the limit. Our results here substantially extend the

preliminary results presented in Plumlee and Nelson [2018] and Eckman et al. [2020].

This article introduces the screening framework, details the computational implementation, and provides some numerical examples. In Section 2, we mathematically formulate the problem of screening unacceptable solutions, and in Section 3, we motivate our approach of exploiting available information about the performance function. Section 4 lays out the theoretical underpinning for assessing plausible acceptability and presents an algorithm for constructing a subset that attains asymptotic confidence and a weak form of consistency. We then present an alternative algorithm in Section 5 that can, in certain instances, more efficiently construct a relaxed subset of solutions. In Section 6, we test the algorithms on realistic simulation-optimization problems. We conclude in Section 7 with potential extensions of the framework and open research questions.

2 Setting and Goals

This section describes the setup for evaluating solutions via stochastic simulation, the general definition of acceptable solutions, and the statistical guarantees we desire in screening.

2.1 Stochastic Simulation

We lay out a mathematical framework for screening simulated solutions from a set of candidate solutions $\mathcal{X} \subseteq \mathbb{R}^d$ which can be discrete, countable or uncountable. Each solution $x \in \mathcal{X}$ has an associated scalar quantity of interest labeled $\mu(x)$, which is unknown but can be estimated by sampling replications of a stochastic simulation. We refer to $\mu(x)$ as the *performance* of solution x. For situations in which $\mathcal{X}$ is large, meaning either a large discrete set, an infinite set, or a continuum of solutions, estimating the performances of all candidate solutions is impractical or impossible. Thus, a decision-maker simulates only a subset of k solutions, $\mathsf{X} \equiv \{x_1, x_2, \ldots, x_k\} \subseteq \mathcal{X}$, termed the *experimental set*. While we discuss the experimental set X generically, it may be chosen, for example, to fill $\mathcal{X}$ or to concentrate sampling around a region of interest. We find it convenient to consider the restriction of the function $\mu\colon \mathcal{X} \mapsto \mathbb{R}$ to X, denoted by $\mu(\mathsf{X}) \equiv (\mu(x_1), \ldots, \mu(x_k))^\top$, which is the vector of the performances of the simulated solutions. While not directly observable, this vector can be estimated through simulation on the limited experimental set.

Let $Y_\ell(x)$ denote the (stochastic) output of the ℓth independent and identically distributed (i.i.d.) simulation replication at a solution x, with $\mathbb{E}[Y_\ell(x)] = \mu(x)$ for all $\ell = 1, 2, \ldots$ and all solutions x in $\mathcal{X}$. For any pair of solutions x and x', $Y_\ell(x)$ and $Y_\ell(x')$ are related via a common covariance function $\Sigma\colon \mathcal{X} \times \mathcal{X} \mapsto \mathbb{R}$ described by $\Sigma(x, x') \equiv \mathrm{Cov}(Y_\ell(x), Y_\ell(x'))$. The covariance matrix denoted by $\Sigma(\mathsf{X})$ gives the covariance between outputs for all pairs of solutions in the experimental set, and we assume that $\Sigma(\mathsf{X})$ is positive definite. For a given ℓ, define $\mathbf{Y}_\ell \equiv (Y_\ell(x_1), \ldots, Y_\ell(x_k))^\top$, the vector of outputs from the ℓth simulation replications at each solution in the experimental set. The vectors $\mathbf{Y}_1, \mathbf{Y}_2, \ldots,$ are assumed to be mutually independent and identically distributed.

We consider two ways of simulating replications across solutions:

(S1) *Independent Sampling:* Outputs at different solutions are independent (i.e., $\Sigma(\mathsf{X})$ is diagonal) and the number of replications taken at each solution $x_i \in \mathsf{X}$ is n_i for $i =$

$1, \ldots, k$, possibly unequal.

(S2) *Dependent Sampling:* Outputs at different solutions are dependent—as would be the case if common random numbers (CRN) were used—and an equal number of replications is taken at each solution $x_i \in \mathsf{X}$, i.e., $n_i = n$ for $i = 1, \ldots, k$.

We estimate $\mu(\mathsf{X})$ by $\widehat{\mu} \equiv (\widehat{\mu}_1, \ldots, \widehat{\mu}_k)^\top$ where $\widehat{\mu}_i = {n_i}^{-1} \sum_{\ell=1}^{n_i} Y_\ell(x_i)$ for $i = 1, \ldots, k$ and $\Sigma(\mathsf{X})$ by

$$\widehat{\Sigma} \equiv \begin{cases} \text{diag}(\widehat{\sigma}_1^2, \ldots, \widehat{\sigma}_k^2) \text{ where } \widehat{\sigma}_i^2 = (n_i - 1)^{-1} \sum_{\ell=1}^{n_i} (Y_\ell(x_i) - \widehat{\mu}_i)^2 \text{ for } i = 1, \ldots, k & \text{in (S1)}, \\ [\widehat{\sigma}_{ij}^2]_{k \times k} \text{ where } \widehat{\sigma}_{ij}^2 = (n-1)^{-1} \sum_{\ell=1}^{n} (Y_\ell(x_i) - \widehat{\mu}_i)(Y_\ell(x_j) - \widehat{\mu}_j) \text{ for } i, j = 1, \ldots, k & \text{in (S2)}. \end{cases}$$

We assume that $\widehat{\Sigma}$ is positive definite with probability one.

2.2 Acceptable Solutions

For a given performance function μ, we define $\mathcal{A}$ as the set of solutions deemed *acceptable* by the decision maker, i.e., those whose performances exhibit some quality of interest. Although $\mathcal{A}$ depends on the unknown function μ, we choose to suppress μ from the notation. Different definitions of acceptability arise in a variety of simulation applications and can be illustrated within the setting of production planning, such as semiconductor wafer fabrication [Liu et al., 2011]. Discrete-event simulation models are used in this domain to study the costs associated with a given release plan—a schedule of batch jobs for different product types—subject to stochastic demand for the products. A decision-maker may be interested in finding a release plan x whose expected total costs (defined as the sum of work-in-progress cost, inventory cost, and backlog cost) is within δ dollars of the smallest. Alternatively, the decision-maker may wish to determine whether a given release plan satisfies a service requirement, e.g., that the associated expected backlog cost is below μ_0. It may also be of interest to improve upon a *control* or default release plan x^c, such as the one suggested by a simple model. On the other hand, the decision-maker may be interested in release plans whose expected work in progress is within ϵ units of $\mu^\dagger$ [Spearman et al., 1990].

While we leave $\mathcal{A}$ purposely vague to demonstrate the versatility of the proposed framework, one can describe these common examples of $\mathcal{A}$ mathematically:

- Optimization: $\{x \in \mathcal{X} \colon \mu(x) \leq \min_{x' \in \mathcal{X}} \mu(x') + \delta\}$ for some optimality gap $\delta \geq 0$;
- Feasibility Determination: $\{x \in \mathcal{X} \colon \mu(x) \leq \mu_0\}$ for some threshold μ_0;
- Comparison to a Control: $\{x \in \mathcal{X} \colon \mu(x) \leq \mu(x^c)\}$ for some control solution $x^c \in \mathcal{X}$;
- Comparison to a Target: $\{x \in \mathcal{X} \colon |\mu(x) - \mu^\dagger| \leq \epsilon\}$ for some tolerance $\epsilon \geq 0$ and target $\mu^\dagger$.

In the first three examples, it is assumed without loss of generality that smaller performance is preferable. A common feature is that determining whether a given solution belongs to $\mathcal{A}$ entails checking a (possibly infinite) system of linear inequalities with respect to the candidate solutions' performances. We later leverage this property to develop tractable methods for inferring whether an arbitrary solution is acceptable.

2.3 Statistical Guarantees in Screening

Ideally, the decision-maker seeks to identify the full set of acceptable solutions and no others, to serve as the basis for some decision. In the presence of simulation error, the decision-maker must settle for a subset of solutions having desirable statistical guarantees in terms of *screening*, i.e., inferring which solutions are acceptable [Bechhofer et al., 1995]. Let $\mathcal{S}_{\mathsf{n}}$ denote the subset of solutions returned after obtaining replications at solutions in X as specified by $\mathsf{n} \equiv (n_1, \ldots, n_k)$. Our definitions of statistical guarantees of subsets suppose that the performance function μ belongs to some function space $\mathscr{M}$, which we specify in Section 3.2.

Definition 1 (Finite-sample confidence) *A subset $\mathcal{S}_{\mathsf{n}}$ achieves* finite-sample confidence $1-\alpha$ *for $\alpha \in [0,1]$ if for sufficiently large $\min_{i=1,\ldots,k} n_i$ and any $\mu \in \mathscr{M}$, $\mathbb{P}(x_0 \in \mathcal{S}_{\mathsf{n}}) \geq 1-\alpha$ for all $x_0 \in \mathcal{A}$.*

Finite-sample confidence states that for any performance function in $\mathscr{M}$, each acceptable solution will be correctly screened with marginal probability exceeding $1-\alpha$. For the most part, finite-sample confidence is unattainable unless the random outputs of the simulation replications come from a known family of distributions. A more widely achievable property is asymptotic confidence, which follows from designing methods for normally distributed outputs and applying the Central Limit Theorem.

Definition 2 (Asymptotic confidence) *A subset $\mathcal{S}_{\mathsf{n}}$ achieves* asymptotic confidence $1-\alpha$ *for $\alpha \in [0,1]$ if for any $\mu \in \mathscr{M}$, $\mathbb{P}(x_0 \in \mathcal{S}_{\mathsf{n}}) \gtrsim 1-\alpha$ as $\min_{i=1,\ldots,k} n_i \to \infty$ for all $x_0 \in \mathcal{A}$.*

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

Finite-sample and asymptotic confidence describe a subset's ability to avoid screening out acceptable solutions with high probability, but not its ability to screen out unacceptable solutions, i.e., those that do not belong to $\mathcal{A}$. For this, we require the notion of consistency.

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

Except in special cases, like exhaustive simulation where $\mathsf{X} = \mathcal{X}$, consistency is unachievable since even with direct evaluation of $\mu(\mathsf{X})$, the rest of the performance function is indeterminable. We soon introduce a less exacting form of consistency that accounts for having simulated at only solutions in the experimental set.

3 Screening Using Functional Properties

In this section, we explain how known functional properties of the performance function can be combined with our screening framework. Section 3.3 develops the main ideas behind our methods in a simplified setting in which solutions are simulated without error.

3.1 Functional Properties of the Performance Function

Our goal is to use information obtained from a small experimental set to screen out a massive number of solutions that, given the data, could not plausibly be acceptable. Importantly, we seek the ability to screen out even unsimulated solutions. This task would be impossible without some means of relating an unsimulated solution's performance to those of simulated solutions from the experimental set. Our approach operates under the assumption that the decision-maker possesses known or assumed properties of the performance function μ that enable such comparisons among simulated and unsimulated solutions. Examples include knowledge that μ is convex (likewise concave, strongly convex, or almost convex) over $\mathcal{X}$, Lipschitz continuous (likewise Hölder continuous or second-order Lipschitz continuous) with a known or assumed upper bound on the associated constant, or a polynomial in x with known or assumed degree. This type of information can also be augmented with auxiliary properties on the performances of individual solutions, like bounds on μ or known performances of some solutions.

Our presupposition of a priori functional information is in contrast to the approach of assuming a probabilistic structure for μ, e.g., treating μ as a realization of a Gaussian process on $\mathcal{X}$, as is common in metamodeling [Santner et al., 2003, Salemi et al., 2019b] and Bayesian optimization [Frazier et al., 2009, Scott et al., 2011]. By the same token, our approach differs from that of estimating the posterior probability that μ satisfies certain functional properties, e.g., convexity [Jian and Henderson, 2020]. Our framework therefore does not require a measure over the function space described by the known properties, only a means of checking whether an arbitrary function lies in the space.

Situations in which such knowledge of functional properties is available are not rare. We explicate two numerical examples where Lipschitz continuity or convexity information is present in Section 6. In addition, we call attention to two common, practical techniques for verifying functional properties of simulation models.

Inheritance from sample-path functions. Many properties of sample-path functions are inherited when applying the expectation operator, e.g., convexity [Shaked and Shanthikumar, 1988], continuity [Shapiro and Wardi, 1996], and bounds. Proving that the sample-path functions possess any such property with probability one implies that the performance function does as well [Kim et al., 2015]. Examples include:

> *Stochastic activity networks.* The expected length of the longest path is a convex function in terms of the mean task durations; see Appendix E of Plambeck et al. [1996] for a derivation.
>
> *Tandem production lines with unreliable machines.* The steady-state throughput is a convex function in terms of the cycle times of the machines [Plambeck et al., 1996].
>
> *Inventory stocking under dynamic customer substitution.* The expected profit can be shown to be a Hölder-continuous function in terms of the initial inventory levels [Mahajan and van Ryzin, 2001].

Stochastic orders. Some stochastic orders imply an inequality relating two expected values [Shaked and Shanthikumar, 2007]. This approach can be used to relate $\mu(x)$ and $\mu(x')$

for some $x \neq x'$ or to relate $\mu(x)$ to another expected value that is known, thereby providing a bound on $\mu(x)$. Examples include:

$GI/GI/c$ queueing systems. Many random quantities of interest (e.g., sequential departure times) are stochastically ordered when comparing a $GI/GI/c$ queueing system with a first-in first-out service discipline to another in which arrivals are arbitrarily assigned among c channels, independent of the service process [Wolff, 1977].

Portfolio optimization. A risk-averse decision-maker wishes to assemble a portfolio from a finite collection of assets to maximize the expected return rate while requiring that the portfolio's return rate stochastically dominate a benchmark rate [Dentcheva and Ruszczyński, 2006].

3.2 Spaces of Performance Functions

We incorporate functional properties into our framework by characterizing how they restrict the set of functions to which μ can belong and, in turn, the values its restriction $\mu(\mathsf{X})$ can take. Let $\mathscr{F}$ denote the set of functions mapping from $\mathcal{X}$ to $\mathbb{R}$ and let $\mathscr{M} \subseteq \mathscr{F}$ denote the set of functions that possess the specified functional properties. Furthermore, for a given performance function $m \in \mathscr{M}$, let $\mathcal{A}(m)$ represent the corresponding set of acceptable solutions. Since screening takes place by examining individual solutions $x_0 \in \mathcal{X}$, we define $\mathscr{M}(x_0) \equiv \{m \in \mathscr{M} : x_0 \in \mathcal{A}(m)\}$, the set of functions in $\mathscr{M}$ for which solution x_0 is acceptable.

Recall that $\mu(\mathsf{X})$ represents the performances of the solutions in our experimental set $x_1, \ldots, x_k$. We will similarly use the notation $m(\mathsf{X})$ to denote the values an arbitrary function m takes at those same $x_1, \ldots, x_k$. The projection of $\mathscr{M}(x_0)$ onto $\mathbb{R}^k$—corresponding to solutions in X—is defined as

$$\mathsf{M}(x_0) \equiv \left\{\mathsf{m} \in \mathbb{R}^k : \text{there exists } m \in \mathscr{M}(x_0) \text{ such that } m(\mathsf{X}) = \mathsf{m}\right\},$$

the set of vectors of performances of the solutions $x_1, \ldots, x_k$ for which there exists an interpolating function m belonging to $\mathscr{M}(x_0)$. It follows from these definitions that for a given performance function $\mu \in \mathscr{M}$, its restriction $\mu(\mathsf{X})$ is in $\mathsf{M}(x_0)$ if x_0 is an acceptable solution. The converse, however, does not necessarily hold, since the restriction of μ to X does not determine the performances of solutions in the rest of the solution space. We next demonstrate the central role $\mathsf{M}(x_0)$ plays in screening.

3.3 Screening Solutions without Estimation Error

Temporarily assume that solutions' performances can be calculated directly without simulation, i.e., the decision-maker can directly obtain $\mu(\mathsf{X})$. Given that only solutions in the experimental set have been evaluated, some solutions likely cannot be correctly screened with certainty. A reasonable approach is to classify as belonging to $\mathcal{A}$ any solution x_0 for which there exists a function in $\mathscr{M}(x_0)$ that interpolates $\mu(\mathsf{X})$. We denote the resulting subset of solutions by

$$\mathsf{S}(\mathsf{X}) \equiv \{x_0 \in \mathcal{X} : \mu(\mathsf{X}) \in \mathsf{M}(x_0)\}.$$

This notation reflects the dependence of the subset $\mathsf{S}(\mathsf{X})$ on X; a different experimental set would yield a different subset of solutions that are possibly acceptable. The subset $\mathsf{S}(\mathsf{X})$ is the smallest subset that contains all acceptable solutions having only evaluated solutions in X and only knowing the given properties of μ. If an arbitrary solution x_0 is not in $\mathsf{S}(\mathsf{X})$, we conclude that there does not exist an interpolating function in $\mathscr{M}(x_0)$, hence it is impossible that x_0 is an acceptable solution. Therefore, all acceptable solutions are included in this subset, i.e. $\mathcal{A} \subseteq \mathsf{S}(\mathsf{X})$. In addition, if $\mathsf{X} = \mathcal{X}$, then all solutions can be correctly screened, meaning $\mathsf{S}(\mathsf{X}) = \mathcal{A}$. Thus, the gap between these two subsets of solutions comes from the fact that the experimental set comprises only a subset of the candidate solutions.

Example 1 (Optimizing a Lipschitz Continuous Function) *We apply the formulation above to an optimization problem in which the objective function μ is known to be Lipschitz continuous with constant γ, taking $\mathcal{A} = \{x \in \mathcal{X} \colon \mu(x) \leq \min_{x' \in \mathcal{X}} \mu(x')\}$. Thus,*

$$\mathsf{M}(x_0) = \left\{\mathsf{m} \in \mathbb{R}^k \colon \mathsf{m}_i - \mathsf{m}_j \leq \gamma \min\left\{\|x_i - x_j\|, \|x_i - x_0\|\right\} \text{ for all } i, j = 1, \ldots, k\right\},$$

and

$$\mathsf{S}(\mathsf{X}) = \left\{x_0 \in \mathcal{X} \colon \max_{i=1,\ldots,k} \mu(x_i) - \gamma\|x_i - x_0\| \leq \min_{j=1,\ldots,k} \mu(x_j)\right\},$$

where m_i is the ith component of the vector m and $\|\cdot\|$ is the Euclidean norm; see Appendix A for a complete derivation.

To determine if a solution x_0 belongs to $\mathsf{S}(\mathsf{X})$, one must check whether $\mu(\mathsf{X})$ belongs to $\mathsf{M}(x_0)$. If $\mathsf{M}(x_0)$ can be expressed as a polyhedron with an explicit constraint matrix and right-hand-side vector, as in Example 1, then checking if $\mu(\mathsf{X})$ is in $\mathsf{M}(x_0)$ is straightforward. More generally, if $\mathsf{M}(x_0)$ can be implicitly described as the projection of a polyhedron, then checking if $\mu(\mathsf{X})$ is in $\mathsf{M}(x_0)$ involves solving a linear program. In Section 5, we exploit this fact to devise efficient methods for screening solutions. One can imagine further expanding this framework to nonlinear constraints, but this paper focuses on the potential in polyhedral representations.

As setup for the following sections, we present a relaxed version of consistency featuring $\mathsf{S}(\mathsf{X})$.

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

As $\mathcal{A}$ is a subset of $\mathsf{S}(\mathsf{X})$, we conclude that $\mathsf{S}(\mathsf{X})$ consistency holds whenever consistency (Definition 3) holds. The property of $\mathsf{S}(\mathsf{X})$ consistency implies that as the simulation effort at solutions in X increases to infinity, the probability that a given solution is in $\mathcal{S}_\mathsf{n}$ goes to zero for any solution that could be screened out if $\mu(\mathsf{X})$ were known. In other words, an $\mathsf{S}(\mathsf{X})$-consistent subset asymptotically screens out all solutions that—given the limited experimental set and known functional properties of the performance function—cannot possibly be acceptable.

4 Plausible Screening

In this section, we give an overview of our method of accounting for simulation error alongside theoretical results that justify its use.

4.1 Overview

When solutions in the experimental set are simulated without error, as in Section 3.3, a natural subset to return is $\mathsf{S}(\mathsf{X})$, which consists of all solutions x_0 for which $\mu(\mathsf{X}) \in \mathsf{M}(x_0)$. However, when there is simulation error, naively plugging in the estimator $\widehat{\mu}$ for the unknown $\mu(\mathsf{X})$ and retaining all solutions x_0 for which $\widehat{\mu} \in \mathsf{M}(x_0)$ will not produce a subset achieving confidence and $\mathsf{S}(\mathsf{X})$ consistency. Since the probability that $\widehat{\mu} \in \mathsf{M}(x_0)$ is not well controlled, this will likely result in a set that eliminates too many solutions, thus violating the confidence guarantee through undercoverage. We properly account for the uncertainty about $\mu(\mathsf{X})$ by developing a subset comprising solutions x_0 for which $\widehat{\mu}$ is sufficiently close to $\mathsf{M}(x_0)$, where the precise meaning of "sufficiently close" ensures our guarantees of confidence and $\mathsf{S}(\mathsf{X})$ consistency are delivered.

To measure the distance between $\widehat{\mu}$ and $\mathsf{M}(x_0)$, we first introduce the *standardized discrepancy* between $\widehat{\mu}$ and a performance vector $\mathsf{m} = (\mathsf{m}_1, \ldots, \mathsf{m}_k)$, denoted by $d_{\mathsf{n}}(\mathsf{m}, \widehat{\mu}, \widehat{\Sigma})$. The vector of sample sizes, n, and sample covariance matrix, $\widehat{\Sigma}$, appear in the standardized discrepancy for the purpose of scaling differences between performance vectors in line with the estimation error; specific examples will be given below. We hereafter assume that the standardized discrepancy satisfies the following condition:

(C1) $d_{\mathsf{n}}(\mathsf{m}, \widehat{\mu}, \widehat{\Sigma}) \geq 0$ for all $\mathsf{m} \in \mathbb{R}^k$ and $d_{\mathsf{n}}(\widehat{\mu}, \widehat{\mu}, \widehat{\Sigma}) = 0$ with probability one.

Additional conditions are introduced in Section 4.2 that are necessary for maintaining confidence and $\mathsf{S}(\mathsf{X})$ consistency.

Minimizing the standardized discrepancy over performance vectors in $\mathsf{M}(x_0)$ gives the *minimum standardized discrepancy* of x_0,

$$D_{\mathsf{n}}(x_0, \widehat{\mu}, \widehat{\Sigma}) \equiv \min_{\mathsf{m} \in \mathsf{M}(x_0)} d_{\mathsf{n}}(\mathsf{m}, \widehat{\mu}, \widehat{\Sigma}), \tag{1}$$

which can be interpreted as the distance between the sample mean vector $\widehat{\mu}$ and the set $\mathsf{M}(x_0)$. The minimum standardized discrepancy is an indication of how likely it is that, given the sample data, the true performance function μ belongs to $\mathscr{M}(x_0)$, the space of functions that possess the known functional properties and for which solution x_0 is acceptable. A smaller value of $D_{\mathsf{n}}(x_0, \widehat{\mu}, \widehat{\Sigma})$ indicates stronger evidence that x_0 is an acceptable solution, while a larger value of $D_{\mathsf{n}}(x_0, \widehat{\mu}, \widehat{\Sigma})$ indicates stronger evidence that x_0 is an unacceptable solution.

We say that a function m is *plausible* with respect to an arbitrary solution x_0 if it belongs to $\mathscr{M}(x_0)$ and its restriction $m(\mathsf{X})$ is sufficiently close to $\widehat{\mu}$ in terms of the standardized discrepancy between the two vectors. From Definition (1), a solution x_0 admits a plausible function if and only if its minimum standardized discrepancy $D_{\mathsf{n}}(x_0, \widehat{\mu}, \widehat{\Sigma})$ is sufficiently small. Our screening method, which we refer to as Plausible Screening (PS), returns the subset comprising solutions x_0 for which there exists a plausible function. To be precise, the PS subset $\mathcal{S}_{\mathsf{n}}^{\mathrm{PS}}$ consists of solutions x_0 for which $\widehat{\mu}$ is within a distance D of $\mathsf{M}(x_0)$, i.e.,

$$\mathcal{S}_{\mathsf{n}}^{\mathrm{PS}} \equiv \left\{ x_0 \in \mathcal{X} \colon D_{\mathsf{n}}(x_0, \widehat{\mu}, \widehat{\Sigma}) \leq \mathsf{D} \right\}.$$

Equivalently, $\mathcal{S}_{\mathsf{n}}^{\mathrm{PS}}$ can be defined as all $x_0 \in \mathcal{X}$ such that $\widehat{\mu} \in \mathsf{R}(x_0)$, where

$$\mathsf{R}(x_0) \equiv \left\{ \widetilde{\mathsf{m}} \in \mathbb{R}^k \colon D_{\mathsf{n}}(x_0, \widetilde{\mathsf{m}}, \widehat{\Sigma}) \leq \mathsf{D} \right\}$$

is the set of performance vectors that are within a distance D to $\mathsf{M}(x_0)$. Just as $\mathsf{M}(x_0)$ is the performance set for which x_0 is possibly acceptable when $\mu(\mathsf{X})$ is directly observed, its random relaxation $\mathsf{R}(x_0)$ can be viewed as a performance set for which x_0 is plausibly acceptable, in light of the uncertainty about $\mu(\mathsf{X})$.

4.2 Statistical Guarantees

From the definition of $\mathcal{S}_{\mathsf{n}}^{\mathrm{PS}}$, we can see that choosing D as the $1-\alpha$ quantile of the minimum standardized discrepancy $D_{\mathsf{n}}(x_0, \widehat{\mu}, \widehat{\Sigma})$ leads to finite-sample confidence. However, the distribution of the minimum standardized discrepancy depends on the unknown quantities $\mu(\mathsf{X})$ and $\Sigma(\mathsf{X})$ in addition to sample sizes, functional constraints and the definition of acceptability. For the cases we investigate, namely Lipschitz continuity and convexity of μ, the associated quantile cannot be evaluated numerically or by Monte Carlo. We circumvent this by considering the statistic $d_{\mathsf{n}}(\mu(\mathsf{X}), \widehat{\mu}, \widehat{\Sigma})$, which first-order stochastically dominates the minimum standardized discrepancy because when x_0 is an acceptable solution, $\mu(\mathsf{X}) \in \mathsf{M}(x_0)$, and hence $d_{\mathsf{n}}(\mu(\mathsf{X}), \widehat{\mu}, \widehat{\Sigma}) \geq \min_{\mathsf{m} \in \mathsf{M}(x_0)} d_{\mathsf{n}}(\mathsf{m}, \widehat{\mu}, \widehat{\Sigma}) = D_{\mathsf{n}}(x_0, \widehat{\mu}, \widehat{\Sigma})$. We introduce standardized discrepancies for which $d_{\mathsf{n}}(\mu(\mathsf{X}), \widehat{\mu}, \widehat{\Sigma})$ is pivotal under a normality assumption; i.e., its distribution is independent of $\mu(\mathsf{X})$ and $\Sigma(\mathsf{X})$. Its distribution is also independent of $\mathsf{M}(x_0)$, since setting $\mathsf{m} = \mu(\mathsf{X})$ avoids the minimization in Definition (1). This simplification allows us to derive a deterministic, uniform cutoff D that ensures $\mathcal{S}_{\mathsf{n}}^{\mathrm{PS}}$ has the desired statistical properties.

We require that the pairing of $d_{\mathsf{n}}(\cdot, \widehat{\mu}, \widehat{\Sigma})$ and D satisfy three conditions for all $\mu(\mathsf{X}) \in \mathbb{R}^k$ and $\Sigma(\mathsf{X}) \in \mathbb{R}^{k \times k}$ positive definite:

(C2) $\mathbb{P}\left(d_{\mathsf{n}}(\mu(\mathsf{X}), \widehat{\mu}, \widehat{\Sigma}) \leq \mathsf{D}\right) \geq 1 - \alpha$;

(C3) $\mathbb{P}\left(d_{\mathsf{n}}(\mu(\mathsf{X}), \widehat{\mu}, \widehat{\Sigma}) \leq \mathsf{D}\right) \to 1 - \alpha \quad \text{as} \min_{i=1,\ldots,k} n_i \to \infty$; and

(C4) $\max_{\mathsf{m} \in \mathbb{R}^k} \left\{ \|\widehat{\mu} - \mathsf{m}\| \colon d_{\mathsf{n}}(\mathsf{m}, \widehat{\mu}, \widehat{\Sigma}) \leq \mathsf{D} \right\} \overset{w.p.1}{\to} 0$ as $\min_{i=1,\ldots,k} n_i \to \infty$,

where $\|\cdot\|$ again denotes the Euclidean norm. Although the choice of D satisfying the conditions above depends on the values of k, n, and α, we choose to suppress this dependence in the notation.

Conditions (C2) and (C3) relate to finite-sample and asymptotic confidence, respectively, ensuring that D is sufficiently large. Condition (C4), on the other hand, relates to consistency. It ensures that D remains sufficiently small as the sample sizes increase, so that for a solution to be included in $\mathcal{S}_{\mathsf{n}}^{\mathrm{PS}}$, the restriction of the best-fitting model to the solutions in the experimental set must more closely align with the observed sample means.

Theorems 1 and 2 establish that, under Conditions (C2)–(C4), $\mathcal{S}_{\mathsf{n}}^{\mathrm{PS}}$ possesses the desired properties of confidence and $\mathsf{S}(\mathsf{X})$ consistency; proofs appear in Appendix C.

Theorem 1 *If $d_{\mathsf{n}}(\cdot, \widehat{\mu}, \widehat{\Sigma})$ and D satisfy Conditions (C2) and (C3), then $\mathcal{S}_{\mathsf{n}}^{\mathrm{PS}}$ achieves finite-sample confidence and asymptotic confidence.*

Theorem 2 *If $d_{\mathsf{n}}(\cdot, \widehat{\mu}, \widehat{\Sigma})$ and D satisfy Condition (C4), then $\mathcal{S}_{\mathsf{n}}^{\mathrm{PS}}$ achieves $\mathsf{S}(\mathsf{X})$ consistency.*

4.3 Standardized Discrepancies

Our screening framework can easily accommodate different choices of standardized discrepancies and cutoffs. We present several examples that satisfy Conditions (C1)–(C4) and provide a representative proof in Appendix C. Condition (C2) is established under a normality assumption: in Setting (S1), $Y_\ell(x_i) \sim \mathcal{N}(\mu(x_i), \Sigma(x_i, x_i))$ for all $\ell = 1, 2, \ldots, n_i$ and $i = 1, \ldots, k$, and in Setting (S2), $\mathbf{Y}_\ell \sim \mathcal{N}(\mu(\mathsf{X}), \Sigma(\mathsf{X}))$ for all $\ell = 1, 2, \ldots, n$.

In Setting (S1), Conditions (C1)–(C4) are satisfied by

$$d_{\mathsf{n}}^1(\mathsf{m}, \widehat{\mu}, \widehat{\Sigma}) \equiv \sum_{i=1}^{k} \frac{\sqrt{n_i}}{\widehat{\sigma}_i} |\widehat{\mu}_i - \mathsf{m}_i|$$

with D^1 defined as the $1-\alpha$ quantile of the sum of the absolute value of k independent t-distributed random variables, each with degrees of freedom $n_1 - 1, n_2 - 1, \ldots, n_k - 1$, respectively; by

$$d_{\mathsf{n}}^2(\mathsf{m}, \widehat{\mu}, \widehat{\Sigma}) \equiv \sum_{i=1}^{k} \frac{n_i}{\widehat{\sigma}_i^2} (\widehat{\mu}_i - \mathsf{m}_i)^2$$

with D^2 defined as the $1-\alpha$ quantile of the sum of k independent F-distributed random variables, each with numerator degrees of freedom 1 and denominator degrees of freedom $n_1 - 1, n_2 - 1, \ldots, n_k - 1$, respectively; and by

$$d_{\mathsf{n}}^\infty(\mathsf{m}, \widehat{\mu}, \widehat{\Sigma}) \equiv \max_{i=1,\ldots,k} \frac{\sqrt{n_i}}{\widehat{\sigma}_i} |\widehat{\mu}_i - \mathsf{m}_i|$$

with D^∞ defined as the $1-\alpha$ quantile of the maximum of the absolute value of k independent t-distributed random variables, each with $n_i - 1$ degrees of freedom. In our discussion, we find it convenient to refer to these standardized discrepancies by the shorthand d_{n}^1, d_{n}^2, and d_{n}^∞. Plumlee and Nelson [2018] focused on the choice of d_{n}^2 and D^2 and Eckman et al. [2020] explored connections to existing screening methods, such as the Screen-to-the-Best procedure [Nelson et al., 2001].

In Setting (S2), Conditions (C1)–(C4) are satisfied by

$$d_{\mathsf{n}}^{\mathrm{CRN}}(\mathsf{m}, \widehat{\mu}, \widehat{\Sigma}) \equiv n(\widehat{\mu} - \mathsf{m})^\top \widehat{\Sigma}^{-1} (\widehat{\mu} - \mathsf{m})$$

with $\mathsf{D}^{\mathrm{CRN}}$ defined as $k(n-1)/(n-k)$ times the $1-\alpha$ quantile of an F-distributed random variable with numerator degrees of freedom k and denominator degrees of freedom $n-k$. For $\widehat{\Sigma}$ to be invertible in Setting (S2), a minimum of $k+1$ replications must be obtained from each solution, i.e., $n \geq k+1$.

As can be seen from these examples, a uniform cutoff D can be specified as the $1-\alpha$ quantile of the pivotal statistic $d_{\mathsf{n}}(\mu(\mathsf{X}), \widehat{\mu}, \widehat{\Sigma})$. The given cutoffs D^1, D^2, D^∞, and $\mathsf{D}^{\mathrm{CRN}}$ are the tightest uniform cutoffs for their respective standardized discrepancies that deliver finite-sample confidence irrespective of the properties of μ. To see this, consider the case in which the decision-maker has complete knowledge of the performances of the solutions in the experimental set, i.e., $\mathsf{M}(x_0) = \{\mu(\mathsf{X})\}$. Thus $D_{\mathsf{n}}(x_0, \widehat{\mu}, \widehat{\Sigma}) = d_{\mathsf{n}}(\mu(\mathsf{X}), \widehat{\mu}, \widehat{\Sigma})$ and the specified cutoffs are exactly the $1-\alpha$ quantiles of the minimum standardized discrepancies. The coverage of any acceptable solution x_0 is therefore exactly $1-\alpha$.

5 Computational Considerations and Relaxed Screening

Constructing $\mathcal{S}_{\mathsf{n}}^{\mathrm{PS}}$ entails repeatedly solving the optimization problem described in Definition (1) and comparing its optimal value, $D_{\mathsf{n}}(x_0, \widehat{\mu}, \widehat{\Sigma})$, to the cutoff D for each $x_0 \in \mathcal{X}$. Depending on the difficulty of the optimization problem and the number of candidate solutions, constructing $\mathcal{S}_{\mathsf{n}}^{\mathrm{PS}}$ in this manner could be computationally expensive. In this section, we present an alternative subset consisting of solutions for which $\widehat{\mu}$ belongs to a polyhedral relaxation of $\mathsf{R}(x_0)$. Screening a solution therefore involves solving a linear program which, in certain cases, can be substantially cheaper. Compared to the subset $\mathcal{S}_{\mathsf{n}}^{\mathrm{PS}}$, this approach results in a more conservative subset in the sense that it contains all of the solutions in $\mathcal{S}_{\mathsf{n}}^{\mathrm{PS}}$ and possibly more.

5.1 Polyhedral Relaxation of $\mathsf{R}(x_0)$ via a Relaxation of $\mathsf{M}(x_0)$

We demonstrate this conservative approach for the situation in which $\mathsf{M}(x_0)$ can be described as the projection of a polyhedron.

Assumption 1 *For each solution $x_0 \in \mathcal{X}$,*

$$\mathsf{M}(x_0) = \left\{\mathsf{m} \in \mathbb{R}^k \colon \textit{there exists } \mathsf{w} \in \mathbb{R}^q \textit{ such that } A\mathsf{m} + C\mathsf{w} \leq b\right\},$$

for some $A \in \mathbb{R}^{p \times k}$, $C \in \mathbb{R}^{p \times q}$, $b \in \mathbb{R}^p$, where A, C, and b may depend on x_0 and X.

For A, C, and b in Assumption 1, we suppress x_0 and X for notational convenience.

Assumption 1 depends on both the choice of function space and the definition of the set of acceptable solutions. This assumption holds for most combinations discussed in this paper, e.g., finding the global minima of a convex or Lipschitz continuous function. In Example 1, for instance, $\mathsf{M}(x_0)$ was explicitly expressed as a polyhedron in $\mathbb{R}^k$. In Example 2 below, we demonstrate that Assumption 1 also holds for the convex case; see Appendix A for a complete derivation.

Example 2 (Optimizing a Convex Function) *For the problem of optimizing a convex function, one formulation of $\mathsf{M}(x_0)$ is*

$$\begin{aligned}\mathsf{M}(x_0) = \{\mathsf{m} \in \mathbb{R}^k \colon \textit{there exists } \mathsf{m}_0 \in \mathbb{R} \textit{ and } \xi_1, \ldots, \xi_k \in \mathbb{R}^d \textit{ such that}\\ \mathsf{m}_i - \mathsf{m}_j - (x_i - x_j)^\top \xi_i \leq 0 \textit{ for all } i, j = 1, \ldots, k\\ \mathsf{m}_i - \mathsf{m}_0 - (x_i - x_0)^\top \xi_i \leq 0 \textit{ for all } i = 1, \ldots, k\\ -\mathsf{m}_i + \mathsf{m}_0 \leq 0 \textit{ for all } i = 1, \ldots, k\}.\end{aligned}$$

From this representation of $\mathsf{M}(x_0)$, it is easy to identify A, C, and b as defined in Assumption 1. The components of $\mathsf{w} = (\mathsf{m}_0, \xi_1^\top, \ldots, \xi_k^\top)^\top$ represent the performance of solution x_0 and subgradients at solutions $x_1, \ldots, x_k$.

To explain our approach, define the polyhedron $\mathsf{P} \equiv \{(\mathsf{m}, \mathsf{w}) \in \mathbb{R}^k \times \mathbb{R}^q \colon A\mathsf{m} + C\mathsf{w} \leq b\}$, such that the projection of P onto $\mathbb{R}^k$ is $\mathsf{M}(x_0)$. Our approach is to relax P by increasing its right-hand-side vector b and project the enlarged polyhedron onto $\mathbb{R}^k$. To compensate for the uncertainty about $\mu(\mathsf{X})$, we offset b by defining

$$b'_j = b_j + \max_{\mathsf{m} \in \mathbb{R}^k} \left\{ a_j^\top (\widehat{\mu} - \mathsf{m}) \colon d_{\mathsf{n}}(\mathsf{m}, \widehat{\mu}, \widehat{\Sigma}) \leq \mathsf{D} \right\} \text{ for all } j = 1, \ldots, p,$$

where a_j is the jth row of A, expressed as a column vector. Condition (C1) implies that for any $a \in \mathbb{R}^k$, $\max_{\mathsf{m} \in \mathbb{R}^k} \left\{ a^\top (\widehat{\mu} - \mathsf{m}) \colon d_{\mathsf{n}}(\mathsf{m}, \widehat{\mu}, \widehat{\Sigma}) \leq \mathsf{D} \right\} \geq 0$; thus $b'_j \geq b_j$ with probability one for all $j = 1, \ldots, p$.

For the four standardized discrepancies outlined in Section 4.2,

$$\max_{\mathsf{m} \in \mathbb{R}^k} \left\{ a_j^\top (\widehat{\mu} - \mathsf{m}) \colon d^1_{\mathsf{n}}(\mathsf{m}, \widehat{\mu}, \widehat{\Sigma}) \leq \mathsf{D}^1 \right\} = \mathsf{D}^1 \max_{i=1,\ldots,k} \frac{\widehat{\sigma}_i}{\sqrt{n_i}} |a_{ji}|,$$

$$\max_{\mathsf{m} \in \mathbb{R}^k} \left\{ a_j^\top (\widehat{\mu} - \mathsf{m}) \colon d^2_{\mathsf{n}}(\mathsf{m}, \widehat{\mu}, \widehat{\Sigma}) \leq \mathsf{D}^2 \right\} = \sqrt{\mathsf{D}^2 \sum_{i=1}^k \frac{\widehat{\sigma}_i^2}{n_i} a_{ji}^2},$$

$$\max_{\mathsf{m} \in \mathbb{R}^k} \left\{ a_j^\top (\widehat{\mu} - \mathsf{m}) \colon d^\infty_{\mathsf{n}}(\mathsf{m}, \widehat{\mu}, \widehat{\Sigma}) \leq \mathsf{D}^\infty \right\} = \mathsf{D}^\infty \sum_{i=1}^k \frac{\widehat{\sigma}_i}{\sqrt{n_i}} |a_{ji}|, \text{ and}$$

$$\max_{\mathsf{m} \in \mathbb{R}^k} \left\{ a_j^\top (\widehat{\mu} - \mathsf{m}) \colon d^{\mathrm{CRN}}_{\mathsf{n}}(\mathsf{m}, \widehat{\mu}, \widehat{\Sigma}) \leq \mathsf{D}^{\mathrm{CRN}} \right\} = \sqrt{\frac{\mathsf{D}^{\mathrm{CRN}}}{n} a_j^\top \widehat{\Sigma} a_j},$$

for all $j = 1, \ldots, p$; derivations appear in Appendix B. In the case of $d^{\mathrm{CRN}}_{\mathsf{n}}$ and $\mathsf{D}^{\mathrm{CRN}}$, adjusting the right-hand-side vector in this way follows the approach of Anderson [1984] for constructing simultaneous confidence intervals for linear combinations of the components of $\mu(\mathsf{X})$; see Equation (15) on pages 166–167 therein. From the expressions above, it is apparent that $b' \equiv (b'_1, \ldots, b'_p)^\top$ is a random vector whose components are functions of $\widehat{\Sigma}$ and n, but not $\widehat{\mu}$.

5.2 Relaxed Plausible Screening

The projection of the relaxation of P above onto $\mathbb{R}^k$ is given by

$$\mathsf{R}'(x_0) \equiv \left\{ \mathsf{m} \in \mathbb{R}^k \colon \text{there exists } \mathsf{w} \in \mathbb{R}^q \text{ such that } A\mathsf{m} + C\mathsf{w} \leq b' \right\}.$$

The polyhedron $\mathsf{R}'(x_0)$ is a random relaxation of $\mathsf{M}(x_0)$, and Lemma 1 further shows that it is also a relaxation of $\mathsf{R}(x_0)$.

Lemma 1 *If Assumption 1 holds, then* $\mathsf{R}(x_0) \subseteq \mathsf{R}'(x_0)$ *with probability one for all* $x_0 \in \mathcal{X}$.

Our more conservative screening method, which we refer to as Relaxed Plausible Screening (RPS), returns a subset $\mathcal{S}^{\mathrm{RPS}}_{\mathsf{n}}$ defined as

$$\mathcal{S}^{\mathrm{RPS}}_{\mathsf{n}} \equiv \{x_0 \in \mathcal{X} \colon \widehat{\mu} \in \mathsf{R}'(x_0)\},$$

the conservatism of which is is made clear in Corollary 1.

Corollary 1 *If Assumption 1 holds, then* $\mathcal{S}_{\mathsf{n}}^{\mathrm{PS}} \subseteq \mathcal{S}_{\mathsf{n}}^{\mathrm{RPS}}$ *with probability one.*

Theorems 3 and 4 establish that, under Conditions (C2)–(C4), $\mathcal{S}_{\mathsf{n}}^{\mathrm{RPS}}$ possesses the desired properties of confidence and $\mathsf{S}(\mathsf{X})$ consistency.

Theorem 3 *If* $d_{\mathsf{n}}(\cdot, \widehat{\mu}, \widehat{\Sigma})$ *and* D *satisfy Conditions (C2) and (C3), then* $\mathcal{S}_{\mathsf{n}}^{\mathrm{RPS}}$ *achieves finite-sample confidence and asymptotic confidence.*

Theorem 4 *If* $d_{\mathsf{n}}(\cdot, \widehat{\mu}, \widehat{\Sigma})$ *and* D *satisfy Condition (C4), then* $\mathcal{S}_{\mathsf{n}}^{\mathrm{RPS}}$ *achieves* $\mathsf{S}(\mathsf{X})$ *consistency.*

The relaxation $\mathsf{R}'(x_0)$ that is used to construct $\mathcal{S}_{\mathsf{n}}^{\mathrm{RPS}}$ depends on the representation of $\mathsf{M}(x_0)$ in Assumption 1. Hence a different representation of $\mathsf{M}(x_0)$—meaning a different choice of A, C, and b—can result in a different relaxation $\mathsf{R}'(x_0)$ and thus different solutions being included in $\mathcal{S}_{\mathsf{n}}^{\mathrm{RPS}}$. The extreme case of this would be the elimination of C altogether, as shown in Theorem 5. This result demonstrates that the representation $\mathsf{M}(x_0) = \left\{\mathsf{m} \in \mathbb{R}^k \colon \bar{A}\mathsf{m} \leq \bar{b}\right\}$ for some $\bar{A} \in \mathbb{R}^{\bar{p} \times k}$ and $\bar{b} \in \mathbb{R}^{\bar{p}}$ yields a tighter polyhedral relaxation of $\mathsf{R}(x_0)$ and thus a smaller subset.

Theorem 5 *Suppose Assumption 1 holds and that for a fixed* $x_0 \in \mathcal{X}$*,*

$$\mathsf{M}(x_0) = \left\{\mathsf{m} \in \mathbb{R}^k \colon \textit{there exists } \mathsf{w} \in \mathbb{R}^q \textit{ such that } A\mathsf{m} + C\mathsf{w} \leq b\right\} = \left\{\mathsf{m} \in \mathbb{R}^k \colon \bar{A}\mathsf{m} \leq \bar{b}\right\},$$

for some $A \in \mathbb{R}^{p \times k}$*,* $C \in \mathbb{R}^{p \times q}$*,* $b \in \mathbb{R}^p$*,* $\bar{A} \in \mathbb{R}^{\bar{p} \times k}$*, and* $\bar{b} \in \mathbb{R}^{\bar{p}}$*. Then for any* $\mu \in \mathscr{M}$*,*

$$\bar{\mathsf{R}}'(x_0) \equiv \left\{\mathsf{m} \in \mathbb{R}^k \colon \bar{A}\mathsf{m} \leq \bar{b}'\right\} \subseteq \mathsf{R}'(x_0) \textit{ with probability one,}$$

where

$$\bar{b}'_j = \bar{b}_j + \max_{\mathsf{m} \in \mathbb{R}^k} \left\{\bar{a}_j^\top (\widehat{\mu} - \mathsf{m}) \colon d_{\mathsf{n}}(\mathsf{m}, \widehat{\mu}, \widehat{\Sigma}) \leq \mathsf{D}\right\} \textit{ for all } j = 1, \ldots, \bar{p}.$$

In some cases, e.g., optimizing a Lipschitz continuous function, deriving an explicit polyhedral representation of $\mathsf{M}(x_0)$ is relatively straightforward, while in other cases, e.g., optimizing a convex function, it is challenging. Projecting out some or all components of w has the potential to yield a less conservative subset $\mathcal{S}_{\mathsf{n}}^{\mathrm{RPS}}$, but can come at the cost of an increase in the number of constraints implicitly describing $\mathsf{M}(x_0)$. While classical techniques for eliminating variables, e.g., Fourier-Motzkin elimination, can cause an explosion in the number of constraints, many of them redundant, recent advances are more promising [Jing et al., 2018].

Remark 1 *Both* $\mathcal{S}_{\mathsf{n}}^{\mathrm{PS}}$ *and* $\mathcal{S}_{\mathsf{n}}^{\mathrm{RPS}}$ *exhibit an appealing, intuitive trait: given the same observed simulation outputs, knowing additional functional properties of* μ *leads to a smaller subset. That is, adding constraints that further shrink* $\mathsf{M}(x_0)$ *results in more solutions being screened out. This assertion is made mathematically precise in Theorems 6 and 7, which appear in Appendix C.*

Table 1: Properties of the Plausible Screening and Relaxed Plausible Screening optimization problems; k is the number of solutions in X, while p and q are the number of constraints and extra variables in the description of $\mathsf{M}(x_0)$ as in Assumption 1.

Subset	Discrepancy	Linear/Quadratic	# Decision Variables	# Constraints
$\mathcal{S}_{\mathsf{n}}^{\mathrm{PS}}$	d_{n}^{1}	Linear	$2k+q$	$p+2k$
	d_{n}^{2}	Quadratic	$k+q$	p
	d_{n}^{∞}	Linear	$k+q+1$	$p+2k$
	$d_{\mathsf{n}}^{\mathrm{CRN}}$	Quadratic	$k+q$	p
$\mathcal{S}_{\mathsf{n}}^{\mathrm{RPS}}$	All	Linear	$q+1$	p

5.3 Optimization Problems

Checking whether $\widehat{\mu} \in \mathsf{R}'(x_0)$ amounts to checking the feasibility of a system of linear equations—namely, does there exist a $\mathsf{w} \in \mathbb{R}^q$ such that $C\mathsf{w} \leq b' - A\widehat{\mu}$? This is equivalent to determining the sign of the optimal value of a related linear program:

$$z_{\mathsf{n}} \equiv \max_{\mathsf{w},\eta} \eta \text{ s.t. } C\mathsf{w} + \eta \mathbf{1}_p \leq b' - A\widehat{\mu}, \tag{2}$$

where $\mathbf{1}_p$ is a p-vector of ones. The notation z_{n} reflects the dependence of the parameters of the optimization problem on the sample sizes; it is also convenient in the proofs of the asymptotic guarantees delivered by $\mathcal{S}_{\mathsf{n}}^{\mathrm{RPS}}$. If $z_{\mathsf{n}} \geq 0$, the solution x_0 is included in $\mathcal{S}_{\mathsf{n}}^{\mathrm{RPS}}$, otherwise it is excluded.

On the other hand, constructing $\mathcal{S}_{\mathsf{n}}^{\mathrm{PS}}$ requires evaluating $D_{\mathsf{n}}(x_0, \widehat{\mu}, \widehat{\Sigma}) \equiv \min_{(\mathsf{m},\mathsf{w})\in\mathsf{P}} d_{\mathsf{n}}(\mathsf{m}, \widehat{\mu}, \widehat{\Sigma})$. Definition (2) therefore reduces the number of decision variables by roughly k, relative to optimizing over P. If $\mathsf{M}(x_0)$ can be expressed as a projection with few extra variables (small q), then solving the problem in Definition (2) may be appreciably faster than solving $\min_{(\mathsf{m},\mathsf{w})\in\mathsf{P}} d_{\mathsf{n}}(\mathsf{m}, \widehat{\mu}, \widehat{\Sigma})$, with greater savings as the size of the experimental set increases. Furthermore, if a large number of solutions are to be screened, the computational savings from working with $\mathcal{S}_{\mathsf{n}}^{\mathrm{RPS}}$ can be substantial. Table 1 summarizes properties of the optimization problems associated with screening solutions via the PS and RPS methods for the four standardized discrepancies.

Remark 2 *Example 1 featured an explicit polyhedral representation of* $\mathsf{M}(x_0)$*, i.e.,* $q = 0$*. Thus, for Lipschitz performance functions,* $\mathcal{S}_{\mathsf{n}}^{\mathrm{RPS}}$ *can be constructed without optimization by simply checking whether* $A\widehat{\mu} \leq b'$ *for each solution.*

6 Numerical Experiments

To illuminate the theoretical developments thus far in a more practical light, we implemented the PS and RPS approaches on two problems. Both examples illustrate how prior knowledge of functional properties can assist in screening out swathes of unacceptable solutions using only a limited experimental set. Our first example in Section 6.1 illustrates how the behavior of PS varies depending on the standardized discrepancy and demonstrates the advantages over subset-selection procedures. In a much larger example described in Section 6.2, PS and

RPS screen out hundreds of thousands of solutions using an experimental set consisting of only a hundred solutions.

We implemented our methods in MATLAB using the software's built-in optimization algorithms with their default settings: `linprog` (dual-simplex method) for linear programs and `quadprog` (interior-point method) for quadratic programs. Source code is available at `https://github.com/daveckman/plausible-screening`. We ran our experiments on a high-performance computing cluster using eight cores on a compute node with 256GB of RAM. For the first example, we ran independent macroreplications of our methods in parallel to study the differences between methods, while for the larger second example, we classified solutions in parallel to mirror a reasonable implementation in practice.

6.1 Newsvendor Problem

The first problem is a modified version of the classical newsvendor problem [Porteus, 1990]. Here, a vendor orders inventory of a given product in discrete quantities at a per-unit order cost c_{order}, observes a realization of stochastic demand ξ for a continuous quantity of the product, and sells it at a per-unit sales price p_{sales}. For example, consider a gas-station operator who orders gasoline in truckloads, but sells it in continuous quantities at the pump. At the end of the sales period, leftover inventory is salvaged at a per-unit price $p_{salvage}$ and unmet demand incurs a fixed per-unit cost of $c_{shortage}$.

The vendor's objective is to determine the order quantity that maximizes the expected profit or, equivalently, minimizes the expected loss over the following sales period. For a fixed realization of demand, ξ, the loss associated with an order quantity x is given by

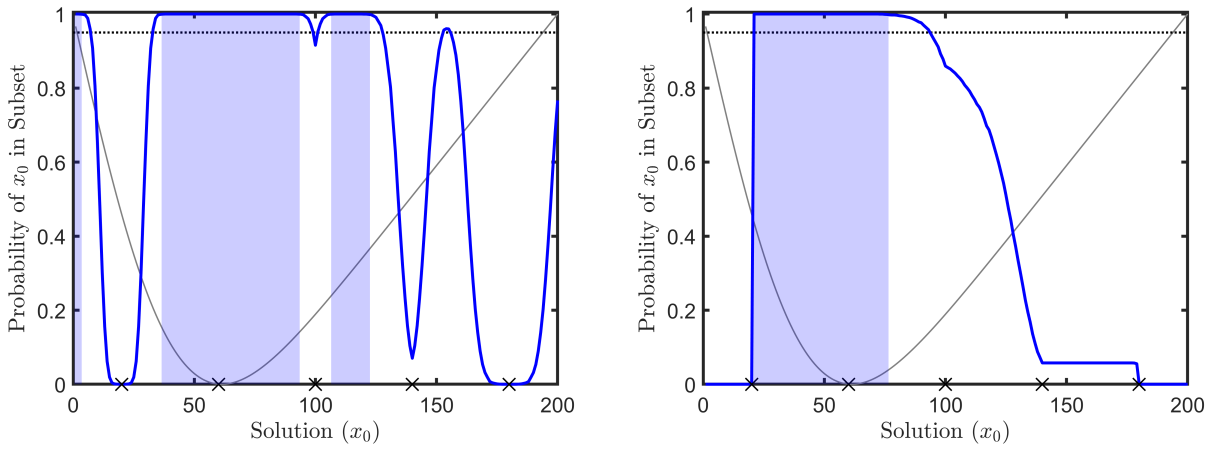
$$Y(x,\xi) = c_{order}x - p_{sales}\min\{\xi, x\} - p_{salvage}\max\{x-\xi, 0\} + c_{shortage}\max\{\xi - x, 0\}. \quad (3)$$

The sample-path function $Y(\cdot,\xi)$ is convex in x provided $p_{sales} \geq p_{salvage}$. Furthermore, $Y(\cdot,\xi)$ is γ-Lipschitz continuous with constant $\gamma = \max\{p_{sales} + c_{shortage} - c_{order}, c_{order} - p_{salvage}\}$. The expected loss function $\mu(x) := \mathbb{E}_\xi[Y(x,\xi)]$ inherits these properties from the sample-path functions, as discussed in Section 3.1. In our experiments, we set $c_{order} = 3$, $p_{sales} = 9$, $p_{salvage} = 1$, and $c_{shortage} = 1$ with ξ being Weibull distributed with scale parameter 50 and shape parameter 2.

We considered a feasible region $\mathcal{X} = \{1, 2, \ldots, 200\}$ and tested our methods by simulating at 5 evenly spaced solutions (20, 60, 100, 140, 180) with a total sample size of 400 replications. Though not presented in this article, we varied the experimental set and arrived at similar conclusions as the ones presented in this article. We tested the PS method with the d_{n}^1, d_{n}^2, and d_{n}^∞ standardized discrepancies and $1-\alpha = 0.95$ when either exploiting the properties that μ is convex or Lipschitz continuous with $\gamma = 7$. As a benchmark we applied the Screen-to-the-Best (STB) subset-selection procedure of Nelson et al. [2001], which takes an equal number of i.i.d. replications from all solutions in $\mathcal{X}$ and achieves both finite-sample confidence (under the normality assumption) and asymptotic confidence. Given the same total sample size of 400, the STB procedure took two replications at each feasible solution. We ran 3000 macroreplications of each procedure.

With a small total sample size spread thinly over the solution space, the STB procedure struggled to eliminate solutions, failing to screen out any solutions on 93.8% of the

macroreplications and never screening out more than four solutions. In addition, each feasible solution was retained on at least 99.5% of the macroreplications. Figure 1 shows the empirical probability that individual solutions were included in $\mathcal{S}_{\mathsf{n}}^{\mathrm{PS}}$ for d_{n}^2; curves for d_{n}^1 and d_{n}^∞ were similar. In both the Lipschitz and convex cases, our method retained the optimal solution, $x^* = 61$, on all macroreplications, indicating conservatism. The two instances of functional properties led to interesting features in the geometry of the retained solutions. In the Lipschitz case, PS screened out solutions near clearly suboptimal solutions in the experimental set, namely $x = 40$, $x = 140$, and $x = 180$, while in the convex case it screened out those on the periphery of the feasible region. Because the probability of being in $\mathcal{S}_{\mathsf{n}}^{\mathrm{PS}}$ is neither 0 nor 1 for many solutions, the composition of $\mathcal{S}_{\mathsf{n}}^{\mathrm{PS}}$ varied from macroreplication to macroreplication even with 80 replications taken at each solution in X. The subset $\mathcal{S}_{\mathsf{n}}^{\mathrm{PS}}$ also differed from $\mathsf{S}(\mathsf{X})$, the subset of solutions that would be returned by an oracle who can observe $\mu(\mathsf{X})$ without simulation error, implying that more solutions could be screened out if the number of replications were increased.

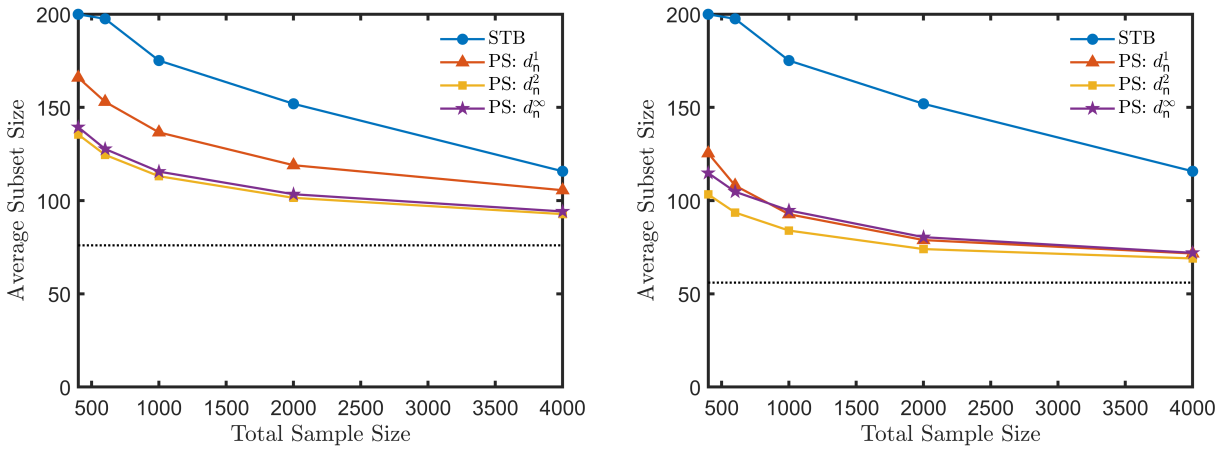


(a) Lipschitz (b) Convex

Figure 1: Empirical probability of including individual solutions in $\mathcal{S}_{\mathsf{n}}^{\mathrm{PS}}$ for the d_{n}^2 standardized discrepancy with 80 replications taken at $k = 5$ equally spaced solutions when separately using knowledge that the objective function is Lipschitz continuous or convex. The thin gray line depicts the (shifted and scaled) objective function, the black dotted line indicates the desired coverage of $1 - \alpha = 0.95$, the black Xs indicate the solutions in the experimental set, and the shaded blue regions indicate the solutions in $\mathsf{S}(\mathsf{X})$.

We also varied the total sample size, testing budgets of 400, 600, 1000, 2000, and 4000 replications. Figure 2 shows the average subset sizes for the four procedures when fixing $k = 5$ and increasing the total sample size. All methods returned smaller subsets on average when taking more samples, with STB reducing the gap relative to PS. This is a consequence of the limited inference PS can make, having simulated only a fixed experimental set. Specifically, as the total sample size increases, $\mathcal{S}_{\mathsf{n}}^{\mathrm{PS}}$ achieves $\mathsf{S}(\mathsf{X})$ consistency—the cardinality of which is shown in Figure 2—while STB will eventually screen out all strictly suboptimal solutions. Figure 2 demonstrates that knowing μ is convex leads to more powerful screening than knowing a universal Lipschitz constant. In both cases, PS screened out anywhere from

15–65% of the feasible solutions on average while simulating only 2.5% of them, and more solutions could be screened out if the decision-maker were willing to accept more risk as represented by the nominal confidence level. A separate analysis measuring the average average-optimality gap of the solutions in the returned subsets yielded the same conclusions.

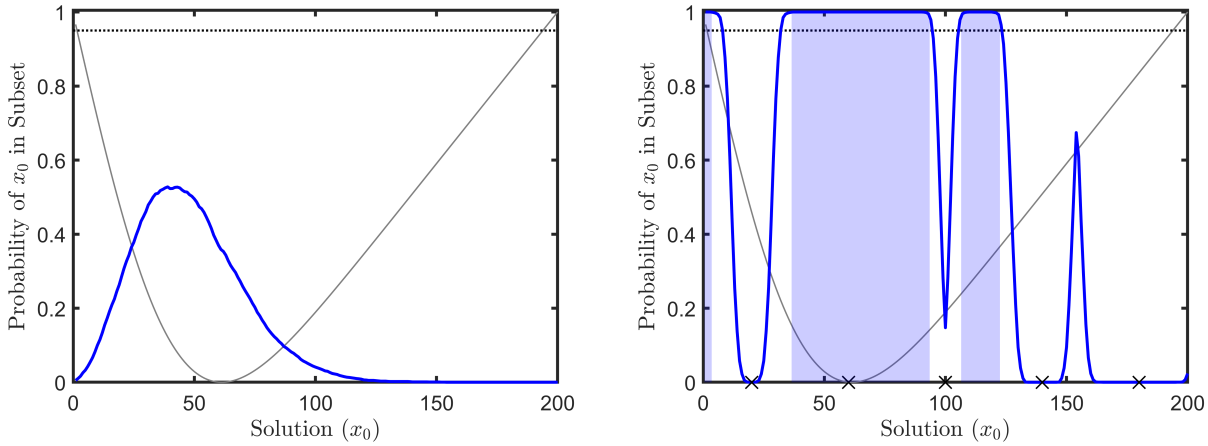


(a) Lipschitz (b) Convex

Figure 2: Average subset sizes for the STB procedure and PS with the d_{n}^1, d_{n}^2, and d_{n}^∞ standardized discrepancies for $k = 5$ and different total sample sizes. The black dotted line indicates the cardinality of $\mathsf{S}(\mathsf{X})$. All average sample sizes are individually precise to within ± 1 with 95% confidence.

We also compared PS with the $d_{\mathsf{n}}^{\mathrm{CRN}}$ standardized discrepancy to a version of the STB procedure that accommodates the use of CRN; see Section 3 of Nelson et al. [2001] for STB details. We again took $k = 5$ with a total sample size of 400 replications and each procedure generated its replications using CRN across solutions. Figure 3 shows the empirical probability that individual solutions were included in the returned subset for STB with CRN and PS, when exploiting knowledge that μ is Lipschitz continuous. The STB procedure with CRN was more liberal in screening out solutions—returning an average subset of size 28—but severely undercovered the optimal solution, retaining it on only 36% of the macroreplications. This behavior is a consequence of the severe nonnormality of the outputs and the use of CRN with a small sample size per solution. To be precise, the STB procedure with CRN obtains two sample-path functions $Y(\cdot, \xi_1)$ and $Y(\cdot, \xi_2)$ and performs pairwise comparisons based on the variance of $Y(x, \xi_1) - Y(x', \xi_1)$ and $Y(x, \xi_2) - Y(x', \xi_2)$ for solutions $x, x' \in \mathcal{X}$. From Equation (3), it can be seen that for $x, x' \notin [\min\{\xi_1, \xi_2\}, \max\{\xi_1, \xi_2\}]$, the variance of the two differences is zero, implying that any solution $x_0 \notin [\min\{\xi_1, \xi_2\}, \max\{\xi_1, \xi_2\}]$ will be screened out. Since the mode of the Weibull distribution from which ξ_1 and ξ_2 is generated is about 35.4, the STB subsets are biased to the left of $x^* = 61$.

PS with $d_{\mathsf{n}}^{\mathrm{CRN}}$ screened out similar solutions to its counterparts that use independent sampling, but returned somewhat smaller subsets with an average size of 104 solutions. (In the convex case, PS similarly returned smaller subsets when using CRN, with an average size of 70.) This additional screening power should be weighed against the increased difficulty of the underlying optimization problems, i.e., the need to solve quadratic programs with dense



(a) STB with CRN (b) PS with $d_{\mathsf{n}}^{\mathrm{CRN}}$

Figure 3: Empirical probability of including individual solutions in the STB subset and $\mathcal{S}_{\mathsf{n}}^{\mathrm{PS}}$ with the $d_{\mathsf{n}}^{\mathrm{CRN}}$ standardized discrepancy with 80 replications taken at $k = 5$ equally spaced solutions when using knowledge that the objective function is Lipschitz continuous.

Hessian matrices.

6.2 Tandem Production Line Problem

The second problem is a resource-allocation problem for a production line with manufacturing blocking (e.g., buffers), adapted from Plambeck et al. [1996]. The decision-maker is tasked with allocating discrete resources across five single-server stations arranged in a tandem (serial) configuration. Each station processes products using a first-in first-out service discipline. If Station i is allocated a_i resources, its cycle (processing) time for a given product is assumed to be exponentially distributed with rate parameter $\rho_i = \bar{\rho}_i(1 + a_i)$, where $\bar{\rho}_i$ is a base processing rate. We set $\bar{\rho}_1 = 3$, $\bar{\rho}_2 = 5$, $\bar{\rho}_3 = 2$, $\bar{\rho}_4 = 5$, and $\bar{\rho}_5 = 1$.

There is a buffer in front of each machine for products awaiting processing. If the buffer is full, upstream stations can become blocked, whereas if it is empty, downstream stations can become starved. We assume that there is an infinite supply of products immediately available to process at Station 1 and an infinite-capacity buffer in front of that station, i.e., there is no external arrival process. The buffer capacities in front of Stations 2–5 are fixed at 4, 6, 8, and 4, respectively.

The decision-maker's objective is to allocate 50 resources to minimize the expected completion time of the 100th product. Under the assumptions above, the objective function is convex in the allocation $x \equiv (a_1, a_2, a_3, a_4, a_5)$; see Section IV.B of Shanthikumar and Yao [1989] for a complete derivation. We restrict attention to solutions that allocate all available resources, i.e., $a_1 + a_2 + a_3 + a_4 + a_5 = 50$, resulting in a total of $316,251$ feasible solutions. Because of this tight constraint, the feasible region can be reduced to a four-dimensional space.

Shanthikumar and Yao [1989] provide dynamic recursion equations for simulating the completion times of all products, thereby avoiding the need to run a full discrete-event sim-

ulation of the system. Even so, we consider this problem to be representative of large-scale simulation-optimization problems for which simulating all feasible solutions is impractical, but properties of the objective function may be known. In such cases, the available computational budget may permit only a small fraction of feasible solutions to be simulated. We fixed a total sample size of 10,000 replications, which is enough to simulate one replication from about 3% of the feasible solutions. We ran a single macroreplication of the PS and RPS methods with d_{n}^1, d_{n}^2, and d_{n}^∞. An experimental set consisting of $k = 100$ reasonably space-filling solutions was determined via k-means clustering, hence 100 replications were generated at each solution in X. For our formulation of convexity, the underlying optimization problems for PS and RPS featured about 500 decision variables and 10,000 constraints.

Screening and timing results for each method are given in Table 2. All three versions of PS screened out more than 60% of the feasible solutions while simulating only 0.03% of them. The efficacy of RPS varied depending on the standardized discrepancy. For d_{n}^∞, the same subset of solutions was returned by PS and RPS (i.e., $\mathcal{S}_{\mathsf{n}}^{\mathrm{PS}} = \mathcal{S}_{\mathsf{n}}^{\mathrm{RPS}}$), yet for d_{n}^1, no solutions were screened out by RPS.

Table 2: Times and subset sizes for a single macroreplication on the tandem production line problem.

Method and Discrepancy	Time per Solution (s)	$\lvert\mathcal{S}_{\mathsf{n}}^{\mathrm{PS}}\rvert$ / $\lvert\mathcal{S}_{\mathsf{n}}^{\mathrm{RPS}}\rvert$	Fraction Screened
PS with d_{n}^1 / RPS with d_{n}^1	0.08 / 0.07	123,904 / 316,251	60.8% / 0%
PS with d_{n}^2 / RPS with d_{n}^2	1.63 / 0.09	69,198 / 83,748	78.1% / 73.5%
PS with d_{n}^∞ / RPS with d_{n}^∞	0.40 / 0.09	61,897 / 61,897	80.4% / 80.4%

Remark 3 *In all of our experiments for PS and RPS with d_{n}^∞, we observed that on all macroreplications, $\mathcal{S}_{\mathsf{n}}^{\mathrm{PS}} = \mathcal{S}_{\mathsf{n}}^{\mathrm{RPS}}$ for both the Lipschitz and convex cases. Theorem 8 in Appendix D formalizes this observation and proves that it holds with probability one for the Lipschitz case. We were unable to prove an analogous result for the convex case.*

All together, the results in Table 2 illustrate the diverse performance of the various methods. PS with d_{n}^2, which required the solution of quadratic programs, was the most computationally intensive procedure. At the other extreme, PS with d_{n}^1 was roughly 20 times faster, but retained about twice as many solutions. The most effective and efficient procedure was RPS with d_{n}^∞; it removed over 80% of the feasible solutions with an overall run time of about 8 core hours. As a practical recommendation, for either d_{n}^2 or d_{n}^∞, the faster RPS method can be run first, followed by the PS method on the solutions in the returned subset $\mathcal{S}_{\mathsf{n}}^{\mathrm{RPS}}$. This approach notably does not requiring splitting α to preserve the statistical guarantee, as is sometimes the case with multi-stage selection procedures [Nelson et al., 2001].

Figure 4 shows the sorted minimum standardized discrepancies of the feasible solutions, $D_{\mathsf{n}}(x_0, \widehat{\mu}, \widehat{\Sigma})$, relative to the cutoff, D, for the three versions of PS. The minimum standardized discrepancies were divided by the cutoffs and log-transformed to produce a clear, standardized comparison. The flat stretches on the left-hand side of Figure 4 correspond to solutions x_0 for which there exists an x_0-optimal convex function that coincides with the best-fitting convex function (with respect to the standardized discrepancy) at solutions

in X. More solutions can be screened out if a tighter statistically valid cutoff value were used, especially for the d_{n}^1 standardized discrepancy—the potential gains for the d_{n}^2 and d_{n}^∞ standardized discrepancies are more limited.

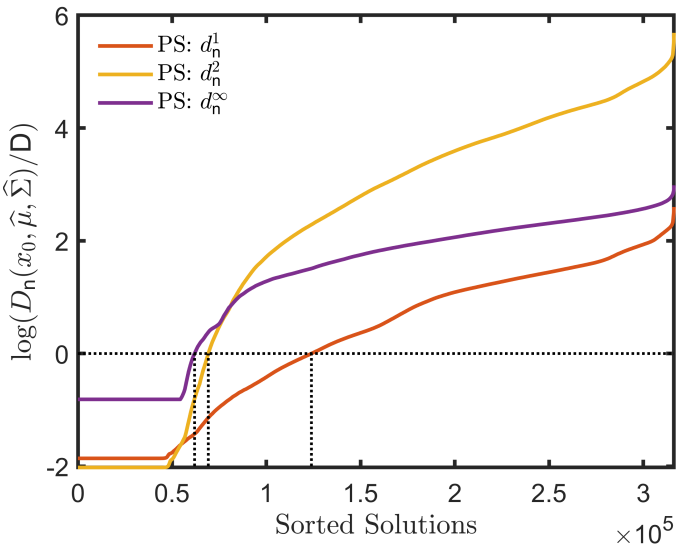


Figure 4: Sorted logarithm of scaled minimum standardized discrepancies of feasible solutions for a single macroreplication of PS with the d_{n}^1, d_{n}^2, and d_{n}^∞ standardized discrepancies. The horizontal black dotted line differentiates solutions that are retained (below) and screened out (above). The vertical black dotted lines indicate subset sizes.

Without an oracle for evaluating the true objective function, we took 500 replications at each feasible solution (using CRN) and estimated the optimality gaps based on the sample means. Figure 5 shows the optimality gaps for the feasible solutions, as well as those in $\mathcal{S}_{\mathsf{n}}^{\mathrm{PS}}$ and $\mathcal{S}_{\mathsf{n}}^{\mathrm{RPS}}$ with d_{n}^2. The results demonstrate that PS and RPS can screen out a large portion of the inferior solutions, while retaining high-quality solutions.

7 Conclusions and Discussions

This article describes a novel but nascent framework for screening solutions whose performances could be evaluated via stochastic simulation. In contrast to traditional subset-selection procedures, our methods can screen out unsimulated solutions, making them appealing statistical inference techniques for large-scale simulation-optimization problems on which such procedures are otherwise unworkable. For the Plausible Screening method, solutions are screened by minimizing a standardized discrepancy—a function measuring the distance between the sample means and a given vector—over a feasible region characterized by known properties of the expected response function. For the Relaxed Plausible Screening method, solutions are screened by checking the feasibility of a system of linear equations. Both methods return subsets of solutions that attain typical statistical properties of confidence and consistency.

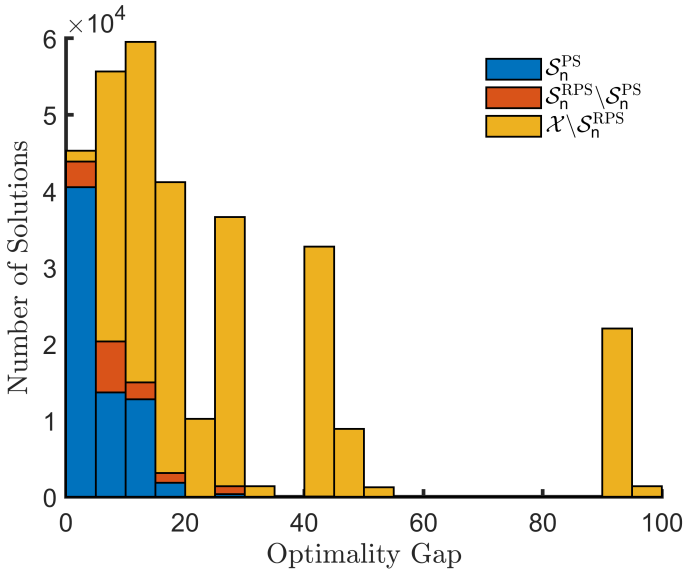


Figure 5: Histogram of the optimality gaps of solutions retained in $\mathcal{S}_{\mathsf{n}}^{\mathrm{PS}}$ (blue), the optimality gaps of additional solutions in $\mathcal{S}_{\mathsf{n}}^{\mathrm{RPS}}$ (red), and the optimality gaps of all remaining solutions (orange) for the d_{n}^2 standardized discrepancy.

Experimental results demonstrate the power of exploiting known functional properties, with varying degrees of effectiveness for different standardized discrepancies. In the absence of any specialized insight into the structure of the underlying optimization problems, we recommend the d_{n}^{∞} standardized discrepancy as an efficient and powerful choice.

The proposed methodology can be extended well beyond the initial treatment in this paper. Other, more sophisticated, forms of functional properties can also be incorporated, such as estimated first-order information like stochastic gradients. One could also imagine employing local function information, e.g. a local Lipschitz constant or local convexity. Answering the question of how one acquires functional information is critical to convert this idea into a practical tool. One direction could pair this methodology with existing tests for functional properties [Juditsky and Nemirovski, 2002]. Another tact is to explicitly leverage our minimal discrepancy to test for functional properties of expected response functions, though we have not fully developed these ideas. We conjecture that there are many classes of simulation problems with functional information available upon careful examination.

Another area of future research is how the choice of the experimental set, X, and the number of simulation replications allocated to solutions in it, n, dictate the effectiveness of our methods. There are many relevant practical questions that can be addressed: Given a fixed budget, is it better to obtain few replications at many solutions or more replications at fewer solutions? Given the known properties of μ, how should the solutions in X be spread over $\mathcal{X}$? The answers to these questions might be informed by an asymptotic analysis of our methods as k and n increase together.

Extending our methods to allow for sequential experimentation has great potential. Adaptively identifying solutions in X at which to obtain more replications or new solutions to add to X can lead to more efficient and powerful screening. However, preserving the

statistical guarantees of such procedures will require careful attention.

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

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

Bruce Ankenman, Barry L. Nelson, and Jeremy Staum. Stochastic kriging for simulation metamodeling. *Operations Research*, 58(2):371–382, 2010.

RE Bechhofer, TJ Santner, and DM Goldsman. *Designing Experiments for Statistical Selection, Screening, and Multiple Comparisons*. Wiley & Sons, New York, 1995.

Bert Bettonvil and Jack PC Kleijnen. Searching for important factors in simulation models with many factors: Sequential bifurcation. *European Journal of Operational Research*, 96 (1):180–194, 1997.

Justin Boesel, Barry L Nelson, and Seong-Hee Kim. Using ranking and selection to "clean up" after simulation optimization. *Operations Research*, 51(5):814–825, 2003.

Stephen Boyd and Lieven Vandenberghe. *Convex Optimization*. Cambridge University Press, Cambridge, 2004.

Kuo-Hao Chang, L Jeff Hong, and Hong Wan. Stochastic trust-region response-surface method (STRONG)—A new response-surface framework for simulation optimization. *INFORMS Journal on Computing*, 25(2):230–243, 2013.

Darinka Dentcheva and Andrzej Ruszczyński. Portfolio optimization with stochastic dominance constraints. *Journal of Banking and Finance*, 30(2):433–451, 2006.

David J. Eckman, Matthew Plumlee, and Barry L. Nelson. Revisiting subset selection. In K.-H. Bae, B. Feng, S. Kim, S. Lazarova-Molnar, Z. Zheng, T. Roeder, and R. Thiesing, editors, *Proceedings of the 2020 Winter Simulation Conference*, Piscataway, New Jersey, 2020. Institute of Electrical and Electronics Engineers, Inc. Forthcoming.

Peter Frazier, Warren Powell, and Savas Dayanik. The knowledge-gradient policy for correlated normal beliefs. *INFORMS journal on Computing*, 21(4):599–613, 2009.

Shanti S Gupta. On some multiple decision (selection and ranking) rules. *Technometrics*, 7 (2):225–245, 1965.

Nanjing Jian and Shane G. Henderson. Estimating the probability that a function observed with noise is convex. *INFORMS Journal on Computing*, 32(2):376–389, 2020.

Rui-Juan Jing, Marc Moreno Maza, and Delaram Talaashrafi. Complexity estimates for Fourier-Motzkin elimination. *CoRR*, abs/1811.01510, 2018. URL `http://arxiv.org/abs/1811.01510`.

Anatoli Juditsky and Arkadi Nemirovski. On nonparametric tests of positivity/monotonicity/convexity. *Annals of Statistics*, 30(2):498–527, 2002.

Sujin Kim, Raghu Pasupathy, and Shane G Henderson. A guide to sample average approximation. In Michael Fu, editor, *Handbook of Simulation Optimization*, pages 207–243. Springer-Verlag, New York, 2015.

J. P. C. Kleijnen. *Design and Analysis of Simulation Experiments*. Springer, Cham, Switzerland, 2015.

Manuel Laguna. Optquest: Optimization of complex systems. Technical report, 2011. URL `https://www.opttek.com/sites/default/files/pdfs/OptQuest-Optimization%20of%20Complex%20Systems.pdf`.

Eunji Lim and Peter W. Glynn. Consistency of multidimensional convex regression. *Operations Research*, 60(1):196–208, 2012.

Jingang Liu, Chihui Li, Feng Yang, Hong Wan, and Reha Uzsoy. Production planning for semiconductor manufacturing via simulation optimization. In S. Jain, R. R. Creasey, J. Himmelspach, K. P. White, and M. Fu, editors, *Proceedings of the 2011 Winter Simulation Conference*, pages 3612–3622, Piscataway, NJ, 2011. Institute of Electrical and Electronics Engineers.

Siddharth Mahajan and Garrett van Ryzin. Stocking retail assortments under dynamic consumer substitution. *Operations Research*, 49(3):334–351, 2001.

Kazuo Murota. Discrete convex analysis. *Mathematical Programming*, 83(1-3):313–371, 1998.

Barry L Nelson, Julie Swann, David Goldsman, and Wheyming Song. Simple procedures for selecting the best simulated system when the number of alternatives is large. *Operations Research*, 49(6):950–963, 2001.

George Nemhauser and Laurence Wolsey. *Integer and Combinatorial Optimization*. John Wiley & Sons, New York, 1999.

Eric C Ni, Shane G Henderson, and Susan R Hunter. A comparison of two parallel ranking and selection procedures. In *Proceedings of the 2014 Winter Simulation Conference*, pages 3761–3772, 2014.

Erica L. Plambeck, Bor-Ruey Fu, Stephen M. Robinson, and Rajan Suri. Sample-path optimization of convex stochastic performance functions. *Mathematical Programming*, 75 (2):137–176, 1996.

Matthew Plumlee and Barry L. Nelson. Plausible optima. In Markus Rabe, Angel A. Juan, Navonil Mustafee, Anders Skoogh, Sanjay Jain, and Björn Johansson, editors, *Proceedings of the 2018 Winter Simulation Conference*, pages 1981–1992, Piscataway, New Jersey, 2018. Institute of Electrical and Electronics Engineers, Inc.

Evan L. Porteus. Stochastic inventory theory. In D. P. Heyman and M. J. Sobel, editors, *Stochastic Models*, volume 2 of *Handbooks in Operations Research and Management Science*, chapter 12, pages 605–652. Elsevier, New York, 1990.

Peter Salemi, Eunhye Song, Barry L. Nelson, and Jeremy Staum. Gaussian Markov random fields for discrete optimization via simulation: Framework and algorithms. *Operations Research*, 67(1):250–266, 2019a.

Peter Salemi, Jeremy Staum, and Barry L. Nelson. Generalized integrated Brownian fields for simulation metamodeling. *Operations Research*, 67(3):874–891, 2019b.

Thomas J. Santner, Brian J. Williams, and William I. Notz. *The Design and Analysis of Computer Experiments*. Springer, New York, 2nd edition, 2003.

Warren Scott, Peter Frazier, and Warren Powell. The correlated knowledge gradient for simulation optimization of continuous parameters using Gaussian process regression. *SIAM Journal on Optimization*, 21(3):996–1026, 2011.

Moshe Shaked and J. George Shanthikumar. Stochastic convexity and its applications. *Advances in Applied Probability*, 20(2):427–446, 1988.

Moshe Shaked and J. George Shanthikumar. *Stochastic Orders*. Springer Science+Business Media, New York, 2007.

J. George Shanthikumar and David D. Yao. Second-order stochastic properties in queueing systems. *Proceedings of the IEEE*, 77(1):162–170, 1989.

A. Shapiro and Y. Wardi. Convergence analysis of gradient descent stochastic algorithms. *Journal of Optimization Theory and Applications*, 91(2):439–454, 1996.

Sara Shashaani, Fatemeh S. Hashemi, and Raghu Pasupathy. ASTRO-DF: A class of adaptive sampling trust-region algorithms for derivative-free stochastic optimization. *SIAM Journal on Optimization*, 28(4):3145–3176, 2018.

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

Mark L. Spearman, David L. Woodruff, and Wallace J. Hopp. CONWIP: A pull alternative to kanban. *International Journal of Production Research*, 28(5):879–894, 1990.

Fang Wan, Wei Liu, Frank Bretz, and Yang Han. Confidence sets for optimal factor levels of a response surface. *Biometrics*, 72(4):1285–1293, 2016.

Hong Wan, Bruce E Ankenman, and Barry L Nelson. Controlled sequential bifurcation: A new factor-screening method for discrete-event simulation. *Operations Research*, 54(4): 743–755, 2006.

Ronald W. Wolff. An upper bound for multi-channel queues. *Journal of Applied Probability*, 14(4):884–888, 1977.

Appendices

A Derivations of $\mathsf{M}(x_0)$ and $\mathsf{S}(\mathsf{X})$ for Examples 1 and 2

A.1 Derivations for Example 1

Portions of this proof are adapted from that of Theorem 5 of Plumlee and Nelson [2018].

For a given μ, let $\mathcal{A} = \{x \in \mathcal{X} \colon \mu(x) \leq \min_{x' \in \mathcal{X}} \mu(x')\}$ and suppose that μ is known to be γ-Lipschitz. From the standard definition of a γ-Lipschitz function,

$$\mathscr{M} = \{m \in \mathscr{F} \colon |m(x) - m(x')| \leq \gamma \|x - x'\| \text{ for all } x, x' \in \mathcal{X}\}.$$

Adding the constraint that $m(x_0) \leq m(x)$ for all $x \in \mathcal{X}$, we have that

$$\begin{aligned}\mathscr{M}(x_0) = \{m \in \mathscr{F} \colon |m(x) - m(x')| \leq \gamma \|x - x'\| \text{ for all } x, x' \in \mathcal{X} \\ \text{and } m(x_0) \leq m(x) \text{ for all } x \in \mathcal{X}\},\end{aligned}$$

the set of γ-Lipschitz functions for which solution x_0 is optimal.

We show that the projection of $\mathscr{M}(x_0)$ onto $\mathbb{R}^k$ is given by

$$\begin{aligned}\mathsf{M}(x_0) = \{\mathsf{m} \in \mathbb{R}^k \colon &\text{there exists } \mathsf{m}_0 \in \mathbb{R} \text{ such that} \\ &|\mathsf{m}_i - \mathsf{m}_j| \leq \gamma \|x_i - x_j\| \text{ for all } i, j = 1, \ldots, k \\ &|\mathsf{m}_i - \mathsf{m}_0| \leq \gamma \|x_i - x_0\| \text{ for all } i = 1, \ldots, k \\ &-\mathsf{m}_i + \mathsf{m}_0 \leq 0 \text{ for all } i = 1, \ldots, k\}.\end{aligned}$$

We first prove that for any function $m \in \mathscr{M}(x_0)$, there exists $\mathsf{m} \in \mathsf{M}(x_0)$ such that $\mathsf{m}_i = m(x_i)$ for $i = 1, \ldots, k$. Fix an arbitrary function $m \in \mathscr{M}(x_0)$ and define $\mathsf{m}_i = m(x_i)$ for $i = 0, 1, \ldots, k$. Since m is a γ-Lipschitz function, $|\mathsf{m}_i - \mathsf{m}_j| = |m(x_i) - m(x_j)| \leq \gamma \|x_i - x_j\|$ for all $i, j = 1, \ldots, k$ and $|\mathsf{m}_i - \mathsf{m}_0| = |m(x_i) - m(x_0)| \leq \gamma \|x_i - x_0\|$ for all $i = 1, \ldots, k$. And since m is an x_0-optimal function, $\mathsf{m}_0 = m(x_0) \leq m(x_i) = \mathsf{m}_i$ for all $i = 1, \ldots, k$. Hence, $\mathsf{m} = (\mathsf{m}_1, \ldots, \mathsf{m}_k)^\top \in \mathsf{M}(x_0)$.

We next prove that for any $\mathsf{m} \in \mathsf{M}(x_0)$, there exists a function $\overline{m} \in \mathscr{M}(x_0)$ such that $\overline{m}(x_i) = \mathsf{m}_i$ for $i = 1, \ldots, k$. Fix an arbitrary vector $\mathsf{m} \in \mathsf{M}(x_0)$. From the definition of $\mathsf{M}(x_0)$, there exists an $\mathsf{m}_0 \in \mathbb{R}$ such that $|\mathsf{m}_i - \mathsf{m}_0| \leq \gamma \|x_i - x_0\|$ for all $i = 1, \ldots, k$ and $\mathsf{m}_0 \leq \mathsf{m}_i$ for all $i = 1, \ldots, k$. Let

$$\begin{aligned}m^+(x) &= \min_{i=0,1,\ldots,k} \mathsf{m}_i + \gamma \|x - x_i\|, \\ i^+(x) &= \arg \min_{i=0,1,\ldots,k} \mathsf{m}_i + \gamma \|x - x_i\|, \\ m^-(x) &= \max_{i=0,1,\ldots,k} \mathsf{m}_i - \gamma \|x - x_i\|, \text{ and} \\ i^-(x) &= \arg \max_{i=0,1,\ldots,k} \mathsf{m}_i - \gamma \|x - x_i\|,\end{aligned}$$

for all $x \in \mathcal{X}$ and consider $\overline{m}(x) \equiv (m^+(x) + m^-(x))/2$.

From the definition of $\mathsf{M}(x_0)$, for any $i = 0, 1, \ldots, k$, $\mathsf{m}_i \leq \mathsf{m}_j + \gamma \|x_i - x_j\|$ for all $j = 0, 1, \ldots, k$. Therefore $m^+(x_i) = \mathsf{m}_i$ for all $i = 0, 1, \ldots, k$. Likewise, since $\mathsf{m}_i \geq \mathsf{m}_j - \gamma \|x_i - x_j\|$ for all $j = 0, 1, \ldots, k$, it follows that $m^-(x_i) = \mathsf{m}_i$. Thus $\overline{m}(x_i) = \mathsf{m}_i$ for all $i = 0, 1, \ldots, k$.

One can also verify that $\overline{m}$ is γ-Lipschitz. We first show that both m^+ and m^- are γ-Lipschitz. For any $x, x' \in \mathcal{X}$,

$$\begin{aligned} m^+(x) - m^+(x') &\leq m_{i^+(x')} + \gamma\|x - x_{i^+(x')}\| - m_{i^+(x')} - \gamma\|x' - x_{i^+(x')}\| \\ &\leq \gamma\|x - x_{i^+(x')}\| - \gamma\|x' - x_{i^+(x')}\| \leq \gamma\|x - x'\|, \end{aligned}$$

and

$$\begin{aligned} m^+(x) - m^+(x') &\geq m_{i^+(x)} + \gamma\|x - x_{i^+(x)}\| - m_{i^+(x)} - \gamma\|x' - x_{i^+(x)}\| \\ &\geq \gamma\|x - x_{i^+(x)}\| - \gamma\|x' - x_{i^+(x)}\| \geq -\gamma\|x - x'\|. \end{aligned}$$

Thus $|m^+(x) - m^+(x')| \leq \gamma\|x - x'\|$. Likewise,

$$\begin{aligned} m^-(x) - m^-(x') &\leq m_{i^-(x)} - \gamma\|x - x_{i^-(x)}\| - m_{i^-(x)} + \gamma\|x' - x_{i^-(x)}\| \\ &\leq -\gamma\|x - x_{i^-(x)}\| + \gamma\|x' - x_{i^-(x)}\| \leq \gamma\|x - x'\|, \end{aligned}$$

and

$$\begin{aligned} m^-(x) - m^-(x') &\geq m_{i^-(x')} - \gamma\|x - x_{i^-(x')}\| - m_{i^-(x')} + \gamma\|x' - x_{i^-(x')}\| \\ &\geq -\gamma\|x - x_{i^-(x')}\| + \gamma\|x' - x_{i^-(x')}\| \geq -\gamma\|x - x'\|. \end{aligned}$$

Thus $|m^-(x) - m^-(x')| \leq \gamma\|x - x'\|$. By substitution,

$$\begin{aligned} |\overline{m}(x) - \overline{m}(x')| &= \frac{1}{2}|m^+(x) + m^-(x) - m^+(x') - m^-(x')| \\ &\leq \frac{1}{2}|m^+(x) - m^+(x')| + \frac{1}{2}|m^-(x) - m^-(x')| \\ &\leq \frac{1}{2}\gamma\|x - x'\| + \frac{1}{2}\gamma\|x - x'\| \\ &= \gamma\|x - x'\|, \end{aligned}$$

i.e., $\overline{m}$ is γ-Lipschitz.

We next demonstrate that $\overline{m}$ is x_0-optimal. By construction, for any $x \in \mathcal{X}$, $\overline{m}(x_{i^+(x)}) = \mathsf{m}_{i^+(x)} \geq \mathsf{m}_0 = \overline{m}(x_0)$. In addition, from the definition of $i^-(x)$,

$$\overline{m}(x_{i^-(x)}) - \gamma\|x - x_{i^-(x)}\| \geq \overline{m}(x_{i^+(x)}) - \gamma\|x - x_{i^+(x)}\|.$$

Therefore

$$\overline{m}(x) = \frac{1}{2}\left(\overline{m}(x_{i^+(x)}) + \gamma\|x - x_{i^+(x)}\| + \overline{m}(x_{i^-(x)}) - \gamma\|x - x_{i^-(x)}\|\right) \geq \overline{m}(x_{i^+(x)}) \geq \overline{m}(x_0).$$

Altogether, we have that $\overline{m}(x_i) = \mathsf{m}_i$ for $i = 1, \ldots, k$ and $\overline{m} \in \mathscr{M}(x_0)$.

These two implications prove that $\mathsf{M}(x_0)$ is the projection of $\mathscr{M}(x_0)$ onto $\mathbb{R}^k$.

We next show that by projecting out m_0, $\mathsf{M}(x_0)$ can be explicitly expressed as

$$\mathsf{M}(x_0) = \left\{\mathsf{m} \in \mathbb{R}^k \colon \mathsf{m}_i - \mathsf{m}_j \leq \gamma \min\left\{\|x_i - x_j\|, \|x_i - x_0\|\right\} \text{ for all } i, j = 1, \ldots, k\right\}.$$

Applying Fourier-Motzkin elimination, we rearrange the inequalities defining $\mathsf{M}(x_0)$ to obtain

$$\begin{aligned}
\mathsf{M}(x_0) = \{\mathsf{m} \in \mathbb{R}^k\colon\ & \text{there exists } \mathsf{m}_0 \in \mathbb{R} \quad \text{such that} \\
& \mathsf{m}_i - \mathsf{m}_j \le \gamma\|x_i - x_j\| \text{ for all } i,j = 1,\dots,k \\
& \mathsf{m}_0 \ge \mathsf{m}_i - \gamma\|x_i - x_0\| \text{ for all } i = 1,\dots,k \\
& \mathsf{m}_0 \le \mathsf{m}_i + \gamma\|x_i - x_0\| \text{ for all } i = 1,\dots,k \\
& \mathsf{m}_0 \le \mathsf{m}_i \text{ for all } i = 1,\dots,k\} \\
= \{\mathsf{m} \in \mathbb{R}^k\colon\ & \text{there exists } \mathsf{m}_0 \in \mathbb{R} \quad \text{such that} \\
& \mathsf{m}_i - \mathsf{m}_j \le \gamma\|x_i - x_j\| \text{ for all } i,j = 1,\dots,k \\
& \mathsf{m}_0 \ge \mathsf{m}_i - \gamma\|x_i - x_0\| \text{ for all } i = 1,\dots,k \\
& \mathsf{m}_0 \le \mathsf{m}_i \text{ for all } i = 1,\dots,k\} \\
= \{\mathsf{m} \in \mathbb{R}^k\colon\ & \mathsf{m}_i - \mathsf{m}_j \le \gamma\|x_i - x_j\| \text{ for all } i,j = 1,\dots,k \\
& \mathsf{m}_i - \gamma\|x_i - x_0\| \le \mathsf{m}_j \text{ for all } i,j = 1,\dots,k\} \\
= \{\mathsf{m} \in \mathbb{R}^k\colon\ & \mathsf{m}_i - \mathsf{m}_j \le \gamma \min\{\|x_i - x_j\|, \|x_i - x_0\|\} \text{ for all } i,j = 1,\dots,k\}.
\end{aligned}$$

Lastly, we derive $\mathsf{S}(\mathsf{X})$. For any fixed $\mu \in \mathscr{M}$,

$$\begin{aligned}
\mathsf{S}(\mathsf{X}) &= \{x_0 \in \mathcal{X}\colon \mu(\mathsf{X}) \in \mathsf{M}(x_0)\} \\
&= \{x_0 \in \mathcal{X}\colon \mu(x_i) - \mu(x_j) \le \gamma \min\{\|x_i - x_j\|, \|x_i - x_0\|\} \text{ for all } i,j = 1,\dots,k\} \\
&= \{x_0 \in \mathcal{X}\colon \mu(x_i) - \mu(x_j) \le \gamma\|x_i - x_0\| \text{ for all } i,j = 1,\dots,k\} \\
&= \{x_0 \in \mathcal{X}\colon \mu(x_i) - \gamma\|x_i - x_0\| \le \mu(x_j) \text{ for all } i,j = 1,\dots,k\} \\
&= \left\{x_0 \in \mathcal{X}\colon \max_{i=1,\dots,k} \mu(x_i) - \gamma\|x_i - x_0\| \le \min_{j=1,\dots,k} \mu(x_j)\right\},
\end{aligned}$$

where in the third equality we drop the constraints $\mu(x_i) - \mu(x_j) \le \gamma\|x_i - x_j\|$ since μ is known to be γ-Lipschitz. $\square$

A.2 Derivations for Example 2

For a given μ, let $\mathcal{A} = \{x \in \mathcal{X}\colon \mu(x) \le \min_{x' \in \mathcal{X}} \mu(x')\}$ and suppose that μ is known to be convex over $\mathcal{X}$. To account for the possibility that $\mathcal{X}$ may be discrete, we adopt the following definition of convexity: a function m is convex over $\mathcal{X} \subseteq \mathbb{R}^d$ if at each $x \in \mathcal{X}$, there exists a subgradient $\xi(x) \in \mathbb{R}^d$ such that $m(x') \ge m(x) - (x - x')^\top \xi(x)$ for all $x' \in \mathcal{X}$ [Murota, 1998]. In terms of our notation,

$$\begin{aligned}
\mathscr{M} = \big\{m \in \mathscr{F}\colon & \text{for all } x \in \mathcal{X}, \text{ there exists } \xi(x) \in \mathbb{R}^d \text{ such that} \\
& m(x) - m(x') \le (x - x')^\top \xi(x) \text{ for all } x' \in \mathcal{X}\big\}.
\end{aligned}$$

Adding the constraint that $m(x_0) \le m(x)$ for all $x \in \mathcal{X}$, we have that

$$\begin{aligned}
\mathscr{M}(x_0) = \big\{m \in \mathscr{F}\colon & \text{for all } x \in \mathcal{X}, \text{ there exists } \xi(x) \in \mathbb{R}^d \text{ such that} \\
& m(x) - m(x') \le (x - x')^\top \xi(x) \text{ for all } x' \in \mathcal{X} \text{ and } m(x_0) \le m(x)\big\},
\end{aligned}$$

the set of convex functions for which solution x_0 is optimal.

We show that the projection of $\mathscr{M}(x_0)$ onto $\mathbb{R}^k$ is given by

$$\begin{aligned}\mathsf{M}(x_0) = \{\mathsf{m} \in \mathbb{R}^k \colon \text{there exists } \mathsf{m}_0 \in \mathbb{R} \text{ and } \xi_1, \ldots, \xi_k \in \mathbb{R}^d \text{ such that} \\ \mathsf{m}_i - \mathsf{m}_j - (x_i - x_j)^\top \xi_i \leq 0 \text{ for all } i, j = 1, \ldots, k \\ \mathsf{m}_i - \mathsf{m}_0 - (x_i - x_0)^\top \xi_i \leq 0 \text{ for all } i = 1, \ldots, k \\ -\mathsf{m}_i + \mathsf{m}_0 \leq 0 \text{ for all } i = 1, \ldots, k\}.\end{aligned}$$

We first prove that for any function $m \in \mathscr{M}(x_0)$, there exists $\mathsf{m} \in \mathsf{M}(x_0)$ such that $\mathsf{m}_i = m(x_i)$ for $i = 1, \ldots, k$. Fix an arbitrary function $m \in \mathscr{M}(x_0)$ and define $\mathsf{m}_i = m(x_i)$ for $i = 0, 1, \ldots, k$. Since $m \in \mathscr{M}(x_0)$, there exists $\xi(x_1), \ldots, \xi(x_k) \in \mathbb{R}^d$ such that $m(x_i) - m(x') \leq (x_i - x')^\top \xi(x_i)$ for all $x' \in \mathcal{X}$. Let $\xi_i = \xi(x_i)$ for $i = 1, \ldots, k$ so that $\mathsf{m}_i - \mathsf{m}_j = m(x_i) - m(x_j) \leq (x_i - x_j)^\top \xi(x_i) = (x_i - x_j)^\top \xi_i$ for all $j = 0, 1, \ldots, k$. Since m is an x_0-optimal function, $\mathsf{m}_0 = m(x_0) \leq m(x_i) = \mathsf{m}_i$ for all $i = 1, \ldots, k$. Hence, for these choices of m_0 and $\xi_1, \ldots, \xi_k$, $\mathsf{m} = (\mathsf{m}_1, \ldots, \mathsf{m}_k)^\top \in \mathsf{M}(x_0)$.

We next prove that for any $\mathsf{m} \in \mathsf{M}(x_0)$, there exists a function $\overline{m} \in \mathscr{M}(x_0)$ such that $\overline{m}(x_i) = \mathsf{m}_i$ for $i = 1, \ldots, k$. Fix an arbitrary vector $\mathsf{m} \in \mathsf{M}(x_0)$. From the definition of $\mathsf{M}(x_0)$, there exists an $\mathsf{m}_0 \in \mathbb{R}$ and $\xi_1, \ldots, \xi_k \in \mathbb{R}^d$ such that for all $i = 1, \ldots, k$, $\mathsf{m}_i - \mathsf{m}_j \leq (x_i - x_j)^\top \xi_i$ for all $j = 0, 1, \ldots, k$ and $\mathsf{m}_0 \leq \mathsf{m}_i$. Define $\xi_0 = \mathbf{0}_d$, a d-vector of all zeros, and consider

$$\overline{m}(x) \equiv \max_{i=0,1,\ldots,k} \mathsf{m}_i + (x - x_i)^\top \xi_i,$$

for all $x \in \mathcal{X}$. For $i = 1, \ldots, k$,

$$\overline{m}(x_i) = \max_{j=0,1,\ldots,k} \mathsf{m}_j + (x_i - x_j)^\top \xi_j = \mathsf{m}_i,$$

from the first and second sets of inequalities describing $\mathsf{M}(x_0)$. Since the function $\overline{m}$ is the maximum of $k+1$ convex functions, it is convex. Also, $\overline{m}$ is x_0-optimal because for all $x \in \mathcal{X}$,

$$\overline{m}(x) = \max_{i=0,1,\ldots,k} \mathsf{m}_i + (x - x_i)^\top \xi_i \geq \mathsf{m}_0 + (x - x_0)^\top \mathbf{0}_d = \mathsf{m}_0.$$

Altogether, we have that $\overline{m}(x_i) = \mathsf{m}_i$ for $i = 1, \ldots, k$ and $\overline{m} \in \mathscr{M}(x_0)$.
These two implications prove that $\mathsf{M}(x_0)$ is the projection of $\mathscr{M}(x_0)$ onto $\mathbb{R}^k$.
By definition, for any fixed $\mu \in \mathscr{M}$,

$$\begin{aligned}\mathsf{S}(\mathsf{X}) &= \{x_0 \in \mathcal{X} \colon \mu(\mathsf{X}) \in \mathsf{M}(x_0)\} \\ &= \big\{x_0 \in \mathcal{X} \colon \text{there exists } \mathsf{m}_0 \in \mathbb{R} \text{ and } \xi_1, \ldots, \xi_k \in \mathbb{R}^d \text{ such that} \\ &\qquad - (x_i - x_j)^\top \xi_i \leq -\mu(x_i) + \mu(x_j) \text{ for all } i, j = 1, \ldots, k \\ &\qquad - \mathsf{m}_0 - (x_i - x_0)^\top \xi_i \leq -\mu(x_i) \text{ for all } i = 1, \ldots, k \\ &\qquad \mathsf{m}_0 \leq \mu(x_i) \text{ for all } i = 1, \ldots, k\big\}. \quad \square\end{aligned}$$

B Derivations for Relaxed Plausible Screening

In this appendix, we verify the equations stated in Section 5.1 when describing how we offset the right-hand-side vector b that appears in Assumption 1. These equations are rewritten

below in terms of a general vector $\mathsf{v} = (v_1, \ldots, v_k)^\top$:

$$\max_{\mathsf{m}\in\mathbb{R}^k} \left\{ \mathsf{v}^\top(\widehat{\mu} - \mathsf{m}) \colon d_{\mathsf{n}}^1(\mathsf{m}, \widehat{\mu}, \widehat{\Sigma}) \le \mathsf{D}^1 \right\} = \mathsf{D}^1 \max_{i=1,\ldots,k} \frac{\widehat{\sigma}_i}{\sqrt{n_i}} |v_i|, \tag{4}$$

$$\max_{\mathsf{m}\in\mathbb{R}^k} \left\{ \mathsf{v}^\top(\widehat{\mu} - \mathsf{m}) \colon d_{\mathsf{n}}^2(\mathsf{m}, \widehat{\mu}, \widehat{\Sigma}) \le \mathsf{D}^2 \right\} = \sqrt{\mathsf{D}^2 \sum_{i=1}^k \frac{\widehat{\sigma}_i^2}{n_i} v_i^2}, \tag{5}$$

$$\max_{\mathsf{m}\in\mathbb{R}^k} \left\{ \mathsf{v}^\top(\widehat{\mu} - \mathsf{m}) \colon d_{\mathsf{n}}^\infty(\mathsf{m}, \widehat{\mu}, \widehat{\Sigma}) \le \mathsf{D}^\infty \right\} = \mathsf{D}^\infty \sum_{i=1}^k \frac{\widehat{\sigma}_i}{\sqrt{n_i}} |v_i|, \text{ and} \tag{6}$$

$$\max_{\mathsf{m}\in\mathbb{R}^k} \left\{ \mathsf{v}^\top(\widehat{\mu} - \mathsf{m}) \colon d_{\mathsf{n}}^{\mathrm{CRN}}(\mathsf{m}, \widehat{\mu}, \widehat{\Sigma}) \le \mathsf{D}^{\mathrm{CRN}} \right\} = \sqrt{\frac{\mathsf{D}^{\mathrm{CRN}}}{n} \mathsf{v}^\top \widehat{\Sigma} \mathsf{v}}. \tag{7}$$

In the proofs that follow, we define

$$\mathsf{c} \equiv (\sqrt{n_1}/\widehat{\sigma}_1, \ldots, \sqrt{n_k}/\widehat{\sigma}_k)^\top \quad \text{and} \quad \bar{\mathsf{c}} \equiv (\widehat{\sigma}_1/\sqrt{n_1}, \ldots, \widehat{\sigma}_k/\sqrt{n_k})^\top,$$

and let $\mathsf{y} \odot \mathsf{z}$ denote the element-wise multiplication of vectors $\mathsf{y}, \mathsf{z} \in \mathbb{R}^k$. Clearly, $\mathsf{c} \odot \bar{\mathsf{c}} = \mathbf{1}_k$ where $\mathbf{1}_k$ is a k-vector of all ones. For $p \ge 1$, let $\|\mathsf{y}\|_p \equiv \left(\sum_{i=1}^k |y_i|^p\right)^{1/p}$ denote the p-norm of a vector $\mathsf{y} \in \mathbb{R}^k$. We will make use of the following well-known result in mathematical optimization; see Appendix A.1.6 of Boyd and Vandenberghe [2004].

Lemma 2 (Dual norm of p-norm) *For any $\mathsf{y} \in \mathbb{R}^k$ and $p \ge 1$,*

$$\max_{\mathsf{z}\in\mathbb{R}^k} \left\{ \mathsf{y}^\top \mathsf{z} : \|\mathsf{z}\|_p \le 1 \right\} = \|\mathsf{y}\|_q \text{ where } q \text{ satisfies } 1/p + 1/q = 1.$$

We are now prepared to prove Equations (4)–(6).

B.1 Proof of Equation (4)

Fix an arbitrary $\mathsf{v} \in \mathbb{R}^k$. Then

$$\begin{aligned}
\max_{\mathsf{m}\in\mathbb{R}^k} \left\{ \mathsf{v}^\top(\widehat{\mu} - \mathsf{m}) \colon d_{\mathsf{n}}^1(\mathsf{m}, \widehat{\mu}, \widehat{\Sigma}) \le \mathsf{D}^1 \right\} &= \max_{\mathsf{m}\in\mathbb{R}^k} \left\{ \mathsf{v}^\top(\widehat{\mu} - \mathsf{m}) \colon \sum_{i=1}^k \frac{\sqrt{n_i}}{\widehat{\sigma}_i} |\widehat{\mu}_i - \mathsf{m}_i| \le \mathsf{D}^1 \right\} \\
&= \max_{\mathsf{m}\in\mathbb{R}^k} \left\{ \mathsf{v}^\top(\widehat{\mu} - \mathsf{m}) \colon \sum_{i=1}^k \frac{1}{\mathsf{D}^1} \frac{\sqrt{n_i}}{\widehat{\sigma}_i} |\widehat{\mu}_i - \mathsf{m}_i| \le 1 \right\} \\
&= \max_{\mathsf{m}\in\mathbb{R}^k} \left\{ \mathsf{v}^\top(\widehat{\mu} - \mathsf{m}) \colon \left\| \frac{\mathsf{c} \odot (\widehat{\mu} - \mathsf{m})}{\mathsf{D}^1} \right\|_1 \le 1 \right\} \\
&= \max_{\mathsf{m}\in\mathbb{R}^k} \left\{ \left(\mathsf{D}^1 (\mathsf{v} \odot \bar{\mathsf{c}})\right)^\top \left(\frac{\mathsf{c} \odot (\widehat{\mu} - \mathsf{m})}{\mathsf{D}^1} \right) : \left\| \frac{\mathsf{c} \odot (\widehat{\mu} - \mathsf{m})}{\mathsf{D}^1} \right\|_1 \le 1 \right\} \\
&= \left\| \mathsf{D}^1 (\mathsf{v} \odot \bar{\mathsf{c}}) \right\|_\infty \\
&= \mathsf{D}^1 \max_{i=1,\ldots,k} \frac{\widehat{\sigma}_i}{\sqrt{n_i}} |v_i|. \quad \square
\end{aligned}$$

B.2 Proof of Equation (5)

Fix an arbitrary $\mathsf{v} \in \mathbb{R}^k$. Then

$$\begin{aligned}
\max_{\mathsf{m}\in\mathbb{R}^k}\left\{\mathsf{v}^\top(\widehat{\mu}-\mathsf{m})\colon d_{\mathsf{n}}^2(\mathsf{m},\widehat{\mu},\widehat{\Sigma})\le \mathsf{D}^2\right\} &= \max_{\mathsf{m}\in\mathbb{R}^k}\left\{\mathsf{v}^\top(\widehat{\mu}-\mathsf{m})\colon \sum_{i=1}^k \frac{n_i}{\widehat{\sigma}_i^2}(\widehat{\mu}_i-\mathsf{m}_i)^2\le \mathsf{D}^2\right\}\\
&= \max_{\mathsf{m}\in\mathbb{R}^k}\left\{\mathsf{v}^\top(\widehat{\mu}-\mathsf{m})\colon \sum_{i=1}^k \frac{1}{\mathsf{D}^2}\frac{n_i}{\widehat{\sigma}_i^2}(\widehat{\mu}_i-\mathsf{m}_i)^2\le 1\right\}\\
&= \max_{\mathsf{m}\in\mathbb{R}^k}\left\{\mathsf{v}^\top(\widehat{\mu}-\mathsf{m})\colon \left\|\frac{\mathsf{c}\odot(\widehat{\mu}-\mathsf{m})}{\sqrt{\mathsf{D}^2}}\right\|_2\le 1\right\}\\
&= \max_{\mathsf{m}\in\mathbb{R}^k}\left\{\left(\sqrt{\mathsf{D}^2}\,(\mathsf{v}\odot\bar{\mathsf{c}})\right)^\top\left(\frac{\mathsf{c}\odot(\widehat{\mu}-\mathsf{m})}{\sqrt{\mathsf{D}^2}}\right)\colon \left\|\frac{\mathsf{c}\odot(\widehat{\mu}-\mathsf{m})}{\sqrt{\mathsf{D}^2}}\right\|_2\le 1\right\}\\
&= \left\|\sqrt{\mathsf{D}^2}\,(\mathsf{v}\odot\bar{\mathsf{c}})\right\|_2\\
&= \sqrt{\mathsf{D}^2\sum_{i=1}^k \frac{\widehat{\sigma}_i^2}{n_i}v_i^2}. \quad \square
\end{aligned}$$

B.3 Proof of Equation (6)

Fix an arbitrary $\mathsf{v} \in \mathbb{R}^k$. Then

$$\begin{aligned}
\max_{\mathsf{m}\in\mathbb{R}^k}\left\{\mathsf{v}^\top(\widehat{\mu}-\mathsf{m})\colon d_{\mathsf{n}}^\infty(\mathsf{m},\widehat{\mu},\widehat{\Sigma})\le \mathsf{D}^\infty\right\} &= \max_{\mathsf{m}\in\mathbb{R}^k}\left\{\mathsf{v}^\top(\widehat{\mu}-\mathsf{m})\colon \max_{i=1,\ldots,k}\frac{\sqrt{n_i}}{\widehat{\sigma}_i}|\widehat{\mu}_i-\mathsf{m}_i|\le \mathsf{D}^\infty\right\}\\
&= \max_{\mathsf{m}\in\mathbb{R}^k}\left\{\mathsf{v}^\top(\widehat{\mu}-\mathsf{m})\colon \max_{i=1,\ldots,k}\frac{1}{\mathsf{D}^\infty}\frac{\sqrt{n_i}}{\widehat{\sigma}_i}|\widehat{\mu}_i-\mathsf{m}_i|\le 1\right\}\\
&= \max_{\mathsf{m}\in\mathbb{R}^k}\left\{\mathsf{v}^\top(\widehat{\mu}-\mathsf{m})\colon \left\|\frac{\mathsf{c}\odot(\widehat{\mu}-\mathsf{m})}{\mathsf{D}^\infty}\right\|_\infty\le 1\right\}\\
&= \max_{\mathsf{m}\in\mathbb{R}^k}\left\{\left(\mathsf{D}^\infty(\mathsf{v}\odot\bar{\mathsf{c}})\right)^\top\left(\frac{\mathsf{c}\odot(\widehat{\mu}-\mathsf{m})}{\mathsf{D}^\infty}\right)\colon \left\|\frac{\mathsf{c}\odot(\widehat{\mu}-\mathsf{m})}{\mathsf{D}^\infty}\right\|_\infty\le 1\right\}\\
&= \left\|\mathsf{D}^\infty(\mathsf{v}\odot\bar{\mathsf{c}})\right\|_1\\
&= \mathsf{D}^\infty\sum_{i=1}^k\frac{\widehat{\sigma}_i}{\sqrt{n_i}}|v_i|.\square
\end{aligned}$$

For Equation (7) we require another dual norm result. For a $k\times k$ positive definite matrix A, let $\|\mathsf{y}\|_A \equiv \sqrt{\mathsf{y}^\top A\mathsf{y}}$ denote the norm of a vector $\mathsf{y}\in\mathbb{R}^k$ induced by the inner product $\langle \mathsf{y},\mathsf{z}\rangle_A \equiv \mathsf{y}^\top A\mathsf{z}$.

Lemma 3 *For any $\mathsf{y}\in\mathbb{R}^k$ and $A\in\mathbb{R}^{k\times k}$ positive definite,*

$$\max_{\mathsf{z}\in\mathbb{R}^k}\left\{\mathsf{y}^\top\mathsf{z}:\|\mathsf{z}\|_A\le 1\right\} = \|\mathsf{y}\|_{A^{-1}}.$$

B.4 Proof of Lemma 3

The Cholesky decomposition allows us to write $A = LL^\top$ for some $k \times k$ lower-diagonal matrix L. Note that for Cholesky decomposition, there are only nonzero entries along the diagonal of L and thus it is invertible. Therefore,

$$\begin{aligned}
\max_{\mathsf{z}\in\mathbb{R}^k} \left\{\mathsf{y}^\top \mathsf{z} : \|\mathsf{z}\|_A \le 1\right\} &= \max_{\mathsf{z}\in\mathbb{R}^k} \left\{\mathsf{y}^\top \mathsf{z} : \sqrt{\mathsf{z}^\top A \mathsf{z}} \le 1\right\} \\
&= \max_{\mathsf{z}\in\mathbb{R}^k} \left\{\mathsf{y}^\top \mathsf{z} : \sqrt{\mathsf{z}^\top L L^\top \mathsf{z}} \le 1\right\} \\
&= \max_{\mathsf{z}\in\mathbb{R}^k} \left\{\mathsf{y}^\top \mathsf{z} : \|L^\top \mathsf{z}\|_2 \le 1\right\} \\
&= \max_{\mathsf{z}\in\mathbb{R}^k} \left\{\mathsf{y}^\top (L^\top)^{-1} L^\top \mathsf{z} : \|L^\top \mathsf{z}\|_2 \le 1\right\} \\
&= \max_{\mathsf{z}\in\mathbb{R}^k} \left\{(L^{-1}\mathsf{y})^\top L^\top \mathsf{z} : \|L^\top \mathsf{z}\|_2 \le 1\right\} \\
&= \|L^{-1}\mathsf{y}\|_2 \\
&= \sqrt{\mathsf{y}^\top (L^{-1})^\top L^{-1} \mathsf{y}} \\
&= \sqrt{\mathsf{y}^\top A^{-1} \mathsf{y}} \\
&= \|\mathsf{y}\|_{A^{-1}}. \quad \square
\end{aligned}$$

B.5 Proof of Equation (7)

Fix an arbitrary $\mathsf{v} \in \mathbb{R}^k$. Then

$$\begin{aligned}
&\max_{\mathsf{m}\in\mathbb{R}^k} \left\{\mathsf{v}^\top(\widehat{\mu} - \mathsf{m}) \colon d_{\mathsf{n}}^{\mathrm{CRN}}(\mathsf{m}, \widehat{\mu}, \widehat{\Sigma}) \le \mathsf{D}^{\mathrm{CRN}}\right\} \\
&\quad = \max_{\mathsf{m}\in\mathbb{R}^k} \left\{\mathsf{v}^\top(\widehat{\mu} - \mathsf{m}) \colon n(\widehat{\mu} - \mathsf{m})^\top \widehat{\Sigma}^{-1} (\widehat{\mu} - \mathsf{m}) \le \mathsf{D}^{\mathrm{CRN}}\right\} \\
&\quad = \max_{\mathsf{m}\in\mathbb{R}^k} \left\{\mathsf{v}^\top(\widehat{\mu} - \mathsf{m}) \colon \frac{1}{\mathsf{D}^{\mathrm{CRN}}} n(\widehat{\mu} - \mathsf{m})^\top \widehat{\Sigma}^{-1} (\widehat{\mu} - \mathsf{m}) \le 1\right\} \\
&\quad = \max_{\mathsf{m}\in\mathbb{R}^k} \left\{\mathsf{v}^\top(\widehat{\mu} - \mathsf{m}) \colon \left\|\sqrt{\frac{n}{\mathsf{D}^{\mathrm{CRN}}}}(\widehat{\mu} - \mathsf{m})\right\|_{\widehat{\Sigma}^{-1}} \le 1\right\} \\
&\quad = \max_{\mathsf{m}\in\mathbb{R}^k} \left\{\left(\sqrt{\frac{\mathsf{D}^{\mathrm{CRN}}}{n}}\mathsf{v}\right)^\top \left(\sqrt{\frac{n}{\mathsf{D}^{\mathrm{CRN}}}}(\widehat{\mu} - \mathsf{m})\right) : \left\|\sqrt{\frac{n}{\mathsf{D}^{\mathrm{CRN}}}}(\widehat{\mu} - \mathsf{m})\right\|_{\widehat{\Sigma}^{-1}} \le 1\right\} \\
&\quad = \left\|\sqrt{\frac{\mathsf{D}^{\mathrm{CRN}}}{n}}\mathsf{v}\right\|_{\widehat{\Sigma}} \\
&\quad = \sqrt{\frac{\mathsf{D}^{\mathrm{CRN}}}{n}\mathsf{v}^\top \widehat{\Sigma} \mathsf{v}}. \quad \square
\end{aligned}$$

C Proofs of Theoretical Results

C.1 Proof of Conditions (C1)–(C4) for the d_{n}^2 standardized discrepancy with D^2

Condition (C1) is clearly satisfied for

$$d_{\mathsf{n}}^2(\mathsf{m}, \widehat{\mu}, \widehat{\Sigma}) = \sum_{i=1}^{k} \frac{n_i}{\widehat{\sigma}_i^2} \left(\widehat{\mu}_i - \mathsf{m}_i\right)^2.$$

Under the normality assumption for Setting (S1), $(\widehat{\mu}_i - \mu(x_i))/(\widehat{\sigma}_i/\sqrt{n_i}) \sim t_{n_i-1}$ for $i = 1, \ldots, k$, hence

$$d_{\mathsf{n}}^2(\mu(\mathsf{X}), \widehat{\mu}, \widehat{\Sigma}) = \sum_{i=1}^{k} \frac{n_i}{\widehat{\sigma}_i^2} (\widehat{\mu}_i - \mu(x_i))^2 \stackrel{d}{=} \sum_{i=1}^{k} F_{1,n_i-1}.$$

where t_ν denotes a t-distributed random variable with ν degrees of freedom and F_{ν_1,ν_2} denotes an F-distributed random variables with numerator degrees of freedom ν_1 and denominator degrees of freedom ν_2. The described value of D^2 is precisely the $1-\alpha$ quantile of $d_{\mathsf{n}}^2(\mu(\mathsf{X}), \widehat{\mu}, \widehat{\Sigma})$, i.e., Condition (C2) is satisfied.

By the Central Limit Theorem and the Continuous Mapping Theorem, $(\widehat{\mu}_i - \mu(x_i))^2/(\widehat{\sigma}_i^2/n_i) \sim \chi_1^2$ as $n_i \to \infty$ for $i = 1, \ldots, k$ where χ_ν^2 denotes a chi-squared random variable with ν degree of freedom. From our construction, D^2 converges to the $1-\alpha$ quantile of a χ_k^2 random variable. Therefore,

$$\mathbb{P}\left(d_{\mathsf{n}}^2(\mu(\mathsf{X}), \widehat{\mu}, \widehat{\Sigma}) \le \mathsf{D}^2\right) \to 1-\alpha \quad \text{as} \min_{i=1,\ldots,k} n_i \to \infty,$$

satisfying Condition (C3).

Lastly, we have that

$$\begin{aligned}
\max_{\mathsf{m}\in\mathbb{R}^k} \left\{ \|\widehat{\mu} - \mathsf{m}\| \colon d_{\mathsf{n}}^2(\mathsf{m}, \widehat{\mu}, \widehat{\Sigma}) \le \mathsf{D}^2 \right\} &= \max_{\mathsf{m}\in\mathbb{R}^k} \left\{ \sqrt{\sum_{i=1}^{k} \left(\widehat{\mu}_i - \mathsf{m}_i\right)^2} \colon \sum_{i=1}^{k} \frac{n_i}{\widehat{\sigma}_i^2} \left(\widehat{\mu}_i - \mathsf{m}_i\right)^2 \le \mathsf{D}^2 \right\} \\
&= \sqrt{\max_{\mathsf{m}\in\mathbb{R}^k} \left\{ \sum_{i=1}^{k} \left(\widehat{\mu}_i - \mathsf{m}_i\right)^2 \colon \sum_{i=1}^{k} \frac{n_i}{\widehat{\sigma}_i^2} \left(\widehat{\mu}_i - \mathsf{m}_i\right)^2 \le \mathsf{D}^2 \right\}} \\
&\le \sqrt{\sum_{i=1}^{k} \max_{\mathsf{m}_i\in\mathbb{R}} \left\{ \left(\widehat{\mu}_i - \mathsf{m}_i\right)^2 \colon \frac{n_i}{\widehat{\sigma}_i^2} \left(\widehat{\mu}_i - \mathsf{m}_i\right)^2 \le \mathsf{D}^2 \right\}} \\
&= \sqrt{\mathsf{D}^2 \sum_{i=1}^{k} \frac{\widehat{\sigma}_i^2}{n_i}} \stackrel{w.p.1}{\to} 0 \text{ as } \min_{i=1,\ldots,k} n_i \to \infty,
\end{aligned}$$

since $\widehat{\sigma}_i^2 \stackrel{w.p.1}{\to} \sigma_i^2 < \infty$ and D^2 converges to a constant. By Condition (C1),

$$\max_{\mathsf{m}\in\mathbb{R}^k} \left\{ \|\widehat{\mu} - \mathsf{m}\| \colon d_{\mathsf{n}}^2(\mathsf{m}, \widehat{\mu}, \widehat{\Sigma}) \le \mathsf{D}^2 \right\} \ge 0 \text{ with probability 1.}$$

It follows that

$$\max_{\mathsf{m}\in\mathbb{R}^k}\left\{\|\widehat{\mu}-\mathsf{m}\|\colon d_{\mathsf{n}}^2(\mathsf{m},\widehat{\mu},\widehat{\Sigma})\le \mathsf{D}^2\right\}\overset{w.p.1}{\to} 0 \text{ as } \min_{i=1,\dots,k} n_i\to\infty,$$

satisfying Condition (C4). □

C.2 Proof of Theorem 1

Fix arbitrary $\mu\in\mathscr{M}$. For any $x_0\in\mathcal{A}$, it follows by definition that $\mu\in\mathscr{M}(x_0)$, hence $\mu(\mathsf{X})\in\mathsf{M}(x_0)$. Then for any fixed n for which $\min_{i=1,\dots,k} n_i$ is sufficiently large,

$$\mathbb{P}(x_0\in\mathcal{S}_{\mathsf{n}}^{\mathrm{PS}})=\mathbb{P}\left(D_{\mathsf{n}}(x_0,\widehat{\mu},\widehat{\Sigma})\le\mathsf{D}\right)=\mathbb{P}\left(\min_{\mathsf{m}\in\mathsf{M}(x_0)} d_{\mathsf{n}}(\mathsf{m},\widehat{\mu},\widehat{\Sigma})\le\mathsf{D}\right)\ge\mathbb{P}\left(d_{\mathsf{n}}(\mu(\mathsf{X}),\widehat{\mu},\widehat{\Sigma})\le\mathsf{D}\right).$$

By Condition (C2),

$$\mathbb{P}\left(d_{\mathsf{n}}(\mu(\mathsf{X}),\widehat{\mu},\widehat{\Sigma})\le\mathsf{D}\right)=1-\alpha,$$

implying that $\mathcal{S}_{\mathsf{n}}^{\mathrm{PS}}$ achieves finite-sample confidence. Similarly, by Condition (C3),

$$\mathbb{P}\left(d_{\mathsf{n}}(\mu(\mathsf{X}),\widehat{\mu},\widehat{\Sigma})\le\mathsf{D}\right)\to 1-\alpha \quad \text{as } \min_{i=1,\dots,k} n_i\to\infty,$$

implying that $\mathbb{P}(x_0\in\mathcal{S}_{\mathsf{n}}^{\mathrm{PS}})\gtrsim 1-\alpha$ as $\min_{i=1,\dots,k} n_i\to\infty$. Hence $\mathcal{S}_{\mathsf{n}}^{\mathrm{PS}}$ achieves asymptotic confidence. □

C.3 Proof of Theorem 2

The result follows immediately from Corollary 1 and Theorem 4. For any $\mu\in\mathscr{M}$ and $x_0\notin\mathsf{S}(\mathsf{X})$,

$$\mathbb{P}(x_0\in\mathcal{S}_{\mathsf{n}}^{\mathrm{PS}})\le\mathbb{P}(x_0\in\mathcal{S}_{\mathsf{n}}^{\mathrm{RPS}})\to 0 \quad \text{as } \min_{i=1,\dots,k} n_i\to\infty. \quad \square$$

C.4 Proof of Lemma 1

Fix a suitable $\mathscr{M}$ and $\mathcal{A}$ satisfying Assumption 1, a performance function $\mu\in\mathscr{M}$, a solution $x_0\in\mathcal{X}$, and a vector $\widetilde{\mathsf{m}}\in\mathsf{R}(x_0)$. Consider the vector $\mathsf{m}^*\equiv\arg\min_{\mathsf{m}\in\mathsf{M}(x_0)} d_{\mathsf{n}}(\mathsf{m},\widetilde{\mathsf{m}},\widehat{\Sigma})$, for which $d_{\mathsf{n}}(\mathsf{m}^*,\widetilde{\mathsf{m}},\widehat{\Sigma})\le\mathsf{D}$, by definition of $\mathsf{R}(x_0)$. Since $\mathsf{m}^*\in\mathsf{M}(x_0)$, there is an associated $\mathsf{w}^*\in\mathbb{R}^q$ such that $A\mathsf{m}^*+C\mathsf{w}^*\le b$. Then for this w^*,

$$A\widetilde{\mathsf{m}}+C\mathsf{w}^*=A(\widetilde{\mathsf{m}}-\mathsf{m}^*)+A\mathsf{m}^*+C\mathsf{w}^*\le A(\widetilde{\mathsf{m}}-\mathsf{m}^*)+b\le b',$$

since for each $j=1,\dots p$,

$$a_j(\widetilde{\mathsf{m}}-\mathsf{m}^*)+b_j\le\max_{\mathsf{m}\in\mathbb{R}^k}\left\{a_j^\top(\widetilde{\mathsf{m}}-\mathsf{m})\colon d_{\mathsf{n}}(\mathsf{m},\widetilde{\mathsf{m}},\widehat{\Sigma})\le\mathsf{D}\right\}+b_j=b_j'.$$

Therefore $\widetilde{\mathsf{m}}\in\mathsf{R}'(x_0)$ and all together $\mathsf{R}(x_0)\subseteq\mathsf{R}'(x_0)$ with probability one. □

C.5 Proof of Corollary 1

Fix a suitable $\mathscr{M}$ and $\mathcal{A}$ satisfying Assumption 1, a performance function $\mu \in \mathscr{M}$, and a solution $x_0 \in \mathcal{S}_{\mathsf{n}}^{\mathrm{PS}}$. From the definition of $\mathcal{S}_{\mathsf{n}}^{\mathrm{PS}}$, $\widehat{\mu} \in \mathsf{R}(x_0)$ and by Lemma 1, $\widehat{\mu} \in \mathsf{R}'(x_0)$. Thus $x_0 \in \mathcal{S}_{\mathsf{n}}^{\mathrm{RPS}}$. Since the choice of x_0 was arbitrary, $\mathcal{S}_{\mathsf{n}}^{\mathrm{PS}} \subseteq \mathcal{S}_{\mathsf{n}}^{\mathrm{RPS}}$ with probability one. $\square$

C.6 Proof of Theorem 3

The result follows immediately from Corollary 1 and Theorem 1. For any $\mu \in \mathscr{M}$ and $x_0 \in \mathcal{A}$,

$$\mathbb{P}(x_0 \in \mathcal{S}_{\mathsf{n}}^{\mathrm{RPS}}) \geq \mathbb{P}(x_0 \in \mathcal{S}_{\mathsf{n}}^{\mathrm{PS}}) \geq 1 - \alpha,$$

and

$$\mathbb{P}(x_0 \in \mathcal{S}_{\mathsf{n}}^{\mathrm{RPS}}) \geq \mathbb{P}(x_0 \in \mathcal{S}_{\mathsf{n}}^{\mathrm{PS}}) \gtrsim 1 - \alpha \quad \text{as} \min_{i=1,\ldots,k} n_i \to \infty.$$

Thus $\mathcal{S}_{\mathsf{n}}^{\mathrm{RPS}}$ achieves finite-sample and asymptotic confidence. $\square$

C.7 Proof of Theorem 4

Fix arbitrary $\mu \in \mathscr{M}$. For any $x_0 \notin \mathsf{S}(\mathsf{X})$, it follows by definition that $\mu(\mathsf{X}) \notin \mathsf{M}(x_0)$, meaning that there does not exist a $\mathsf{w} \in \mathbb{R}^q$ s.t. $C\mathsf{w} \leq b - A\mu(\mathsf{X})$. By Farkas' Lemma, the optimal value

$$z^* \equiv \min_y (b - A\mu(\mathsf{X}))^\top y \quad \text{s.t. } C^\top y = 0 \text{ and } y \in \mathbb{R}^p_+ \tag{8}$$

is strictly negative, where $\mathbb{R}^p_+$ denotes the nonnegative orthant of $\mathbb{R}^p$. Thus, there exists $\bar{y} \in \mathbb{R}^p_+$ such that $C^\top \bar{y} = 0$ and $(b - A\mu(\mathsf{X}))^\top \bar{y} < 0$.

Recall that an arbitrary solution $x_0 \in \mathcal{X}$ belongs to $\mathcal{S}_{\mathsf{n}}^{\mathrm{RPS}}$ if and only if there exists a $\mathsf{w} \in \mathbb{R}^q$ such that $C\mathsf{w} \leq b' - A\widehat{\mu}$. As discussed in Section 5.3, this is equivalent to the event that $z_{\mathsf{n}} \geq 0$ where

$$z_{\mathsf{n}} \equiv \max_{\mathsf{w},\eta} \eta \text{ s.t. } C\mathsf{w} + \eta \mathbf{1}_p \leq b' - A\widehat{\mu}.$$

By Farkas' Lemma, it is also equivalent to the event that $z'_{\mathsf{n}} \geq 0$ where

$$z'_{\mathsf{n}} \equiv \min_y (b' - A\widehat{\mu})^\top y \text{ s.t. } C^T y = 0 \text{ and } y \in \mathbb{R}^p_+. \tag{9}$$

We proceed to show that the probability of the event that $z'_{\mathsf{n}} \geq 0$ goes to zero as the minimum sample size increases to infinity.

By the Cauchy-Schwarz Inequality and Condition (C4),

$$\begin{aligned} b'_j &= b_j + \max_{\mathsf{m} \in \mathbb{R}^k} \left\{ a_j^\top (\widehat{\mu} - \mathsf{m}) \colon d_{\mathsf{n}}(\mathsf{m}, \widehat{\mu}, \widehat{\Sigma}) \leq \mathsf{D} \right\} \\ &\leq b_j + \|a_j\| \max_{\mathsf{m} \in \mathbb{R}^k} \left\{ \|\widehat{\mu} - \mathsf{m}\| \colon d_{\mathsf{n}}(\mathsf{m}, \widehat{\mu}, \widehat{\Sigma}) \leq \mathsf{D} \right\} \overset{a.s.}{\to} b_j \text{ as } \min_{i=1,\ldots k} n_i \to \infty, \end{aligned}$$

for all $j = 1, \ldots, p$. Then since $b'_j \geq b_j$, we have that $b'_j \overset{a.s.}{\to} b_j$ for all $j = 1, \ldots, p$. In addition, $\widehat{\mu} \overset{a.s.}{\to} \mu(\mathsf{X})$ as $\min_{i=1,\ldots,k} n_i \to \infty$. The objective function in Definition (9) therefore converges

pointwise to that in Definition (8), while the feasible regions are the same: $\{y \in \mathbb{R}^p_+ \colon C^\top y = 0\}$.

For the optimization problem in Definition (9), consider the feasible solution $y_0 = \mathbf{0}_p$, whose associated objective function value is zero. This implies that to have $x_0 \in \mathcal{S}^{\text{RPS}}_{\mathsf{n}}$, we must have $z'_{\mathsf{n}} = 0$ and therefore y_0 would have to be an optimal solution to the optimization problem in Definition (9). The probability of this event goes to zero as $\min_{i=1,\ldots,k} n_i \to \infty$ because $(b' - A\widehat{\mu})^\top \bar{y} \overset{a.s.}{\to} (b - A\mu(\mathsf{X}))^\top \bar{y} < 0$, implying that the probability that $\bar{y}$ has a better objective function value than y_0 goes to one.

Combining these results,

$$\begin{aligned}
\mathbb{P}(x_0 \in \mathcal{S}^{\text{RPS}}_{\mathsf{n}}) &= \mathbb{P}(z_{\mathsf{n}} \geq 0) \\
&= \mathbb{P}(z'_{\mathsf{n}} \geq 0) \\
&= \mathbb{P}(y_0 = \mathbf{0}_p \text{ is an optimal solution to the problem in Definition (9)}) \\
&\to 0 \text{ as } \min_{i=1,\ldots,k} n_i \to \infty. \quad \square
\end{aligned}$$

C.8 Proof of Theorem 5

Fix a suitable $\mathscr{M}$, $\mu \in \mathscr{M}$, and $\mathcal{A}$ satisfying Assumption 1 and an arbitrary $x_0 \in \mathcal{X}$. Theorem 4.10 of Nemhauser and Wolsey [1999] implies that $\bar{a}_j = \nu_j A$ and $\bar{b}_j = \nu_j b$ for $j = 1, \ldots, \bar{p}$ where $\nu_1, \ldots, \nu_{\bar{p}}$ are the extreme rays of $Q \equiv \{\nu \in \mathbb{R}^p_+ \colon \nu C = 0\}$. Since $\mathsf{R}'(x_0) = \text{proj}_{\mathsf{m}}(\mathsf{P}')$ where $\mathsf{P}' \equiv \left\{(\mathsf{m}, \mathsf{w}) \in \mathbb{R}^k \times \mathbb{R}^q \colon A\mathsf{m} + C\mathsf{w} \leq b'\right\}$, the same result also implies that

$$\mathsf{R}'(x_0) = \left\{\mathsf{m} \in \mathbb{R}^k \colon \bar{A}\mathsf{m} \leq \bar{\bar{b}}\right\},$$

where $\bar{\bar{b}}_j = \nu_j b'$.

Fix arbitrary $\widetilde{\mathsf{m}} \in \overline{\mathsf{R}}'(x_0)$. For all $j = 1, \ldots, \bar{p}$,

$$\begin{aligned}
\bar{a}_j \widetilde{\mathsf{m}} &\leq \bar{b}'_j \\
&= \bar{b}_j + \max_{\mathsf{m} \in \mathbb{R}^k} \left\{\bar{a}_j^\top (\widehat{\mu} - \mathsf{m}) \colon d_{\mathsf{n}}(\mathsf{m}, \widehat{\mu}, \widehat{\Sigma}) \leq \mathsf{D}\right\} \\
&= \nu_j b + \max_{\mathsf{m} \in \mathbb{R}^k} \left\{\nu_j A(\widehat{\mu} - \mathsf{m}) \colon d_{\mathsf{n}}(\mathsf{m}, \widehat{\mu}, \widehat{\Sigma}) \leq \mathsf{D}\right\} \\
&= \nu_j b + \max_{\mathsf{m} \in \mathbb{R}^k} \left\{\sum_{\ell=1}^p \nu_{j\ell} a_\ell^\top (\widehat{\mu} - \mathsf{m}) \colon d_{\mathsf{n}}(\mathsf{m}, \widehat{\mu}, \widehat{\Sigma}) \leq \mathsf{D}\right\} \\
&\leq \nu_j b + \sum_{\ell=1}^p \max_{\mathsf{m} \in \mathbb{R}^k} \left\{\nu_{j\ell} a_\ell^\top (\widehat{\mu} - \mathsf{m}) \colon d_{\mathsf{n}}(\mathsf{m}, \widehat{\mu}, \widehat{\Sigma}) \leq \mathsf{D}\right\} \\
&= \nu_j b + \sum_{\ell=1}^p \nu_{j\ell} \max_{\mathsf{m} \in \mathbb{R}^k} \left\{a_\ell^\top (\widehat{\mu} - \mathsf{m}) \colon d_{\mathsf{n}}(\mathsf{m}, \widehat{\mu}, \widehat{\Sigma}) \leq \mathsf{D}\right\} \\
&= \nu_j b' = \bar{\bar{b}}_j.
\end{aligned}$$

The third-to-last equality follows from the fact that $\nu_j \in \mathbb{R}^p_+$ for all $j = 1, \ldots, \bar{p}$. Therefore $\widetilde{\mathsf{m}} \in \mathsf{R}'(x_0)$ and since $\widetilde{\mathsf{m}}$ was arbitrary, $\overline{\mathsf{R}}'(x_0) \subseteq \mathsf{R}'(x_0)$ with probability one. $\square$

Theorem 6 *For any $\mathscr{M}, \overline{\mathscr{M}} \in \mathscr{F}$ with $\overline{\mathscr{M}} \subseteq \mathscr{M}$, define*

$$\overline{\mathsf{M}}(x_0) \equiv \left\{\mathsf{m} \in \mathbb{R}^k \colon \textit{there exists } m \in \overline{\mathscr{M}} \textit{ such that } x_0 \in \mathcal{A}(m) \textit{ and } m(\mathsf{X}) = \mathsf{m}\right\}.$$

Then

$$\overline{\mathcal{S}}_{\mathsf{n}}^{\mathrm{PS}} \equiv \left\{x_0 \in \mathcal{X} \colon \overline{D}_{\mathsf{n}}(x_0, \widehat{\mu}, \widehat{\Sigma}) \leq \mathsf{D}\right\} \subseteq \mathcal{S}_{\mathsf{n}}^{\mathrm{PS}} \textit{ with probability one,}$$

where

$$\overline{D}_{\mathsf{n}}(x_0, \widehat{\mu}, \widehat{\Sigma}) \equiv \min_{\mathsf{m} \in \overline{\mathsf{M}}(x_0)} d_{\mathsf{n}}(\mathsf{m}, \widehat{\mu}, \widehat{\Sigma}).$$

C.9 Proof of Theorem 6

Fix arbitrary $\mathscr{M}, \overline{\mathscr{M}} \in \mathscr{F}$ with $\overline{\mathscr{M}} \subseteq \mathscr{M}$, arbitrary $\mu \in \overline{\mathscr{M}}$, and arbitrary $\mathcal{A}$. As the projections of intersections of function spaces $\overline{\mathsf{M}}(x_0) \subseteq \mathsf{M}(x_0)$. Hence for any solution $x_0 \in \mathcal{X}$,

$$\overline{D}_{\mathsf{n}}(x_0, \widehat{\mu}, \widehat{\Sigma}) = \min_{\mathsf{m} \in \overline{\mathsf{M}}(x_0)} d_{\mathsf{n}}(\mathsf{m}, \widehat{\mu}, \widehat{\Sigma}) \geq \min_{\mathsf{m} \in \mathsf{M}(x_0)} d_{\mathsf{n}}(\mathsf{m}, \widehat{\mu}, \widehat{\Sigma}) = D_{\mathsf{n}}(x_0, \widehat{\mu}, \widehat{\Sigma}),$$

implying that

$$\overline{\mathcal{S}}_{\mathsf{n}}^{\mathrm{PS}} = \left\{x_0 \in \mathcal{X} \colon \overline{D}_{\mathsf{n}}(x_0, \widehat{\mu}, \widehat{\Sigma}) \leq \mathsf{D}\right\} \subseteq \left\{x_0 \in \mathcal{X} \colon D_{\mathsf{n}}(x_0, \widehat{\mu}, \widehat{\Sigma}) \leq \mathsf{D}\right\} = \mathcal{S}_{\mathsf{n}}^{\mathrm{PS}} \text{ with probability one.} \quad \square$$

Theorem 7 *For any $\mathscr{M}, \overline{\mathscr{M}} \in \mathscr{F}$ with $\overline{\mathscr{M}} \subseteq \mathscr{M}$, define*

$$\overline{\mathsf{P}} \equiv \left\{(\mathsf{m}, \mathsf{w}, \mathsf{z}) \in \mathbb{R}^k \times \mathbb{R}^q \times \mathbb{R}^{\bar{q}} : A\mathsf{m} + C\mathsf{w} \leq b \textit{ and } \bar{A}\mathsf{m} + \overline{C}\mathsf{w} + \overline{E}\mathsf{z} \leq \bar{b}\right\},$$

for some $\bar{A} \in \mathbb{R}^{\bar{p} \times k}$, $\overline{C} \in \mathbb{R}^{\bar{p} \times q}$, $\overline{E} \in \mathbb{R}^{\bar{p} \times \bar{q}}$, and $\bar{b} \in \mathbb{R}^{\bar{p}}$. Then

$$\begin{aligned}\overline{\mathcal{S}}_{\mathsf{n}}^{\mathrm{RPS}} \equiv \big\{&x_0 \in \mathcal{X} \colon \textit{there exists } (\mathsf{w}, \mathsf{z}) \in \mathbb{R}^q \times \mathbb{R}^{\bar{q}} \textit{ such that } A\widehat{\mu} + C\mathsf{w} \leq b' \textit{ and } \bar{A}\widehat{\mu} + \overline{C}\mathsf{w} + \overline{E}\mathsf{z} \leq \bar{b}'\big\} \\ &\subseteq \mathcal{S}_{\mathsf{n}}^{\mathrm{RPS}} \textit{ with probability one,}\end{aligned}$$

where

$$\bar{b}'_j = \bar{b}_j + \max_{\mathsf{m} \in \mathbb{R}^k} \left\{\bar{a}_j^\top (\widehat{\mu} - \mathsf{m}) \colon d_{\mathsf{n}}(\mathsf{m}, \widehat{\mu}, \widehat{\Sigma}) \leq \mathsf{D}\right\} \textit{ for all } j = 1, \ldots, \bar{p},$$

where $\bar{a}_j$ is the jth row of $\bar{A}$, expressed as a column vector.

C.10 Proof of Theorem 7

Fix arbitrary $\mathscr{M}, \overline{\mathscr{M}} \in \mathscr{F}$ with $\overline{\mathscr{M}} \subseteq \mathscr{M}$, arbitrary $\mu \in \overline{\mathscr{M}}$, arbitrary $\mathcal{A}$, and a solution $x_0 \in \overline{\mathcal{S}}_{\mathsf{n}}^{\mathrm{RPS}}$. There exists an associated $\mathsf{w}^* \in \mathbb{R}^q$ and $\mathsf{z}^* \in \mathbb{R}^{\bar{q}}$ such that $A\widehat{\mu} + C\mathsf{w}^* \leq b'$ and $\bar{A}\widehat{\mu} + \overline{C}\mathsf{w}^* + \overline{E}\mathsf{z}^* \leq \bar{b}'$. Since, $A\widehat{\mu} + C\mathsf{w}^* \leq b'$, it follows that $\widehat{\mu} \in \mathsf{R}'(x_0)$, implying that $x_0 \in \mathcal{S}_{\mathsf{n}}^{\mathrm{RPS}}$. Because the choice of x_0 was arbitrary, $\overline{\mathcal{S}}_{\mathsf{n}}^{\mathrm{RPS}} \subseteq \mathcal{S}_{\mathsf{n}}^{\mathrm{RPS}}$ with probability one. $\square$

D Tightness of $\mathcal{S}_{\mathsf{n}}^{\mathrm{RPS}}$ for d_{n}^{∞} Standardized Discrepancy

Theorem 8 states that for the problem of minimizing a Lipschitz continuous function, Plausible Screening and Relaxed Plausible Screening return the same subset of solutions when using the d_{n}^{∞} standardized discrepancy.

Theorem 8 *For any $\mathcal{X}$, $\mathsf{X} \subseteq \mathcal{X}$, and fixed $0 < \gamma < \infty$, let*

$$\mathscr{M} = \{m \in \mathscr{F} \colon |m(x) - m(x')| \leq \gamma \|x - x'\| \text{ for all } x, x' \in \mathcal{X}\},$$

and for any $m \in \mathscr{M}$ let $\mathcal{A}(m) = \{x \in \mathcal{X} : m(x) \leq \min_{x' \in \mathcal{X}} m(x')\}$, so that for any $x_0 \in \mathcal{X}$,

$$\mathsf{M}(x_0) = \left\{\mathsf{m} \in \mathbb{R}^k \colon \mathsf{m}_i - \mathsf{m}_j \leq \gamma \min\left\{\|x_i - x_j\|, \|x_i - x_0\|\right\} \text{ for all } i, j = 1, \ldots, k\right\}.$$

Then for the $d_{\mathsf{n}}^{\infty}(\mathsf{m}, \widehat{\mu}, \widehat{\Sigma})$ standardized discrepancy, $\mathcal{S}_{\mathsf{n}}^{\mathrm{PS}} = \mathcal{S}_{\mathsf{n}}^{\mathrm{RPS}}$ with probability one.

D.1 Proof of Theorem 8

Our approach to establishing that $\mathcal{S}_{\mathsf{n}}^{\mathrm{PS}} \equiv \{x_0 \in \mathcal{X} \colon \widehat{\mu} \in \mathsf{R}(x_0)\}$ equals $\mathcal{S}_{\mathsf{n}}^{\mathrm{PS}} \equiv \{x_0 \in \mathcal{X} \colon \widehat{\mu} \in \mathsf{R}'(x_0)\}$ with probability one is to show that for each $x_0 \in \mathcal{X}$, $\mathsf{R}(x_0) = \mathsf{R}'(x_0)$ with probability one.

Fix $\mathcal{X}$, X, and an arbitrary $x_0 \in \mathcal{X}$ and $\widehat{\Sigma}$. For the d_{n}^{∞} standardized discrepancy,

$$\begin{aligned}
\mathsf{R}(x_0) &= \left\{\mathsf{m} \in \mathbb{R}^k \colon D_{\mathsf{n}}^{\infty}(x_0, \mathsf{m}, \widehat{\Sigma}) \leq \mathsf{D}^{\infty}\right\} \\
&= \left\{\mathsf{m} \in \mathbb{R}^k \colon \min_{\widetilde{\mathsf{m}} \in \mathsf{M}(x_0)} \max_{i=1,\ldots,k} \frac{\sqrt{n_i}}{\widehat{\sigma}_i} |\mathsf{m}_i - \widetilde{\mathsf{m}}_i| \leq \mathsf{D}^{\infty}\right\} \\
&= \left\{\mathsf{m} \in \mathbb{R}^k \colon \text{there exists } \widetilde{\mathsf{m}} \in \mathsf{M}(x_0) \text{ such that } \max_{i=1,\ldots,k} \frac{\sqrt{n_i}}{\widehat{\sigma}_i} |\mathsf{m}_i - \widetilde{\mathsf{m}}_i| \leq \mathsf{D}^{\infty}\right\} \\
&= \Big\{\mathsf{m} \in \mathbb{R}^k \colon \text{there exists } (\widetilde{\mathsf{m}}, \mathsf{w}) \in \mathbb{R}^k \times \mathbb{R}^q \text{ such that } A\widetilde{\mathsf{m}} + C\mathsf{w} \leq b \text{ and} \\
&\qquad \max_{i=1,\ldots,k} \frac{\sqrt{n_i}}{\widehat{\sigma}_i} |\mathsf{m}_i - \widetilde{\mathsf{m}}_i| \leq \mathsf{D}^{\infty}\Big\} \qquad (10) \\
&= \Big\{\mathsf{m} \in \mathbb{R}^k \colon \text{ there exists } (\Delta, \mathsf{w}) \in \mathbb{R}^k \times \mathbb{R}^q \text{ such that} \\
&\qquad A\mathsf{m} + C\mathsf{w} + A\Delta \leq b \text{ and } \frac{\sqrt{n_i}}{\widehat{\sigma}_i}|\Delta_i| \leq \mathsf{D}^{\infty} \text{ for all } i = 1, \ldots, k\Big\}, \qquad (11)
\end{aligned}$$

and

$$\begin{aligned}
\mathsf{R}'(x_0) &= \left\{\mathsf{m} \in \mathbb{R}^k \colon \text{ there exists } \mathsf{w} \in \mathbb{R}^q \text{ such that } A\mathsf{m} + C\mathsf{w} \leq b'\right\} \\
&= \left\{\mathsf{m} \in \mathbb{R}^k \colon \text{ there exists } \mathsf{w} \in \mathbb{R}^q \text{ such that } A\mathsf{m} + C\mathsf{w} \leq b + \mathsf{D}^{\infty}\underline{\mathrm{a}}\right\}, \qquad (12)
\end{aligned}$$

where $\underline{\mathrm{a}} = (\underline{\mathrm{a}}_1, \ldots, \underline{\mathrm{a}}_p)$ and

$$\underline{\mathrm{a}}_j = \sum_{i=1}^{k} \frac{\widehat{\sigma}_i}{\sqrt{n_i}} |a_{ji}| \text{ for all } j = 1, \ldots, p,$$

and p is the number of rows in A and C.

We proceed to show that $\mathsf{R}(x_0) = \mathsf{R}'(x_0)$ for this fixed $\widehat{\Sigma}$ by projecting out Δ from Equation (11). We use Fourier-Motzkin elimination to iteratively project out the variables $\Delta_1, \ldots, \Delta_k$, in any order, leading to an expression for $\mathsf{R}(x_0)$ in terms of only m and w that matches Equation (12).

Before carrying out the Fourier-Motzkin elimination procedure, we introduce some useful notation. For a given vector $v \in \mathbb{R}^k$ and subset $\mathcal{I} \subseteq \{1, \ldots, k\}$, let $v(\mathcal{I})$ represent the truncated vector whose elements correspond to indices in $\mathcal{I}$, i.e., the components with indices in $\{1, \ldots, k\} \backslash \mathcal{I}$ have been removed. And for a given subset $\mathcal{O} \subseteq \{1, \ldots, k\}$, let $\underline{\mathsf{a}}_j(\mathcal{O}) = \sum_{i \in \mathcal{O}} (\widehat{\sigma}_i / \sqrt{n_i}) |a_{ji}|$ for all $j = 1, \ldots, p$. Furthermore, for a given index $i^* = 1, \ldots, k$, define $\mathcal{J}^+(i^*) \equiv \{j = 1, \ldots, p \colon a_{ji^*} > 0\}$, $\mathcal{J}^-(i^*) \equiv \{j = 1, \ldots, p \colon a_{ji^*} < 0\}$, and $\mathcal{J}^0(i^*) \equiv \{j = 1, \ldots, p \colon a_{ji^*} = 0\}$. The subsets $\mathcal{J}^+(i^*)$, $\mathcal{J}^-(i^*)$, and $\mathcal{J}^0(i^*)$ form a partition of the indices $\{1, \ldots, p\}$ of the constraints $A\mathsf{m} + C\mathsf{w} + A\Delta \leq b$, classifying those for which Δ_{i^*} has a positive, negative, or zero coefficient, respectively. When carrying out Fourier-Motzkin elimination to project out Δ_{i^*}, these subsets will characterize which constraints can be re-expressed as lower bounds and upper bounds on Δ_{i^*}.

We prove via induction that iteratively projecting out the variables $\Delta_1, \ldots, \Delta_k$ from Equation (11) (in any order) using the Fourier-Motzkin elimination procedure will produce the desired result. For a more concise presentation, we find it easier to verify the two steps of an inductive proof in the opposite order: We first show a recursion for the case of projecting out an arbitrary variable Δ_{i^*} during the repeated procedure, assuming that after having projected out a subset of components of Δ, the current set of constraints describing $\mathsf{R}(x_0)$ has a particular representation. We then show that for the base case, projecting out the first of the k variables yields that same representation of constraints assumed in the inductive result.

Suppose we are on an arbitrary iteration of the Fourier-Motzkin elimination procedure where a subset of components of Δ have been projected out. Let $\mathcal{O}$ denote the subset of indices of these components and let $\mathcal{I} = \{1, \ldots, k\} \backslash \{\mathcal{O}\}$ denote the set of indices of components of Δ that have yet to be projected out. We assume that at this stage in the procedure, $\mathsf{R}(x_0)$ can be represented as

$$\begin{aligned} \mathsf{R}(x_0) = \Big\{ & \mathsf{m} \in \mathbb{R}^k \colon \text{ there exists } (\Delta(\mathcal{I}), \mathsf{w}) \in \mathbb{R}^{|\mathcal{I}|} \times \mathbb{R}^q \text{ such that} \\ & a_j^\top \mathsf{m} + c_j^\top \mathsf{w} + a_j^\top(\mathcal{I}) \Delta(\mathcal{I}) \leq b_j + \mathsf{D}^\infty \underline{\mathsf{a}}_j(\mathcal{O}) \text{ for all } j = 1, \ldots, p \text{ and} \\ & \frac{\sqrt{n_i}}{\widehat{\sigma}_i} |\Delta_i| \leq \mathsf{D}^\infty \text{ for all } i \in \mathcal{I} \Big\}. \end{aligned} \tag{13}$$

Suppose that for some $i^* \in \mathcal{I}$, Δ_{i^*} is the next variable to be projected out of Equation (13). Remove i^* from $\mathcal{I}$, i.e, set $\mathcal{I} \leftarrow \mathcal{I} \backslash \{i^*\}$. With this updated definition of $\mathcal{I}$, we rearrange the constraints in Equation (13) pertaining to Δ_{i^*} into either upper bounds

$$\begin{aligned} \Delta_{i^*} &\leq \frac{b_{j^+} + \mathsf{D}^\infty \underline{\mathsf{a}}_{j^+}(\mathcal{O}) - a_{j^+}^\top \mathsf{m} - c_{j^+}^\top \mathsf{w} - a_{j^+}(\mathcal{I})^\top \Delta(\mathcal{I})}{a_{j^+ i^*}} \text{ for all } j^+ \in \mathcal{J}^+(i^*) \\ \Delta_{i^*} &\leq \frac{\widehat{\sigma}_{i^*}}{\sqrt{n_{i^*}}} \mathsf{D}^\infty \end{aligned} \tag{14}$$

or lower bounds

$$\frac{b_{j^-} + \mathsf{D}^\infty \underline{\mathsf{a}}_{j^-}(\mathcal{O}) - a_{j^-}^\top \mathsf{m} - c_{j^-}^\top \mathsf{w} - a_{j^-}(\mathcal{I})^\top \Delta(\mathcal{I})}{a_{j^- i^*}} \leq \Delta_{i^*} \text{ for all } j^- \in \mathcal{J}^-(i^*)$$

$$-\frac{\widehat{\sigma}_{i^*}}{\sqrt{n_{i^*}}} \mathsf{D}^\infty \leq \Delta_{i^*}. \tag{15}$$

Those constraints in Equation (13) that do not feature Δ_i, namely

$$a_j^\top \mathsf{m} + c_j^\top \mathsf{w} + a_j^\top(\mathcal{I}) \Delta(\mathcal{I}) \leq b_j + \mathsf{D}^\infty \underline{\mathsf{a}}_j(\mathcal{O}) \text{ for all } j \in \mathcal{J}^0(i^*) \tag{16}$$

$$\frac{\sqrt{n_i}}{\widehat{\sigma}_i} |\Delta_i| \leq \mathsf{D}^\infty \text{ for all } i \in \mathcal{I},$$

are temporarily set aside, but will be included in the representation of $\mathsf{R}(x_0)$ at the end of the iteration.

Fourier Motzkin elimination replaces Constraints (14) and (15) with the constraints

$$-\frac{\widehat{\sigma}_{i^*}}{\sqrt{n_{i^*}}} \mathsf{D}^\infty \leq \frac{b_{j^+} + \mathsf{D}^\infty \underline{\mathsf{a}}_{j^+}(\mathcal{O}) - a_{j^+}^\top \mathsf{m} - c_{j^+}^\top \mathsf{w} - a_{j^+}(\mathcal{I})^\top \Delta(\mathcal{I})}{a_{j^+ i^*}} \text{ for all } j^+ \in \mathcal{J}^+(i^*) \tag{17}$$

$$\frac{b_{j^-} + \mathsf{D}^\infty \underline{\mathsf{a}}_{j^-}(\mathcal{O}) - a_{j^-}^\top \mathsf{m} - c_{j^-}^\top \mathsf{w} - a_{j^-}(\mathcal{I})^\top \Delta(\mathcal{I})}{a_{j^- i^*}} \leq \frac{\widehat{\sigma}_{i^*}}{\sqrt{n_{i^*}}} \mathsf{D}^\infty \text{ for all } j^- \in \mathcal{J}^-(i^*) \tag{18}$$

$$\frac{b_{j^-} + \mathsf{D}^\infty \underline{\mathsf{a}}_{j^-}(\mathcal{O}) - a_{j^-}^\top \mathsf{m} - c_{j^-}^\top \mathsf{w} - a_{j^-}(\mathcal{I})^\top \Delta(\mathcal{I})}{a_{j^- i^*}} \leq \frac{b_{j^+} + \mathsf{D}^\infty \underline{\mathsf{a}}_{j^+}(\mathcal{O}) - a_{j^+}^\top \mathsf{m} - c_{j^+}^\top \mathsf{w} - a_{j^+}(\mathcal{I})^\top \Delta(\mathcal{I})}{a_{j^+ i^*}}$$

$$\text{for all } j^+ \in \mathcal{J}^+(i^*) \text{ and } j^- \in \mathcal{J}^-(i^*). \tag{19}$$

Due to the positive sign of $a_{j^+ i^*}$ and the negative sign of $a_{j^- i^*}$, Constraints (17) and (18) can be succinctly written as

$$a_j^\top \mathsf{m} + c_j^\top \mathsf{w} + a_j(\mathcal{I})^\top \Delta(\mathcal{I}) \leq b_j + \mathsf{D}^\infty \underline{\mathsf{a}}_j(\mathcal{O} \cup \{i^*\}) \text{ for all } j \in \mathcal{J}^+(i^*) \cup \mathcal{J}^-(i^*).$$

We next show that Constraints (19) are redundant for the setup of minimizing a Lipschitz continuous function. Fix an arbitrary $j^+ \in \mathcal{J}^+(i^*)$ and $j^- \in \mathcal{J}^-(i^*)$. From the formulation

$$\mathsf{M}(x_0) = \left\{ \mathsf{m} \in \mathbb{R}^k \colon \mathsf{m}_i - \mathsf{m}_j \leq \gamma \min\left\{ \|x_i - x_j\|, \|x_i - x_0\| \right\} \text{ for all } i, j = 1, \ldots, k \right\},$$

the corresponding constraint among Constraints (19) can be expressed as

$$-\gamma \min\{\|x_\ell - x_{i^*}\|, \|x_\ell - x_0\|\} - \mathsf{D}^\infty \frac{\widehat{\sigma}_\ell}{\sqrt{n_\ell}} \mathbf{1}\{\ell \in \mathcal{O}\} + \mathsf{m}_\ell - \mathsf{m}_{i^*} + \Delta_\ell \mathbf{1}\{\ell \in \mathcal{I}\}$$

$$\leq \gamma \min\{\|x_{i^*} - x_{\ell'}\|, \|x_{i^*} - x_0\|\} + \mathsf{D}^\infty \frac{\widehat{\sigma}_{\ell'}}{\sqrt{n_{\ell'}}} \mathbf{1}\{\ell' \in \mathcal{O}\} - \mathsf{m}_{i^*} + \mathsf{m}_{\ell'} + \Delta_{\ell'} \mathbf{1}\{\ell' \in \mathcal{I}\} \tag{20}$$

for some $\ell, \ell' \in \{1, \ldots, k\} \backslash \{i^*\}$. Inequality (20) reduces to

$$\begin{aligned}
&\mathsf{m}_\ell - \mathsf{m}_{\ell'} + \Delta_\ell \mathbf{1}\{\ell \in \mathcal{I}\} - \Delta_{\ell'} \mathbf{1}\{\ell' \in \mathcal{I}\} \\
&\leq \gamma \min\{\|x_\ell - x_{i^*}\|, \|x_\ell - x_0\|\} + \gamma \min\{\|x_{i^*} - x_{\ell'}\|, \|x_{i^*} - x_0\|\} \\
&\quad + \mathsf{D}^\infty \frac{\widehat{\sigma}_\ell}{\sqrt{n_\ell}} \mathbf{1}\{\ell \in \mathcal{O}\} + \mathsf{D}^\infty \frac{\widehat{\sigma}_{\ell'}}{\sqrt{n_{\ell'}}} \mathbf{1}\{\ell' \in \mathcal{O}\}. \qquad (21)
\end{aligned}$$

We next show that Inequality (21) is implied by a constraint

$$a_{j^*}^\top \mathsf{m} + c_{j^*}^\top \mathsf{w} + a_{j^*}(\mathcal{I})^\top \Delta(\mathcal{I}) \leq b_{j^*} + \mathsf{D}^\infty \underline{\mathrm{a}}_{j^*}(\mathcal{O} \cup \{i^*\}) \qquad (22)$$

for some $j^* \in \mathcal{J}^0(i^*)$, i.e., one of the constraints in (16). In particular, we consider j^* corresponding to the constraint

$$\mathsf{m}_\ell - \mathsf{m}_{\ell'} \leq \gamma \min\{\|x_\ell - x_{\ell'}\|, \|x_\ell - x_0\|\}$$

in the formulation of $\mathsf{M}(x_0)$. For this choice of j^*, Constraint (22) can be written as

$$\begin{aligned}
\mathsf{m}_\ell - \mathsf{m}_{\ell'} + \Delta_\ell \mathbf{1}\{\ell \in \mathcal{I}\} - \Delta_{\ell'} \mathbf{1}\{\ell' \in \mathcal{I}\} &\leq \gamma \min\{\|x_\ell - x_{\ell'}\|, \|x_\ell - x_0\|\} \\
&\quad + \mathsf{D}^\infty \frac{\widehat{\sigma}_\ell}{\sqrt{n_\ell}} \mathbf{1}\{\ell \in \mathcal{O}\} + \mathsf{D}^\infty \frac{\widehat{\sigma}_{\ell'}}{\sqrt{n_{\ell'}}} \mathbf{1}\{\ell' \in \mathcal{O}\}. \qquad (23)
\end{aligned}$$

By the triangle inequality,

$$\gamma \min\{\|x_\ell - x_{\ell'}\|, \|x_\ell - x_0\|\} \leq \gamma \min\{\|x_\ell - x_{i^*}\|, \|x_\ell - x_0\|\} + \gamma \min\{\|x_{i^*} - x_{\ell'}\|, \|x_{i^*} - x_0\|\},$$

hence Constraint (23) implies Constraint (21). Constraint (21) is therefore redundant and can be dropped from the formulation.

Since the choice of j^+ and j^- were arbitrary, we have altogether that

$$\begin{aligned}
\mathsf{R}(x_0) = \Bigg\{ &\mathsf{m} \in \mathbb{R}^k\text{: there exists } (\Delta(\mathcal{I}), \mathsf{w}) \in \mathbb{R}^{|\mathcal{I}|} \times \mathbb{R}^q \text{ such that} \\
&a_j^\top \mathsf{m} + c_j^\top \mathsf{w} + a_j^\top(\mathcal{I}) \Delta(\mathcal{I}) \leq b_j + \mathsf{D}^\infty \underline{\mathrm{a}}_j(\mathcal{O}) \text{ for all } j \in \mathcal{J}^0(i^*), \\
&a_j^\top \mathsf{m} + c_j^\top \mathsf{w} + a_j(\mathcal{I})^\top \Delta(\mathcal{I}) \leq b_j + \mathsf{D}^\infty \underline{\mathrm{a}}_j(\mathcal{O} \cup \{i^*\}) \text{ for all } j \in \mathcal{J}^+(i) \cup \mathcal{J}^-(i) \text{ and} \\
&\frac{\sqrt{n_i}}{\widehat{\sigma}_i} |\Delta_i| \leq \mathsf{D}^\infty \text{ for all } i \in \mathcal{I} \Bigg\}.
\end{aligned}$$

By adding i^* to $\mathcal{O}$, i.e., setting $\mathcal{O} \leftarrow \mathcal{O} \cup \{i^*\}$, we can represent $\mathsf{R}(x_0)$ by

$$\begin{aligned}
\mathsf{R}(x_0) = \Bigg\{ &\mathsf{m} \in \mathbb{R}^k\text{: there exists } (\Delta(\mathcal{I}), \mathsf{w}) \in \mathbb{R}^{|\mathcal{I}|} \times \mathbb{R}^q \text{ such that} \\
&a_j^\top \mathsf{m} + c_j^\top \mathsf{w} + a_j^\top(\mathcal{I}) \Delta(\mathcal{I}) \leq b_j + \mathsf{D}^\infty \underline{\mathrm{a}}_j(\mathcal{O}) \text{ for all } j = 1, \ldots, p \text{ and} \\
&\frac{\sqrt{n_i}}{\widehat{\sigma}_i} |\Delta_i| \leq \mathsf{D}^\infty \text{ for all } i \in \mathcal{I} \Bigg\},
\end{aligned}$$

which matches Equation (13). Therefore we have proven the induction step.

The base case is then easily established by observing that for $\mathcal{I} = \{1, \ldots, k\}$ and $\mathcal{O} = \emptyset$, Equation (13) is precisely Equation (11). Additionally, for $\mathcal{I} = \emptyset$ and $\mathcal{O} = \{1, \ldots, k\}$, corresponding to having projected out all components of Δ, Equation (13) is precisely Equation (12). Since $\widehat{\Sigma}$ was fixed arbitrarily, $\mathsf{R}(x_0) = \mathsf{R}'(x_0)$ with probability one, and since the choice of x_0 was also arbitrary, $\mathcal{S}_{\mathsf{n}}^{\mathrm{PS}} = \mathcal{S}_{\mathsf{n}}^{\mathrm{RPS}}$ with probability one. $\square$