Submitted to *Operations Research*
manuscript (Please, provide the manuscript number!)

Optimization-based Calibration of Simulation Input Models

Aleksandrina Goeva
Broad Institute, Cambridge, MA 02142, USA. agoeva@broadinstitute.org

Henry Lam
Department of Industrial Engineering and Operations Research, Columbia University, New York, NY 10027, USA. henry.lam@columbia.edu

Huajie Qian
Department of Industrial Engineering and Operations Research, Columbia University, New York, NY 10027, USA. hq2157@columbia.edu

Bo Zhang
IBM Research AI, Yorktown Heights, NY 10598, USA. zhangbo@us.ibm.com

Studies on simulation input uncertainty often built on the availability of input data. In this paper, we investigate an inverse problem where, given only the availability of output data, we nonparametrically calibrate the input models and other related performance measures of interest. We propose an optimization-based framework to compute statistically valid bounds on input quantities. The framework utilizes constraints that connect the statistical information of the real-world outputs with the input-output relation via a simulable map. We analyze the statistical guarantees of this approach from the view of data-driven distributionally robust optimization, and show how they relate to the function complexity of the constraints arising in our framework. We investigate an iterative procedure based on a stochastic quadratic penalty method to approximately solve the resulting optimization. We conduct numerical experiments to demonstrate our performances in bounding the input models and related quantities.

Key words: model calibration; distributionally robust optimization; stochastic simulation; input modeling

1. Introduction

Stochastic simulation takes in input models and generates random outputs for subsequent performance analyses. The accuracy of these input model assumptions is critical to the analyses' credibility. In the conventional premise in studying stochastic simulation, these input models are conferred either through physical implication or expert opinions, or observable via input data. In this paper, we answer a converse question: Given only *output* data from a stochastic system, can one infer about the input model?

The main motivation for asking this question is that, in many situations, a simulation modeler plainly may not have the luxury of direct data or knowledge about the input. The only way to gain such knowledge could be data from other sources that are at the output level. For instance, one of the authors has experienced such complication when building a simulation model for a contract fulfillment center, where service agents work on a variety of processing tasks and, despite the abundant transaction data stored in the center's IT system, there is no record on the start, completion, or service times spent by each agent on each particular task. Similarly, in clinic operations, patients often receive service in multiple phases such as initial checkup, medical tests and doctor's consultation. Patients' check-in and check-out times could be accurately noted, but the "service" times provided by the medical staff could very well be unrecorded. Clearly, these service time distributions are needed to build a simulation model, if an analyst wants to use the model for sensitivity analysis or system optimization purposes.

The problem of inferring an input model from output data is sometimes known as *model calibration*. In the simulation literature, this is often treated as a refinement process that occurs together with iterative comparisons between simulation reports and real-world output data (a task known as *model validation*; Sargent (2005), Kleijnen (1995)). If simulation reports differ significantly from output data, the simulation model is re-calibrated (which can involve both the input distributions and system specifications), re-compared, and the process is iterated. Suggested approaches to compare simulation with real-world data include conducting statistical tests such as two-sample

mean-difference tests (Balci and Sargent (1982)) and the Schruben-Turing test (Schruben (1980)). Beyond that, inferring input from output seems to be an important problem that has not been widely discussed in the stochastic simulation literature (Nelson (2016)).

The setting we consider can be briefly described as follows. We assume an input model is missing and make no particular assumptions on the form of its probability distribution. We assume, however, that a certain output random variable from a well-specified system is observable with some data. Our task is to nonparametrically infer the input distribution, or other quantities related to this input distribution (e.g., a second output measure driven by the same input distribution). One distinction between our setting and model calibration in other literature (e.g., computer experiments) is the intrinsic probabilistic structure of the system. Namely, the input and the output in stochastic simulation are represented as probability distributions, or in other words, the relation that links the observed and the to-be-calibrated objects is a (simulable) map between the spaces of distributions. Our calibration method will be designed to take such a relation into account.

Specifically, we study an optimization-based framework for model calibration, where the optimization, on a high level, entails an objective function associated with the "input" and constraints associated with the "output". The decision variable in this optimization is the unknown input distribution. The constraints comprise a confidence region on the the output distribution that is compiled from the observed output statistics. By expressing the region in terms of the input distributions via the simulable map, the optimization objective, which is set to be some target input quantity, will then give rise to statistically valid confidence bounds on this target. Advantageously, this approach leads to valid bounds even if the input model is *non-identifiable*, i.e., there exist more than one input model that give rise to the same observable output pattern, which may occur since the simulable map is typically highly complicated. The tightness of the bounds in turn depends on the degree of non-identifiability (which also leads to a notion of *identifiability gap* that we will discuss). The idea of utilizing a confidence region as the constraint is inspired by distributionally robust optimization (DRO). However, in the conventional DRO literature, the constraints (often

called collectively as the uncertainty set or the ambiguity set) are constructed based on direct observation of data. On the other hand, our constraints here serve as a tool to integrate the input-output relation, in addition to the output-level statistical noise, to effectively calibrate the input model. This leads to several new methodological challenges and solution approaches.

Under this general framework, we propose a concrete optimization formulation that balances statistical validity and the required computational efforts. Specifically, we use a nonparametric statistic, namely the Kolmogorov-Smirnov (KS) statistic, to construct the output-level confidence region. This formulation has the strengths of being statistically consistent (implied by the KS statistic) and expressible as expectation-type constraints that can be effectively solved by our subsequent algorithms. It also has an interesting additional benefit in terms of controlling the dimension of the optimization. Because of computational capacity, the decision variable, which is the unknown input distribution and potentially infinite-dimensional, needs to be suitably discretized by randomly generating a finite number of support points. A consistent statistic typically induces a large number of constraints, and one may need to use a large number of support points to retain the discretization error. However, as will be seen, it turns out that the KS constraints allow us to use a moderate support size without compromising the asymptotic statistical guarantees, thanks to their low complexity as measured by the so-called bracketing number in the empirical process theory (Van Der Vaart and Wellner (1996), Arcones and Gine (1993)). This thus leads us to an optimization problem with both a controllable number of decision variables and statistical validity.

Next, due to the sophisticated input-output map, the optimization programs generally involve non-convex stochastic (i.e., simulation-based) constraints. We propose and analyze a stochastic quadratic penalty method, by adding a growing penalty on the squared constraint violation. This method borrows from the quadratic penalty method used in deterministic nonlinear programming. However, while the deterministic version suggests solving a nonlinear program at each particular value of the penalty coefficient and letting the coefficient grows, the stochastic method we analyze involves a stochastic approximation (SA) that runs updates of the solution, slack variables and

the penalty coefficient simultaneously. This is motivated from the typical challenge of finding good stopping times for SA, which are needed for each SA run at each penalty coefficient value if one were to mimic the deterministic procedure. Simultaneous updates of all the quantities, however, only need one SA run. We analyze the convergence guarantee of this algorithm and provide guidance on the step sizes of all the constituent updates. Our SA update uses a mirror descent stochastic approximation (MDSA) (Nemirovski et al. (2009)), in particular the entropic descent (Beck and Teboulle (2003)).

The remainder of the paper is organized as follows. Section 2 reviews the related literature. Section 3 introduces the problem setting and presents our general optimization-based framework. Section 4 refines our framework with the KS-based formulations and demonstrates the statistical guarantees. Section 5 presents and analyzes our optimization algorithm. Section 6 reports numerical results. Section 7 concludes. The Appendix contains all the proofs.

2. Related literature

We organize the literature review in two aspects, one related to the model calibration problem, and one related to our optimization approach.

2.1. Literature Related to Our Problem Setting

Input modeling and uncertainty quantification in the stochastic simulation focus mostly on the input level. Barton (2012) and Song et al. (2014), e.g., review some major methods in quantifying the statistical errors from finite input data. These methods include the delta or two-point method (Cheng and Holland (1998, 2004)), Bayesian methodology and model averaging (Chick (2001), Zouaoui and Wilson (2004)) and resampling methods (Barton and Schruben (2001), Barton et al. (2013)). Our problem is more related to model calibration. In the simulation literature, this is often considered together with model validation (Sargent (2005), Kleijnen (1995)). Conventional approaches compare simulation data with real-world historical output data according to statistical or Turing tests (Balci and Sargent (1982), Schruben (1980)), conduct re-calibration, and repeat the process until the data are successfully validated (Banks et al. (2009), Kelton and Law (2000)).

The model calibration problem is also known as the *inverse problem* (Tarantola (2005)) in the literature of other fields. It generally refers to the identification of parameters or functions that can only be inferred from transformed outputs. In the context where the parameters are probability distributions, Kraan and Bedford (2005) demonstrates theoretically the characterization of a distribution that leads to the smallest relative entropy with a reference measure, and proposes an entropy maximization to calibrate the distribution from output data. Our work relates to Kraan and Bedford (2005) as we also utilize a probabilistic input-output map, but we focus on maps that are evaluable only by simulation, and aim to compute confidence bounds on the true distribution instead of attempting to recover the maximum entropy distribution.

The inverse problem also appeared in many other contexts. In signal processing, the linear inverse problem (e.g., Csiszár (1991), Donoho et al. (1992)) reconstructs signals from measurements of linear transformations. Common approaches consist of least-square minimization and the use of penalty such as the entropy. In computer experiments (Santner et al. (2013)), surrogate models built on complex physical laws require the calibration of physical parameters. Such models have wide scientific applications such as weather prediction, oceanography, nuclear physics, and acoustics (e.g., Wunsch (1996), Shirangi (2014)). Bayesian and Gaussian process methodologies are commonly used (e.g., Kennedy and O'Hagan (2001), Currin et al. (1991)). We point out that Bayesian methods could be a potential alternative to the approach considered in this paper, but because of the nature of discrete-event systems, one might need to resort to sophisticated techniques such as approximate Bayesian computation (Marjoram et al. (2003)). Other related literature include experimental design to optimize inference for input parameters (e.g., Chick and Ng (2002)) and calibrating financial option prices (e.g., Avellaneda et al. (2001), Glasserman and Yu (2005)).

Also related to our work is the body of research on inference problems in the context of queueing systems. The first stream, similar to our paper, aims at inferring the constituent probability distributions of a queueing model based on its output data, e.g., queue length or waiting time data, collected either continuously or at discrete time points. This stream of papers focuses on

systems whose structures allow closed-form analyses or are amenable to analytic approximations via, for instance, the diffusion limit. The majority of them assume that the inferred distribution(s) comes from a parametric family and use maximum likelihood estimators (Basawa et al. (1996), Pickands III and Stine (1997), Basawa et al. (2008), Fearnhead (2004), Wang et al. (2006), Ross et al. (2007), Heckmüller and Wolfinger (2009), Whitt (2012)). Others work on nonparametric inference by exploiting specific queueing system structures (Bingham and Pitts (1999), Hall and Park (2004), Moulines et al. (2007), Feng et al. (2014)). A related stream of literature studies point process approximation (see Section 4.7 of Cooper (1972), Whitt (1981, 1982), and the references therein), based on a parametric approach and is motivated from traffic pattern modeling in communication networks. Finally, there are also a number of studies inspired by the "queue inference engine" by Larson (1990). But, instead of inferring the input models, many of these studies use transaction data to estimate the performance of a queueing system directly and hence do not take on the form of an inverse problem (see Mandelbaum and Zeltyn (1998) for a good survey of the earlier literature and Frey and Kaplan (2010) and its references for more recent progress). Several papers estimate both the queueing operational performance and the constituent input models (e.g., Daley and Servi (1998), Kim and Park (2008), Park et al. (2011)), and can be considered to belong to both this stream and the aforementioned first stream of literature.

2.2. Literature Related to Our Methodology

Our formulation uses ideas from robust optimization (e.g., Bertsimas et al. (2011), Ben-Tal et al. (2009)), which studies optimization under uncertain parameters and suggests to obtain decisions that optimize the worst-case scenarios, subject to a set of constraints on the belief/uncertainty that is often known as the ambiguity set or the uncertainty set. Of particular relevance to us is the setting of distributionally robust optimization (DRO), where the uncertainty is on the probability distribution in a stochastic optimization problem (e.g., Delage and Ye (2010), Wiesemann et al. (2014), Ben-Tal et al. (2013)). This approach has been applied in many disciplines such as stochastic control (e.g., Petersen et al. (2000), Xu and Mannor (2012), Iyengar (2005)), economics (Hansen

and Sargent (2008)), finance (Glasserman and Xu (2013)), queueing (Jain et al. (2010)), inventory control (Xin and Goldberg (2015)), power systems (Zhang et al. (2017), Xie and Ahmed (2018), Zhao and Jiang (2018)) and dynamic pricing (Lim and Shanthikumar (2007)). Its connection to machine learning and statistics has also been investigated (Shafieezadeh-Abadeh et al. (2015), Blanchet et al. (2016)). In the DRO literature, common choices of the uncertainty set are based on moments (Delage and Ye (2010), Goh and Sim (2010), Wiesemann et al. (2014), Bertsimas and Popescu (2005), Smith (1995), Bertsimas and Natarajan (2007)), distances from nominal distributions (Ben-Tal et al. (2013), Bayraksan and Love (2015), Blanchet and Murthy (2016), Esfahani and Kuhn (2015), Gao and Kleywegt (2016)), and shape conditions (Popescu (2005), Lam and Mottet (2017), Li et al. (2016), Hanasusanto et al. (2017)). The literature of data-driven DRO further addresses the question of calibrating these sets, using for instance confidence regions or hypothesis testing (Bertsimas et al. (2014)), empirical likelihood (Lam and Zhou (2017), Duchi et al. (2016), Lam (2016a)), relatedly the Wasserstein profile inference (Blanchet et al. (2016)), and Bayesian perspectives (Gupta (2015)).

For DRO in the simulation context, Hu et al. (2012) studies the computation of robust bounds under Gaussian model assumptions, Glasserman and Xu (2014), Glasserman and Yang (2016) study distance-based constraints to address model risks, Lam (2016b, 2017) study asymptotic approximations for related formulations, and Ghosh and Lam (2015c) studies performance guarantees and solution techniques in quantifying simulation input uncertainty. Fan et al. (2013), Ryzhov et al. (2012) study the use of robust optimization in simulation-based decision-making. Our framework in particular follows the concept in using confidence region such that the uncertainty set covers the true distribution with high probability. However, it also involves the simulation map between input and output that serves as the key in our model calibration goal.

Our optimization procedure builds on the quadratic penalty method (Bertsekas (1999)), which is a deterministic nonlinear programming technique that reformulates the constraints as squared penalty and sequentially tunes the penalty coefficient to approach optimality. Different from the

deterministic technique, our procedure in solving the stochastic quadratic penalty formulation sequentially updates the penalty parameter simultaneously together with the solution and slack variables. This involves a specialized version of MDSA proposed by Nemirovski et al. (2009), who analyzed convergence guarantees on convex programs with stochastic objectives. Lan and Zhou (2017), Yu et al. (2017) investigated convex stochastic constraints, and Ghadimi and Lan (2013, 2015), Dang and Lan (2015), Ghadimi et al. (2016) studied related schemes for nonconvex and nonsmooth objectives. Wang and Spall (2008) introduced a quadratic penalty method for stochastic objectives with deterministic constraints. The particular scheme of MDSA we consider uses entropic penalty, and is known as the entropic descent algorithm (Beck and Teboulle (2003)).

3. Proposed Framework

Consider a generic input variate X with an input probability distribution P_X. We let $\mathbf{X} = (X_1, \ldots, X_T)$, where $X_t \in \mathcal{X}$, be an i.i.d. sequence of input variates each distributed under P_X over a time horizon T. We denote the function $h(\cdot) \in \mathbb{R}$ as the system logic from the input sequence $\mathbf{X}$ to the output $h(\mathbf{X})$. We assume that h is completely specified and is computable, even though it may not be writable in closed-form, i.e. we can evaluate the output given $\mathbf{X}$. For example, $\mathbf{X}$ can denote the sequence of interarrival or service times for the customers in a queue, and $h(\mathbf{X})$ is an average queue length seen by the T customers. Note that we can work in a more general framework where h depends on both $\mathbf{X}$ and other independent input sequences, denoted collectively as $\mathbf{W}$, that possess known or observable distributions. In other words, we can have $h(\mathbf{X}, \mathbf{W})$ as the output. Our developments can readily handle this case, but for expositional convenience we will assume the absence of these auxiliary input sequences most of the time, and will indicate the modifications of our developments in handling them at various suitable places.

Consider the situation that only $h(\mathbf{X})$ can be observed via data. Let $D = \{y_1, \ldots, y_n\}$ be n observations of $h(\mathbf{X})$. Our task is to calibrate some quantities related to P_X, which we call $\psi(P_X)$. Two types of target quantities we will consider are:

1. Restricting X to real value, we consider the distribution function of P_X, denoted $F_X(x)$, where x can take a range of values. Note that, obviously, $F_X(x) = E_{P_X}[I(X \leq x)]$ where $E_{P_X}[\cdot]$ denotes the expectation with respect to P_X and $I(\cdot)$ denotes the indicator function.
2. We consider a performance measure $E_{P_X}[g(\mathbf{X})]$ where $E_{P_X}[\cdot]$ here denotes the expectation with respect to the product measure induced by the i.i.d. sequence $\mathbf{X} = (X_1, \ldots, X_S)$ over a time horizon S. The function $g(\mathbf{X})$ can denote another output of interest different from $h(\mathbf{X})$ that is unobservable, and requires information about $\mathbf{X}$. This case includes the first target quantity above (by choosing $g(\mathbf{X}) = I(X_1 \leq x)$ when X is real-valued), as well as other statistics of X such as power moments (by choosing $g(\mathbf{X}) = X_1^k$ for some k).

To describe our framework, we denote $P_Y = P_{h(\mathbf{X})}$ as the probability distribution of the output $Y = h(\mathbf{X})$. Since P_Y is completely identified by P_X, we can view P_Y as a transformation of P_X, i.e., $P_Y = \gamma(P_X)$ for some map γ between probability distributions. We denote $\mathcal{P}_X$ and $\mathcal{P}_Y$ as the spaces of all possible input and output distributions respectively.

On an abstract level, we use the optimization formulations

$$\begin{aligned} &\max && \psi(P_X) \\ &\text{subject to} && P_Y \in \mathcal{U} \end{aligned} \tag{1}$$

and

$$\begin{aligned} &\min && \psi(P_X) \\ &\text{subject to} && P_Y \in \mathcal{U} \end{aligned} \tag{2}$$

where the decision variable is the unknown $P_X \in \mathcal{P}_X$, and $\mathcal{U} \subset \mathcal{P}_Y$ is an "uncertainty set" that covers a set of possibilities for P_Y. The objective function $\psi(P_X)$ refers to either $F_X(x)$ in case 1 or $E_{P_X}[g(\mathbf{X})]$ in case 2 above.

An important element in formulations (1) and (2) is that the constraints represented by $\mathcal{U}$ are cast on the output level. Since we have available output data, $\mathcal{U}$ can be constructed using these observations in a statistically valid manner (e.g., by using the confidence region on the output statistic). By expressing $P_Y = \gamma(P_X)$, the region $\mathcal{U}$ can be viewed as a region on P_X, given by $\{P_X \in \mathcal{P}_X : \gamma(P_X) \in \mathcal{U}\}$. The following result summarizes the confidence guarantee for the optimal values of (1) and (2) in bounding $\psi(P_X)$ when $\mathcal{U}$ is chosen suitably:

PROPOSITION 1. *Let $P_X^0 \in \mathcal{P}_X$ and $P_Y^0 \in \mathcal{P}_Y$ be the true input and output distributions. Suppose $\mathcal{U}$ is a $(1-\alpha)$-level confidence region for P_Y^0, i.e.,*

$$\mathbb{P}_D(P_Y^0 \in \mathcal{U}) \geq 1-\alpha \tag{3}$$

where $\mathbb{P}_D(\cdot)$ denotes the probability with respect to the data D. Let $\overline{Z}$ and $\underline{Z}$ be the optimal values of (1) *and* (2) *respectively. Then we have*

$$\mathbb{P}_D(\underline{Z} \leq \psi(P_X^0) \leq \overline{Z}) \geq 1-\alpha$$

Similar statements hold if the confidence is approximate, i.e., if

$$\liminf_{n\to\infty} \mathbb{P}_D(P_Y^0 \in \mathcal{U}) \geq 1-\alpha$$

then

$$\liminf_{n\to\infty} \mathbb{P}_D(\underline{Z} \leq \psi(P_X^0) \leq \overline{Z}) \geq 1-\alpha$$

It is worth pointing out that the same guarantee holds, without any statistical adjustment, if one solves (1) and (2) simultaneously for different $\psi(\cdot)$, say $\psi_l(\cdot), l = 1, \ldots, L$, i.e., supposing that (3) holds, then the confidence statement

$$\mathbb{P}_D(\underline{Z}_l \leq \psi_l(P_X^0) \leq \overline{Z}_l,\ l = 1, \ldots, L) \geq 1-\alpha$$

holds, so does a similar statement for the limiting counterpart. We provide this extended version of Proposition 1 in the appendix (Proposition EC.1). This allows us to obtain bounds for multiple quantities about the input model at the same time. Note that, in conventional statistical methods, simultaneous estimation like this sort often requires Bonferroni correction or more advanced techniques, but these are not needed in our approach.

We mention an important feature of our framework related to the issue of *non-identifiability* (e.g., Tarantola (2005)). When there are more than one input model P_X that leads to the same

output distribution, it is statistically impossible to recover exactly the true P_X, and methods that attempt to do so may result in ill-posed problems. Our framework, however, gets around this issue by focusing on computing bounds instead of full model recovery. Even though P_X can be non-identifiable, our optimization always produces valid bounds for it. One special case of interest is when we use $\mathcal{U} = \{P_Y^0\}$, i.e., the true output distribution is exactly known. In this case, (1) and (2) will provide the best bounds for $\psi(P_X)$ given the output. If $\underline{Z} < \overline{Z}$, then P_X cannot be exactly identified, implying an issue of non-identifiabilty, but our outputs would still be valid. In fact, the difference $\overline{Z} - \underline{Z}$ can be viewed as an *identifiability gap* with respect to ψ.

4. Kolmogorov-Smirnov-based Constraints

We will now choose a specific $\mathcal{U}$ that is statistically consistent on the output level, i.e., $\mathcal{U}$ shrinks to $\{P_Y^0\}$ as $n \to \infty$ (in a suitable sense). In particular, we use $\mathcal{U}$ implied by the Kolmogorov-Smirnov (KS) statistic, and discuss how this choice enjoys benefits balancing statistical consistency and computation.

4.1. Statistical Confidence Guarantee

It is known that the empirical distribution for continuous i.i.d. data D, denoted $\hat{F}_Y(y)$, satisfies $\sqrt{n}\|\hat{F}_Y - F_Y^0\|_\infty \Rightarrow \sup_{u\in[0,1]} BB(u)$ where F_Y^0 is the true distribution function of Y, $\|\cdot\|_\infty$ denotes the sup norm over $\mathbb{R}$, $BB(\cdot)$ is a standard Brownian bridge, and $\Rightarrow$ denotes weak convergence. This implies that the KS-statistic $\sqrt{n}\|\hat{F}_Y - F_Y^0\|_\infty$ satisfies

$$\lim_{n\to\infty} P\left(\|\hat{F}_Y - F_Y^0\|_\infty \le \frac{q_{1-\alpha}}{\sqrt{n}}\right) = 1-\alpha$$

where $q_{1-\alpha}$ is the $(1-\alpha)$-quantile of $\sup_{u\in[0,1]} BB(u)$. Therefore, setting

$$\mathcal{U} = \left\{P_Y \in \mathcal{P}_Y : \|F_Y - \hat{F}_Y\|_\infty \le \frac{q_{1-\alpha}}{\sqrt{n}}\right\} \tag{4}$$

ensures that (3) holds and subsequently the conclusion in Proposition 1. As n increases, the size of (4) shrinks to zero.

The following result states precisely the implication of this construction, and moreover, describes how this leads to a more tractable optimization formulation:

THEOREM 1. *Let $\overline{Z}$ and $\underline{Z}$ be the optimal values of the optimization programs*

$$\begin{aligned} \max \quad & \psi(P_X) \\ \textit{subject to } & \|F_Y - \hat{F}_Y\|_\infty \le \tfrac{q_{1-\alpha}}{\sqrt{n}} \\ & P_X \in \mathcal{P}_X \end{aligned} \tag{5}$$

and

$$\begin{aligned} \min \quad & \psi(P_X) \\ \textit{subject to } & \|F_Y - \hat{F}_Y\|_\infty \le \tfrac{q_{1-\alpha}}{\sqrt{n}} \\ & P_X \in \mathcal{P}_X \end{aligned} \tag{6}$$

where $q_{1-\alpha}$ is the $(1-\alpha)$-quantile of $\sup_{u\in[0,1]} BB(u)$, and $\hat{F}_Y$ is the empirical distribution of i.i.d. output data. Supposing the true output distribution is continuous, we have

$$\liminf_{n\to\infty} \mathbb{P}_D(\underline{Z} \le \psi(P_X^0) \le \overline{Z}) \ge 1-\alpha \tag{7}$$

where P_X^0 is the true distribution of the input variate X. Moreover, (5) *and* (6) *are equivalent to*

$$\begin{aligned} \max \quad & \psi(P_X) \\ \textit{subject to } & \hat{F}_Y(y_j+) - \tfrac{q_{1-\alpha}}{\sqrt{n}} \le E_{P_X}[I(h(\mathbf{X}) \le y_j)] \le \hat{F}_Y(y_j-) + \tfrac{q_{1-\alpha}}{\sqrt{n}}, j = 1,\dots,n \\ & P_X \in \mathcal{P}_X \end{aligned} \tag{8}$$

and

$$\begin{aligned} \min \quad & \psi(P_X) \\ \textit{subject to } & \hat{F}_Y(y_j+) - \tfrac{q_{1-\alpha}}{\sqrt{n}} \le E_{P_X}[I(h(\mathbf{X}) \le y_j)] \le \hat{F}_Y(y_j-) + \tfrac{q_{1-\alpha}}{\sqrt{n}}, j = 1,\dots,n \\ & P_X \in \mathcal{P}_X \end{aligned} \tag{9}$$

respectively, where $\hat{F}_Y(y_j+)$ and $\hat{F}_Y(y_j-)$ refer to the right- and left-limits of the empirical distributions $\hat{F}_Y$ at y_j, and $E_{P_X}[\cdot]$ denotes the expectation taken with respect to the T-fold product measure of P_X.

A merit of using the depicted KS-based uncertainty set, seen by Theorem 1, is that it can be reformulated into linear constraints in terms of the expectations $E_{P_X}[\cdot]$ of certain "moments" of $h(\mathbf{X})$. These constraints constitute precisely n interval-type conditions, and the moment functions

are the indicator functions of $h(\mathbf{X})$ falling under the thresholds y_j's. The derivation leading to the reformulation result in Theorem 1 has been used conventionally in computing the KS-statistic. Similar reformulations have also appeared in recent work in approximating stochastic optimization via robust optimization (Bertsimas et al. (2014)).

The asymptotic of the KS-statistic is more complicated if the output distribution is discrete (this happens if the outputs we look at are for instance the queue length). In such cases, the critical values are generally smaller than those for the continuous distribution (Lehmann and Romano (2006)). Consequently, using $q_{1-\alpha}/\sqrt{n}$ to calibrate the size of the uncertainty set as in (4) is still valid, but could be conservative, i.e., we still have $P\left(\|\hat{F}_Y - F_Y^0\|_\infty \leq \frac{q_{1-\alpha}}{\sqrt{n}}\right)$ asymptotically at least $1-\alpha$, but possibly strictly higher. As a remedy, one can use bootstrapping to calibrate the size of a tighter set. Moreover, the constraint of the form $\|\hat{F}_Y - F_Y^0\|_\infty \leq q$ will now be written as

$$\hat{F}_Y(w_j) - q \leq E_{P_X}[I(h(\mathbf{X}) \leq w_j)] \leq \hat{F}_Y(w_j) + q, j = 1, \ldots, K \tag{10}$$

where $w_j, j = 1, \ldots$ are the ordered support points of Y, with $K = \min\{j : \hat{F}_Y(w_j) = 1\}$. These are the points where jumps occur (and the constraints put on the first K of them automatically ensure the rest). If the support size is small, an alternative is to impose constraints on each probability mass, i.e.,

$$\hat{P}(Y = w_j) - q \leq E_{P_X}[I(h(\mathbf{X}) = w_j)] \leq \hat{P}(Y = w_j) + q, j = 1, \ldots, K \tag{11}$$

where $\hat{P}(Y = w_j)$ is the observed proportions of Y being w_j, and q can be calibrated by a standard binomial quantile and the Bonferroni correction.

The KS-statistic has several advantages over other types of uncertainty sets in our considered settings. Alternatives like χ^2 goodness-of-fit tests could be used, but the resulting formulations would not come as handy when expressed in terms of P_X or $h(\mathbf{X})$, which would affect the efficiency of the gradient estimator that we will discuss in Section 5.2.1. Another advantage of using KS-statistic relates to the statistical property of a discretization that is needed to feed into an implementable optimization procedure, which we shall discuss next.

4.2. Randomizing the Decision Space

Note that optimization programs (8) and (9) involve decision variable P_X that is potentially infinite-dimensional, e.g., when X is a continuous variable. This can cause algorithmic and storage issues. One could appropriately discretize the decision variable by randomly sampling a finite set of support points on $\mathcal{X}$. Once these support points are realized, the optimization is imposed on the probability weights on these points, or in other words on a discrete input distribution.

Our next result shows that as the support points are generated from a suitably chosen distribution, and the number of these points grows at an appropriate rate relative to the output data size, the discretized KS-implied optimization will retain the confidence guarantee as the original formulation:

THEOREM 2. *Suppose we sample* $\{z_i\}_{i=1,\dots,m}$ *in the space* $\mathcal{X}$ *from a distribution* Q*. Suppose that* P_X^0*, the true distribution of* X*, is absolutely continuous with respect to* Q *and* $\|dP_X^0/dQ\|_\infty \le C$ *for some* $C>0$*, where* dP_X^0/dQ *is the likelihood ratio calculated from the Radon-Nikodym derivative of* P_X^0 *with respect to* Q*, and* $\|\cdot\|_\infty$ *denotes the essential supremum. Using the notations as in Theorem 1, we solve*

$$\begin{aligned} \max \quad & \psi(P_X) \\ \textit{subject to } & \hat{F}_Y(y_j+) - \tfrac{q_{1-\alpha}}{\sqrt{n}} \le E_{P_X}[I(h(\mathbf{X}) \le y_j)] \le \hat{F}_Y(y_j-) + \tfrac{q_{1-\alpha}}{\sqrt{n}}, j=1,\dots,n \\ & P_X \in \hat{\mathcal{P}}_X \end{aligned} \tag{12}$$

and

$$\begin{aligned} \min \quad & \psi(P_X) \\ \textit{subject to } & \hat{F}_Y(y_j+) - \tfrac{q_{1-\alpha}}{\sqrt{n}} \le E_{P_X}[I(h(\mathbf{X}) \le y_j)] \le \hat{F}_Y(y_j-) + \tfrac{q_{1-\alpha}}{\sqrt{n}}, j=1,\dots,n \\ & P_X \in \hat{\mathcal{P}}_X \end{aligned} \tag{13}$$

where $\hat{\mathcal{P}}_X$ *denotes the set of distributions with support points* $\{z_i\}_{i=1,\dots,m}$*. Let* $\hat{\overline{Z}}$ *and* $\hat{\underline{Z}}$ *be the optimal values of* (12) *and* (13)*.*

Denote $\mathbb{P}$ *as the probability taken with respect to both the output data and the support generation for* X*. Suppose that* $\psi(P_X)$ *takes the form* $E_{P_X}[g(\mathbf{X})]$ *(which subsumes both types of target*

measures discussed in Section 3) where $E_{P_X^0}[g(X_{i_1},\ldots,X_{i_T})^2]<\infty$ *for any* $1\le i_1,\ldots,i_T\le T$. *Also suppose that the true output distribution is continuous and that* $\mathbb{P}($*for any* $P_X\in\hat{\mathcal{P}}_X, supp(\gamma(P_X))\cap\{y_j\}_{j=1,\ldots,n}\neq\emptyset)=0$ *where* $supp(\gamma(P_X))$ *denotes the support of the distribution* $\gamma(P_X)$. *Then, we have*

$$\liminf_{n\to\infty,m/n\to\infty}\mathbb{P}\left(\hat{\underline{Z}}+O_p\left(\frac{1}{\sqrt{m}}\right)\le\psi(P_X^0)\le\hat{\overline{Z}}+O_p\left(\frac{1}{\sqrt{m}}\right)\right)\ge 1-\alpha$$

The error term $O_p(1/\sqrt{m})$ represents a random variable of stochastic order $1/\sqrt{m}$, i.e., $a_m=O_p(1/\sqrt{m})$ if for any $\epsilon>0$, there exists $M,N>0$ such that $P(|\sqrt{m}a_m|\le N)>1-\epsilon$ for $m>M$.

Theorem 2 guarantees that by solving the finite-dimensional optimization problems (12) and (13), we obtain confidence bounds for the true quantity of interest $\psi(P_X^0)$, up to an error of order $O_p(1/\sqrt{m})$. Note that the conclusion holds with the numbers of constraints in (12) and (13) growing in the data size n. One significance of the result is that, despite this growth, as long as one generates the supports of X from a distribution with a heavier tail than the true distribution, and with a size m of order higher than n, the confidence guarantee is approximately retained. A key element in explaining this behavior lies in the low complexity of the function class $I(h(\cdot)\le y)$ (parametrized by y) appearing in the constraints and interplayed with the likelihood ratio dP_X^0/dQ, as measured by the bracketing number. This number captures the richness of the involved function class with the counts of neighborhoods, each formed by an upper and a lower bounding function that is known as a bracket, in covering the whole class (see the discussion in Appendix EC.6.1). A slowly growing (e.g., polynomial in our case) bracketing number turns out to allow the statistic on the output performance measure to be approximated uniformly well with a discretized input distribution, by invoking the empirical process theory for so-called U-statistics (Arcones and Gine (1993)). On the other hand, using other moment functions (implied by other test statistics) may not preserve this behavior. This connection to function complexity, which informs the usefulness of sampling-based procedures when integrating with output data, is the first of such kind in the model calibration literature as far as we know.

We have focused on a continuous output distribution in Theorem 2. The assumption $\mathbb{P}(\text{for any } P_X \in \hat{\mathcal{P}}_X, \text{supp}(\gamma(P_X)) \cap \{y_j\}_{j=1,\ldots,n} \neq \emptyset) = 0$ is a technical condition that ensures the distribution of $h(\mathbf{X})$ under $P_X \in \hat{\mathcal{P}}_X$ does not have overlapping support points as y_j's, which allows us to reduce the KS-implied constraint into the n interval constraints depicted in the theorem. This assumption holds in almost every discrete-event performance measure provided that the considered P_X and P_Y are continuous. On the other hand, if P_Y is discrete, then the theorem holds with the first constraints in (12) and (13) replaced by (10) or (11) (with q suitably calibrated as discussed there), without needing the assumption $\mathbb{P}(\text{for any } P_X \in \hat{\mathcal{P}}_X, \text{supp}(\gamma(P_X)) \cap \{y_j\}_{j=1,\ldots,n} \neq \emptyset) = 0$.

We mention that Ghosh and Lam (2015c) provides a similar guarantee for robust optimization problems designed for quantifying input uncertainty. In particular, their analysis allows to give confidence bounds on output performance measures. However, they do not consider the asymptotic confidence guarantee in relation to the data size and the randomized support size. As a consequence, they do not need considering the complexity of the constraints. Moreover, since they handle input uncertainty, the uncertainty sets are more elementary, in contrast to ours which serve as a tool to invert the input-output relation.

We note that, like Proposition 1, all the results in this section can be similarly extended to a simultaneous guarantee when solving L optimization problems, where each problem has a different objective function $\psi_l(P_X)$. For instance, under the same assumptions as Theorem 2 with L different objectives in (12) and (13), and using the same generated set of support points across all optimization problems, we would obtain that

$$\liminf_{n\to\infty, m/n\to\infty} \mathbb{P}\left(\hat{\underline{Z}}_l + O_p\left(\frac{1}{\sqrt{m}}\right) \leq \psi_l(P_X^0) \leq \hat{\overline{Z}}_l + O_p\left(\frac{1}{\sqrt{m}}\right), l = 1, \ldots, L\right) \geq 1 - \alpha$$

where $\hat{\underline{Z}}_l, \hat{\overline{Z}}_l$ are the minimum and maximum values of the discretized optimization with objective $\psi_l(P_X)$, and each $O_p(1/\sqrt{m})$ is the error term corresponding to each optimization program.

Lastly, we point out that all the results in Sections 3 and 4 hold when we consider $h(\mathbf{X}, \mathbf{W})$ and $g(\mathbf{X}, \mathbf{W})$, where $\mathbf{W}$ consist of other input variate sequences independent from $\mathbf{X}$ with known probability distributions. This is as long as we treat all the expectations $E_{P_X}[\cdot]$ as taken jointly

under both the product measure of P_X and the known distribution of $\mathbf{W}$. We provide further remarks in the appendix.

5. Optimization Procedure

This section presents our optimization strategy for (locally) solving (12) and (13). Without loss of generality, we only focus on the minimization problem (13) since maximization can be converted to minimization by negating the objective. Section 5.1 first discusses the transformation of the stochastic constrained program into a sequence of programs with deterministic convex constraints, using the quadratic penalty method in nonlinear programming. Section 5.2 then investigates how this transformation can be utilized effectively in a fully iterative stochastic algorithm using MDSA. Section 5.3 provides a convergence theorem. In the appendix, we also provide an alternate approach that has a similar convergence guarantee but differs in the implementation details.

5.1. A Stochastic Quadratic Penalty Method

When restricted to distributions with support points $\{z_i\}_{i=1,\ldots,m}$, the candidate input distribution P_X can be identified by an m-dimensional vector $\mathbf{p} = (p_1, \ldots, p_m)$ on the probability simplex $\mathcal{P} := \{\mathbf{p} : \sum_{i=1}^m p_i = 1, p_i \geq 0 \text{ for each } i\}$, where the subscript X is suppressed with no ambiguity. By the vector $\mathbf{p}$, we mean the distribution that assigns probability p_i to the point z_i. The optimization program (13) can thus be rewritten as

$$\begin{aligned} \min \quad & \psi(\mathbf{p}) \\ \text{subject to } & \hat{F}_Y(y_j+) - \tfrac{q_{1-\alpha}}{\sqrt{n}} \leq E_{\mathbf{p}}[I(h(\mathbf{X}) \leq y_j)] \leq \hat{F}_Y(y_j-) + \tfrac{q_{1-\alpha}}{\sqrt{n}}, j = 1, \ldots, n \\ & \mathbf{p} \in \mathcal{P}. \end{aligned} \tag{14}$$

Note that the constraints in (14) are in general non-convex because the i.i.d. input sequence means that the expectation $E_{\mathbf{p}}[I(h(\mathbf{X}) \leq y_j)]$ is a high-dimensional polynomial in $\mathbf{p}$. Moreover, this polynomial can involve a huge number of terms and hence its evaluation requires simulation approximation. As far as we know, the literature on dealing with stochastic non-convex constraints

is very limited. To overcome this difficulty, we first introduce the quadratic penalty method (Bertsekas (1999)) to transform program (14) into a sequence of penalized programs with deterministic convex constraints

$$\begin{aligned} \min \quad & \lambda\psi(\mathbf{p}) + \textstyle\sum_{j=1}^{n}(E_{\mathbf{p}}[I(h(\mathbf{X}) \le y_j)] - s_j)^2 \\ \text{subject to } & \hat{F}_Y(y_j+) - \tfrac{q_{1-\alpha}}{\sqrt{n}} \le s_j \le \hat{F}_Y(y_j-) + \tfrac{q_{1-\alpha}}{\sqrt{n}}, j = 1, \ldots, n \\ & \mathbf{p} \in \mathcal{P} \end{aligned} \tag{15}$$

where $\mathbf{s} = (s_1, \ldots, s_n)$ are slack variables and $\lambda > 0$ is an inverse measure of the cost/penalty of infeasibility. A related scheme is also used by Wang and Spall (2008) in the context of nonconvex stochastic objectives (with deterministic constraints). As $\lambda \to 0$, there is an increasing cost of violating the stochastic constraints, therefore the optimal solution of (15) converges to that of (14), as stated in the following proposition.

PROPOSITION 2. *Suppose* (14) *has at least one feasible solution. Let* $(\mathbf{p}^*(\lambda), \mathbf{s}^*(\lambda))$ *be an optimal solution of* (15) *indexed at* λ*. As* λ *decreases to* 0*, every limit point of the sequence* $\{\mathbf{p}^*(\lambda)\}$ *is an optimal solution of* (14)*.*

As suggested in the proof of Proposition 2, a mathematically equivalent reformulation of (15) with the slack variables optimized is

$$\begin{aligned} \min \quad & \lambda\psi(\mathbf{p}) + \textstyle\sum_{j=1}^{n}(E_{\mathbf{p}}[I(h(\mathbf{X}) \le y_j)] - \Pi_j(E_{\mathbf{p}}[I(h(\mathbf{X}) \le y_j)]))^2 \\ \text{subject to } & \mathbf{p} \in \mathcal{P} \end{aligned} \tag{16}$$

where each Π_j is the projection onto the interval $[F_Y(y_j+) - \frac{q_{1-\alpha}}{\sqrt{n}}, F_Y(y_j-) + \frac{q_{1-\alpha}}{\sqrt{n}}]$ defined as

$$\Pi_j(x) = \begin{cases} \hat{F}_Y(y_j+) - \frac{q_{1-\alpha}}{\sqrt{n}} & \text{if } x < \hat{F}_Y(y_j+) - \frac{q_{1-\alpha}}{\sqrt{n}} \\ \hat{F}_Y(y_j-) + \frac{q_{1-\alpha}}{\sqrt{n}} & \text{if } x > \hat{F}_Y(y_j-) + \frac{q_{1-\alpha}}{\sqrt{n}} \\ x & \text{otherwise.} \end{cases} \tag{17}$$

5.2. Constrained Stochastic Approximation

Although the formulations (15), (16) are still non-convex, their constraints are convex and deterministic, which can be handled more easily using SA than in the original formulation (14). This

section investigates the design and analysis of an MDSA algorithm for finding local optima of (14) by solving (15) with decreasing values of λ. The appendix would illustrate another algorithm that uses formulation (16) instead of (15).

To describe the algorithm, MD finds the next iterate via optimizing the objective function linearized at the current iterate, together with a penalty on the distance of movement of the iterate. When the objective function is only accessible via simulation, the linearized objective function, or the gradient, at each iteration can only be estimated with noise, in which case the procedure becomes MDSA (Nemirovski et al. (2009)). More precisely, when applied to the formulation (15) with slack variables, MDSA solves the following optimization given a current iterate $(\mathbf{p}^k, \mathbf{s}^k)$

$$\begin{aligned} \min \quad & \gamma^k(\lambda\hat{\mathbf{\Psi}}^k + \hat{\boldsymbol{\phi}}^k_{\mathbf{p}})'(\mathbf{p}-\mathbf{p}^k) + \beta^k\hat{\boldsymbol{\phi}}^{k\prime}_{\mathbf{s}}(\mathbf{s}-\mathbf{s}^k) + V(\mathbf{p}^k,\mathbf{p}) + \tfrac{1}{2}\|\mathbf{s}-\mathbf{s}^k\|_2^2 \\ \text{subject to } & \hat{F}_Y(y_j+) - \tfrac{q_{1-\alpha}}{\sqrt{n}} \le s_j \le \hat{F}_Y(y_j-) + \tfrac{q_{1-\alpha}}{\sqrt{n}}, j=1,\dots,n \\ & \mathbf{p}\in\mathcal{P} \end{aligned} \tag{18}$$

where $\hat{\mathbf{\Psi}}^k$ carries the gradient information of the target performance measure ψ at $\mathbf{p}^k$, while $\hat{\boldsymbol{\phi}}^k_{\mathbf{p}}$ and $\hat{\boldsymbol{\phi}}^k_{\mathbf{s}}$ contain the gradient information of the penalty function in (15) with respect to $\mathbf{p}, \mathbf{s}$ respectively. The sum $V(\mathbf{p}^k,\mathbf{p}) + \frac{1}{2}\|\mathbf{s}-\mathbf{s}^k\|_2^2$ serves as the penalty on the movement of the iterate, where $\|\cdot\|_2$ denotes the standard Euclidean distance, and $V(\cdot,\cdot)$ defined as

$$V(\mathbf{p},\mathbf{q}) = \sum_{i=1}^{n} q_i \log\frac{q_i}{p_i} \tag{19}$$

is the KL divergence between two probability measures. This particular choice of V has been shown (Nemirovski et al. (2009)) to have superior performance to other choices like the Euclidean distance when the decision space is the probability simplex. Different from traditional SA, the step sizes γ^k and β^k, used for updating $\mathbf{p}$ and $\mathbf{s}$ in (18), are different, the rationale for which shall be discussed in Section 5.3.

However, iterations in the form of (18) can only find optima of (15) for a particular penalty coefficient λ while retrieving the optimal solution of the original problem (14) through (15) hinges on sending λ to 0. Literature on deterministic optimization suggests solving the penalized optimization repeatedly for a set of decreasing values of λ, but it could be difficult to tell when to

stop decreasing the λ in our stochastic case. In order to output the optimal solution in one single run, we decrease λ together with the step size from one iteration to the next, hence arrive at the following sequential joint solution-and-penalty-updating routine

$$\begin{aligned} \min \quad & \gamma^k(\lambda^k\hat{\boldsymbol{\Psi}}^k + \hat{\boldsymbol{\phi}}^k_{\mathbf{p}})'(\mathbf{p}-\mathbf{p}^k) + \beta^k\hat{\boldsymbol{\phi}}^{k\prime}_{\mathbf{s}}(\mathbf{s}-\mathbf{s}^k) + V(\mathbf{p}^k,\mathbf{p}) + \tfrac{1}{2}\|\mathbf{s}-\mathbf{s}^k\|_2^2 \\ \text{subject to } & \hat{F}_Y(y_j+) - \tfrac{q_{1-\alpha}}{\sqrt{n}} \le s_j \le \hat{F}_Y(y_j-) + \tfrac{q_{1-\alpha}}{\sqrt{n}}, j=1,\dots,n \\ & \mathbf{p}\in\mathcal{P} \end{aligned} \tag{20}$$

where λ^k is appropriately chosen in conjunction with γ^k, β^k, and decreases to 0. To implement the fully sequential scheme, we need to investigate: 1) how to obtain $\hat{\boldsymbol{\Psi}}^k, \hat{\boldsymbol{\phi}}^k_{\mathbf{p}}$ and $\hat{\boldsymbol{\phi}}^k_{\mathbf{s}}$, 2) efficient solution method for program (20), and 3) how to select the parameters γ^k, β^k and λ^k. The next two subsections present the first two investigations respectively, while Section 5.3 will analyze the convergence of the algorithm in relation to the parameter choices.

5.2.1. Gradient Estimation and Restricted Programs. Denote by $W(\mathbf{p})$ the penalty function in (16), and by $W_s(\mathbf{p},\mathbf{s})$ the quadratic penalty in (15) where the subscript s refers to "slack variable". These are functions of variables on the probability simplex, for which naive differentiation may not lead to simulable object since an arbitrary perturbation may shoot out of the simplex. Ghosh and Lam (2015a) and Ghosh and Lam (2015b) have used the idea of Gateaux derivative (in the sense described in Chapter 6 of Serfling (2009)) to obtain simulable representations of gradients of expectation-type performance measures. We generalize their result to sums of functions of expectations:

PROPOSITION 3. *We have:*

1. *Suppose $\psi, W, W_s(\cdot,\mathbf{s})$ are differentiable in the probability simplex $\mathcal{P}$, then*

$$\nabla\psi(\mathbf{p})'(\mathbf{q}-\mathbf{p}) = \boldsymbol{\Psi}(\mathbf{p})'(\mathbf{q}-\mathbf{p}) \tag{21}$$

$$\nabla W(\mathbf{p})'(\mathbf{q}-\mathbf{p}) = \boldsymbol{\phi}(\mathbf{p})'(\mathbf{q}-\mathbf{p}) \tag{22}$$

$$\nabla_{\mathbf{p}} W_s(\mathbf{p},\mathbf{s})'(\mathbf{q}-\mathbf{p}) = \boldsymbol{\phi}_{\mathbf{p}}(\mathbf{p},\mathbf{s})'(\mathbf{q}-\mathbf{p}) \tag{23}$$

for any $\mathbf{p},\mathbf{q} \in \mathcal{P}$, *where the Gateaux derivatives* $\boldsymbol{\Psi}(\mathbf{p}) = (\Psi_1(\mathbf{p}),\ldots,\Psi_m(\mathbf{p}))'$, $\boldsymbol{\phi}(\mathbf{p}) = (\phi_1(\mathbf{p}),\ldots,\phi_m(\mathbf{p}))'$, $\boldsymbol{\phi}_{\mathbf{p}}(\mathbf{p},\mathbf{s}) = (\phi_{\mathbf{p},1}(\mathbf{p},\mathbf{s}),\ldots,\phi_{\mathbf{p},m}(\mathbf{p},\mathbf{s}))'$, *and*

$$\Psi_i(\mathbf{p}) = \frac{d}{d\epsilon}\psi((1-\epsilon)\mathbf{p}+\epsilon\mathbf{1}_i)\Big|_{\epsilon=0^+} \tag{24}$$

$$\phi_i(\mathbf{p}) = \frac{d}{d\epsilon}W((1-\epsilon)\mathbf{p}+\epsilon\mathbf{1}_i)\Big|_{\epsilon=0^+} \tag{25}$$

$$\phi_{\mathbf{p},i}(\mathbf{p},\mathbf{s}) = \frac{d}{d\epsilon}W_s((1-\epsilon)\mathbf{p}+\epsilon\mathbf{1}_i,\mathbf{s})\Big|_{\epsilon=0^+} \tag{26}$$

2. *Assume* $\mathbf{p} = (p_1,\ldots,p_m)$ *where each* $p_i > 0$. *Then the Gateaux derivatives* (24)(25)(26) *are finite and can be expressed as*

$$\Psi_i(\mathbf{p}) = E_{\mathbf{p}}[g(\mathbf{X})S_i(\mathbf{X};\mathbf{p})] \tag{27}$$

$$\phi_i(\mathbf{p}) = 2\sum_{j=1}^{n}(E_{\mathbf{p}}[I(h(\mathbf{X})\le y_j)] - \Pi_j(E_{\mathbf{p}}[I(h(\mathbf{X})\le y_j)]))E_{\mathbf{p}}[I(h(\mathbf{X})\le y_j)S_i(\mathbf{X};\mathbf{p})] \tag{28}$$

$$\phi_{\mathbf{p},i}(\mathbf{p},\mathbf{s}) = 2\sum_{j=1}^{n}(E_{\mathbf{p}}[I(h(\mathbf{X})\le y_j)] - s_j)E_{\mathbf{p}}[I(h(\mathbf{X})\le y_j)S_i(\mathbf{X};\mathbf{p})] \tag{29}$$

where

$$S_i(\mathbf{x};\mathbf{p}) = \sum_{t=1}^{S}\frac{I_i(x_t)}{p_i} - S \text{ for (27), and } \sum_{t=1}^{T}\frac{I_i(x_t)}{p_i} - T \text{ for (28)(29).}$$

Here $I_i(x) = 1$ *if* $x = z_i$ *and 0 otherwise, and* $\mathbf{X}$ *is the i.i.d. input process generated under* $\mathbf{p}$.

The representations (27) and (29) suggest the following unbiased estimators for the gradient of ψ, $\boldsymbol{\Psi}(\mathbf{p}) = (\Psi_i(\mathbf{p}))_{i=1}^m$, and the gradient of the penalty function, $\boldsymbol{\phi}_{\mathbf{p}}(\mathbf{p},\mathbf{s}) = (\phi_{\mathbf{p},i}(\mathbf{p},\mathbf{s}))_{i=1}^m$

$$\hat{\Psi}_i(\mathbf{p}) = \frac{1}{M_3}\sum_{r=1}^{M_3} g(\mathbf{X}^{(r)})S_i(\mathbf{X}^{(r)};\mathbf{p}) \tag{30}$$

$$\hat{\phi}_{\mathbf{p},i}(\mathbf{p},\mathbf{s}) = 2\sum_{j=1}^{n}\frac{1}{M_1}\sum_{r=1}^{M_1}(I(h(\mathbf{X}^{(r)})\le y_j) - s_j)\frac{1}{M_2}\sum_{r=1}^{M_2} I(h(\tilde{\mathbf{X}}^{(r)})\le y_j)S_i(\tilde{\mathbf{X}}^{(r)};\mathbf{p}) \tag{31}$$

where $\mathbf{X}^{(r)}$ and $\tilde{\mathbf{X}}^{(r)}$ are independent copies of the i.i.d. input process generated under $\mathbf{p}$ and are used simultaneously for all i,j. By direct differentiation, a straightforward unbiased estimator for $\boldsymbol{\phi}_{\mathbf{s}}(\mathbf{p},\mathbf{s}) = (\phi_{\mathbf{s},j}(\mathbf{p},\mathbf{s}))_{j=1}^n$, the gradient of the penalty function with respect to $\mathbf{s}$, is

$$\hat{\phi}_{\mathbf{s},j}(\mathbf{p},\mathbf{s}) = \frac{-2}{M_1}\sum_{r=1}^{M_1}(I(h(\mathbf{X}^{(r)})\le y_j) - s_j). \tag{32}$$

Our main procedure (shown in Algorithm 2 momentarily) uses the above gradient estimators, while an alternate MDSA depicted in Algorithm 3 in the appendix solves (16) using a biased estimator of $\boldsymbol{\phi}(\mathbf{p})$ conferred by (28).

Note that the above gradient estimators are available thanks to the KS-implied constraints we introduced. By the reformulation in Theorem 1, the constraints in (8) and (9) become (T-fold) expectation-type constraints. Thus, when differentiating the squared expectation in the quadratic penalty, the gradient becomes the product of two T-fold expectations, one with the extra factor $S_i(\cdot;\mathbf{p})$ which can be interpreted as a score function. This then allows unbiased estimation of the gradient by generating two independent batches of simulation runs each for one of the expectations. Using other statistics to induce the constraints may not lead to such a convenient form.

Note that the $S_i(\cdot;\mathbf{p})$ in the gradient estimators (30) and (31) contains p_i at the denominator, so a small p_i can blow up the variances of the estimators and in turn adversely affect the convergence of MDSA. To ensure convergence, we make an adjustment to our procedure and solve the following restricted version of (14)

$$\begin{aligned} \min \quad & \psi(\mathbf{p}) \\ \text{subject to } & \hat{F}_Y(y_j+) - \tfrac{q_{1-\alpha}}{\sqrt{n}} \le E_{\mathbf{p}}[I(h(\mathbf{X}) \le y_j)] \le \hat{F}_Y(y_j-) + \tfrac{q_{1-\alpha}}{\sqrt{n}}, j = 1,\dots,n \\ & \mathbf{p} \in \mathcal{P}(\epsilon) \end{aligned} \tag{33}$$

where the restricted probability simplex $\mathcal{P}(\epsilon) := \{\mathbf{p} \in \mathcal{P} : p_i \ge \epsilon \text{ for each } i\}$. Accordingly, the full simplex $\mathcal{P}$ in the penalized program (15) and stepwise subproblem (20) has to be replaced by $\mathcal{P}(\epsilon)$.

To maintain the statistical guarantee provided by Theorem 2 when solving the restricted programs, the shrinking size ϵ has to be appropriately chosen. Theorem 3 below indicates that it suffices to choose ϵ smaller than $1/(m\sqrt{n})$ in case of bounded $g(\mathbf{X})$.

THEOREM 3. *Denote by $\hat{\overline{Z}}_\epsilon$ and $\hat{\underline{Z}}_\epsilon$ the maximum and minimum of $\psi(\mathbf{p})$ in the feasible set of* (33). *In addition to the conditions of Theorem 2, further assume that $g(\mathbf{X})$ is bounded. If ϵ is chosen such that $\epsilon = o\big(\frac{1}{m\sqrt{n}}\big)$ then we have*

$$\liminf_{n\to\infty, m/n\to\infty} \mathbb{P}\left(\hat{\underline{Z}}_\epsilon + O_p\left(m\epsilon + \frac{1}{\sqrt{m}}\right) \le \psi(P_X^0) \le \hat{\overline{Z}}_\epsilon + O_p\left(m\epsilon + \frac{1}{\sqrt{m}}\right)\right) \ge 1-\alpha.$$

In particular, the first type of target quantities we consider has a bounded $g(\mathbf{X})$. Note that the original optimization itself already poses an error of size $O_p(1/\sqrt{m})$ in the confidence bounds (Theorem 2), so to keep the error at the same level one can use an $\epsilon = O(1/m^{\frac{3}{2}})$ (recall that $m/n \to \infty$). Since the variances of our gradient estimators (30)(31) can be shown inversely proportional to the components p_i (Ghosh and Lam (2015c)), such an ϵ gives rise to variances of order $O(m^{\frac{3}{2}})$. We point out that this is only slightly worse than the best attainable order $O(m)$, which results from the fact that the average size of $\mathbf{p}$ in the m-dimensional probability simplex is $1/m$.

5.2.2. Solving Stepwise Subproblem in MDSA. Since we are now solving the restricted version of subproblem (20), consider the following generic form

$$\begin{aligned} \min \quad & \boldsymbol{\xi}'(\mathbf{q}-\mathbf{p}) + \boldsymbol{\eta}'(\mathbf{t}-\mathbf{s}) + V(\mathbf{p},\mathbf{q}) + \tfrac{1}{2}\|\mathbf{t}-\mathbf{s}\|_2^2 \\ \text{subject to } & \hat{F}_Y(y_j+) - \tfrac{q_{1-\alpha}}{\sqrt{n}} \le t_j \le \hat{F}_Y(y_j-) + \tfrac{q_{1-\alpha}}{\sqrt{n}}, j = 1,\dots,n \\ & \mathbf{q} \in \mathcal{P}(\epsilon). \end{aligned} \tag{34}$$

Because the objective and the feasible set are both separable in $\mathbf{q}$ and $\mathbf{t}$, the above program can be decomposed into two independent programs. One is

$$\begin{aligned} \min \quad & \boldsymbol{\xi}'(\mathbf{q}-\mathbf{p}) + V(\mathbf{p},\mathbf{q}) \\ \text{subject to } & \mathbf{q} \in \mathcal{P}(\epsilon) \end{aligned} \tag{35}$$

and the other is

$$\begin{aligned} \min \quad & \boldsymbol{\eta}'(\mathbf{t}-\mathbf{s}) + \tfrac{1}{2}\|\mathbf{t}-\mathbf{s}\|_2^2 \\ \text{subject to } & \hat{F}_Y(y_j+) - \tfrac{q_{1-\alpha}}{\sqrt{n}} \le t_j \le \hat{F}_Y(y_j-) + \tfrac{q_{1-\alpha}}{\sqrt{n}}, j = 1,\dots,n. \end{aligned} \tag{36}$$

Program (36) is exactly the step-wise routine that appears in the standard gradient descent whose solution takes the form

$$t_j^* = \Pi_j(s_j - \eta_j)$$

where Π_j is the projection defined in (17).

The solution of program (35) has a semi-explicit expression as shown in the following proposition.

PROPOSITION 4. *The optimal solution of the stepwise subproblem* (35) *with* $0 \le \epsilon < 1/m$ *is*

$$q_i^* = \frac{\max\{\eta^*, p_i e^{-\xi_i}\}}{\sum_{i=1}^m \max\{\eta^*, p_i e^{-\xi_i}\}} \tag{37}$$

where $\eta^* \in [0, \max_i p_i e^{-\xi_i})$ *solves the equation*

$$\epsilon = \mu(\eta^*) := \frac{\eta^*}{\sum_{i=1}^m \max\{\eta^*, p_i e^{-\xi_i}\}}. \tag{38}$$

Proposition 4 suggests a procedure for solving (35) that involves a root-finding problem (38). To design an efficient root-finding routine, note that the function $\mu(\eta)$ is strictly increasing in η. More importantly, it consists of at most m smooth pieces, and on the i-th piece it takes the form

$$\mu(\eta) = \frac{\eta}{i\eta + \sum_{i'=i+1}^m p_{(i')} e^{-\xi_{(i')}}}, \text{ if } p_{(i)} e^{-\xi_{(i)}} \le \eta \le p_{(i+1)} e^{-\xi_{(i+1)}}$$

where $(p_{(1)} e^{-\xi_{(1)}}, \ldots, p_{(m)} e^{-\xi_{(m)}})$ is obtained by sorting $(p_1 e^{-\xi_1}, \ldots, p_m e^{-\xi_m})$ in ascending order. Thus one can first locate which piece the root η^* lies on by comparing the values of μ with ϵ at the points $p_{(i)} e^{-\xi_{(i)}}$ and then compute η^* in closed form from the above expression on that piece. This efficient sort-and-search procedure is described in Algorithm 1, whose proof follows from straightforward algebraic verification and hence is omitted.

Algorithm 1 Sort-and-search for solving (35) with $0 \le \epsilon < 1/m$

1. Sort $(p_1 e^{-\xi_1}, \ldots, p_m e^{-\xi_m})$ into ascending order $(p_{(1)} e^{-\xi_{(1)}}, \ldots, p_{(m)} e^{-\xi_{(m)}})$, and let $p_{(0)} e^{-\xi_{(0)}} = 0$

2. Search for the i^* from 0 to $m-1$ such that

$$\frac{p_{(i^*)} e^{-\xi_{(i^*)}}}{i^* p_{(i^*)} e^{-\xi_{(i^*)}} + \sum_{i=i^*+1}^m p_{(i)} e^{-\xi_{(i)}}} \le \epsilon < \frac{p_{(i^*+1)} e^{-\xi_{(i^*+1)}}}{(i^*+1) p_{(i^*+1)} e^{-\xi_{(i^*+1)}} + \sum_{i=i^*+2}^m p_{(i)} e^{-\xi_{(i)}}}$$

3. Output q_i^* according to (37) with

$$\eta^* = \frac{\epsilon \sum_{i=i^*+1}^m p_{(i)} e^{-\xi_{(i)}}}{1 - \epsilon i^*}$$

5.3. Convergence Analysis

We depict our MDSA procedure in Algorithm 2. Steps 1, 2 and 3 of the procedure estimate the gradients using the estimators proposed in Section 5.2.1, and Step 4 updates the decision variable with step size γ^k and the slack variables with step size β^k. Steps 1-4 combined are in effect solving the stepwise subproblem (20) with $\mathcal{P}$ replaced by $\mathcal{P}(\epsilon)$. Therefore by iterating with decreasing penalty coefficient λ^k, Algorithm 2 searches for the optimum of the restricted formulation (33).

To provide convergence guarantee for Algorithm 2, we assume the following:

ASSUMPTION 1. *The restricted program* (33) *has a unique optimal solution* $\mathbf{p}^*_\epsilon \in \mathcal{P}(\epsilon)$ *such that for any feasible* $\mathbf{p} \in \mathcal{P}(\epsilon)$ *and* $\mathbf{p} \neq \mathbf{p}^*_\epsilon$ *it holds* $\boldsymbol{\Psi}(\mathbf{p})'(\mathbf{p} - \mathbf{p}^*_\epsilon) > 0$*, and for any infeasible* $\mathbf{p} \in \mathcal{P}(\epsilon)$ *it holds* $\boldsymbol{\phi}(\mathbf{p})'(\mathbf{p} - \mathbf{p}^*_\epsilon) > 0$*, where* $\boldsymbol{\Psi}, \boldsymbol{\phi}$ *are respectively the Gateaux derivatives of the target quantity* ψ *and the quadratic penalty function* $\sum_{j=1}^n (E_{\mathbf{p}}[I(h(\mathbf{X}) \leq y_j)] - \Pi_j(E_{\mathbf{p}}[I(h(\mathbf{X}) \leq y_j)]))^2$ *in* (16).

ASSUMPTION 2. *There is some threshold* $\lambda_\epsilon > 0$ *such that*

1. for any $\lambda \in (0, \lambda_\epsilon]$ *the optimization problem* (16) *with* $\mathcal{P}$ *replaced by* $\mathcal{P}(\epsilon)$ *has a unique optimal solution* $\mathbf{p}^*_\epsilon(\lambda) \in \mathcal{P}(\epsilon)$ *such that for any* $\mathbf{p} \in \mathcal{P}(\epsilon)$ *it holds* $(\lambda \boldsymbol{\Psi}(\mathbf{p}) + \boldsymbol{\phi}(\mathbf{p}))'(\mathbf{p} - \mathbf{p}^*_\epsilon(\lambda)) \geq 0$

2. $\mathbf{p}^*_\epsilon(\lambda)$ *as a function of* $\lambda \in (0, \lambda_\epsilon]$ *has finite total variation, meaning that there exists a constant* $M_\epsilon > 0$ *such that* $\sum_{i=0}^{K-1} \|\mathbf{p}^*_\epsilon(\lambda_i) - \mathbf{p}^*_\epsilon(\lambda_{i+1})\| \leq M_\epsilon$ *for any* $0 < \lambda_K < \cdots < \lambda_1 < \lambda_0 \leq \lambda_\epsilon$ *and* K.

ASSUMPTION 3. $\|\mathbf{p}^*_\epsilon(\lambda) - \mathbf{p}^*_\epsilon\| = O(\lambda)$ *as* $\lambda \to 0$.

The condition $(\lambda \boldsymbol{\Psi}(\mathbf{p}) + \boldsymbol{\phi}(\mathbf{p}))'(\mathbf{p} - \mathbf{p}^*_\epsilon(\lambda)) \geq 0$ in Assumption 2 is a weakened version of the general convexity criterion that has appeared in online learning (e.g., Bottou (1998)) and SA (e.g., Benveniste et al. (2012), Broadie et al. (2011)) literature. For a minimization problem with objective $f(x)$ and minimizer x^*, this criterion refers to the condition that $\nabla f(x)'(x - x^*) > 0$ for any $x \neq x^*$. A geometric interpretation of it is that the opposite of the gradient direction always points to the optimum. Part 1 of Assumption 2 stipulates that the criterion holds weakly for the penalized program (16) when the penalty coefficient λ lies in a small neighborhood of zero. Assumption 1 can be viewed as the same criterion for the limit case $\lambda = 0$. To explain, at a feasible solution $\mathbf{p}$ of (33)

Algorithm 2 MDSA for solving (15)

Input: A small parameter $\epsilon > 0$, initial solution $\mathbf{p}^1 \in \mathcal{P}(\epsilon) = \{\mathbf{p} : \sum_{i=1}^m p_i = 1, p_i \geq \epsilon \text{ for } i = 1, \ldots, m\}$ and $\mathbf{s}^1 \in [\hat{F}_Y(y_1+) - \frac{q_{1-\alpha}}{\sqrt{n}}, \hat{F}_Y(y_1-) + \frac{q_{1-\alpha}}{\sqrt{n}}] \times \cdots \times [\hat{F}_Y(y_n+) - \frac{q_{1-\alpha}}{\sqrt{n}}, \hat{F}_Y(y_n+) - \frac{q_{1-\alpha}}{\sqrt{n}}]$, a step size sequence γ^k for $\mathbf{p}$, a penalty sequence λ^k, a step size sequence β^k for $\mathbf{s}$, and sample sizes M_1, M_2, M_3.

Iteration: For $k = 1, 2, \ldots$, do the following: Given $\mathbf{p}^k, \mathbf{s}^k$,

1. Estimate $\hat{\boldsymbol{\phi}}^k_{\mathbf{p}} = (\hat{\phi}^k_{\mathbf{p},1}, \ldots, \hat{\phi}^k_{\mathbf{p},m})$, the gradient of the penalty term with respect to $\mathbf{p}$, with

$$\hat{\phi}^k_{\mathbf{p},i} = 2\sum_{j=1}^{n} \frac{1}{M_1}\sum_{r=1}^{M_1}(I(h(\mathbf{X}^{(r)}) \leq y_j) - s^k_j)\frac{1}{M_2}\sum_{r=1}^{M_2} I(h(\tilde{\mathbf{X}}^{(r)}) \leq y_j) S_i(\tilde{\mathbf{X}}^{(r)}; \mathbf{p}^k)$$

where $\mathbf{X}^{(r)}, \tilde{\mathbf{X}}^{(r)}$ are M_1 and M_2 independent copies of the input process generated under $\mathbf{p}^k$.

2. Estimate $\hat{\boldsymbol{\Psi}}^k = (\hat{\Psi}^k_1, \ldots, \hat{\Psi}^k_m)$, the gradient of $E_{\mathbf{p}}[g(\mathbf{X})]$, with

$$\hat{\Psi}^k_i = \frac{1}{M_3}\sum_{r=1}^{M_3} g(\tilde{\tilde{\mathbf{X}}}^{(r)}) S_i(\tilde{\tilde{\mathbf{X}}}^{(r)}; \mathbf{p}^k)$$

where $\tilde{\tilde{\mathbf{X}}}^{(r)}$ are another M_3 independent copies of the input process generated under $\mathbf{p}^k$.

3. Estimate $\hat{\boldsymbol{\phi}}^k_{\mathbf{s}} = (\hat{\phi}^k_{\mathbf{s},1}, \ldots, \hat{\phi}^k_{\mathbf{s},n})$, the gradient of the penalty term with respect to $\mathbf{s}$, with

$$\hat{\phi}^k_{\mathbf{s},j} = -\frac{2}{M_1 + M_2}\Big(\sum_{r=1}^{M_1}(I(h(\mathbf{X}^{(r)}) \leq y_j) - s^k_j) + \sum_{r=1}^{M_2}(I(h(\tilde{\mathbf{X}}^{(r)}) \leq y_j) - s^k_j)\Big)$$

where $\mathbf{X}^{(r)}, \tilde{\mathbf{X}}^{(r)}$ are the same replications used in Step 1.

4. Compute $\mathbf{p}^{k+1} = (p^{k+1}_1, \ldots, p^{k+1}_m)$ by running Algorithm 1 with $p_i = p^k_i$ and $\xi_i = \gamma^k(\lambda^k \hat{\Psi}^k_i + \hat{\phi}^k_{\mathbf{p},i})$, and compute $\mathbf{s}^{k+1} = (s^{k+1}_1, \ldots, s^{k+1}_n)$ by

$$s^{k+1}_j = \Pi_j(s^k_j - \beta^k \hat{\phi}^k_{\mathbf{s},j}).$$

---

the derivative $\boldsymbol{\phi}(\mathbf{p})$ vanishes hence the criterion in Assumption 2 reduces to $\boldsymbol{\Psi}(\mathbf{p})'(\mathbf{p} - \mathbf{p}^*_\epsilon(\lambda)) \geq 0$ when $\lambda > 0$, which in the limit $\lambda \to 0$ forces $\boldsymbol{\Psi}(\mathbf{p})'(\mathbf{p} - \mathbf{p}^*_\epsilon) \geq 0$ since $\mathbf{p}^*_\epsilon(\lambda) \to \mathbf{p}^*_\epsilon$. On the other hand, for an infeasible solution $\mathbf{p}$ the derivative $\boldsymbol{\phi}(\mathbf{p})$ is non-zero, and thus the criterion becomes $\boldsymbol{\phi}(\mathbf{p})'(\mathbf{p} - \mathbf{p}^*_\epsilon) \geq 0$ as $\lambda \to 0$ because $\lambda\boldsymbol{\Psi}(\mathbf{p}) \to \mathbf{0}$ and $\mathbf{p}^*_\epsilon(\lambda) \to \mathbf{p}^*_\epsilon$. Note that Assumption 1 further requires the two inequalities to hold strictly.

Part 2 of Assumption 2 and Assumption 3 impose mild regularity conditions on the solution path of (16) parametrized by λ. In fact, the solution path is expected to be continuously differentiable in λ, a stronger property than the assumptions. The reason is that the optimal solution $\mathbf{p}^*_\epsilon(\lambda)$ has to satisfy the set of KKT conditions which is smooth in the decision variable $\mathbf{p}$ and the penalty coefficient λ, hence an application of the implicit function theorem reveals the continuous differentiability of $\mathbf{p}^*_\epsilon(\lambda)$ in λ.

When the target quantity $\psi(\mathbf{p}) = \mathbf{c}'\mathbf{p}$ for some $\mathbf{c} \in \mathbb{R}^m$, which includes the first type of target quantities we consider in Section 3, the condition $\boldsymbol{\Psi}(\mathbf{p})'(\mathbf{p} - \mathbf{p}^*_\epsilon) > 0$ in Assumption 1 is guaranteed to hold. To explain, note that the feasible set of program (33) is supported by the hyperplane $\{\mathbf{p} : \mathbf{c}'\mathbf{p} = \mathbf{c}'\mathbf{p}^*_\epsilon\}$ at the optimum $\mathbf{p}^*_\epsilon$ even if the feasible set is non-convex, and any non-optimal solution $\mathbf{p}$ will lie in the strict half-space $\{\mathbf{p} : \mathbf{c}'\mathbf{p} > \mathbf{c}'\mathbf{p}^*_\epsilon\}$ which is exactly the condition in Assumption 1. However, the second condition $\boldsymbol{\phi}(\mathbf{p})'(\mathbf{p} - \mathbf{p}^*_\epsilon) > 0$ could still be hard to verify because of the nonlinearity of the constraint functions $E_{\mathbf{p}}[I(h(\mathbf{X}) \le y_j)]$. In our numerical experiments, we investigate the use of multi-start and show that our procedure appears to perform well empirically.

Our convergence guarantee of Algorithm 2 is stated in Theorem 4, whose proof follows the framework in Blum (1954) that considers SA on unconstrained problems.

THEOREM 4. *Under Assumptions 1, 2 and 3, if the step size sequences* $\{\gamma^k\}, \{\beta^k\}$ *and the penalty sequence* $\{\lambda^k\}$ *of Algorithm 2 are chosen as*

$$\begin{aligned}
\gamma^k &= \frac{a}{k^{\alpha_1}},\ \frac{3}{4} < \alpha_1 \le 1 \\
\beta^k &= \frac{b}{k^{\alpha_2}},\ 2 - 2\alpha_1 < \alpha_2 < 2\alpha_1 - 1 \\
\lambda^k &= \begin{cases} \frac{c}{k^{\alpha_3}},\ 0 < \alpha_3 \le 1 - \alpha_1 & \text{if } \frac{3}{4} < \alpha_1 < 1 \\ \frac{c}{\log k} & \text{if } \alpha_1 = 1 \end{cases}
\end{aligned} \tag{39}$$

then $\mathbf{p}^k$ *generated in Algorithm 2 converges to* $\mathbf{p}^*_\epsilon$ *a.s..*

Here γ^k and β^k are chosen in such a way that the slack variables $\mathbf{s}^k$ is guaranteed to stay close to the projections $\Pi_j(E_{\mathbf{p}^k}[I(h(\mathbf{X}) \le y_j)])$ and hence the MDSA is effectively solving (16). Note that

the choice of penalty coefficient λ^k only depends on the step size γ^k. The rule of thumb is that $\gamma^k\lambda^k$ should sum up to ∞, as indicated by the relation between α_1 and α_3 in (39). This ensures sufficient exploration of the feasible region of (33), the rationale of which will be further elaborated in Appendix EC.4.

Finally, we mention that in the presence of a collection of auxiliary input sequences $\mathbf{W}$ with known distribution that is independent of $\mathbf{X}$, namely that we now have $h(\mathbf{X},\mathbf{W})$ instead of $h(\mathbf{X})$ and $g(\mathbf{X},\mathbf{W})$ instead of $g(\mathbf{X})$, all the results in this section hold by viewing $E_{\mathbf{p}}[\cdot]$ as taken jointly with respect to the product measure of $\mathbf{p}$ and the true distribution of $\mathbf{W}$. In Algorithm 2 (and also the other algorithms in the appendix), one only needs to simulate the independent $\mathbf{W}$ in conjunction with $\mathbf{X}$ in each replication, e.g., $h(\mathbf{X}^{(r)},\mathbf{W}^{(r)})$ instead of $h(\mathbf{X}^{(r)})$. Appendix EC.3 provides further discussion.

6. Numerical Results

This section provides numerical illustration of our methodology. We focus on a stylized M/G/1 queue, where we assume known i.i.d. unit rate exponential interarrival times. Our goal is to calibrate the unknown i.i.d. service time distribution P_X given the output data. Here, we assume the collection of data for the averaged wait time of the first 20 customers, starting from the empty state. Say these observations are i.i.d. (e.g., among different days or work cycles), denoted $y_1,\ldots,y_n$. The data size n varies from 30 to 100 in our experiments.

We consider two target quantities of interest $\psi(P_X)$: 1) the expected averaged queue length seen by the first 20 customers. This performance measure, though related to the waiting time data, is not directly observable and depends on the unknown service time distribution; 2) the distribution function of the service time. We also consider two different "true" service time distributions, first one is exponential with rate 1.2, and second one is a mixture of beta distributions that has a bimodal shape. We set the confidence level to be 95%, i.e., $\alpha = 5\%$.

Since the input distribution of interest and the output distribution are both continuous, we use optimization programs (12) and (13) to infer the confidence bounds on $\psi(P_X^0)$. From Theorem 2,

we first randomly sample m support points from some "safe" input distribution (i.e., distribution believed to have heavier tail than the truth), where m varies from 100 to 500 in our experiments. Then we implement Algorithm 2. In our implementation we choose $M_1 = M_2 = M_3 = 100$, $\gamma^k = a/k^{0.8}$, $\beta^k = b/k^{0.5}$, $\lambda^k = c/k^{0.2}$, in which the constants a, b, c will be determined slightly different in different cases. The iteration stops when $\|\mathbf{p}^{k+1} - \mathbf{p}^k\|_\infty \leq 0.0005$.

6.1. Inferring the Average Queue Length

We first consider inferring the average queue length $E_{P_X}[g(\mathbf{X})]$, and consider a small output data size $n = 30$ for the average waiting time. In this setting, the true service time distribution is set as exponential with rate 1.2. We generate the input support points with a lognormal distribution with parameter $\mu = 0$ and standard deviation $\sigma = 1$. In light of Theorem 2, we choose $m = 100$ to make m bigger than n. Figure 1 shows the trend of the objective value $E_{\mathbf{p}}[g(\mathbf{X})]$ when we apply Algorithm 2 to the max and the min problems. The algorithm appears to converge fairly quickly (within about 10 iterations). The jitter of the trend is due to the evaluation of the objective values, for each of whom we use $100,000$ simulation runs. The minimization stops at 0.622 and the maximization stops at 0.688 according to our stopping criterion described above. This gives us an interval $[0.622, 0.688]$. The true value in this case is $E_{\mathbf{p}}[g(\mathbf{X})] = 0.636$ (from running 1 million simulation using the true service time distribution), thus demonstrating that the confidence interval we obtained covers the truth. Moreover, the interval we obtained is encouragingly tight.

We also investigate the shape of the input distribution when the algorithm stops. This is shown in Figure 2. We observe that both the obtained maximal and minimal distributions place more masses on the lower value than the upper, roughly following the true exponential distribution. We should mention, however, that the shapes of the obtained optimal distributions are not indicative of the performance of our method, as the latter intends to compute valid bounds for a target quantity, namely the average queue length in this example, instead of direct recovery of the input distribution. The shapes in Figure 2 should be interpreted as the worst-case distributions that give rise to the lower and upper bounds for the queue length. The resemblance of these distributions to

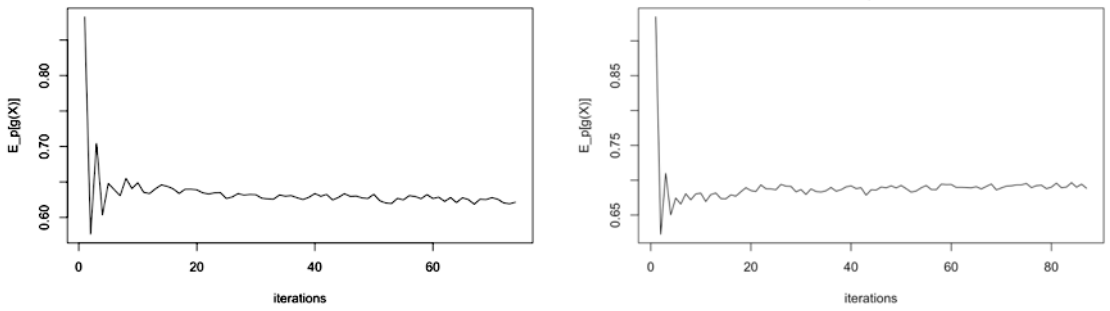

Figure 1 Objective value of the minimization (left) and maximization (right) for the expected queue length using Algorithm 2 against the iteration number; $n = 30, m = 100$; true service time distribution is exponential

the true one leads us to conjecture that the service time distribution could be close to identifiable with the waiting time data.

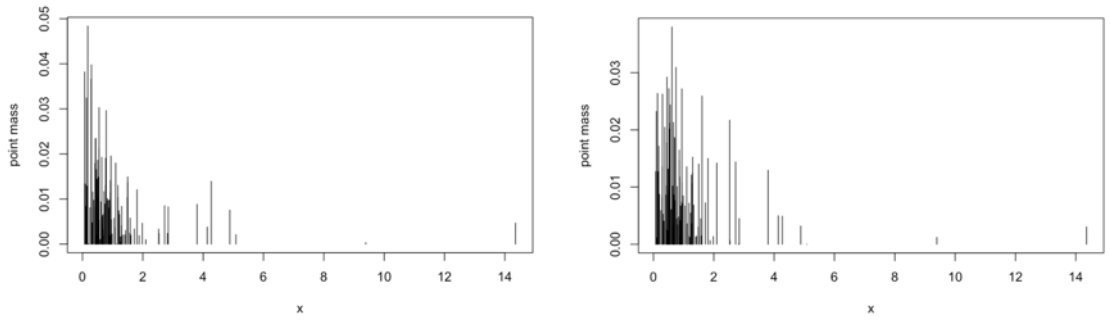

Figure 2 Minimal (left) and maximal (right) distribution of the service time for bounding the expected queue length; $n = 30, m = 100$; true service time distribution is exponential

Next we increase our support size m to 200, keeping the output data size n fixed at 30. Like the previous case, we show the trend of the objective value as the algorithm progresses, in Figure 3. Compared to the case $m = 100$, the algorithm appears to stabilize faster, at around 5 iteration, and exhibit a more monotonic trend (which could be due to our initialization). The minimization stops at 0.622 and the maximization stops at 0.647. This gives us an interval $[0.622, 0.647]$ which again covers the true value 0.636, and is shorter than the one obtained when $m = 100$. Finally, the obtained maximal and minimal distributions, shown in Figure 4, show a pattern even closer to the exponential distribution.

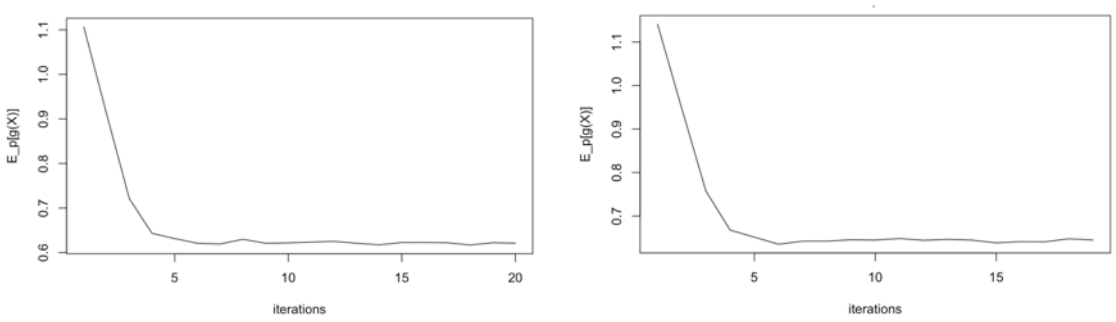

Figure 3 Objective value of the minimization (left) and maximization (right) for the expected queue length using Algorithm 2 against the iteration number; $n = 30, m = 200$; true service time distribution is exponential

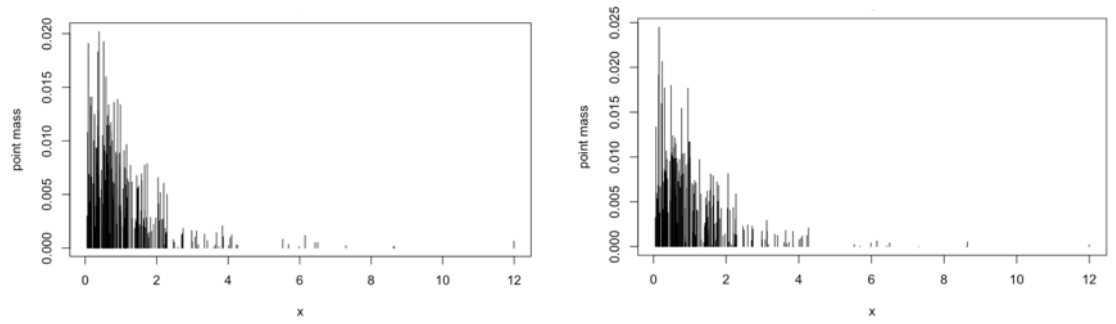

Figure 4 Minimal (left) and maximal (right) distribution of the service time for bounding the expected queue length; $n = 30, m = 200$; true service time distribution is exponential

We increase the support size m further to 300 or the data size n to 100. Table 1 shows the obtained optimal values. These runs provide valid lower and upper bounds for the true value 0.636, except when $m = 300$ and $n = 30$ that misses marginally. The interval lengths do not seem to vary much; all are around $0.03 - 0.06$. Nonetheless, comparing between the cases $n = 30$ and $n = 100$, when $m = 100$, we see that the resulting interval becomes tighter as n increases, which matches our theoretical implication that our bounds should get tighter with a larger data size.

The selection of a, b, c in γ^k, β^k, λ^k depends on m and n. We have selected $a = 0.2$ when $m = 100$ and $n = 30$, $a = 0.1$ and 0.075 when $m = 200$ and 300 while $n = 30$, and $a = 0.1$ when $m = 100$ and $n = 100$. We always choose $b = 0.2$ and $c = 1$. These choices appear to work well. Regarding running times, when $m = 100$ and $n = 30$, each iteration takes about 40 seconds. The running time seems to increase linearly as m and n increase.

m	n	min value	max value
100	30	0.622	0.688
200	30	0.622	0.647
300	30	0.593	0.629
100	100	0.627	0.652

Table 1 Optimal values for bounding the expected queue length under different combinations of n and m; true service time distribution is exponential

Next we check how the initialization of the probability weights in the algorithm affects the obtained optimal values. This is especially important since our algorithm is only guaranteed local convergence. We randomly generate 34 initial distributions of $\mathbf{p}$ from a Dirichlet distribution to run the algorithm. Figure 5 shows the boxplot of the obtained optimal values under different initial distributions. The minimum value varies from 0.621 to 0.635, whereas the maximum value varies from 0.648 to 0.665. The differences among the initial distributions seem to be quite small compared to the gap between the minimum and maximum values, and the true value 0.636 is always covered. This shows that the algorithm tends to converge to the same optimal solution or solutions that have similar objective values.

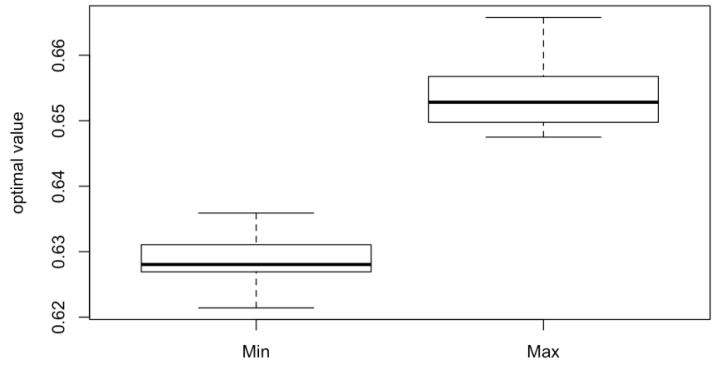

Figure 5 Minimum and maximum values for the expected queue length under different initializations; $n = 30, m = 100$; true service time distribution is exponential

We then test the coverage of our obtained bounds. For this, we repeatedly sample new output data set of size $n = 30$ for 100 times. For each data set, we generate new support points of size $m = 100$. Then we run Algorithm 2. Out of 100 intervals we obtained, 95 of them cover the true expected queue length. This gives us a 95% confidence interval for the coverage probability $[0.91, 0.99]$, which is consistent with the theoretical guarantee provided by Theorem 2.

We have also tested the use of randomized stochastic projected gradient (RSPG), proposed by Ghadimi et al. (2016), that has been shown to perform well theoretically and empirically for problems with non-convex stochastic objectives. Specifically, we adapt the algorithm in Section 4.1 and 4.2 of Ghadimi et al. (2016) heuristically for the current problem we face that has stochastic non-convex constraints. Algorithm 4 in the appendix shows the adaptation of a single run procedure, and Algorithm 5 shows the adaptation of a post-optimization step to boost the final performance. In our algorithmic specification, we choose $N = 30$, $S = 5$, $M = 500$, $M' = 500$, $\bar{\gamma} = 0.03$, and we fix λ at 0.03. We run Algorithm 5 for two realizations of data and support generation when the true service time distribution is exponential, with $n = 30$ and $m = 100$. For each realization, we also run Algorithm 2 for comparison. For the first realization, we obtained $[0.622, 0.640]$ using RSPG, compared with $[0.626, 0.658]$ using Algorithm 2. For the second realization, we obtained $[0.616, 0.644]$ using RSPG, compared with $[0.621, 0.660]$ using Algorithm 2. The RSPG thus appears to perform very similarly as our procedure, at least for this particular setup (which shows that RSPG could be an alternative for future investigation).

We test the sensitivity of the algorithm with respect to the bounds in the constraints provided by the KS statistic. More concretely, in Algorithm 2, we increase the number $q_{1-\alpha}/\sqrt{n}$ in the constraint interval by a small δ. Table 2 shows that the obtained bounds are quite stable and do not show significant changes.

Finally, we test with a more "challenging" service time distribution that is an equally weighted mixture of two beta distributions with parameters $\alpha = 9, \beta = 3$ and $\alpha = 3, \beta = 9$. This bimodal distribution has highest masses around 0.2 and 0.8, with a shape shown in Figure 6.

perturbation size	min value	max value
0.01	0.625	0.649
0.02	0.628	0.649
0.03	0.624	0.643
0.05	0.621	0.646

Table 2 Effect on optimal values for bounding the expected queue length when perturbing the interval in the optimization constraint; $n = 30, m = 100$; true service time distribution is exponential

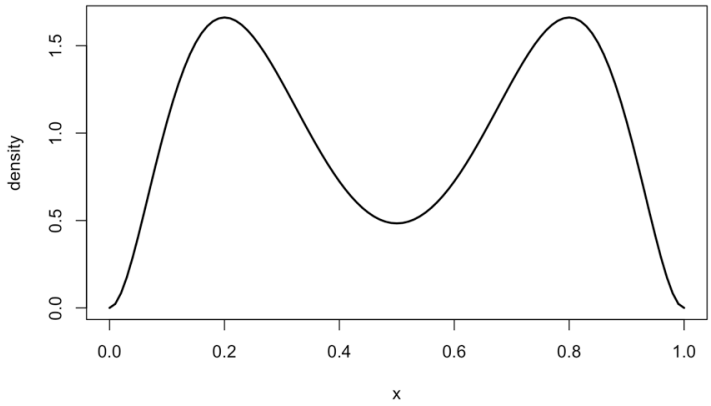

Figure 6 Density of a mixture of two beta distributions

We consider the setting with $n = 50$ output observations. We randomly select $m = 100$ input support points from uniform distribution in $[0, 1]$, and run Algorithm 2, using the same specifications as in the previous setup. The minimization stops at the value 0.242 and the maximization stops at 0.284. These cover the true value 0.274 (from running 1 million simulation using the true service time distribution). Thus our method appears to continue working in this case.

Figure 7 shows the minimal and maximal distributions from Algorithm 2. The distributions are quite spread out throughout the support, with the minimal distribution showing an apparent noisy bimodal pattern. As we have discussed before, the shapes of these distributions should be interpreted as the worst-case distributions giving rise to the bounds instead of indicators of the performance of our approach. Nonetheless, the spread over the entire support, in contrast to the

exponential-like pattern in Figure 4, suggests that our approach indeed captures the general shape of the true distribution.

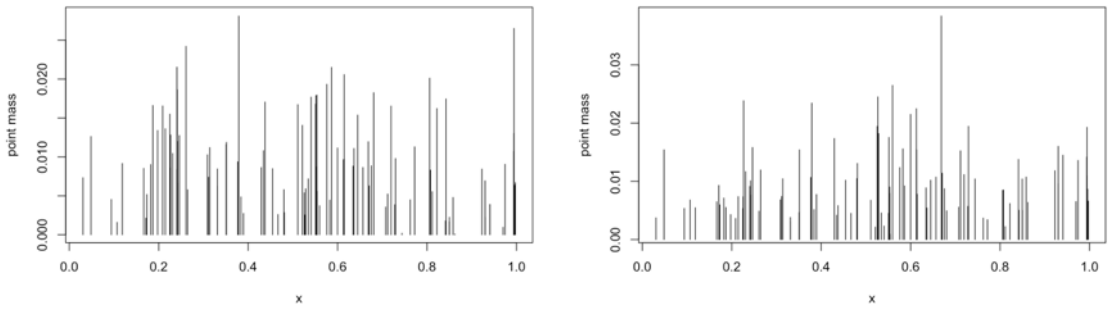

Figure 7 Minimal (left) and maximal (right) distribution of the service time for bounding the expected queue length; $n = 50, m = 100$; true service time distribution is mixture of betas

6.2. Inferring the Input Distribution Function

We now consider inferring the distribution function of the service time, i.e., $P_X(X \leq a)$ for a range of values a. We first use a true service time distribution that is exponential with rate 1.2. We consider a collection of $n = 50$ observations from the average waiting time. We randomly generate $m = 100$ support points from a lognormal distribution with $\mu = 0$ and $\sigma^2 = 1$. We use Algorithm 2 with parameters $\gamma^k = 0.1/k^{0.8}$, $\beta^k = 0.1/k^{0.5}$, $\lambda^k = 1/k^{0.2}$.

Table 3 shows the obtained maximum and minimum values compared with the true distribution function evaluated at values a ranging from 0.3 to 1.2. Figure 8 further plots the trends of these values. The dashed lines represent the maximum and minimum values, and the solid line represents the true values. Note that Proposition EC.1, and the analogous extension of Theorem 2 to multiple objective functions discussed at the end of Section 4.2, allow us to compute the bounds for different a values simultaneously with little sacrifice of statistical accuracy. In Table 3 and Figure 8, the obtained optimal values cover the truth at all points except the leftmost $a = 0.3$. This could be due to the challenge in inferring the tail (either left or right), stemming from perhaps the observed

a	min value	max value	true value
0.3	0.118	0.250	0.302
0.4	0.302	0.441	0.381
0.5	0.398	0.464	0.451
0.6	0.435	0.565	0.513
0.7	0.506	0.579	0.568
0.8	0.601	0.673	0.617
0.9	0.636	0.735	0.660
1	0.699	0.741	0.699
1.1	0.723	0.756	0.733
1.2	0.756	0.798	0.763

Table 3 Minimum, maximum and true values of the distribution function $P_X(X \leq a)$ of the service time across a; $n = 50, m = 100$; true service time distribution that is exponential

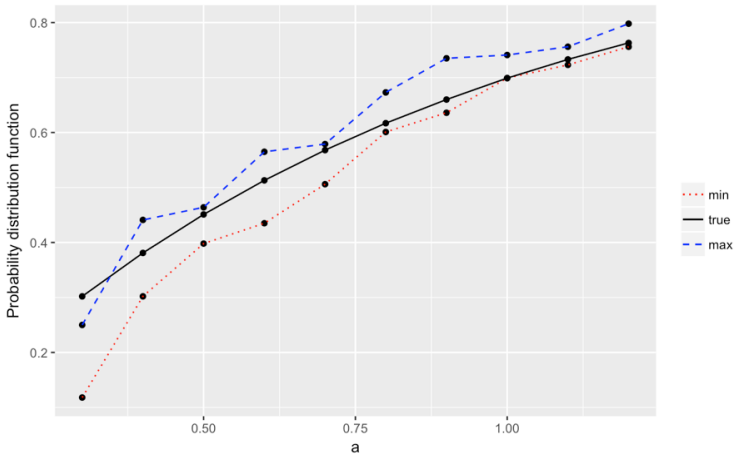

Figure 8 Bounds and true distribution function values for the service time, when the true service time distribution is exponential; $n = 50, m = 100$

output we use (i.e., the waiting time) or the statistic we use to form our uncertainty set (i.e., the KS-statistic, which is known to be quite insensitive to the tail of a distribution).

Figure 9 shows the minimal and maximal distributions for bounding $P_X(X \leq 0.5)$ when the algorithm terminates. We see that the shapes of both distributions resemble exponential, hinting that the service time distribution is close to identifiable in this case.

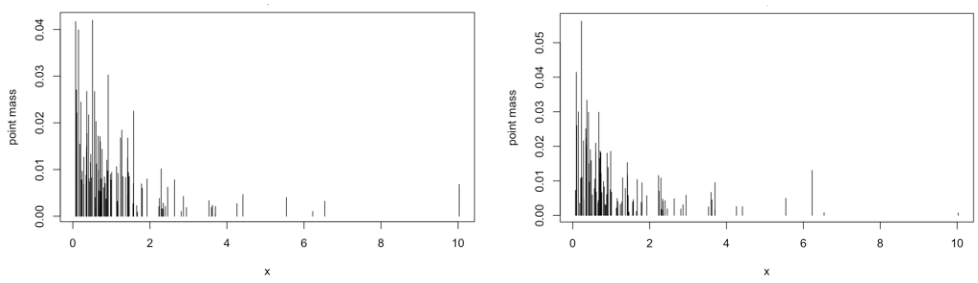

Figure 9 Minimal (left) and maximal (right) distribution of the service time for bounding $P_X(X \leq 0.5)$, when the true service time distribution is exponential; $n = 50, m = 100$

Next, we investigate the case when the true service time distribution is a mixture of two beta distributions with parameters $\alpha = 9, \beta = 3$ and $\alpha = 3, \beta = 9$. We consider a collection of $n = 50$ observations from the average waiting time. We randomly generate $m = 100$ support points from a uniform distribution on $[0, 1]$.

Like in the previous case, Table 4 shows the maximum and minimum values from Algorithm 2, against the true values of $P_X(X \leq a)$ at different a values. Figure 10 further plots the trends of these values. Here, the obtained optimal values all cover the truth except at $a = 0.35$. The latter could be attributed to the statistical noise when running the many optimization procedures. The point $a = 0.35$ is also one that could be "difficult" to infer intuitively, as it is in between the two modes. Nonetheless, our procedure appears to be reliable in general in bounding the distribution function across the domain of the service time.

Figure 11 shows the minimal and maximal distributions for bounding $P_X(X \leq 0.5)$ when the algorithm terminates. The shapes of these distributions are now considerably noisier than the exponential case in Figure 9. Nonetheless, there is a rough bimodal pattern (around 0.2 and 0.7).

7. Conclusion

We have studied an optimization-based framework to calibrate input quantities in stochastic simulation with only the availability of output data. Our approach uses an output-level uncertainty set, inspired by the DRO literature, to represent the statistical noise of the output data. By expressing the output distribution in terms of a simulable map of the input distribution, we can

a	min value	max value	true value
0.2	0.129	0.231	0.188
0.25	0.208	0.266	0.267
0.3	0.262	0.358	0.337
0.35	0.296	0.395	0.393
0.4	0.362	0.413	0.435
0.45	0.389	0.464	0.466
0.5	0.416	0.503	0.491
0.55	0.504	0.577	0.516
0.6	0.509	0.594	0.548
0.65	0.573	0.611	0.591
0.7	0.628	0.679	0.649
0.75	0.678	0.736	0.722
0.8	0.724	0.834	0.805

Table 4 Minimum, maximum and true values of the distribution function $P_X(X \leq a)$ of the service time across a, under a true service time distribution that is mixture of betas; $n = 50, m = 100$

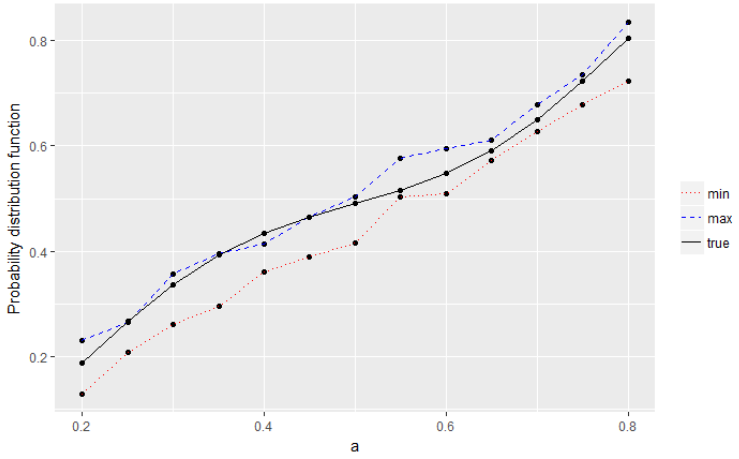

Figure 10 Bounds and true distribution function values for the service time, when the true service time distribution is mixture of betas; $n = 50, m = 100$

set up optimization programs cast over the input distribution that infers valid confidence bounds on the input quantities of interest.

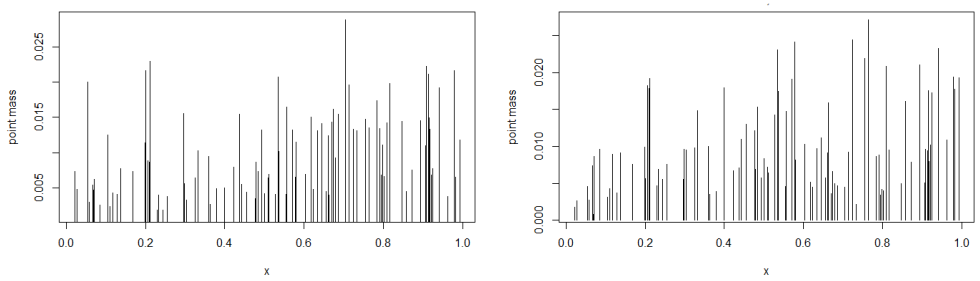

Figure 11 Minimal (left) and maximal (right) distribution of the service time for bounding $P_X(X \leq 0.5)$, when the true service time distribution is mixture of betas; $n = 50, m = 100$

We propose in particular an output-level uncertainty set based on the KS statistic, which exhibits advantages in computation (thanks to reformulation) and statistical accuracy (thanks to a controllable discretization scale needed to retain the confidence guarantee). We have shown these advantages via looking at the complexity of the resulting constraints and invoking the empirical process theory for U-statistics. We also study a stochastic quadratic penalty method to solve the resulting optimization problems, including a convergence analysis that informs the suitable tuning of the parameters. Our numerical results demonstrate how our method could provide valid bounds for input quantities such as the input distribution function and other performance measures that rely on the input.

Acknowledgments

A preliminary conference version of this work has appeared in Goeva et al. (2014). We gratefully acknowledge support from the National Science Foundation under grants CMMI-1542020, CMMI-1523453 and CAREER CMMI-1653339. We also thank Peter Haas for suggesting the use of quantile-based moments, and Russell Barton, Shane Henderson and Barry Nelson for other helpful suggestions.

References

Arcones MA, Gine E (1993) Limit theorems for u-processes. *The Annals of Probability* 1494–1542.

Avellaneda M, Buff R, Friedman C, Grandechamp N, Kruk L, Newman J (2001) Weighted Monte Carlo: a new technique for calibrating asset-pricing models. *International Journal of Theoretical and Applied Finance* 4(01):91–119.

Balci O, Sargent RG (1982) Some examples of simulation model validation using hypothesis testing. *Proceedings of the 14th Winter Simulation conference*, volume 2, 621–629 (Winter Simulation Conference).

Banks J, Carson J, Nelson B, Nicol D (2009) *Discrete-Event System Simulation* (Prentice Hall Englewood Cliffs, NJ, USA), 5th edition edition.

Barton RR (2012) Tutorial: Input uncertainty in outout analysis. *Proceedings of the 2012 Winter Simulation Conference (WSC)*, 1–12 (IEEE).

Barton RR, Nelson BL, Xie W (2013) Quantifying input uncertainty via simulation confidence intervals. *INFORMS Journal on Computing* 26(1):74–87.

Barton RR, Schruben LW (2001) Resampling methods for input modeling. *Proceedings of the 2001 Winter Simulation Conference*, volume 1, 372–378 (IEEE).

Basawa I, Bhat U, Zhou J (2008) Parameter estimation using partial information with applications to queueing and related models. *Statistics & Probability Letters* 78(12):1375–1383.

Basawa IV, Bhat UN, Lund R (1996) Maximum likelihood estimation for single server queues from waiting time data. *Queueing systems* 24(1-4):155–167.

Bayraksan G, Love DK (2015) Data-driven stochastic programming using phi-divergences. *The Operations Research Revolution*, 1–19 (INFORMS).

Beck A, Teboulle M (2003) Mirror descent and nonlinear projected subgradient methods for convex optimization. *Operations Research Letters* 31(3):167–175.

Ben-Tal A, Den Hertog D, De Waegenaere A, Melenberg B, Rennen G (2013) Robust solutions of optimization problems affected by uncertain probabilities. *Management Science* 59(2):341–357.

Ben-Tal A, El Ghaoui L, Nemirovski A (2009) *Robust optimization* (Princeton University Press).

Benveniste A, Métivier M, Priouret P (2012) *Adaptive Algorithms and Stochastic Approximations*, volume 22 (Springer Science & Business Media).

Bertsekas DP (1999) *Nonlinear programming* (Athena Scientific).

Bertsekas DP, Nedi A, Ozdaglar AE, et al. (2003) Convex analysis and optimization .

Bertsimas D, Brown DB, Caramanis C (2011) Theory and applications of robust optimization. *SIAM review* 53(3):464–501.

Bertsimas D, Gupta V, Kallus N (2014) Robust saa. *arXiv preprint arXiv:1408.4445* .

Bertsimas D, Natarajan K (2007) A semidefinite optimization approach to the steady-state analysis of queueing systems. *Queueing Systems* 56(1):27–39.

Bertsimas D, Popescu I (2005) Optimal inequalities in probability theory: A convex optimization approach. *SIAM Journal on Optimization* 15(3):780–804.

Bingham N, Pitts SM (1999) Non-parametric estimation for the M/G/∞ queue. *Annals of the Institute of Statistical Mathematics* 51(1):71–97.

Blanchet J, Kang Y, Murthy K (2016) Robust wasserstein profile inference and applications to machine learning. *arXiv preprint arXiv:1610.05627* .

Blanchet J, Murthy K (2016) Quantifying distributional model risk via optimal transport .

Blum JR (1954) Multidimensional stochastic approximation methods. *The Annals of Mathematical Statistics* 737–744.

Bottou L (1998) Online learning and stochastic approximations. *On-line learning in neural networks* 17(9):142.

Broadie M, Cicek D, Zeevi A (2011) General bounds and finite-time improvement for the Kiefer-Wolfowitz stochastic approximation algorithm. *Operations Research* 59(5):1211–1224.

Cheng RC, Holland W (1998) Two-point methods for assessing variability in simulation output. *Journal of Statistical Computation Simulation* 60(3):183–205.

Cheng RC, Holland W (2004) Calculation of confidence intervals for simulation output. *ACM Transactions on Modeling and Computer Simulation (TOMACS)* 14(4):344–362.

Chick SE (2001) Input distribution selection for simulation experiments: accounting for input uncertainty. *Operations Research* 49(5):744–758.

Chick SE, Ng SH (2002) Simulation input analysis: joint criterion for factor identification and parameter estimation. *Proceedings of the 34th Winter Simulation Conference*, 400–406 (Winter Simulation Conference).

Cooper RB (1972) Introduction to queueing theory .

Csiszár I (1991) Why least squares and maximum entropy? an axiomatic approach to inference for linear inverse problems. *The Annals of Statistics* 19(4):2032–2066.

Currin C, Mitchell T, Morris M, Ylvisaker D (1991) Bayesian prediction of deterministic functions, with applications to the design and analysis of computer experiments. *Journal of the American Statistical Association* 86(416):953–963.

Daley D, Servi L (1998) Moment estimation of customer loss rates from transactional data. *International Journal of Stochastic Analysis* 11(3):301–310.

Dang CD, Lan G (2015) Stochastic block mirror descent methods for nonsmooth and stochastic optimization. *SIAM Journal on Optimization* 25(2):856–881.

Delage E, Ye Y (2010) Distributionally robust optimization under moment uncertainty with application to data-driven problems. *Operations Research* 58(3):595–612.

Donoho DL, Johnstone IM, Hoch JC, Stern AS (1992) Maximum entropy and the nearly black object. *Journal of the Royal Statistical Society. Series B (Methodological)* 41–81.

Duchi J, Glynn P, Namkoong H (2016) Statistics of robust optimization: A generalized empirical likelihood approach. *arXiv preprint arXiv:1610.03425* .

Durrett R (2010) *Probability: Theory and Examples* (Cambridge university press).

Esfahani PM, Kuhn D (2015) Data-driven distributionally robust optimization using the wasserstein metric: Performance guarantees and tractable reformulations. *arXiv preprint arXiv:1505.05116* .

Fan W, Hong LJ, Zhang X (2013) Robust selection of the best. *Proceedings of the 2013 Winter Simulation Conference: Simulation: Making Decisions in a Complex World*, 868–876 (IEEE Press).

Fearnhead P (2004) Filtering recursions for calculating likelihoods for queues based on inter-departure time data. *Statistics and Computing* 14(3):261–266.

Feng H, Dube P, Zhang L (2014) Estimating life-time distribution by observing population continuously. *Performance Evaluation* 79:182–197.

Frey JC, Kaplan EH (2010) Queue inference from periodic reporting data. *Operations Research Letters* 38(5):420–426.

Gao R, Kleywegt AJ (2016) Distributionally robust stochastic optimization with wasserstein distance. *arXiv preprint arXiv:1604.02199* .

Ghadimi S, Lan G (2013) Stochastic first-and zeroth-order methods for nonconvex stochastic programming. *SIAM Journal on Optimization* 23(4):2341–2368.

Ghadimi S, Lan G (2015) Accelerated gradient methods for nonconvex nonlinear and stochastic programming. *Mathematical Programming* 1–41.

Ghadimi S, Lan G, Zhang H (2016) Mini-batch stochastic approximation methods for nonconvex stochastic composite optimization. *Mathematical Programming* 155(1-2):267–305.

Ghosh S, Lam H (2015a) Computing worst-case input models in stochastic simulation. *Available at http://arxiv.org/pdf/1507.05609v1.pdf* .

Ghosh S, Lam H (2015b) Mirror descent stochastic approximation for computing worst-case stochastic input models. *Proceedings of the 2015 Winter Simulation Conference*, 425–436 (IEEE Press).

Ghosh S, Lam H (2015c) Robust analysis in stochastic simulation: Computation and performance guarantees. *arXiv preprint arXiv:1507.05609* .

Glasserman P, Xu X (2013) Robust portfolio control with stochastic factor dynamics. *Operations Research* 61(4):874–893.

Glasserman P, Xu X (2014) Robust risk measurement and model risk. *Quantitative Finance* 14(1):29–58.

Glasserman P, Yang L (2016) Bounding wrong-way risk in cva calculation. *Mathematical Finance* .

Glasserman P, Yu B (2005) Large sample properties of weighted Monte Carlo estimators. *Operations Research* 53(2):298–312.

Goeva A, Lam H, Zhang B (2014) Reconstructing input models via simulation optimization. *Proceedings of the 2014 Winter Simulation Conference*, 698–709 (IEEE Press).

Goh J, Sim M (2010) Distributionally robust optimization and its tractable approximations. *Operations Research* 58(4-Part-1):902–917.

Gupta V (2015) Near-optimal ambiguity sets for distributionally robust optimization. *Preprint* .

Hall P, Park J (2004) Nonparametric inference about service time distribution from indirect measurements. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 66(4):861–875.

Hanasusanto GA, Roitch V, Kuhn D, Wiesemann W (2017) Ambiguous joint chance constraints under mean and dispersion information. *Operations Research* 65(3):751–767.

Hansen LP, Sargent TJ (2008) *Robustness* (Princeton university press).

Heckmüller S, Wolfinger BE (2009) Reconstructing arrival processes to G/D/1 queueing systems and tandem networks. *International Symposium on Performance Evaluation of Computer & Telecommunication Systems, 2009. SPECTS 2009.*, volume 41, 361–368 (IEEE).

Hu Z, Cao J, Hong LJ (2012) Robust simulation of global warming policies using the dice model. *Management science* 58(12):2190–2206.

Iyengar GN (2005) Robust dynamic programming. *Mathematics of Operations Research* 30(2):257–280.

Jain A, Lim A, Shanthikumar J (2010) On the optimality of threshold control in queues with model uncertainty. *Queueing Systems* 65:157–174.

Kelton WD, Law AM (2000) *Simulation Modeling and Analysis* (McGraw Hill Boston).

Kennedy MC, O'Hagan A (2001) Bayesian calibration of computer models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 63(3):425–464.

Kim YB, Park J (2008) New approaches for inference of unobservable queues. *Proceedings of the 40th Conference on Winter Simulation*, 2820–2825 (Winter Simulation Conference).

Kleijnen JP (1995) Verification and validation of simulation models. *European Journal of Operational Research* 82(1):145–162.

Kraan B, Bedford T (2005) Probabilistic inversion of expert judgments in the quantification of model uncertainty. *Management Science* 51(6):995–1006.

Lam H (2016a) Recovering best statistical guarantees via the empirical divergence-based distributionally robust optimization. *Forthcoming in Operations Research. Available at arXiv preprint arXiv:1605.09349* .

Lam H (2016b) Robust sensitivity analysis for stochastic systems. *Mathematics of Operations Research* 41(4):1248–1275.

Lam H (2017) Sensitivity to serial dependency of input processes: A robust approach. *Management Science* .

Lam H, Mottet C (2017) Tail analysis without parametric models: A worst-case perspective. *Operations Research* 65(6):1696–1711.

Lam H, Zhou E (2017) The empirical likelihood approach to quantifying uncertainty in sample average approximation. *Operations Research Letters* 45(4):301–307.

Lan G, Zhou Z (2017) Algorithms for stochastic optimization with expectation constraints. *arXiv preprint arXiv:1604.03887* .

Larson RC (1990) The queue inference engine: Deducing queue statistics from transactional data. *Management Science* 36(5):586–601.

Lehmann EL, Romano JP (2006) *Testing statistical hypotheses* (Springer Science & Business Media).

Li B, Jiang R, Mathieu JL (2016) Ambiguous risk constraints with moment and unimodality information. *Available at Optimization Online* .

Lim AEB, Shanthikumar JG (2007) Relative entropy, exponential utility, and robust dynamic pricing. *Operations Research* 55(2):198–214.

Mandelbaum A, Zeltyn S (1998) Estimating characteristics of queueing networks using transactional data. *Queueing systems* 29(1):75–127.

Marjoram P, Molitor J, Plagnol V, Tavaré S (2003) Markov chain Monte Carlo without likelihoods. *Proceedings of the National Academy of Sciences* 100(26):15324–15328.

Moulines E, Roueff F, Souloumiac A, Trigano T (2007) Nonparametric inference of photon energy distribution from indirect measurement. *Bernoulli* 13(2):365–388.

Nelson B (2016) 'Some tactical problems in digital simulation' for the next 10 years. *Journal of Simulation* 10(1):2–11.

Nemirovski A, Juditsky A, Lan G, Shapiro A (2009) Robust stochastic approximation approach to stochastic programming. *SIAM Journal on Optimization* 19(4):1574–1609.

Park J, Kim YB, Willemain TR (2011) Analysis of an unobservable queue using arrival and departure times. *Computers & Industrial Engineering* 61(3):842–847.

Petersen I, James M, Dupuis P (2000) Minimax optimal control of stochastic uncertain systems with relative entropy constraints. *IEEE Transactions on Automatic Control* 45(3):398–412.

Pickands III J, Stine RA (1997) Estimation for an M/G/∞ queue with incomplete information. *Biometrika* 84(2):295–308.

Popescu I (2005) A semidefinite programming approach to optimal-moment bounds for convex classes of distributions. *Mathematics of Operations Research* 30(3):632–657.

Ross JV, Taimre T, Pollett PK (2007) Estimation for queues from queue length data. *Queueing Systems* 55(2):131–138.

Ryzhov IO, Defourny B, Powell WB (2012) Ranking and selection meets robust optimization. *Proceedings of the Winter Simulation Conference*, 48 (Winter Simulation Conference).

Santner TJ, Williams BJ, Notz WI (2013) *The Design and Analysis of Computer Experiments* (Springer Science & Business Media).

Sargent RG (2005) Verification and validation of simulation models. *Proceedings of the 37th Winter Simulation Conference*, 130–143 (Winter Simulation Conference).

Schruben LW (1980) Establishing the credibility of simulations. *Simulation* 34(3):101–105.

Serfling RJ (2009) *Approximation Theorems of Mathematical Statistics*, volume 162 (John Wiley & Sons).

Shafieezadeh-Abadeh S, Esfahani PM, Kuhn D (2015) Distributionally robust logistic regression. *Advances in Neural Information Processing Systems*, 1576–1584.

Shirangi MG (2014) History matching production data and uncertainty assessment with an efficient TSVD parameterization algorithm. *Journal of Petroleum Science and Engineering* 113:54–71.

Smith JE (1995) Generalized Chebyshev inequalities: theory and applications in decision analysis. *Operations Research* 43(5):807–825.

Song E, Nelson BL, Pegden CD (2014) Advanced tutorial: Input uncertainty quantification. *Proceedings of the 2014 Winter Simulation Conference*, 162–176 (IEEE Press).

Tarantola A (2005) *Inverse Problem Theory and Methods for Model Parameter Estimation* (SIAM).

Van Der Vaart AW, Wellner JA (1996) *Weak convergence and empirical processes* (Springer).

Wang IJ, Spall JC (2008) Stochastic optimisation with inequality constraints using simultaneous perturbations and penalty functions. *International Journal of Control* 81(8):1232–1238.

Wang TY, Ke JC, Wang KH, Ho SC (2006) Maximum likelihood estimates and confidence intervals of an M/M/R queue with heterogeneous servers. *Mathematical Methods of Operations Research* 63(2):371–384.

Whitt W (1981) Approximating a point process by a renewal process: The view through a queue, an indirect approach. *Management Science* 27(6):619–636.

Whitt W (1982) Approximating a point process by a renewal process, I: Two basic methods. *Operations Research* 30(1):125–147.

Whitt W (2012) Fitting birth-and-death queueing models to data. *Statistics & Probability Letters* 82(5):998–1004.

Wiesemann W, Kuhn D, Sim M (2014) Distributionally robust convex optimization. *Operations Research* 62(6):1358–1376.

Wunsch C (1996) *The Ocean Circulation Inverse Problem* (Cambridge University Press).

Xie W, Ahmed S (2018) Distributionally robust chance constrained optimal power flow with renewables: A conic reformulation. *IEEE Transactions on Power Systems* 33(2):1860–1867.

Xin L, Goldberg DA (2015) Distributionally robust inventory control when demand is a martingale. *arXiv preprint arXiv:1511.09437* .

Xu H, Mannor S (2012) Distributionally robust markov decision processes. *Mathematics of Operations Research* 37(2):288–300.

Yu H, Neely M, Wei X (2017) Online convex optimization with stochastic constraints. Guyon I, Luxburg UV, Bengio S, Wallach H, Fergus R, Vishwanathan S, Garnett R, eds., *Advances in Neural Information Processing Systems 30*, 1427–1437 (Curran Associates, Inc.).

Zhang Y, Shen S, Mathieu JL (2017) Distributionally robust chance-constrained optimal power flow with uncertain renewables and uncertain reserves provided by loads. *IEEE Transactions on Power Systems* 32(2):1378–1388.

Zhao C, Jiang R (2018) Distributionally robust contingency-constrained unit commitment. *IEEE Transactions on Power Systems* 33(1):94–102.

Zouaoui F, Wilson JR (2004) Accounting for input-model and input-parameter uncertainties in simulation. *IIE Transactions* 36(11):1135–1151.

Supplementary Materials

EC.1. Proofs and Additional Results for Section 3

Proof of Proposition 1. Note that if $P_Y^0 = \gamma(P_X^0) \in \mathcal{U}$, then P_X^0 must be a feasible solution for programs (1) and (2), and consequently $\underline{Z} \le \psi(P_X^0) \le \overline{Z}$. This implies that

$$\mathbb{P}_D(\underline{Z} \le \psi(P_X^0) \le \overline{Z}) \ge \mathbb{P}_D(P_Y^0 \in \mathcal{U})$$

concluding the proposition. □

Proposition EC.1. *Let P_X^0 and P_Y^0 be the true input and output distributions. Consider a collection of quantities $\psi_l(P_X), l = 1, \dots, L$ and the collection of optimization programs*

$$\begin{aligned} &\max && \psi_l(P_X) \\ &\textit{subject to } && P_Y \in \mathcal{U} \end{aligned} \tag{EC.1}$$

and

$$\begin{aligned} &\min && \psi_l(P_X) \\ &\textit{subject to } && P_Y \in \mathcal{U} \end{aligned} \tag{EC.2}$$

for $l = 1, \dots, L$. Suppose $\mathcal{U}$ is a confidence region for P_Y^0, i.e.,

$$\mathbb{P}_D(P_Y^0 \in \mathcal{U}) = 1 - \alpha$$

where $\mathbb{P}_D(\cdot)$ denotes the probability with respect to the data D. Let $\overline{Z}_l, \underline{Z}_l, l = 1, \dots, L$ be the set of optimal values of (EC.1) *and* (EC.2) *respectively. Then we have*

$$\mathbb{P}_D(\underline{Z}_l \le \psi_l(P_X^0) \le \overline{Z}_l, l = 1, \dots, L) \ge 1 - \alpha$$

Similar statements hold if the confidence is approximate, i.e., if

$$\liminf_{n \to \infty} \mathbb{P}_D(P_Y^0 \in \mathcal{U}) \ge 1 - \alpha$$

then

$$\liminf_{n \to \infty} \mathbb{P}_D(\underline{Z}_l \le \psi_l(P_X^0) \le \overline{Z}_l, l = 1, \dots, L) \ge 1 - \alpha$$

Proof of Proposition EC.1. The proof follows similarly from that of Proposition 1. If $P_Y^0 = \gamma(P_X^0) \in \mathcal{U}$, then P_X^0 must be a feasible solution for programs (EC.1) and (EC.2), and consequently $\underline{Z}_l \le \psi_l(P_X^0) \le \overline{Z}_l$, simultaneously for $l = 1, \ldots, L$. Therefore

$$\mathbb{P}_D(\underline{Z}_l \le \psi_l(P_X^0) \le \overline{Z}_l, l = 1, \ldots, L) \ge \mathbb{P}_D(P_Y^0 \in \mathcal{U})$$

This concludes the proposition. □

EC.2. Proofs for Section 4

Proof of Theorem 1. Note that the first constraints in (5) and (6) can be readily replaced by $P_Y \in \mathcal{U}$ for $\mathcal{U}$ defined in (4). We have $\lim_{n\to\infty} \mathbb{P}_D(P_Y^0 \in \mathcal{U}) = 1 - \alpha$, where P_Y^0 is the true output distribution, as a consequence of the KS statistic asymptotic. By using Proposition 1, we arrive at the guarantee (7).

The second conclusion comes from a reformulation of (4). Note that

$$\|F_Y - \hat{F}_Y\|_\infty \le \frac{q_{1-\alpha}}{\sqrt{n}}$$

is equivalent to

$$\sup_{y \in \mathbb{R}} |E_{P_X}[I(h(\mathbf{X}) \le y)] - \hat{F}_Y(y)| \le \frac{q_{1-\alpha}}{\sqrt{n}}$$

By the monotonicity of distribution functions, this is further equivalent to the set of constraints

$$\hat{F}_Y(y_j+) - \frac{q_{1-\alpha}}{\sqrt{n}} \le E_{P_X}[I(h(\mathbf{X}) \le y_j)] \le \hat{F}_Y(y_j-) + \frac{q_{1-\alpha}}{\sqrt{n}}, j = 1, \ldots, n \tag{EC.3}$$

which gives (8) and (9). □

Proof of Theorem 2. We will show the conclusion when (12) and (13) are replaced by

$$\begin{aligned} \max \quad & \psi(P_X) \\ \text{subject to} \quad & \|E_{P_X}[I(h(\mathbf{X}) \le \cdot)] - \hat{F}_Y(\cdot)\|_\infty \le \tfrac{q_{1-\alpha}}{\sqrt{n}} \\ & P_X \in \hat{\mathcal{P}}_X \end{aligned} \tag{EC.4}$$

and

$$\begin{aligned} \min \quad & \psi(P_X) \\ \text{subject to} \quad & \|E_{P_X}[I(h(\mathbf{X}) \le \cdot)] - \hat{F}_Y(\cdot)\|_\infty \le \tfrac{q_{1-\alpha}}{\sqrt{n}} \\ & P_X \in \hat{\mathcal{P}}_X \end{aligned} \tag{EC.5}$$

Then by the assumption that the true output distribution is continuous and that $\mathbb{P}$(for any $P_X \in \hat{\mathcal{P}}_X, \text{supp}(\gamma(P_X)) \cap \{y_j\}_{j=1,\ldots,n} \neq \emptyset) = 0$, we can use the same argument as in Theorem 1 to deduce that the constraints in (EC.4) and (EC.5) are equivalent to those in (12) and (13) with probability 1, from which we conclude the theorem.

Denote $L = dP_X^0/dQ$. Denote $\hat{P}_X(\cdot)$ as the empirical distribution on $\{z_j\}$ given by

$$\hat{P}_X(\cdot) = \frac{1}{m}\sum_{i=1}^{m} \delta_{z_i}(\cdot)$$

where $\delta_{z_j}(\cdot)$ is the delta mass on z_j. Consider

$$\tilde{P}_X(\cdot) = \sum_{i=1}^{m} \frac{L(z_i)}{\sum_{j=1}^{m} L(z_j)} \delta_{z_i}(\cdot)$$

i.e., $\tilde{P}_X$ is a discrete probability distribution with mass $L(z_i)/\sum_{j=1}^{m} L(z_j)$ on each generated support point z_i of X. Consider, for any $y \in \mathbb{R}$,

$$\begin{aligned} & E_{\tilde{P}_X}[I(h(\mathbf{X}) \le y)] - E_{P_X^0}[I(h(\mathbf{X}) \le y)] \\ = & \left(E_{\tilde{P}_X}[I(h(\mathbf{X}) \le y)] - E_{\bar{P}_X}[I(h(\mathbf{X}) \le y)]\right) + \left(E_{\bar{P}_X}[I(h(\mathbf{X}) \le y)] - E_{P_X^0}[I(h(\mathbf{X}) \le y)]\right) \end{aligned} \tag{EC.6}$$

where $\bar{P}_X(\cdot)$ is a measure (not necessarily a probability) given by

$$\bar{P}_X(\cdot) = \frac{1}{m}\sum_{i=1}^{m} L(z_i)\delta_{z_i}(\cdot)$$

and the expectation $E_{\bar{P}_X}[I(h(\mathbf{X}) \le y)]$ is defined in a general sense as the T-fold integral of $I(h(\mathbf{X}) \le y)$ with respect to $\bar{P}_X$. We consider both terms in (EC.6). Denoting $\mathbf{x} = (x_1, \ldots, x_T)$, we can write the first term as

$$\begin{aligned} & \int \cdots \int I(h(\mathbf{x}) \le y) \prod_{t=1}^{T} d\tilde{P}_X(x_t) - \int \cdots \int I(h(\mathbf{x}) \le y) \prod_{t=1}^{T} d\bar{P}_X(x_t) \\ = & \int \cdots \int I(h(\mathbf{x}) \le y) \frac{\prod_{t=1}^{T} L(x_t) d\hat{P}_X(x_t)}{\left(\frac{1}{m}\sum_{j=1}^{m} L(z_j)\right)^T} - \int \cdots \int I(h(\mathbf{x}) \le y) \prod_{t=1}^{T} L(x_t) d\hat{P}_X(x_t) \\ = & \int \cdots \int I(h(\mathbf{x}) \le y) \prod_{t=1}^{T} L(x_t) d\hat{P}_X(x_t) \left(\frac{1}{\left(\frac{1}{m}\sum_{j=1}^{m} L(z_j)\right)^T} - 1\right) \end{aligned} \tag{EC.7}$$

Since $Var_Q(L) < \infty$, and $E_Q[L] = 1$ by the definition of likelihood ratio, we have $\sqrt{m}((1/m)\sum_{j=1}^m L(z_j) - 1) \Rightarrow N(0, Var_Q(L))$ by the central limit theorem. By using the delta method (Chapter 3 in Serfling (2009)), we also have $\sqrt{m}(1/((1/m)\sum_{j=1}^m L(z_j))^T - 1) \Rightarrow N(0, T^2 Var_Q(L))$.

Moreover, $\int \cdots \int I(h(\mathbf{x}) \le y) \prod_{t=1}^T L(x_t) d\hat{P}_X(x_t)$ is bounded by C^T since $\|L\|_\infty \le C$. Hence (EC.7) satisfies

$$\sup_{y\in\mathbb{R}} \left| \int \cdots \int I(h(\mathbf{x}) \le y) \prod_{t=1}^T L(x_t) d\hat{P}_X(x_t) \left(\frac{1}{\left(\frac{1}{m}\sum_{j=1}^m L(z_j)\right)^T} - 1 \right) \right| \le C^T \left| \frac{1}{\left(\frac{1}{m}\sum_{j=1}^m L(z_j)\right)^T} - 1 \right| = O_p\left(\frac{1}{\sqrt{m}}\right) \tag{EC.8}$$

Now consider the second term in (EC.6). We have

$$\begin{aligned} & E_{\bar{P}_X}[I(h(\mathbf{X}) \le y)] - E_{P_X^0}[I(h(\mathbf{X}) \le y)] \\ = \; & E_{\hat{P}_X}\left[I(h(\mathbf{X}) \le y) \prod_{t=1}^T L(X_t) \right] - E_Q\left[I(h(\mathbf{X}) \le y) \prod_{t=1}^T L(X_t) \right] \end{aligned}$$

by the definition of $\bar{P}_X$, $\hat{P}_X$ and Q. Note that $E_{\hat{P}_X}\left[I(h(\mathbf{X}) \le y) \prod_{t=1}^T L(X_t) \right]$ is the average, over all possible selections with replacement of $x_1, \ldots, x_T$ drawn from $\{z_i\}_{i=1,\ldots,m}$, of the multilinear form $I(h(\mathbf{x}) \le y) \prod_{t=1}^T L(x_t)$. This is equivalent to the V-statistic (Serfling (2009) Chapter 5) with kernel $I(h(\mathbf{x}) \le y) \prod_{t=1}^T L(x_t)$.

Define $\mathcal{F}$ as the class of functions from $\mathcal{X}^T$ to $\mathbb{R}$ given by $\mathcal{F} = \{I(h(\mathbf{x}) \le y) \prod_{t=1}^T L(x_t) : y \in \mathbb{R}\}$. Since $I(h(\mathbf{x}) \le y) \prod_{t=1}^T L(x_t)$ is non-decreasing fixing each $\mathbf{x}$, and the envelope of $\mathcal{F}$, namely $\sup_{y\in\mathbb{R}} I(h(\mathbf{x}) \le y) \prod_{t=1}^T L(x_t)$, is bounded by C^T a.s., Problem 3 in Chapter 2.7 of Van Der Vaart and Wellner (1996) (Theorem EC.2 in the appendix) implies that $\mathcal{F}$ has a polynomial bracketing number. Therefore, Theorem 4.10 in Arcones and Gine (1993) (Theorem EC.3 in the appendix; see also the discussion after therein) concludes the convergence

$$\left\{ \sqrt{m} \left(U_m^T \left[I(h(\mathbf{X}) \le y) \prod_{t=1}^T L(X_t) \right] - E_Q\left[I(h(\mathbf{X}) \le y) \prod_{t=1}^T L(X_t) \right] \right) \right\}_{y\in\mathbb{R}} \Rightarrow \{\mathbb{G}(y)\}_{y\in\mathbb{R}} \text{ in } \ell^\infty(\mathcal{F})$$

where U_m^T is the U-operator defined in (EC.38) generated from P_X, and $\mathbb{G}$ is a Gaussian process defined as in (EC.39).

Following the argument of the lemma in Section 5.7.3 in Serfling (2009), we can write the difference between the U-statistic, denoted for simplicity $U_m = U_m^T\left[I(h(\mathbf{X})\le y)\prod_{t=1}^T L(X_t)\right]$, and the V-statistic, denoted $V_m = E_{\hat{P}_X}\left[I(h(\mathbf{X})\le y)\prod_{t=1}^T L(X_t)\right]$, as

$$m^T(U_m - V_m) = (m^T - m_{(T)})(U_m - W_m)$$

where $m_{(T)} = m(m-1)\cdots(m-T+1)$, and W_m is the average of all $I(h(\mathbf{x})\le y)\prod_{t=1}^T L(x_t)$ where $\mathbf{x}$ are drawn from $\{z_i\}_{i=1,\ldots,m}$ with replacement and at least one overlapping selection. As in Serfling (2009), we can verify $m^T - m_{(T)} = O(m^{T-1})$, and since $\|L\|_\infty \le C$, we have $U_m - W_m$ bounded a.s. Hence $E\sup_{t\in\mathbb{R}}|U_m - V_m|^2 = O(1/m^2)$, and so $\sup_{t\in\mathbb{R}}|U_m - V_m| = O_p(1/m)$.

Therefore, we write

$$\begin{aligned}
&\sqrt{m}\left(E_{\hat{P}_X}\left[I(h(\mathbf{X})\le y)\prod_{t=1}^T L(X_t)\right] - E_Q\left[I(h(\mathbf{X})\le y)\prod_{t=1}^T L(X_t)\right]\right)\\
&= \sqrt{m}\left(E_{\hat{P}_X}\left[I(h(\mathbf{X})\le y)\prod_{t=1}^T L(X_t)\right] - U_m^T\left[I(h(\mathbf{X})\le y)\prod_{t=1}^T L(X_t)\right]\right)\\
&\quad + \sqrt{m}\left(U_m^T\left[I(h(\mathbf{X})\le y)\prod_{t=1}^T L(X_t)\right] - E_Q\left[I(h(\mathbf{X})\le y)\prod_{t=1}^T L(X_t)\right]\right)
\end{aligned}$$

where $\sqrt{m}\left(E_{\hat{P}_X}\left[I(h(\mathbf{X})\le y)\prod_{t=1}^T L(X_t)\right] - U_m^T\left[I(h(\mathbf{X})\le y)\prod_{t=1}^T L(X_t)\right]\right) = o_p(1)$ and $\sqrt{m}\left(U_m^T\left[I(h(\mathbf{X})\le y)\prod_{t=1}^T L(X_t)\right] - E_Q\left[I(h(\mathbf{X})\le y)\prod_{t=1}^T L(X_t)\right]\right)$ converges to a Gaussian process. This entails that

$$\sup_{y\in\mathbb{R}}\left|E_{\hat{P}_X}\left[I(h(\mathbf{X})\le y)\prod_{t=1}^T L(X_t)\right] - E_Q\left[I(h(\mathbf{X})\le y)\prod_{t=1}^T L(X_t)\right]\right| = O_p\left(\frac{1}{\sqrt{m}}\right) \tag{EC.9}$$

From (EC.6), and using (EC.8) and (EC.9), we get

$$\begin{aligned}
&\sup_{y\in\mathbb{R}}\left|E_{\tilde{P}_X}[I(h(\mathbf{X})\le y)] - E_{P_X^0}[I(h(\mathbf{X})\le y)]\right|\\
&\le \sup_{y\in\mathbb{R}}\left|E_{\tilde{P}_X}[I(h(\mathbf{X})\le y)] - E_{\bar{P}_X}[I(h(\mathbf{X})\le y)]\right| + \sup_{y\in\mathbb{R}}\left|E_{\bar{P}_X}[I(h(\mathbf{X})\le y)] - E_{P_X^0}[I(h(\mathbf{X})\le y)]\right|\\
&= O_p\left(\frac{1}{\sqrt{m}}\right)
\end{aligned} \tag{EC.10}$$

For the above chosen $\tilde{P}_X$, we now have, for any small enough $\delta > 0$,

$$P\left(\|E_{\tilde{P}_X}[I(h(\mathbf{X})\le\cdot)] - \hat{F}_Y(\cdot)\|_\infty \le \frac{q_{1-\alpha}}{\sqrt{n}}\right)$$

$$
\begin{aligned}
&\geq P\left(\|E_{\tilde{P}_X}[I(h(\mathbf{X})\leq\cdot)]-E_{P_X^0}[I(h(\mathbf{X})\leq\cdot)]\|_\infty+\|E_{P_X^0}[I(h(\mathbf{X})\leq\cdot)]-\hat{F}_Y(\cdot)\|_\infty\leq\frac{q_{1-\alpha}}{\sqrt{n}}\right)\\
&\geq P\left(\|E_{P_X^0}[I(h(\mathbf{X})\leq\cdot)]-\hat{F}_Y(\cdot)\|_\infty\leq\frac{q_{1-\alpha}-\delta}{\sqrt{n}};\ \|E_{\tilde{P}_X}[I(h(\mathbf{X})\leq\cdot)]-E_{P_X^0}[I(h(\mathbf{X})\leq\cdot)]\|_\infty\leq\frac{\delta}{\sqrt{n}}\right)\\
&\geq P\left(\|E_{P_X^0}[I(h(\mathbf{X})\leq\cdot)]-\hat{F}_Y(\cdot)\|_\infty\leq\frac{q_{1-\alpha}-\delta}{\sqrt{n}}\right)-P\left(\|E_{\tilde{P}_X}[I(h(\mathbf{X})\leq\cdot)]-E_{P_X^0}[I(h(\mathbf{X})\leq\cdot)]\|_\infty>\frac{\delta}{\sqrt{n}}\right)\\
&\to 1-\alpha+\zeta(-\delta) \qquad \text{(EC.11)}
\end{aligned}
$$

as $n\to\infty$ and $m/n\to\infty$, where $\zeta(\cdot)$ is a function with $\lim_{x\to 0}\zeta(x)=0$ that satisfies $P(\sup_{u\in[0,1]}|BB(u)|\leq q_{1-\alpha}+\rho)=1-\alpha+\zeta(\rho)$, which exists by the continuity of the distribution of $\sup_{u\in[0,1]}BB(u)$. The convergence (EC.11) follows from the definition that P_X^0 is the true input distribution and hence $E_{P_X^0}[I(h(\mathbf{X})\leq\cdot)]$ is the true output distribution, thus leading to $\sqrt{n}\|E_{P_X^0}[I(h(\mathbf{X})\leq\cdot)]-\hat{F}_Y(\cdot)\|_\infty\Rightarrow\sup_{u\in[0,1]}|BB(u)|$. It also follows from (EC.10) that $P\left(\|E_{\tilde{P}_X}[I(h(\mathbf{X})\leq\cdot)]-E_{P_X^0}[I(h(\mathbf{X})\leq\cdot)]\|_\infty>\frac{\delta}{\sqrt{n}}\right)\to 0$ as $m/n\to\infty$.

Similarly, for any small enough $\delta>0$, we have

$$
\begin{aligned}
&P\left(\|E_{\tilde{P}_X}[I(h(\mathbf{X})\leq\cdot)]-\hat{F}_Y(\cdot)\|_\infty\leq\frac{q_{1-\alpha}}{\sqrt{n}}\right)\\
&\leq P\left(\|E_{P_X^0}[I(h(\mathbf{X})\leq\cdot)]-\hat{F}_Y(\cdot)\|_\infty-\|E_{\tilde{P}_X}[I(h(\mathbf{X})\leq\cdot)]-E_{P_X^0}[I(h(\mathbf{X})\leq\cdot)]\|_\infty\leq\frac{q_{1-\alpha}}{\sqrt{n}}\right)\\
&\leq P\left(\|E_{P_X^0}[I(h(\mathbf{X})\leq\cdot)]-\hat{F}_Y(\cdot)\|_\infty\leq\frac{q_{1-\alpha}+\delta}{\sqrt{n}}\right)+P\left(\|E_{\tilde{P}_X}[I(h(\mathbf{X})\leq\cdot)]-E_{P_X^0}[I(h(\mathbf{X})\leq\cdot)]\|_\infty>\frac{\delta}{\sqrt{n}}\right)\\
&\to 1-\alpha+\zeta(\delta) \qquad \text{(EC.12)}
\end{aligned}
$$

as $n\to\infty$ and $m/n\to\infty$. Since δ is arbitrary, by combining (EC.11) and (EC.12), we have

$$
P\left(\|E_{\tilde{P}_X}[I(h(\mathbf{X})\leq\cdot)]-\hat{F}_Y(\cdot)\|_\infty\leq\frac{q_{1-\alpha}}{\sqrt{n}}\right)\to 1-\alpha
$$

as $n\to\infty$ and $m/n\to\infty$.

Lastly, we argue that the objective function satisfies $E_{\tilde{P}_X}[g(\mathbf{X})]-E_{P_X^0}[g(\mathbf{X})]=O_p(1/\sqrt{m})$. This follows mostly as a special case of the arguments above in showing $\sup_{y\in\mathbb{R}}|E_{\tilde{P}_X}[I(h(\mathbf{X})\leq y)]-E_{P_X^0}[I(h(\mathbf{X})\leq y)]|=O_p(1/\sqrt{m})$, by simply replacing $I(h(\mathbf{X}\leq y)$ with $g(\mathbf{X})$ and without considering the uniformity over $y\in\mathbb{R}$. More precisely, we have

$$
\begin{aligned}
&E_{\tilde{P}_X}[g(\mathbf{X})]-E_{P_X^0}[g(\mathbf{X})]\\
&=\left(E_{\tilde{P}_X}[g(\mathbf{X})]-E_{\bar{P}_X}[g(\mathbf{X})]\right)+\left(E_{\bar{P}_X}[g(\mathbf{X})]-E_{P_X^0}[g(\mathbf{X})]\right) \qquad \text{(EC.13)}
\end{aligned}
$$

similar to (EC.6), where $E_{\tilde{P}_X}[g(\mathbf{X})] - E_{\bar{P}_X}[g(\mathbf{X})] = O_p(1/\sqrt{m})$ similar to (EC.8), and $E_{\bar{P}_X}[g(\mathbf{X})] - E_{P_X^0}[g(\mathbf{X})] = O_p(1/\sqrt{m})$ by using the standard central limit theorem for U-statistic (Theorem A in Section 5.5 in Serfling (2009)) and, with the assumption $E_{P_X^0}[g(X_{i_1},\ldots,X_{i_T})^2] < \infty$ for any $1 \le i_1,\ldots,i_T \le T$, translating it to V-statistic (the lemma in Section 5.7.3 in Serfling (2009))). Therefore, we have $E_{\tilde{P}_X}[g(\mathbf{X})] - E_{P_X^0}[g(\mathbf{X})] = O_p(1/\sqrt{m})$.

In conclusion, we have found a solution $\tilde{P}_X$ that is feasible for (EC.4) and (EC.5) with probability asymptotically $1-\alpha$ as $n \to \infty$ and $m/n \to \infty$. Moreover, $\psi(\tilde{P}_X) - \psi(P_X^0) = O_p(1/\sqrt{m})$. Therefore, we have

$$\begin{aligned}
1-\alpha &\le \lim_{n\to\infty, m/n\to\infty} \mathbb{P}\left(\|E_{\tilde{P}_X}[I(h(\mathbf{X}) \le \cdot)] - \hat{F}_Y(\cdot)\|_\infty \le \frac{q_{1-\alpha}}{\sqrt{n}} \right) \\
&\le \liminf_{n\to\infty, m/n\to\infty} \mathbb{P}\left(\underline{\hat{Z}} \le \psi(\tilde{P}_X) \le \hat{\overline{Z}} \right) \\
&= \liminf_{n\to\infty, m/n\to\infty} \mathbb{P}\left(\underline{\hat{Z}} + O_p\left(\frac{1}{\sqrt{m}}\right) \le \psi(P_X^0) \le \hat{\overline{Z}} + O_p\left(\frac{1}{\sqrt{m}}\right) \right)
\end{aligned}$$

which concludes the theorem. □

We provide some remark on the case where we consider $h(\mathbf{X},\mathbf{W})$ and $g(\mathbf{X},\mathbf{W})$ for some collection of auxiliary input variate sequences $\mathbf{W}$ that is independent of $\mathbf{X}$ and has a known distribution. In this case, the results in Sections 3 and 4 all hold with the $E_{P_X}[\cdot]$ interpreted as the joint expectation taken with respect to both the product measure of P_X and P_W^0, the known distribution of $\mathbf{W}$. In the proofs above, we keep the expectation $E_{P_X}[\cdot]$ as taken under the product measure of P_X only, but we use a conditioning argument, namely we change $I(h(\mathbf{X}) \le y)$ to $P_{P_W^0}(h(\mathbf{X},\mathbf{W}) \le y|\mathbf{X}) = E_{P_W^0}[I(h(\mathbf{X}) \le y)|\mathbf{X}]$ and $g(\mathbf{X})$ to $E_{P_W^0}[g(\mathbf{X},\mathbf{W})|\mathbf{X}]$, where $P_{P_W^0}(\cdot|\mathbf{X})$ and $E_{P_W^0}[\cdot|\mathbf{X}]$ denote the conditional probability and expectation under the true distribution of $\mathbf{W}$ given $\mathbf{X}$. In particular, in the proof of Theorem 2, we have that $P_{P_W^0}(h(\mathbf{X},\mathbf{W}) \le y|\mathbf{X}=\mathbf{x})$ is non-decreasing given any $\mathbf{x}$, and $P_{P_W^0}(h(\mathbf{X},\mathbf{W}) \le y|\mathbf{X}=\mathbf{x}) \le 1$, which, via Problem 3 in Chapter 2.7 of Van Der Vaart and Wellner (1996) again, gives a polynomial bracketing number for the class of functions $\{P_{P_W^0}(h(\mathbf{X},\mathbf{W}) \le y|\mathbf{X}=\mathbf{x}) \prod_{t=1}^T L(x_t) : y \in \mathbb{R}\}$. We also have $E_{P_X^0}[E_{P_W^0}[g(X_{i_1},\ldots,X_{i_T},\mathbf{W})|X_{i_1},\ldots,X_{i_T}]^2] \le E_{P_X^0,P_W^0}[g(X_{i_1},\ldots,X_{i_T},\mathbf{W})^2] < \infty$ for any $1 \le$

$i_1, \ldots, i_T \leq T$, where $E_{P_X^0, P_W^0}[\cdot]$ denotes the joint expectation under the product measure of P_X^0 and P_W^0, so that the central limit theorem for ensuring the approximation of the objective value holds in the proof. Other proofs follow quite trivially.

EC.3. Proofs for Section 5

Proof of Proposition 2 Consider the equivalent reformulation of the program (14)

$$\begin{aligned} \min \quad & \psi(\mathbf{p}) \\ \text{subject to } & E_{\mathbf{p}}[I(h(\mathbf{X}) \leq y_j)] - s_j = 0, j = 1, \ldots, n \\ & \hat{F}_Y(y_j+) - \tfrac{q_{1-\alpha}}{\sqrt{n}} \leq s_j \leq \hat{F}_Y(y_j-) + \tfrac{q_{1-\alpha}}{\sqrt{n}}, j = 1, \ldots, n \\ & \mathbf{p} \in \mathcal{P} \end{aligned} \tag{EC.14}$$

where both $\mathbf{p}$ and $\mathbf{s}$ are viewed as decision variables. An application of the conventional quadratic penalty method (Bertsekas (1999)) for equality constraints yields the following optimization sequence

$$\begin{aligned} \min \quad & \psi(\mathbf{p}) + c \textstyle\sum_{j=1}^n (E_{\mathbf{p}}[I(h(\mathbf{X}) \leq y_j)] - s_j)^2 \\ \text{subject to } & \hat{F}_Y(y_j+) - \tfrac{q_{1-\alpha}}{\sqrt{n}} \leq s_j \leq \hat{F}_Y(y_j-) + \tfrac{q_{1-\alpha}}{\sqrt{n}}, j = 1, \ldots, n \\ & \mathbf{p} \in \mathcal{P} \end{aligned} \tag{EC.15}$$

for $c > 0$, which is equivalent to (15) with $\lambda = 1/c$. Proposition 4.2.1 in Bertsekas (1999) entails that as $c \to \infty$ ($\lambda \to 0$), every limit point $(\mathbf{p}^*, \mathbf{s}^*)$ of the sequence of optimal solutions $\{(\mathbf{p}^*(\lambda), \mathbf{s}^*(\lambda))\}$ to (EC.15) is an optimal solution to (EC.14), given that (EC.14) is feasible. Note that due to optimality, the optimal slack variables $\mathbf{s}^*(\lambda) = (s_1^*(\lambda), \ldots, s_n^*(\lambda))$ must take the following form

$$s_j^*(\lambda) = \Pi_j(E_{\mathbf{p}^*(\lambda)}[I(h(\mathbf{X}) \leq y_j)]), \ j = 1, \ldots, n$$

where each Π_j is the projection defined in (17). Since projections are continuous maps, the operations of taking limit points and coordinate projection are interchangeable, i.e.

$$\begin{aligned} & \{\mathbf{p}^* : \mathbf{p}^* \text{ is a limit point of } \{\mathbf{p}^*(\lambda)\}\} \\ = & \{\mathbf{p}^* : \text{there exists an } \mathbf{s}^* \text{ s.t. } (\mathbf{p}^*, \mathbf{s}^*) \text{ is a limit point of } \{(\mathbf{p}^*(\lambda), \mathbf{s}^*(\lambda))\}\}. \end{aligned}$$

This allows translation of optimality of the limit point of $\{(\mathbf{p}^*(\lambda), \mathbf{s}^*(\lambda))\}$ to optimality of the limit point of $\{\mathbf{p}^*(\lambda)\}$. The desired conclusion follows. $\square$

Proof of Proposition 3. Part 1 and the expression for Ψ_i in part 2 come from a direct application of Ghosh and Lam (2015a) and Ghosh and Lam (2015b). We will prove (28) and (29) in part 2 only, but in the more general setting of differentiable functions of expectations. Let $f(\mathbf{X})$ with $\mathbf{X} = (X_1, \ldots, X_{T_f})$ be a performance function, where T_f is a finite and deterministic time horizon, and $\Phi(y) : \mathbb{R} \to \mathbb{R}$ be any differentiable function. By the chain rule

$$\Phi_i(\mathbf{p}) := \frac{d}{d\epsilon}\Phi(\mathbf{E}_{(1-\epsilon)\mathbf{p}+\epsilon\mathbf{1}_i}[f(\mathbf{X})])\Big|_{\epsilon=0^+} = \frac{d}{dy}\Phi(\mathbf{E}_{\mathbf{p}}[f(\mathbf{X})])\frac{d}{d\epsilon}\mathbf{E}_{(1-\epsilon)\mathbf{p}+\epsilon\mathbf{1}_i}[f(\mathbf{X})]|_{\epsilon=0^+}.$$

Similar to (27) we have

$$\frac{d}{d\epsilon}\mathbf{E}_{(1-\epsilon)\mathbf{p}+\epsilon\mathbf{1}_i}[f(\mathbf{X})]|_{\epsilon=0^+} = E_{\mathbf{p}}[f(\mathbf{X})S_i(\mathbf{X};\mathbf{p})]$$

where

$$S_i(\mathbf{x};\mathbf{p}) = \sum_{t=1}^{T_f} \frac{I_i(x_t)}{p_i} - T_f.$$

Therefore the following expression holds for the derivative

$$\Phi_i(\mathbf{p}) = \frac{d}{dy}\Phi(\mathbf{E}_{\mathbf{p}}[f(\mathbf{X})])E_{\mathbf{p}}[f(\mathbf{X})S_i(\mathbf{X};\mathbf{p})].$$

(28) and (29) follow from applying the above result to $f(\mathbf{X}) = h(\mathbf{X})$, $\Phi(y) = (y - \Pi_j(y))^2$ and $\Phi(y) = (y - s_j)^2$ respectively, together with the linearity of differentiation. Note that $\frac{d}{dy}(y - \Pi_j(y))^2 = 2(y - \Pi_j(y))$. $\square$

Proof of Proposition 4. First note that the function $\mu(\eta)$ is continuous and strictly increasing in the interval $[0, \max_i p_i e^{-\xi_i}]$, and satisfies $\mu(0) = 0, \mu(\max_i p_i e^{-\xi_i}) = 1/m$ at the endpoints. So indeed there exists a unique η^* that solves (38). Then we show (37) is indeed the optimal solution. Consider the Lagrangian

$$L(\mathbf{q}, \lambda, \boldsymbol{\beta}) = \boldsymbol{\xi}'(\mathbf{q} - \mathbf{p}) + V(\mathbf{p}, \mathbf{q}) + \lambda(\sum_{i=1}^{m} q_i - 1) - \sum_{i=1}^{m} \beta_i(q_i - \epsilon)$$

defined for $\beta_i \geq 0$ and $\lambda \in \mathbb{R}$. Since (35) is a convex program with linear constraints and obviously Slater's condition holds, by Proposition 6.2.5 and Proposition 6.4.4 of Bertsekas et al. (2003) it

suffices to find dual variables λ^* and β_i^* such that the solution given by (37) satisfies the set of KKT conditions

$$\frac{\partial L}{\partial q_i} = \xi_i + \log\frac{q_i^*}{p_i} + 1 + \lambda^* - \beta_i^* = 0 \text{ for } i = 1,\dots,m \tag{EC.16}$$

$$\sum_{i=1}^{m} q_i^* = 1,\ q_i^* \geq \epsilon,\ \text{for } i = 1,\dots,m \tag{EC.17}$$

$$\beta_i^* \geq 0,\ \beta_i^*(q_i^* - \epsilon) = 0 \text{ for } i = 1,\dots,m. \tag{EC.18}$$

Equations (EC.17) obviously hold because of equation (38). Equations (EC.16) can be rewritten as

$$q_i^* = p_i e^{-\xi_i - 1 - \lambda^* + \beta_i^*} \text{ for } i = 1,\dots,m$$

which hold if λ^*, β_i^* are chosen such that

$$e^{1+\lambda^*} = \sum_{i=1}^{m} \max\{\eta^*, p_i e^{-\xi_i}\},\ e^{\beta_i^*} = \frac{\max\{\eta^*, p_i e^{-\xi_i}\}}{p_i e^{-\xi_i}}.$$

It is obvious that such chosen $\beta_i^* \geq 0$. To show complementary slackness (EC.18), note that if $q_i^* > \epsilon$ then (38) forces $p_i e^{-\xi_i} > \eta^*$ which results in $\beta_i^* = 0$. $\square$

Proof of Theorem 3. Consider the auxiliary programs obtained from replacing α by some $\alpha' > \alpha$ in (14)

$$\begin{aligned} \max \quad & \psi(\mathbf{p}) \\ \text{subject to } & \hat{F}_Y(y_j+) - \tfrac{q_{1-\alpha'}}{\sqrt{n}} \leq E_{\mathbf{p}}[I(h(\mathbf{X}) \leq y_j)] \leq \hat{F}_Y(y_j-) + \tfrac{q_{1-\alpha'}}{\sqrt{n}}, j = 1,\dots,n \\ & \mathbf{p} \in \mathcal{P} \end{aligned}$$

and

$$\begin{aligned} \min \quad & \psi(\mathbf{p}) \\ \text{subject to } & \hat{F}_Y(y_j+) - \tfrac{q_{1-\alpha'}}{\sqrt{n}} \leq E_{\mathbf{p}}[I(h(\mathbf{X}) \leq y_j)] \leq \hat{F}_Y(y_j-) + \tfrac{q_{1-\alpha'}}{\sqrt{n}}, j = 1,\dots,n \\ & \mathbf{p} \in \mathcal{P}. \end{aligned}$$

Denote by $\mathbf{p}_{\max}^{*'}$ and $\mathbf{p}_{\min}^{*'}$ optimal solutions of the above maximization and minimization programs, which by Theorem 2 satisfy

$$\liminf_{n\to\infty, m/n\to\infty} \mathbb{P}\left(\psi(\mathbf{p}_{\min}^{*'}) + O_p\left(\frac{1}{\sqrt{m}}\right) \leq \psi(P_X^0) \leq \psi(\mathbf{p}_{\max}^{*'}) + O_p\left(\frac{1}{\sqrt{m}}\right)\right) \geq 1 - \alpha'.$$

Now, we try to show that $\underline{\hat{Z}}_\epsilon \leq \psi(\mathbf{p}^{*'}_{\min}) + O(m\epsilon)$ and $\hat{\overline{Z}}_\epsilon \geq \psi(\mathbf{p}^{*'}_{\max}) - O(m\epsilon)$, therefore to conclude that

$$\liminf_{n\to\infty, m/n\to\infty} \mathbb{P}\left(\underline{\hat{Z}}_\epsilon + O_p\left(m\epsilon + \frac{1}{\sqrt{m}}\right) \leq \psi(P_X^0) \leq \hat{\overline{Z}}_\epsilon + O_p\left(m\epsilon + \frac{1}{\sqrt{m}}\right)\right) \geq 1 - \alpha'. \tag{EC.19}$$

To avoid repetition, we only prove the minimization case here. To proceed, let $\mathbf{p}, \mathbf{q} \in \mathcal{P}$ be two arbitrary probability distributions in $\mathcal{P}$, and $\mathbf{p}^S, \mathbf{q}^S$ be the corresponding S-fold product measure, then we have

$$|\psi(\mathbf{p}) - \psi(\mathbf{q})| = |E_{\mathbf{p}}[g(\mathbf{X})] - E_{\mathbf{q}}[g(\mathbf{X})]| \leq 2 \sup_{\mathbf{X}} |g(\mathbf{X})| \cdot \|\mathbf{p}^S - \mathbf{q}^S\|_{TV}$$

where $\|\cdot\|_{TV}$ denotes the total variation distance between the product measures. It is well-known that the total variation distance between product measures can be bounded as (see, e.g. Lemma 3.6.2 of Durrett (2010))

$$\|\mathbf{p}^S - \mathbf{q}^S\|_{TV} \leq S\|\mathbf{p} - \mathbf{q}\|_{TV},$$

therefore

$$|\psi(\mathbf{p}) - \psi(\mathbf{q})| \leq 2S \sup_{\mathbf{X}} |g(\mathbf{X})| \cdot \|\mathbf{p} - \mathbf{q}\|_{TV} = C_1 \|\mathbf{p} - \mathbf{q}\|_{TV}.$$

Similarly for the constraint functions we have

$$|E_{\mathbf{p}}[I(h(\mathbf{X}) \leq y_j)] - E_{\mathbf{q}}[I(h(\mathbf{X}) \leq y_j)]| \leq 2T\|\mathbf{p} - \mathbf{q}\|_{TV} = C_2\|\mathbf{p} - \mathbf{q}\|_{TV}, j = 1, \ldots, n.$$

Consider the total variation ball of radius $m\epsilon$ surrounding $\mathbf{p}^{*'}_{\min}$

$$B_{TV}(\mathbf{p}^{*'}_{\min}, m\epsilon) = \{\mathbf{p} \in \mathcal{P} : \|\mathbf{p}^{*'}_{\min} - \mathbf{p}\|_{TV} \leq m\epsilon\}.$$

It is clear that for all $\mathbf{p} \in B_{TV}(\mathbf{p}^{*'}_{\min}, m\epsilon)$ it holds

$$|\psi(\mathbf{p}) - \psi(\mathbf{p}^{*'}_{\min})| \leq C_1 m\epsilon \tag{EC.20}$$

$$|E_{\mathbf{p}}[I(h(\mathbf{X}) \leq y_j)] - E_{\mathbf{p}^{*'}_{\min}}[I(h(\mathbf{X}) \leq y_j)]| \leq C_2 m\epsilon, j = 1, \ldots, n. \tag{EC.21}$$

Note that $\mathbf{p}^{*'}_{\min}$ is optimal and hence feasible for the program with α', thus the inequality (EC.21) ensures for all $\mathbf{p} \in B_{TV}(\mathbf{p}^{*'}_{\min}, m\epsilon)$

$$\hat{F}_Y(y_j+) - \frac{q_{1-\alpha'}}{\sqrt{n}} - C_2 m\epsilon \le E_{\mathbf{p}}[I(h(\mathbf{X}) \le y_j)] \le \hat{F}_Y(y_j+) + \frac{q_{1-\alpha'}}{\sqrt{n}} + C_2 m\epsilon, j = 1, \dots, n.$$

Since $\epsilon = o(1/(m\sqrt{n}))$, for large enough m, n we have $C_2 m\epsilon \le (q_{1-\alpha} - q_{1-\alpha'})/\sqrt{n}$ which results in

$$\hat{F}_Y(y_j+) - \frac{q_{1-\alpha}}{\sqrt{n}} \le E_{\mathbf{p}}[I(h(\mathbf{X}) \le y_j)] \le \hat{F}_Y(y_j+) + \frac{q_{1-\alpha}}{\sqrt{n}}, j = 1, \dots, n.$$

That is, all $\mathbf{p} \in B_{TV}(\mathbf{p}^{*'}_{\min}, m\epsilon)$ satisfy the first constraint in (33). In view of inequality (EC.20), it remains to show that $B_{TV}(\mathbf{p}^{*'}_{\min}, m\epsilon) \cap \mathcal{P}(\epsilon) \neq \emptyset$ in order to conclude $\hat{\underline{Z}}_\epsilon \le \psi(\mathbf{p}^{*'}_{\min}) + O(m\epsilon)$. Easily one can verify that for any $\mathbf{p} \in \mathcal{P}$ it holds $\inf\{\|\mathbf{p} - \mathbf{q}\|_{TV} : \mathbf{q} \in \mathcal{P}(\epsilon)\} \le (m-1)\epsilon$, and in particular $\inf\{\|\mathbf{p}^{*'}_{\min} - \mathbf{q}\|_{TV} : \mathbf{q} \in \mathcal{P}(\epsilon)\} \le (m-1)\epsilon$ which implies $B_{TV}(\mathbf{p}^{*'}_{\min}, m\epsilon) \cap \mathcal{P}(\epsilon) \neq \emptyset$.

Lastly note that (EC.19) holds true for arbitrary $\alpha' > \alpha$, hence holds for α as well. This concludes the theorem. □

LEMMA EC.1. *For any i, j and $l = 1, 2$, the moments of gradient estimators*

$$E_{\mathbf{p}}\left[(g(\mathbf{X})S_i(\mathbf{X};\mathbf{p}))^l\right], E_{\mathbf{p}}\left[\left(I(h(\mathbf{X}) \le y_j)S_i(\mathbf{X};\mathbf{p})\right)^l\right]$$

are continuous in $\mathcal{P}^o = \{\mathbf{p} \in \mathcal{P} : p_i > 0 \text{ for all } i\}$, the relative interior of $\mathcal{P}$.

Proof of Lemma EC.1. Restricted to $\mathcal{P}^o$, each of the moments can be written as the sum of finitely many terms each of which are smooth in $\mathbf{p}$. A sum of finitely many smooth functions is also smooth, hence continuous. □

LEMMA EC.2. *Let $\{D^k\}_{k=1}^\infty$ be a positive sequence. If for $0 < \alpha_2 < \alpha_1 \le 1$ and constants $C_1, C_2 > 0$ it holds $D^{k+1} \le (1 - \frac{C_1}{k^{\alpha_2}})D^k + C_2(\frac{1}{k^{2\alpha_2}} + \frac{1}{k^{2\alpha_1 - \alpha_2}})$ for all k large enough, then there exits a constant $C > 0$ such that $D^k \le C(\frac{1}{k^{\alpha_2}} + \frac{1}{k^{2(\alpha_1 - \alpha_2)}})$ for all k.*

Proof of Lemma EC.2. Assume $D^k \le C(\frac{1}{k^{\alpha_2}} + \frac{1}{k^{2(\alpha_1-\alpha_2)}})$, then

$$D^{k+1} \le (1 - \frac{C_1}{k^{\alpha_2}})D^k + C_2(\frac{1}{k^{2\alpha_2}} + \frac{1}{k^{2\alpha_1 - \alpha_2}})$$

$$
\begin{aligned}
&\le \frac{C}{k^{\alpha_2}} + \frac{C}{k^{2(\alpha_1-\alpha_2)}} - \frac{C_1C-C_2}{k^{2\alpha_2}} - \frac{C_1C-C_2}{k^{2\alpha_1-\alpha_2}} \\
&\le \frac{C}{(k+1)^{\alpha_2}} + \frac{C\alpha_2}{k^{\alpha_2+1}} + \frac{C}{(k+1)^{2(\alpha_1-\alpha_2)}} + \frac{C\cdot 2(\alpha_1-\alpha_2)}{k^{2(\alpha_1-\alpha_2)+1}} - \frac{C_1C-C_2}{k^{2\alpha_2}} - \frac{C_1C-C_2}{k^{2\alpha_1-\alpha_2}} \\
&\le \frac{C}{(k+1)^{\alpha_2}} + \frac{C\alpha_2 k^{\alpha_2-1}}{k^{2\alpha_2}} + \frac{C}{(k+1)^{2(\alpha_1-\alpha_2)}} + \frac{C\cdot 2(\alpha_1-\alpha_2)k^{\alpha_2-1}}{k^{2\alpha_1-\alpha_2}} - \frac{C_1C-C_2}{k^{2\alpha_2}} - \frac{C_1C-C_2}{k^{2\alpha_1-\alpha_2}} \\
&\le \frac{C}{(k+1)^{\alpha_2}} + \frac{C}{(k+1)^{2(\alpha_1-\alpha_2)}} - \frac{C(C_1-\alpha_2 k^{\alpha_2-1})-C_2}{k^{2\alpha_2}} - \frac{C(C_1-2(\alpha_1-\alpha_2)k^{\alpha_2-1})-C_2}{k^{2\alpha_1-\alpha_2}} \\
&\le \frac{C}{(k+1)^{\alpha_2}} + \frac{C}{(k+1)^{2(\alpha_1-\alpha_2)}}.
\end{aligned}
$$

Note that the above argument goes through when k is large and C is chosen such that $\frac{C_1}{k^{\alpha_2}} < 1$, $C(C_1 - 2(\alpha_1-\alpha_2)k^{\alpha_2-1}) - C_2 \ge 0$ and $C(C_1 - \alpha_2 k^{\alpha_2-1}) - C_2 \ge 0$. By induction $D^k \le C(\frac{1}{k^{\alpha_2}} + \frac{1}{k^{2(\alpha_1-\alpha_2)}})$ holds for all sufficiently large k. By enlarging C one can make it hold for all k. □

Proof of Theorem 4. We borrow from Lemma 2.1 in Nemirovski et al. (2009) the inequality

$$
V(\mathbf{p}^{k+1}, \mathbf{p}^*_\epsilon(\lambda^k)) - V(\mathbf{p}^k, \mathbf{p}^*_\epsilon(\lambda^k)) \le \gamma^k(\lambda^k\hat{\mathbf{\Psi}}^k + \hat{\boldsymbol{\phi}}^k_{\mathbf{p}})'(\mathbf{p}^*_\epsilon(\lambda^k) - \mathbf{p}^k) + \frac{(\gamma^k)^2\|\lambda^k\hat{\mathbf{\Psi}}^k + \hat{\boldsymbol{\phi}}^k_{\mathbf{p}}\|^2_\infty}{2} \quad \text{(EC.22)}
$$

which holds as long as $\mathbf{p}^{k+1}$ is the prox-mapping of $\mathbf{p}^k$. The norm $\|\cdot\|_\infty$ is the supremum norm, the dual of the L_1-norm that is used in the strong convexity property of $\omega(\mathbf{p}) = \sum_{i=1}^m p_i \log p_i$, with $\alpha = 1$. Note that $V(\mathbf{p}^{k+1}, \mathbf{p}^*_\epsilon(\lambda^k)) = \sum_{i=1}^m p^*_i(\lambda^k)(\log p^*_i(\lambda^k) - \log p^{k+1}_i)$ and both $\mathbf{p}^*_\epsilon(\lambda^k), \mathbf{p}^{k+1} \in \mathcal{P}(\epsilon)$, by mean value theorem it holds

$$
V(\mathbf{p}^{k+1}, \mathbf{p}^*_\epsilon(\lambda^{k+1})) - V(\mathbf{p}^{k+1}, \mathbf{p}^*_\epsilon(\lambda^k)) \le C|\log\epsilon|\|\mathbf{p}^*_\epsilon(\lambda^{k+1}) - \mathbf{p}^*_\epsilon(\lambda^k)\| \quad \text{(EC.23)}
$$

where C is an absolute constant. This gives

$$
\begin{aligned}
&V(\mathbf{p}^{k+1}, \mathbf{p}^*_\epsilon(\lambda^{k+1})) - V(\mathbf{p}^k, \mathbf{p}^*_\epsilon(\lambda^k)) \\
\le&\gamma^k(\lambda^k\hat{\mathbf{\Psi}}^k + \hat{\boldsymbol{\phi}}^k_{\mathbf{p}})'(\mathbf{p}^*_\epsilon(\lambda^k) - \mathbf{p}^k) + \frac{(\gamma^k)^2\|\lambda^k\hat{\mathbf{\Psi}}^k + \hat{\boldsymbol{\phi}}^k_{\mathbf{p}}\|^2_\infty}{2} + C|\log\epsilon|\|\mathbf{p}^*_\epsilon(\lambda^{k+1}) - \mathbf{p}^*_\epsilon(\lambda^k)\| \quad \text{(EC.24)}
\end{aligned}
$$

Let $\mathcal{F}^k$ be the filtration generated by $\{\mathbf{p}^1, \mathbf{s}^1, \ldots, \mathbf{p}^k, \mathbf{s}^k\}$. Taking conditional expectation of (EC.24) with respect to $\mathcal{F}^k$, we have

$$
\begin{aligned}
&E[V(\mathbf{p}^{k+1}, \mathbf{p}^*_\epsilon(\lambda^{k+1})) - V(\mathbf{p}^k, \mathbf{p}^*_\epsilon(\lambda^k))|\mathcal{F}^k] \\
\le&\gamma^k(\lambda^k\mathbf{\Psi}(\mathbf{p}^k) + \boldsymbol{\phi}(\mathbf{p}^k))'(\mathbf{p}^*_\epsilon(\lambda^k) - \mathbf{p}^k) + \gamma^k(E[\hat{\boldsymbol{\phi}}^k_{\mathbf{p}}|\mathcal{F}^k] - \boldsymbol{\phi}(\mathbf{p}^k))'(\mathbf{p}^*_\epsilon(\lambda^k) - \mathbf{p}^k) \\
&+ \frac{1}{2}(\gamma^k)^2 E[\|\lambda^k\hat{\mathbf{\Psi}}^k + \hat{\boldsymbol{\phi}}^k_{\mathbf{p}}\|^2_\infty|\mathcal{F}^k] + C|\log\epsilon|\|\mathbf{p}^*_\epsilon(\lambda^{k+1}) - \mathbf{p}^*_\epsilon(\lambda^k)\|. \quad \text{(EC.25)}
\end{aligned}
$$

Note that on the right hand side we are still using $\boldsymbol{\phi}(\mathbf{p}^k)$, the derivative of the quadratic penalty in the formulation (16), rather than $\boldsymbol{\phi}_{\mathbf{p}}(\mathbf{p}^k, \mathbf{s}^k)$.

In order to use the martingale convergence theorem, we examine the following

$$\sum_{k=1}^{\infty} E[E[V(\mathbf{p}^{k+1}, \mathbf{p}_\epsilon^*(\lambda^{k+1})) - V(\mathbf{p}^k, \mathbf{p}_\epsilon^*(\lambda^k))|\mathcal{F}^k]^+] \tag{EC.26}$$
$$\leq \sum_{k=1}^{\infty} O(\gamma^k \sqrt{E[\|E[\hat{\boldsymbol{\phi}}_{\mathbf{p}}^k|\mathcal{F}^k] - \boldsymbol{\phi}(\mathbf{p}^k)\|^2]}) + \sum_{k=1}^{\infty} \frac{1}{2}(\gamma^k)^2 E[\|\lambda^k \hat{\boldsymbol{\Psi}}^k + \hat{\boldsymbol{\phi}}_{\mathbf{p}}^k\|_\infty^2 |\mathcal{F}^k] + \sum_{k=1}^{\infty} C|\log\epsilon|\|\mathbf{p}_\epsilon^*(\lambda^{k+1}) - \mathbf{p}_\epsilon^*(\lambda^k)\|.$$

We need to bound two quantities, $E[\|E[\hat{\boldsymbol{\phi}}_{\mathbf{p}}^k|\mathcal{F}^k] - \boldsymbol{\phi}(\mathbf{p}^k)\|^2]$ and $E[\|\lambda^k \hat{\boldsymbol{\Psi}}^k + \hat{\boldsymbol{\phi}}_{\mathbf{p}}^k\|_\infty^2 |\mathcal{F}^k]$. To bound the first one

$$\begin{aligned}
&E[\|E[\hat{\boldsymbol{\phi}}_{\mathbf{p}}^k|\mathcal{F}^k] - \boldsymbol{\phi}(\mathbf{p}^k))\|^2] \\
=&\sum_{i=1}^{m} E[|E[\hat{\phi}_{\mathbf{p},i}^k|\mathcal{F}^k] - \phi_i(\mathbf{p}^k)|^2] \\
=&4\sum_{i=1}^{m} E\big[\big|\sum_{j=1}^{n}(\Pi_j(E_{\mathbf{p}^k}[I(h(\mathbf{X}) \leq y_j)]) - s_j^k)E_{\mathbf{p}^k}[I(h(\mathbf{X}) \leq y_i)S_i(\mathbf{X};\mathbf{p}^k)]\big|^2\big] \\
\leq&4\sum_{i=1}^{m} E\big[\sum_{j=1}^{n}(\Pi_j(E_{\mathbf{p}^k}[I(h(\mathbf{X}) \leq y_j)]) - s_j^k)^2 \sum_{j=1}^{n}(E_{\mathbf{p}^k}[I(h(\mathbf{X}) \leq y_i)S_i(\mathbf{X};\mathbf{p}^k)])^2\big] \\
\leq&4E\big[\sum_{j=1}^{n}(\Pi_j(E_{\mathbf{p}^k}[I(h(\mathbf{X}) \leq y_j)]) - s_j^k)^2\big] \sum_{i=1}^{m}\sum_{j=1}^{n} \sup_{\mathbf{p}\in\mathcal{P}(\epsilon)} (E_{\mathbf{p}}[I(h(\mathbf{X}) \leq y_i)S_i(\mathbf{X};\mathbf{p})])^2 \\
\leq&Cmn\sum_{j=1}^{n} E[(\Pi_j(E_{\mathbf{p}^k}[I(h(\mathbf{X}) \leq y_j)]) - s_j^k)^2]
\end{aligned} \tag{EC.27}$$

where in the first inequality we use Cauchy Schwartz inequality, and the third inequality holds because each $E_{\mathbf{p}}[I(h(\mathbf{X}) \leq y_i)S_i(\mathbf{X};\mathbf{p})]$ by Lemma EC.1 is continuous in $\mathbf{p}$ and hence by a compactness argument is uniformly bounded in $\mathcal{P}(\epsilon)$. Therefore the key step lies in deriving an upper bound for each $E[(\Pi_j(E_{\mathbf{p}^k}[I(h(\mathbf{X}) \leq y_j)]) - s_j^k)^2]$, for which we need the counterpart of (EC.22) for s_j^k, i.e.

$$\begin{aligned}
&\frac{1}{2}(s_j^{k+1} - \Pi_j(E_{\mathbf{p}^k}[I(h(\mathbf{X}) \leq y_j)]))^2 - \frac{1}{2}(s_j^k - \Pi_j(E_{\mathbf{p}^k}[I(h(\mathbf{X}) \leq y_j)]))^2 \\
\leq&\beta^k \hat{\phi}_{\mathbf{s},j}^k(\Pi_j(E_{\mathbf{p}^k}[I(h(\mathbf{X}) \leq y_j)] - s_j^k) + \frac{1}{2}(\beta^k)^2(\hat{\phi}_{\mathbf{s},j}^k)^2.
\end{aligned}$$

Taking expectation with respect to $\mathcal{F}_k$ gives

$$\frac{1}{2}E[(s_j^{k+1} - \Pi_j(E_{\mathbf{p}^k}[I(h(\mathbf{X}) \leq y_j)]))^2|\mathcal{F}_k] - \frac{1}{2}(s_j^k - \Pi_j(E_{\mathbf{p}^k}[I(h(\mathbf{X}) \leq y_j)]))^2$$

$$\leq -2\beta^k(E_{\mathbf{p}^k}[I(h(\mathbf{X})\leq y_j)]-s_j^k)(\Pi_j(E_{\mathbf{p}^k}[I(h(\mathbf{X})\leq y_j)])-s_j^k)+\frac{1}{2}(\beta^k)^2(2+q_{1-\alpha}/\sqrt{n})^2$$
$$\leq -2\beta^k(\Pi_j(E_{\mathbf{p}^k}[I(h(\mathbf{X})\leq y_j)])-s_j^k)^2+\frac{C}{2}(\beta^k)^2. \tag{EC.28}$$

Note that with step size γ^k we have

$$\begin{aligned}
&E[(s_j^{k+1}-\Pi_j(E_{\mathbf{p}^k}[I(h(\mathbf{X})\leq y_j)]))^2|\mathcal{F}_k]\\
=&E[(s_j^{k+1}-\Pi_j(E_{\mathbf{p}^{k+1}}[I(h(\mathbf{X})\leq y_j)]))^2|\mathcal{F}_k]\\
&+2E[(s_j^{k+1}-\Pi_j(E_{\mathbf{p}^{k+1}}[I(h(\mathbf{X})\leq y_j)]))(\Pi_j(E_{\mathbf{p}^{k+1}}[I(h(\mathbf{X})\leq y_j)])-\Pi_j(E_{\mathbf{p}^k}[I(h(\mathbf{X})\leq y_j)]))|\mathcal{F}_k]\\
&+E[(\Pi_j(E_{\mathbf{p}^{k+1}}[I(h(\mathbf{X})\leq y_j)])-\Pi_j(E_{\mathbf{p}^k}[I(h(\mathbf{X})\leq y_j)]))^2|\mathcal{F}_k]\\
\geq&E[(s_j^{k+1}-\Pi_j(E_{\mathbf{p}^{k+1}}[I(h(\mathbf{X})\leq y_j)]))^2|\mathcal{F}_k]\\
&-2\sqrt{E[(s_j^{k+1}-\Pi_j(E_{\mathbf{p}^{k+1}}[I(h(\mathbf{X})\leq y_j)]))^2|\mathcal{F}_k]}\sqrt{E[(E_{\mathbf{p}^{k+1}}[I(h(\mathbf{X})\leq y_j)]-E_{\mathbf{p}^k}[I(h(\mathbf{X})\leq y_j)])^2|\mathcal{F}_k]}\\
\geq&E[(s_j^{k+1}-\Pi_j(E_{\mathbf{p}^{k+1}}[I(h(\mathbf{X})\leq y_j)]))^2|\mathcal{F}_k]-2(\sqrt{E[(s_j^{k+1}-\Pi_j(E_{\mathbf{p}^{k+1}}[I(h(\mathbf{X})\leq y_j)]))^2|\mathcal{F}_k]}C\gamma^k)\\
\geq&E[(s_j^{k+1}-\Pi_j(E_{\mathbf{p}^{k+1}}[I(h(\mathbf{X})\leq y_j)]))^2|\mathcal{F}_k]-2\beta^kE[(s_j^{k+1}-\Pi_j(E_{\mathbf{p}^{k+1}}[I(h(\mathbf{X})\leq y_j)]))^2|\mathcal{F}_k]-\frac{C^2(\gamma^k)^2}{2\beta^k}
\end{aligned}$$

where the second last inequality follows from

$$\begin{aligned}
|E_{\mathbf{p}^{k+1}}[I(h(\mathbf{X})\leq y_j)]-E_{\mathbf{p}^k}[I(h(\mathbf{X})\leq y_j)]| &\leq \|\mathbf{p}^{k+1}-\mathbf{p}^k\|\cdot\sup_{\mathbf{p}\in\mathcal{P}(\epsilon)}\|\nabla E_{\mathbf{p}}[I(h(\mathbf{X})\leq y_j)]\|\\
&\leq C\|\mathbf{p}^{k+1}-\mathbf{p}^k\|=O(\gamma^k)
\end{aligned}$$

and in the last inequality we use Young's inequality. Substituting the above into (EC.28) gives

$$\begin{aligned}
&E[(s_j^{k+1}-\Pi_j(E_{\mathbf{p}^{k+1}}[I(h(\mathbf{X})\leq y_j)]))^2|\mathcal{F}_k]\\
\leq&\frac{1-4\beta^k}{1-2\beta^k}(s_j^k-\Pi_j(E_{\mathbf{p}^k}[I(h(\mathbf{X})\leq y_j)]))^2+C((\beta^k)^2+\frac{(\gamma^k)^2}{\beta^k}).
\end{aligned}$$

Hence taking full expectation we have the following recursion

$$E[(s_j^{k+1}-\Pi_j(E_{\mathbf{p}^{k+1}}[I(h(\mathbf{X})\leq y_j)]))^2]\leq(1-2\beta^k)E[(s_j^k-\Pi_j(E_{\mathbf{p}^k}[I(h(\mathbf{X})\leq y_j)]))^2]+C((\beta^k)^2+\frac{(\gamma^k)^2}{\beta^k}).$$

Denote by $D_j^k=E[(s_j^k-\Pi_j(E_{\mathbf{p}^k}[I(h(\mathbf{X})\leq y_j)]))^2]$. When the sequences γ^k and β^k are taken to be (39), the recursion reduces to

$$D_j^{k+1}\leq(1-\frac{2b}{k^{\alpha_2}})D_j^k+C(\frac{1}{k^{2\alpha_2}}+\frac{1}{k^{2\alpha_1-\alpha_2}})$$

which by Lemma EC.2 implies that $D_j^k = O(\frac{1}{k^{\alpha_2}} + \frac{1}{k^{2(\alpha_1-\alpha_2)}})$. Therefore from (EC.27) we conclude

$$E[\|E[\hat{\boldsymbol{\phi}}_{\mathbf{p}}^k|\mathcal{F}^k] - \boldsymbol{\phi}(\mathbf{p}^k))\|^2] \le Cmn\sum_{j=1}^{n} D_j^k = O(\frac{1}{k^{\alpha_2}} + \frac{1}{k^{2(\alpha_1-\alpha_2)}}). \tag{EC.29}$$

To bound the term $E[\|\lambda^k\hat{\boldsymbol{\Psi}}^k + \hat{\boldsymbol{\phi}}_{\mathbf{p}}^k\|_\infty^2|\mathcal{F}^k]$, we use Minkowski inequality to get

$$\begin{aligned} E[\|\hat{\boldsymbol{\phi}}_{\mathbf{p}}^k\|_\infty^2|\mathcal{F}^k] &\le E\left[\sum_{i=1}^{m}\left(\hat{\phi}_{\mathbf{p},i}^k\right)^2\middle|\mathcal{F}^k\right] \\ &\le 4n\sum_{i=1}^{m}\sum_{j=1}^{n}(2+q_{1-\alpha}/\sqrt{n})^2 E_{\mathbf{p}^k}[\big(I(h(\mathbf{X})\le y_j)S_i(\mathbf{X};\mathbf{p}^k)\big)^2] \end{aligned}$$

and

$$E[\|\hat{\boldsymbol{\Psi}}^k\|_\infty^2|\mathcal{F}^k] \le E\left[\sum_{i=1}^{m}\left(\hat{\Psi}_i^k\right)^2\middle|\mathcal{F}^k\right] \le 4m\sum_{i=1}^{m}\sum_{j=1}^{n} E_{\mathbf{p}^k}[\big(g(\mathbf{X})S_i(\mathbf{X};\mathbf{p}^k)\big)^2].$$

Again by Proposition EC.1, each expectation in the sum is continuous in $\mathbf{p}^k$, hence uniformly bounded in $\mathcal{P}(\epsilon)$ by compactness. Therefore $E[\|\hat{\boldsymbol{\phi}}_{\mathbf{p}}^k\|_\infty^2|\mathcal{F}^k] \le C$ and $E[\|\hat{\boldsymbol{\Psi}}^k\|_\infty^2|\mathcal{F}^k] \le C$ uniformly holds for some $C>0$. This implies

$$E[\|\lambda^k\hat{\boldsymbol{\Psi}}^k + \hat{\boldsymbol{\phi}}_{\mathbf{p}}^k\|_\infty^2|\mathcal{F}^k] \le 2(E[\|\hat{\boldsymbol{\phi}}_{\mathbf{p}}^k\|_\infty^2|\mathcal{F}^k] + (\lambda^k)^2 E[\|\hat{\boldsymbol{\Psi}}^k\|_\infty^2|\mathcal{F}^k]) \le C. \tag{EC.30}$$

Assumption 2 entails $\gamma^k(\lambda^k\boldsymbol{\Psi}(\mathbf{p}^k) + \boldsymbol{\phi}(\mathbf{p}^k))'(\mathbf{p}_\epsilon^*(\lambda^k) - \mathbf{p}^k) \le 0$. Substituting (EC.29) and (EC.30) into (EC.26) we arrive at

$$\begin{aligned} &\sum_{k=1}^{\infty} E[E[V(\mathbf{p}^{k+1}, \mathbf{p}_\epsilon^*(\lambda^{k+1})) - V(\mathbf{p}^k, \mathbf{p}_\epsilon^*(\lambda^k))|\mathcal{F}^k]^+] \\ \le &\sum_{k=1}^{\infty} O(\frac{1}{k^{\alpha_1+\frac{1}{2}\alpha_2}} + \frac{1}{k^{2\alpha_1-\alpha_2}}) + \sum_{k=1}^{\infty}\frac{1}{2}(\gamma^k)^2 E[\|\lambda^k\hat{\boldsymbol{\Psi}}^k + \hat{\boldsymbol{\phi}}_{\mathbf{p}}^k\|_\infty^2|\mathcal{F}^k] + \sum_{k=1}^{\infty} C|\log\epsilon|\|\mathbf{p}_\epsilon^*(\lambda^{k+1}) - \mathbf{p}_\epsilon^*(\lambda^k)\| \\ \le &C\sum_{k=1}^{\infty}\big(\frac{1}{k^{\alpha_1+\frac{1}{2}\alpha_2}} + \frac{1}{k^{2\alpha_1-\alpha_2}} + \frac{1}{k^{2\alpha_1}} + \|\mathbf{p}_\epsilon^*(\lambda^{k+1}) - \mathbf{p}_\epsilon^*(\lambda^k)\|\big) < \infty. \end{aligned}$$

By martingale convergence theorem (Corollary in Section 3 in Blum (1954), restated in Theorem EC.4 in the Appendix), we have $V(\mathbf{p}^k, \mathbf{p}_\epsilon^*(\lambda^k))$ converges a.s. to some random variable V_∞. Because of $\mathbf{p}_\epsilon^*(\lambda^k) \to \mathbf{p}_\epsilon^* \in \mathcal{P}(\epsilon)$ and inequality (EC.23) which holds uniformly for $\mathbf{p}^{k+1} \in \mathcal{P}(\epsilon)$, we conclude that $V(\mathbf{p}^k, \mathbf{p}_\epsilon^*)$ converges a.s. to the same variable V_∞.

Now we would like to argue that the limit $V_\infty = 0$ a.s.. To this end it suffices to show that a.s. there exists a subsequence of $\mathbf{p}^k$ converging to $\mathbf{p}^*_\epsilon$. Taking expectation and summing up on both sides of (EC.37) and using similar bounding techniques, we have

$$\begin{aligned}&\sum_{k=1}^{\infty} E[\gamma^k(\lambda^k \boldsymbol{\Psi}(\mathbf{p}^k) + \boldsymbol{\phi}(\mathbf{p}^k))'(\mathbf{p}^k - \mathbf{p}^*_\epsilon(\lambda^k))]\\ \le& V(\mathbf{p}^1, \mathbf{p}^*_\epsilon(\lambda^1)) + C\sum_{k=1}^{\infty}\big(\frac{\gamma^k}{\sqrt{M_1^k}} + (\gamma^k)^2 + \|\mathbf{p}^*_\epsilon(\lambda^{k+1}) - \mathbf{p}^*_\epsilon(\lambda^k)\|\big) < \infty.\end{aligned}$$

Since each $(\lambda^k \boldsymbol{\Psi}(\mathbf{p}^k) + \boldsymbol{\phi}(\mathbf{p}^k))'(\mathbf{p}^k - \mathbf{p}^*_\epsilon(\lambda^k)) \ge 0$, it follows that

$$\sum_{k=1}^{\infty} \gamma^k(\lambda^k \boldsymbol{\Psi}(\mathbf{p}^k) + \boldsymbol{\phi}(\mathbf{p}^k))'(\mathbf{p}^k - \mathbf{p}^*_\epsilon(\lambda^k)) < \infty \, a.s..$$

Define the (random) set of feasible-solution indices

$$\mathcal{K}_1 = \{k \ge 1 : \mathbf{p}^k \text{ is feasible for (33)}\}.$$

Note that when $\mathbf{p}^k$ is feasible for (33), it holds $\boldsymbol{\phi}(\mathbf{p}^k) = \mathbf{0}$, hence

$$\sum_{k\in\mathcal{K}_1} \gamma^k \lambda^k \boldsymbol{\Psi}(\mathbf{p}^k)'(\mathbf{p}^k - \mathbf{p}^*_\epsilon(\lambda^k)) < \infty, \, a.s. \tag{EC.31}$$

$$\sum_{k\notin\mathcal{K}_1} \gamma^k(\lambda^k \boldsymbol{\Psi}(\mathbf{p}^k) + \boldsymbol{\phi}(\mathbf{p}^k))'(\mathbf{p}^k - \mathbf{p}^*_\epsilon(\lambda^k)) < \infty, \, a.s. \tag{EC.32}$$

If $\sum_{k\in\mathcal{K}_1} \gamma^k\lambda^k = \infty$, then due to (EC.31) there must exist a subsequence $k_i \in \mathcal{K}_1$ such that $\boldsymbol{\Psi}(\mathbf{p}^{k_i})'(\mathbf{p}^{k_i} - \mathbf{p}^*_\epsilon(\lambda^{k_i})) \to 0$. Since $\mathbf{p}^*_\epsilon(\lambda^k) \to \mathbf{p}^*_\epsilon$, this implies that $\boldsymbol{\Psi}(\mathbf{p}^{k_i})'(\mathbf{p}^{k_i} - \mathbf{p}^*_\epsilon) \to 0$, which by Assumption 1 further implies that $\mathbf{p}^{k_i} \to \mathbf{p}^*_\epsilon$.

Otherwise if $\sum_{k\in\mathcal{K}_1} \gamma^k\lambda^k < \infty$ then it must hold $\sum_{k\notin\mathcal{K}_1} \gamma^k\lambda^k = \infty$ because the parameters stated in the theorem satisfy $\sum_{k=1}^{\infty} \gamma^k\lambda^k = \infty$. Due to (EC.32) there exists a subsequence $k_i \notin \mathcal{K}_1$ such that

$$(\boldsymbol{\Psi}(\mathbf{p}^{k_i}) + \frac{1}{\lambda^{k_i}}\boldsymbol{\phi}(\mathbf{p}^{k_i}))'(\mathbf{p}^{k_i} - \mathbf{p}^*_\epsilon(\lambda^{k_i})) \to 0. \tag{EC.33}$$

By a compactness argument, there exists a subsubsequence $k_i' \notin \mathcal{K}_1$ such that $\mathbf{p}^{k_i'}$ converges to some $\mathbf{q} \in \mathcal{P}(\epsilon)$. First we argue that $\mathbf{q}$ must be feasible for (33). Since $\lambda^k \to 0$ and $\boldsymbol{\Psi}(\mathbf{p}^{k_i'}), \boldsymbol{\phi}(\mathbf{p}^{k_i'})$ are uniformly bounded, it is clear that $(\lambda^{k_i'}\boldsymbol{\Psi}(\mathbf{p}^{k_i'}) + \boldsymbol{\phi}(\mathbf{p}^{k_i'}))'(\mathbf{p}^{k_i'} - \mathbf{p}^*_\epsilon(\lambda^{k_i'})) \to 0$ and $\lambda^{k_i'}\boldsymbol{\Psi}(\mathbf{p}^{k_i'})'(\mathbf{p}^{k_i'} -$

$\mathbf{p}^*_\epsilon(\lambda^{k'_i})) \to 0$ hold. Therefore the difference $\boldsymbol{\phi}(\mathbf{p}^{k'_i})'(\mathbf{p}^{k'_i} - \mathbf{p}^*_\epsilon(\lambda^{k'_i})) \to 0$. On the other hand $\boldsymbol{\phi}(\mathbf{p}^{k'_i})'(\mathbf{p}^{k'_i} - \mathbf{p}^*_\epsilon(\lambda^{k'_i})) \to \boldsymbol{\phi}(\mathbf{q})'(\mathbf{q} - \mathbf{p}^*_\epsilon)$ because $\boldsymbol{\phi}(\cdot)$ is continuous. This means $\boldsymbol{\phi}(\mathbf{q})'(\mathbf{q} - \mathbf{p}^*_\epsilon) = 0$ so $\mathbf{q}$ must be feasible in view of Assumption 1. Then we argue $\mathbf{q} = \mathbf{p}^*_\epsilon$ in fact. If $\mathbf{q} \neq \mathbf{p}^*_\epsilon$ then $\boldsymbol{\Psi}(\mathbf{q})'(\mathbf{q} - \mathbf{p}^*_\epsilon) > 0$ by Assumption 1, and we derive a contradiction as follows. Recall that each $\mathbf{p}^{k'_i}$ is infeasible for (33) and $\boldsymbol{\phi}(\mathbf{p}^{k'_i}) \to \boldsymbol{\phi}(\mathbf{q}) = 0$, where $\boldsymbol{\phi}(\mathbf{q})$ vanishes since $\mathbf{q}$ is feasible. We have

$$
\begin{aligned}
&\liminf_i (\boldsymbol{\Psi}(\mathbf{p}^{k'_i}) + \frac{1}{\lambda^{k'_i}} \boldsymbol{\phi}(\mathbf{p}^{k'_i}))'(\mathbf{p}^{k'_i} - \mathbf{p}^*_\epsilon(\lambda^{k'_i})) \\
=&\liminf_i \left\{ \boldsymbol{\Psi}(\mathbf{p}^{k'_i})'(\mathbf{p}^{k'_i} - \mathbf{p}^*_\epsilon(\lambda^{k'_i})) + \frac{1}{\lambda^{k'_i}} \boldsymbol{\phi}(\mathbf{p}^{k'_i})'(\mathbf{p}^{k'_i} - \mathbf{p}^*_\epsilon) + \frac{1}{\lambda^{k'_i}} \boldsymbol{\phi}(\mathbf{p}^{k'_i})'(\mathbf{p}^*_\epsilon - \mathbf{p}^*_\epsilon(\lambda^{k'_i})) \right\} \\
\geq&\liminf_i \boldsymbol{\Psi}(\mathbf{p}^{k'_i})'(\mathbf{p}^{k'_i} - \mathbf{p}^*_\epsilon(\lambda^{k'_i})) + \liminf_i \frac{1}{\lambda^{k'_i}} \boldsymbol{\phi}(\mathbf{p}^{k'_i})'(\mathbf{p}^{k'_i} - \mathbf{p}^*_\epsilon) + \liminf_i \frac{1}{\lambda^{k'_i}} \boldsymbol{\phi}(\mathbf{p}^{k'_i})'(\mathbf{p}^*_\epsilon - \mathbf{p}^*_\epsilon(\lambda^{k'_i})) \\
\geq&\boldsymbol{\Psi}(\mathbf{q})'(\mathbf{q} - \mathbf{p}^*_\epsilon) + 0 + \liminf_i \frac{1}{\lambda^{k'_i}} o(1) O(\lambda^{k'_i}) = \boldsymbol{\Psi}(\mathbf{q})'(\mathbf{q} - \mathbf{p}^*_\epsilon) > 0
\end{aligned}
$$

which contradicts (EC.33).

The above argument shows that a.s. there exists a subsequence of $\mathbf{p}^k$ converging to $\mathbf{p}^*_\epsilon$, hence the corresponding $V(\mathbf{p}^k, \mathbf{p}^*_\epsilon) \to 0$. Since we have proved above that $V(\mathbf{p}^k, \mathbf{p}^*_\epsilon)$ converges a.s., the limit must be identically 0. Therefore, by Pinsker's inequality, we have $\mathbf{p}^k \to \mathbf{p}^*_\epsilon$ in total variation a.s.. This concludes the theorem. $\square$

As discussed at the end of Section 5, our results and algorithms still hold in the presence of a collection of auxiliary independent input processes $\mathbf{W}$ distributed according to known distributions. Like in Sections 4 and EC.2, all proofs in this section still apply by invoking the same conditioning argument. Specifically, in Proposition 3 the expressions (27),(28),(29) are still valid with $h(\mathbf{X}), g(\mathbf{X})$ replaced by $E_{P^0_W}[h(\mathbf{X}, \mathbf{W})|\mathbf{X}], E_{P^0_W}[g(\mathbf{X}, \mathbf{W})|\mathbf{X}]$, so are the estimators (30),(31). In Lemma EC.1, the continuity of moments of gradient estimators can be similarly established by conditioning. For example, the moment $E_{\mathbf{p}}[(g(\mathbf{X}, \mathbf{W})S_i(\mathbf{X}; \mathbf{p}))^2]$ is equal to $E_{\mathbf{p}}\big[E_{P^0_W}[g^2(\mathbf{X}, \mathbf{W})|\mathbf{X}]S_i^2(\mathbf{X}; \mathbf{p})\big]$, hence the same proof applies viewing $E_{P^0_W}[g^2(\mathbf{X}, \mathbf{W})|\mathbf{X}]$ as the performance measure. Similarly, the boundedness condition in Theorem 3 is made on $E_{P^0_W}[g(\mathbf{X}, \mathbf{W})|\mathbf{X}]$ instead.

EC.4. An Alternate MDSA Algorithm and Some Further Discussion

Algorithm 3 shows an alternate MDSA algorithm that does not use slack variables, but at the expense of increasing the simulation replication size per iteration.

When applied to the (restricted) penalized minimization problem (16), MDSA solves the following optimization given a current iterate $\mathbf{p}^k$

$$\begin{aligned} &\min \qquad \gamma^k(\lambda\hat{\mathbf{\Psi}}^k + \hat{\boldsymbol{\phi}}^k)'(\mathbf{p}-\mathbf{p}^k) + V(\mathbf{p}^k,\mathbf{p}) \\ &\text{subject to } \mathbf{p}\in\mathcal{P}(\epsilon) \end{aligned} \tag{EC.34}$$

where $\hat{\mathbf{\Psi}}^k$ carries the gradient information of the target performance measure ψ at $\mathbf{p}^k$, $\hat{\boldsymbol{\phi}}^k$ contains the gradient information of the quadratic penalty function in (16) at $\mathbf{p}^k$, and $V(\cdot,\cdot)$ is the KL divergence defined in (19). The step-wise subproblem (EC.34) without stochastic noise is also called the entropic descent algorithm (Beck and Teboulle (2003)). To make it a single-run procedure, we decrease the penalty coefficient λ as the iteration goes on, and thereby arrive at the following counterpart of (20)

$$\begin{aligned} &\min \qquad \gamma^k(\lambda^k\hat{\mathbf{\Psi}}^k + \hat{\boldsymbol{\phi}}^k)'(\mathbf{p}-\mathbf{p}^k) + V(\mathbf{p}^k,\mathbf{p}) \\ &\text{subject to } \mathbf{p}\in\mathcal{P}(\epsilon) \end{aligned} \tag{EC.35}$$

Inspired by (28) in Proposition 3, we use the following estimator for the gradient of the penalty function $\boldsymbol{\phi}(\mathbf{p}) = (\phi_i(\mathbf{p}))_{i=1}^m$

$$\hat{\phi}_i(\mathbf{p}) = 2\sum_{j=1}^{n}(u_j - \Pi_j(u_j))\frac{1}{M_2}\sum_{r=1}^{M_2} I(h(\tilde{\mathbf{X}}^{(r)})\le y_j)S_i(\tilde{\mathbf{X}}^{(r)};\mathbf{p}),\ u_j = \frac{1}{M_1}\sum_{r=1}^{M_1} I(h(\mathbf{X}^{(r)})\le y_j) \tag{EC.36}$$

where $\mathbf{X}^{(r)}$ and $\tilde{\mathbf{X}}^{(r)}$ are independent copies of the i.i.d. input process generated under $\mathbf{p}$ and are used simultaneously for all i,j. Since we are using the plug-in estimator $\Pi_j(u_j)$ for the projection, in general (EC.36) has a bias. In particular, the bias can be shown to vanish as slow as $O(1/\sqrt{M_1})$ if $E_{\mathbf{p}}[I(h(\mathbf{X})\le y_j)]$ is close to either $\hat{F}_Y(y_j+) - q_{1-\alpha}/\sqrt{n}$ or $\hat{F}_Y(y_j-) + q_{1-\alpha}/\sqrt{n}$. Due to this biasedness, the batch size M_1 has to grow to ∞ in the course of iteration in order for the algorithm to converge properly.

Like for Algorithm 2, the following provides the convergence guarantee of Algorithm 3:

THEOREM EC.1. *Under Assumptions 1, 2 and 3, if the step size sequence $\{\gamma^k\}$, the penalty sequence $\{\lambda^k\}$ and the sample size sequence $\{M_1^k\}$ of Algorithm 3 are chosen such that*

$$\sum_{k=1}^{\infty}\gamma^k\lambda^k = \infty,\quad \sum_{k=1}^{\infty}(\gamma^k)^2 < \infty,\quad \sum_{k=1}^{\infty}\frac{\gamma^k}{\sqrt{M_1^k}} < \infty,\quad \lambda^k\to 0 \text{ and non-increasing}$$

Algorithm 3 Alternate MDSA for solving (16)

Input: A small parameter $\epsilon > 0$, initial solution $\mathbf{p}^1 \in \mathcal{P}(\epsilon) = \{\mathbf{p} : \sum_{i=1}^m p_i = 1, p_i \geq \epsilon \text{ for } i = 1, \ldots, m\}$, a step size sequence γ^k, a penalty sequence λ^k, a sample size sequences M_1^k, and sample sizes M_2, M_3.

Iteration: For $k = 1, 2, \ldots$, do the following: Given $\mathbf{p}^k$,

1. Estimate the probabilities $E_{\mathbf{p}^k}[I(h(\mathbf{X}) \leq y_j)], j = 1, \ldots, n$ with

$$u_j^k = \frac{1}{M_1^k} \sum_{r=1}^{M_1^k} I(h(\mathbf{X}^{(r)}) \leq y_j)$$

where $\mathbf{X}^{(r)}$ are M_1^k independent copies of the input process generated under $\mathbf{p}^k$.

2. Estimate $\hat{\boldsymbol{\phi}}^k = (\hat{\phi}_1^k, \ldots, \hat{\phi}_m^k)$, the gradient of the penalty term, with

$$\hat{\phi}_i^k = 2 \sum_{j=1}^{n} (u_j^k - \Pi_j(u_j^k)) \frac{1}{M_2} \sum_{r=1}^{M_2} I(h(\tilde{\mathbf{X}}^{(r)}) \leq y_j) S_i(\tilde{\mathbf{X}}^{(r)}; \mathbf{p}^k)$$

where $\mathbf{X}^{(r)}$ are the same set of replications used in Step 1, and $\tilde{\mathbf{X}}^{(r)}$ are another M_2 independent copies of the input process generated under $\mathbf{p}^k$.

3. Estimate $\hat{\boldsymbol{\Psi}}^k = (\hat{\Psi}_1^k, \ldots, \hat{\Psi}_m^k)$, the gradient of $E_{\mathbf{p}}[g(\mathbf{X})]$, with

$$\hat{\Psi}_i^k = \frac{1}{M_3} \sum_{r=1}^{M_3} g(\tilde{\tilde{\mathbf{X}}}^{(r)}) S_i(\tilde{\tilde{\mathbf{X}}}^{(r)}; \mathbf{p}^k)$$

where $\tilde{\tilde{\mathbf{X}}}^{(r)}$ are another M_3 independent copies of the input process generated under $\mathbf{p}^k$.

4. Compute $\mathbf{p}^{k+1} = (p_1^{k+1}, \ldots, p_m^{k+1})$ by running Algorithm 1 with $p_i = p_i^k$ and $\xi_i = \gamma^k(\lambda^k \hat{\Psi}_i^k + \hat{\phi}_i^k)$.

then $\mathbf{p}^k$ generated in Algorithm 3 converges to $\mathbf{p}_\epsilon^*$ a.s.. In particular, when the sequences are chosen as

$$\gamma^k = \frac{a}{k^{\alpha_1}}, \ \frac{1}{2} < \alpha_1 \leq 1$$

$$M_1^k = bk^{\alpha_2}, \ \alpha_2 > 2(1 - \alpha_1)$$

$$\lambda^k = \begin{cases} \frac{c}{k^{\alpha_3}}, \ 0 < \alpha_3 \leq 1 - \alpha_1 & \text{if } \frac{1}{2} < \alpha_1 < 1 \\ \frac{c}{\log k} & \text{if } \alpha_1 = 1 \end{cases}$$

$\mathbf{p}^k$ converges to $\mathbf{p}_\epsilon^*$ a.s..

Here are some discussions on the parameter choices of Algorithm 3. $\sum_{k=1}^{\infty}(\gamma^k)^2<\infty$ is a standard condition in SA which ensures that the effect of stochasticity will vanish eventually, whereas the condition $\sum_{k=1}^{\infty}\gamma^k/\sqrt{M_1^k}<\infty$ is meant to eliminate the effect of biasedness of the gradient estimator (EC.36). What is special about our MDSA is the condition $\sum_{k=1}^{\infty}\gamma^k\lambda^k=\infty$. The rationale for this condition is as follows. When $\mathbf{p}$ is feasible for (33), the gradient of the penalty function vanishes, i.e. $\boldsymbol{\phi}(\mathbf{p})=\mathbf{0}$, hence the effective step size in (EC.35) is $\gamma^k\lambda^k$. Under the condition $\sum_{k=1}^{\infty}\gamma^k\lambda^k=\infty$, the algorithm is able to fully explore the feasible set of (33).

The difference between Algorithm 2 and 3 lies in how the projection $\Pi_j(E_{\mathbf{p}^k}[I(h(\mathbf{X})\le y_j)])$ at the current iterate $\mathbf{p}^k$ is estimated. Algorithm 3 computes the projection by directly simulating $E_{\mathbf{p}^k}[I(h(\mathbf{X})\le y_j)]$ from scratch and substituting into the projection Π_j in each iteration, whereas Algorithm 2 iteratively updates the slack variables s_j^k together with the decision variable in such a way that eventually each s_j^k consistently estimates the projection $\Pi_j(E_{\mathbf{p}^k}[I(h(\mathbf{X})\le y_j)])$.

We point out that both Algorithm 2 and 3 are essentially solving the formulation (16), despite the fact that the design of Algorithm 2 is mostly based on (15). The reason that neither of Algorithm 2 and 3 solves the formulation (15) has to do with the fact that algorithmically the formulation (15) with slack variables in general is not as well behaved as the formulation (16) with the projections, despite their mathematical equivalence. To see this, consider a generic inequality constraint $f(x)\le 0$ where x is some decision variable. It is easy to see that the quadratic penalty $(\max\{f(x),0\})^2$ expressed via projection preserves the convexity of $f(x)$, whereas the one with slack variable $s\le 0$, $(f(x)-s)^2$, can very likely lose convexity even if $f(x)$ itself is convex. In fact, if $(f(x)-s)^2$ is jointly convex in x and s, $(\max\{f(x),0\})^2$ is guaranteed to be convex. This also explains why the general convexity criterion in Assumptions 1 and 2 is imposed on formulation (16).

Proof of Theorem EC.1. The proof resembles that of Theorem 4. Let $\mathcal{F}^k$ be the filtration generated by $\{\mathbf{p}^1,\dots,\mathbf{p}^k\}$. Following the same line of argument, we have the following counterpart of (EC.25)

$$E[V(\mathbf{p}^{k+1},\mathbf{p}_\epsilon^*(\lambda^{k+1}))-V(\mathbf{p}^k,\mathbf{p}_\epsilon^*(\lambda^k))|\mathcal{F}^k]$$

$$\leq \gamma^k(\lambda^k \boldsymbol{\Psi}(\mathbf{p}^k) + \boldsymbol{\phi}(\mathbf{p}^k))'(\mathbf{p}_\epsilon^*(\lambda^k) - \mathbf{p}^k) + \gamma^k(E[\hat{\boldsymbol{\phi}}^k|\mathcal{F}^k] - \boldsymbol{\phi}(\mathbf{p}^k))'(\mathbf{p}_\epsilon^*(\lambda^k) - \mathbf{p}^k)$$
$$+ \frac{1}{2}(\gamma^k)^2 E[\|\lambda^k \hat{\boldsymbol{\Psi}}^k + \hat{\boldsymbol{\phi}}^k\|_\infty^2 |\mathcal{F}^k] + C|\log\epsilon|\|\mathbf{p}_\epsilon^*(\lambda^{k+1}) - \mathbf{p}_\epsilon^*(\lambda^k)\|. \tag{EC.37}$$

We need to bound $E[\hat{\boldsymbol{\phi}}^k|\mathcal{F}^k] - \boldsymbol{\phi}(\mathbf{p}^k)$ and $E[\|\lambda^k \hat{\boldsymbol{\Psi}}^k + \hat{\boldsymbol{\phi}}^k\|_\infty^2|\mathcal{F}^k]$. By independence of $\mathbf{X}^{(r)}$ and $\tilde{\mathbf{X}}^{(r)}$ and conditional Jensen's inequality

$$\begin{aligned}
|E[\hat{\phi}_i^k\big|\mathcal{F}^k] - \phi_i(\mathbf{p}^k)| &= 2\Big|\sum_{j=1}^n (\Pi_j(E_{\mathbf{p}^k}[I(h(\mathbf{X}) \leq y_j)]) - E[\Pi_j(u_j^k)\big|\mathcal{F}^k])E_{\mathbf{p}^k}[I(h(\mathbf{X}) \leq y_j)S_i(\mathbf{X};\mathbf{p}^k)]\Big| \\
&\leq C\sum_{j=1}^n |\Pi_j(E_{\mathbf{p}^k}[I(h(\mathbf{X}) \leq y_j)]) - E[\Pi_j(u_j^k)\big|\mathcal{F}^k]| \\
&\leq C\sum_{j=1}^n E[|\Pi_j(E_{\mathbf{p}^k}[I(h(\mathbf{X}) \leq y_j)]) - \Pi_j(u_j^k)|\big|\mathcal{F}^k] \\
&\leq C\sum_{j=1}^n E[|E_{\mathbf{p}^k}[I(h(\mathbf{X}) \leq y_j)] - u_j^k|\big|\mathcal{F}^k] \\
&\leq C\sum_{j=1}^n \sqrt{E[(E_{\mathbf{p}^k}[I(h(\mathbf{X}) \leq y_j)] - u_j^k)^2\big|\mathcal{F}^k]} = O(\frac{1}{\sqrt{M_1^k}})
\end{aligned}$$

where in the second last inequality we use the contraction property of projection, i.e. $|\Pi_j(a) - \Pi_j(b)| \leq |a-b|$ for any $a, b \in \mathbb{R}$. The first inequality holds because each derivative $E_{\mathbf{p}^k}[I(h(\mathbf{X}) \leq y_j)S_i(\mathbf{X};\mathbf{p}^k)]$ by Proposition EC.1 is continuous in $\mathbf{p}$ and by a compactness argument is hence uniformly bounded in $\mathcal{P}(\epsilon)$. Following the proof of Theorem 4, one can show that

$$E[\|\lambda^k \hat{\boldsymbol{\Psi}}^k + \hat{\boldsymbol{\phi}}^k\|_\infty^2|\mathcal{F}^k] \leq C$$

as the counterpart of (EC.30).

Therefore, taking expectation and summing up on both sides of (EC.37), we have

$$\begin{aligned}
&\sum_{k=1}^\infty E[E[V(\mathbf{p}^{k+1}, \mathbf{p}_\epsilon^*(\lambda^{k+1})) - V(\mathbf{p}^k, \mathbf{p}_\epsilon^*(\lambda^k))|\mathcal{F}^k]^+] \\
\leq &\sum_{k=1}^\infty O(\frac{\gamma^k}{\sqrt{M_1^k}}) + \sum_{k=1}^\infty \frac{1}{2}(\gamma^k)^2 E[\|\lambda^k \hat{\boldsymbol{\Psi}}^k + \hat{\boldsymbol{\phi}}^k\|_\infty^2|\mathcal{F}^k] + \sum_{k=1}^\infty C|\log\epsilon|\|\mathbf{p}_\epsilon^*(\lambda^{k+1}) - \mathbf{p}_\epsilon^*(\lambda^k)\| \\
\leq &C\sum_{k=1}^\infty \big(\frac{\gamma^k}{\sqrt{M_1^k}} + (\gamma^k)^2 + \|\mathbf{p}_\epsilon^*(\lambda^{k+1}) - \mathbf{p}_\epsilon^*(\lambda^k)\|\big) < \infty.
\end{aligned}$$

The rest of the proof is the same as that of Theorem 4. □

EC.5. A Randomized Stochastic Projected Gradient Algorithm for the Comparison in Section 6

We show a randomized stochastic projected gradient (RSPG) algorithm that we compare with in the numerical section. Algorithm 4 shows the procedure for a single run. Algorithm 5 includes a post-optimization step to boost its performance. As a rough guidance, we use $\bar{\gamma} < 1/L$ where L is the Lipschitz constant of the gradient function, and $M = O(Nm)$ (the m here could possibly be removed), where m is the dimension of the decision space. S could be a small number like $5, 10$, and the post-optimization batch size M' is chosen to be some big number. The penalty λ is chosen small and fixed.

EC.6. Auxiliary Results

EC.6.1. Results on Empirical Processes and U-Statistics

We first introduce some definitions. Using Definition 2.1.6 in Van Der Vaart and Wellner (1996), given two functions l and u, the bracket $[l, u]$ is defined as the set of all functions f with $l \le f \le u$. An ϵ-bracket is a bracket $[l, u]$ with $\|l - u\| < \epsilon$ for some norm $\|\cdot\|$. For a class of measurable functions $\mathcal{F}$ on $\mathcal{Y} \to \mathbb{R}$, the bracketing number $N_{[]}(\epsilon, \mathcal{F}, \|\cdot\|)$ is the minimum number of ϵ-brackets needed to cover $\mathcal{F}$. Moreover, define the envelope of $\mathcal{F}$ as $F(\cdot) = \sup_{f \in \mathcal{F}} |f(\cdot)|$.

We have the following theorem:

THEOREM EC.2 **(Problem 3 in Chapter 2.7 of Van Der Vaart and Wellner (1996))**. *Let $\mathcal{F}$ be a class of measurable functions $f(\cdot, r)$ on $\mathcal{Y} \to \mathbb{R}$, indexed by $0 \le r \le 1$, such that $f(x, \cdot)$ is monotone for each x. If the envelope function of $\mathcal{F}$ is square integrable, then the bracketing number of $\mathcal{F}$ is polynomial.*

To introduce the next theorem, we define several additional notions. For any function $f : \mathcal{X}^T \to \mathbb{R}$, and $X_1, \ldots, X_m$ generated i.i.d. from P, define the U-operator U_T^m by

$$U_T^m f = U_T^m(f, P) = \frac{(m-T)!}{m!} \sum_{(i_1, \ldots, i_T) \in I_T^m} f(X_{i_1}, \ldots, X_{i_T}) \tag{EC.38}$$

Algorithm 4 Randomized stochastic projected gradient (RSPG) for solving (15)

Input: A small parameter $\epsilon > 0$, initial solution $\mathbf{p}^1 \in \mathcal{P}(\epsilon) = \{\mathbf{p} : \sum_{i=1}^m p_i = 1, p_i \geq \epsilon \text{ for } i = 1, \ldots, m\}$ and $\mathbf{s}^1 \in [\hat{F}_Y(y_1+) - \frac{q_{1-\alpha}}{\sqrt{n}}, \hat{F}_Y(y_1-) + \frac{q_{1-\alpha}}{\sqrt{n}}] \times \cdots \times [\hat{F}_Y(y_n+) - \frac{q_{1-\alpha}}{\sqrt{n}}, \hat{F}_Y(y_n+) - \frac{q_{1-\alpha}}{\sqrt{n}}]$, step size $\bar{\gamma}$ for both $\mathbf{p}$ and $\mathbf{s}$, penalty λ, batch size M, and number of iterations N.

Generate random stopping time: Draw τ uniformly from $\{1, \ldots, N\}$

Iteration: For $k = 1, \ldots, \tau - 1$ do the following: Given $\mathbf{p}^k, \mathbf{s}^k$,

1. Estimate $\hat{\boldsymbol{\phi}}^k_{\mathbf{p}} = (\hat{\phi}^k_{\mathbf{p},1}, \ldots, \hat{\phi}^k_{\mathbf{p},m})$, the gradient of the penalty term with respect to $\mathbf{p}$, with

$$\hat{\phi}^k_{\mathbf{p},i} = 2\sum_{j=1}^{n} \frac{1}{M}\sum_{r=1}^{M}(I(h(\mathbf{X}^{(r)}) \leq y_j) - s^k_j)\frac{1}{M}\sum_{r=1}^{M} I(h(\tilde{\mathbf{X}}^{(r)}) \leq y_j) S_i(\tilde{\mathbf{X}}^{(r)}; \mathbf{p}^k)$$

where each of $\mathbf{X}^{(r)}, \tilde{\mathbf{X}}^{(r)}$ are M independent copies of the input process generated under $\mathbf{p}^k$.

2. Estimate $\hat{\boldsymbol{\Psi}}^k = (\hat{\Psi}^k_1, \ldots, \hat{\Psi}^k_m)$, the gradient of $E_{\mathbf{p}}[g(\mathbf{X})]$, with

$$\hat{\Psi}^k_i = \frac{1}{M}\sum_{r=1}^{M} g(\tilde{\tilde{\mathbf{X}}}^{(r)}) S_i(\tilde{\tilde{\mathbf{X}}}^{(r)}; \mathbf{p}^k)$$

where $\tilde{\tilde{\mathbf{X}}}^{(r)}$ are another M independent copies of the input process generated under $\mathbf{p}^k$.

3. Estimate $\hat{\boldsymbol{\phi}}^k_{\mathbf{s}} = (\hat{\phi}^k_{\mathbf{s},1}, \ldots, \hat{\phi}^k_{\mathbf{s},n})$, the gradient of the penalty term with respect to $\mathbf{s}$, with

$$\hat{\phi}^k_{\mathbf{s},j} = -\frac{1}{M}\Big(\sum_{r=1}^{M}(I(h(\mathbf{X}^{(r)}) \leq y_j) - s^k_j) + \sum_{r=1}^{M}(I(h(\tilde{\mathbf{X}}^{(r)}) \leq y_j) - s^k_j)\Big)$$

where $\mathbf{X}^{(r)}, \tilde{\mathbf{X}}^{(r)}$ are the same replications used in Step 1.

4. Compute $\mathbf{p}^{k+1} = (p^{k+1}_1, \ldots, p^{k+1}_m)$ by running Algorithm 1 with $\xi_i = \bar{\gamma}(\lambda \hat{\Psi}^k_i + \hat{\phi}^k_{\mathbf{p},i})$ and compute $\mathbf{s}^{k+1} = (s^{k+1}_1, \ldots, s^{k+1}_n)$ by

$$s^{k+1}_j = \Pi_j(s^k_j - \bar{\gamma}\hat{\phi}^k_{\mathbf{s},j})$$

Output: $\mathbf{p}^\tau, \mathbf{s}^\tau$

where $I^m_T = \{(i_1, \ldots, i_m) : 1 \leq i_j \leq m, i_j \neq i_k \text{ if } j \neq k\}$. For convenience we denote $P^T f = E_P[f]$, where $E_P[\cdot]$ is the expectation with respect to the T-fold product measure of P.

We say that a central theorem holds for $\{\sqrt{m}(U^m_T f - P^T f)\}_{f\in\mathcal{F}}$ if

$$\{\sqrt{m}(U^m_T f - P^T f)\}_{f\in\mathcal{F}} \Rightarrow \{\mathbb{G}(f)\}_{f\in\mathcal{F}} \text{ in } \ell^\infty(\mathcal{F}) \tag{EC.39}$$

Algorithm 5 Two-phase RSPG for solving (15)

Input: A small parameter $\epsilon > 0$, initial solution $\mathbf{p}^1 \in \mathcal{P}(\epsilon) = \{\mathbf{p} : \sum_{i=1}^m p_i = 1, p_i \geq \epsilon \text{ for } i = 1, \ldots, m\}$ and $\mathbf{s}^1 \in [\hat{F}_Y(y_1+) - \frac{q_{1-\alpha}}{\sqrt{n}}, \hat{F}_Y(y_1-) + \frac{q_{1-\alpha}}{\sqrt{n}}] \times \cdots \times [\hat{F}_Y(y_n+) - \frac{q_{1-\alpha}}{\sqrt{n}}, \hat{F}_Y(y_n+) - \frac{q_{1-\alpha}}{\sqrt{n}}]$, step size $\bar{\gamma}$ for both $\mathbf{p}$ and $\mathbf{s}$, penalty λ, batch size M, number of RSPG runs S, and number of iterations N per run. Batch size M' in the post-optimization phase.

1. Optimization phase: For $s = 1, \ldots, S$, run Algorithm 4 with initial point $\mathbf{p}^1, \mathbf{s}^1$, step size $\bar{\gamma}$, penalty λ, batch size M, and number of iterations N. Let $\mathbf{p}_s, \mathbf{s}_s$ be the output of the s-th run of Algorithm 4.

2. Post-optimization phase: For $s = 1, \ldots, S$, run one iteration of Step 1,2,3,4 of Algorithm 4 but with batch size M' at $\mathbf{p}_s, \mathbf{s}_s$. Let $\mathbf{p}'_s, \mathbf{s}'_s$ be the output from Step 4 at $\mathbf{p}_s, \mathbf{s}_s$. Then compute

$$(g_{\mathbf{p}}(\mathbf{p}_s, \mathbf{s}_s), g_{\mathbf{s}}(\mathbf{p}_s, \mathbf{s}_s)) = (\frac{1}{\bar{\gamma}}(\mathbf{p}'_s - \mathbf{p}_s), \frac{1}{\bar{\gamma}}(\mathbf{s}'_s - \mathbf{s}_s))$$

Output: the $\mathbf{p}_{s^*}, \mathbf{s}_{s^*}$ where $s^* = \arg\min_s \{\|g_{\mathbf{p}}(\mathbf{p}_s, \mathbf{s}_s)\|_1^2 + \|g_{\mathbf{s}}(\mathbf{p}_s, \mathbf{s}_s)\|_2^2\}$

where $\ell^\infty(\mathcal{F})$ is the space (for functionals on $\mathcal{F}$) defined by

$$\ell^\infty(\mathcal{F}) = \left\{ y : \mathcal{F} \to \mathbb{R} : \sup_{f \in \mathcal{F}} |y(f)| < \infty \right\}$$

$\mathbb{G}(f)$ is a Gaussian process indexed by $\mathcal{F}$ that is centered and has covariance function

$$Cov(\mathbb{G}(f_1), \mathbb{G}(f_2)) = Cov(TP^{T-1}S_T f_1, TP^{T-1}S_T f_2)$$

where P^{T-1} is defined by $P^{T-1}f(x) = \int \cdots \int f(x_1, \ldots, x_{T-1}, x) \prod_{t=1}^{T-1} dP(x_t)$ and

$$S_T f(x_1, \ldots, x_T) = \frac{1}{T!} \sum f(x_{i_1}, \ldots, x_{i_T})$$

where the sum is taken over all permutations $(x_{i_1}, \ldots, x_{i_T})$ of $(x_1, \ldots, x_T)$. Moreover, the process $\mathbb{G}(\cdot)$ is sample continuous with respect to the canonical semi-metric

$$\tau_{P,T}^2(f_1, f_2) = Var(P^{T-1}S_T(f_1 - f_2))$$

where $Var(\cdot)$ is taken with respect to the probability P. These discussions follow from Arcones and Gine (1993). We have ignored some measurability issues; see Van Der Vaart and Wellner (1996) for more details.

We have the following theorem:

THEOREM EC.3 **(Theorem 4.10 in Arcones and Gine (1993))**. *Let $\mathcal{F}$ be a class of functions on $\mathcal{X}^T \to \mathbb{R}$. If*

$$\int_0^1 \sqrt{\log N_{[]}(\epsilon, \mathcal{F}, \|\cdot\|_{P^T,2})} d\epsilon < \infty$$

where $\|\cdot\|_{P^T,2}$ is the norm induced in the L_2-space under P^T, the T-fold product measure of P. Then the central limit theorem holds for $\{\sqrt{m}(U_T^m f - P^T f)\}_{f\in\mathcal{F}}$ in the sense of (EC.39).

From Theorem EC.3, it is immediate that for a class of functions on $\mathcal{X}^T \to \mathbb{R}$, a bracketing number that is polynomial in ϵ implies the central limit theorem (EC.39).

EC.6.2. Results Needed in the Convergence Proofs of the MDSA

THEOREM EC.4 **(Corollary in Section 3 in Blum (1954))**. *Let Y_k be a sequence of integrable random variables that satisfy*

$$\sum_{k=1}^{\infty} E[E[Y_{k+1} - Y_k | Y_1, \ldots, Y_k]^+] < \infty$$

where $x^+ = x$ if $x > 0$ and 0 otherwise, and are bounded below uniformly in k. Then Y_k converges a.s. to a random variable.

LEMMA EC.3 **(Adapted from Lemma 2.1 in Nemirovski et al. (2009))**. *Let V be the KL divergence defined in* (19). *For every $\mathbf{q} \in \mathcal{P}$, $\mathbf{p} \in \mathcal{P}^\circ$, and $\boldsymbol{\xi} \in \mathbb{R}^m$, one has*

$$V(\tilde{\mathbf{p}}, \mathbf{q}) \le V(\mathbf{p}, \mathbf{q}) + \boldsymbol{\xi}'(\mathbf{q} - \mathbf{p}) + \frac{\|\boldsymbol{\xi}\|_\infty^2}{2}$$

where $\tilde{\mathbf{p}} = \arg\min_{\mathbf{u}\in\mathcal{P}} \boldsymbol{\xi}'(\mathbf{u} - \mathbf{p}) + V(\mathbf{p}, \mathbf{u})$, and $\|\cdot\|_\infty$ is the sup norm.