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

Input-Output Uncertainty Comparisons for Discrete Optimization via Simulation

Eunhye Song
Department of Industrial and Manufacturing Engineering, The Pennsylvania State University, University Park, PA 16802, eus358@psu.edu

Barry L. Nelson
Department of Industrial Engineering and Management Sciences, Northwestern University, Evanston, IL 60208, nelsonb@northwestern.edu

When input distributions to a simulation model are estimated from real-world data, they naturally have estimation error causing *input uncertainty* in the simulation output. If an optimization via simulation (OvS) method is applied that treats the input distributions as "correct," then there is a risk of making a suboptimal decision for the real world, which we call *input model risk*. This paper addresses a discrete OvS (DOvS) problem of selecting the real-world optimal from among a finite number of systems when all of them share the same input distributions estimated from common input data. Since input uncertainty cannot be reduced without collecting additional real-world data—which may be expensive or impossible—a DOvS procedure should reflect the limited resolution provided by the simulation model in distinguishing the real-world optimal solution from the others. In light of this, our input-output uncertainty comparisons (IOU-C) procedure focuses on *comparisons rather than selection:* it provides simultaneous confidence intervals for the difference between each system's real-world mean and the best mean of the rest with any desired probability, while accounting for both stochastic and input uncertainty. To make the resolution as high as possible (intervals as short as possible) we exploit the common input data effect to reduce uncertainty in the estimated differences. Under mild conditions we prove that the IOU-C procedure provides the desired statistical guarantee asymptotically as the real-world sample size and simulation effort increase, but it is designed to be effective in finite samples.

Key words: Optimization via simulation under input uncertainty, common input data effect, multiple comparisons with the best

History: First submitted on June 2016; revisions submitted on July 2017 and May 2018.

1. Introduction

Due to the flexibility of simulation, optimization via simulation (OvS) is a widely accepted tool to improve system performance. Real-world problems typically involve stochastic processes, e.g., demand for a new product or arrivals of patients to an emergency room, which are often modeled by probability distributions. Stochastic simulation is driven by random variates generated from these input models to produce outputs that mimic real-world performance. Therefore, when we make decisions based on the simulation outputs, we are subject to the risk of making suboptimal decisions when the input models do not faithfully represent the real-world stochastic processes; this is known as *input model risk*. Most standard OvS methods do not take into account input model risk and instead optimize under the assumption that the input models are accurate representations of the real-world randomness. However, the best system chosen *conditional* on the input models may not be the best system with respect to real-world performance when implemented. We refine this point below and illustrate it further using an inventory management example with estimated input demand distribution in Section 2. Of course, there may also be a logical discrepancy between the simulation model and the real-world system but that is beyond the scope of this paper.

The problem of interest is to compare k systems, where the ith system's performance measure is its simulation output mean, $\mathrm{E}[Y_i(F_i^c)]$, under real-world input distribution F_i^c (c for correct), where $Y_i(\cdot)$ is the stochastic output performance which depends on the chosen input distribution. When there are many input processes in the system, F_i^c represents the joint distribution of all of the input random variables. Our specific goal is to find $\arg\max_i \mathrm{E}[Y_i(F_i^c)]$ (or $\arg\min_i \mathrm{E}[Y_i(F_i^c)]$) with a statistical guarantee (e.g., 95%) that the selected system is the real-world optimal. As mentioned earlier, in most cases $F_1^c, F_2^c, \ldots, F_k^c$ are unknown, which forces us to use estimates, $\widehat{F}_1, \widehat{F}_2, \ldots, \widehat{F}_k$, to run simulations and implicitly target $\mathrm{E}[Y_i(\widehat{F}_i)|\widehat{F}_i]$ instead of $\mathrm{E}[Y(F_i^c)]$ to evaluate the ith system's

performance. Typically, $\widehat{F}_i$ is estimated from finite real-world observations from F_i^c and therefore is subject to estimation error. Input model risk arises as $\mathrm{E}[Y_i(\widehat{F}_i)|\widehat{F}_i]$ depends on random $\widehat{F}_i$, and thus the conditional optimal, $\arg\max_i \mathrm{E}[Y_i(\widehat{F}_i)|\widehat{F}_i]$, may not be the same as $\arg\max_i \mathrm{E}[Y_i(F_i^c)]$. *In this paper we show that it is possible to provide a meaningful statistical guarantee with respect to the real-world optimal, rather than the conditional optimal.*

To accomplish this we first need to understand how much uncertainty in $\mathrm{E}[Y_i(\widehat{F}_i)|\widehat{F}_i]$ is caused by the estimation error in $\widehat{F}_i$. This is referred to *input uncertainty* and formally defined as $\mathrm{Var}(\mathrm{E}[Y_i(\widehat{F}_i)|\widehat{F}_i])$, where the variance is taken with respect to the sampling distribution of $\widehat{F}_i$. Typically, we have only one "observation" of $\widehat{F}_i$ estimated from the real-world data, which makes it difficult to evaluate the variance. Another challenge is that the functional form of $\mathrm{E}[Y_i(\widehat{F}_i)|\widehat{F}_i]$ is generally unknown and can only be estimated via simulations. Several methods have been developed to quantify the *marginal* impact of input uncertainty on a single simulated system; see Barton (2012), Song et al. (2014), and Lam (2016) for surveys.

Unlike simulation stochastic error, which can be reduced by increasing the number of simulation replications, input uncertainty can only be reduced by collecting more real-world data. However, real-world data collection is typically much more expensive than simulation replications, or it may be impossible if an implementation decision has to be made before having another chance to collect data (e.g., logistics decisions for a natural disaster). *Our DOvS procedure is designed to provide statistical inference on the real-world optimal solution in the presence of input model risk that will not be further reduced by collecting more real-world data.*

Optimization under input model risk is more challenging than conditional DOvS since even with an infinite number of simulation replications we may not be able to distinguish the real-world best from the others due to the remaining input uncertainty. But *effective* DOvS under input model risk requires more than just quantifying the *marginal* input uncertainty in each system's simulation output; instead we need to compare how systems are affected *jointly* by input uncertainty.

Recently, several DOvS procedures that incorporate input model risk have been proposed; they can be categorized into three groups in terms of what they promise to deliver: the first group

of procedures selects a system that best hedges input model risk by identifying the worst-case input distributions given real-world data for each system marginally, and then selects the system with the *best worst-case performance*. For a maximization problem this beomes selecting $\arg\max_i \min_{\widehat{F}_i \in \mathcal{U}_i} \mathrm{E}[Y_i(\widehat{F}_i)|\widehat{F}_i]$, where $\mathcal{U}_i$ is the uncertainty set that contains the candidates for F_i^c inferred from the real-world data. Such a formulation is used in the distributionally robust optimization literature (Scarf 1958, Delage and Ye 2010, Ben-Tal et al. 2013). The robust selection of the best procedure of Fan et al. (2013) and the optimal computational budget allocation scheme of Gao et al. (2017) belong in this category. A benefit of this formulation is that we can always select a single solution no matter how large input uncertainty is. However, the selected system may, and often will, perform poorly under the true real-world input distributions. See Section 2.

The second category selects a system with the best performance *averaged over input uncertainty*, i.e., $\arg\max_i \mathrm{E}\left(\mathrm{E}[Y_i(\widehat{F}_i)|\widehat{F}_i]\right)$, where the outer expectation is taken with respect to the sampling or posterior distribution of $\widehat{F}_i$. Corlu and Biller (2015) propose a subset selection procedure that averages both stochastic and input uncertainties to find a subset of optimal/near-optimal systems where $\widehat{F}_i$ is a Bayesian posterior distribution given real-world data. Even if the input uncertainty, $\mathrm{Var}(\mathrm{E}[Y_i(\widehat{F}_i)|\widehat{F}_i])$, is large the variance of an estimate of $\mathrm{E}\left(\mathrm{E}[Y_i(\widehat{F}_i)|\widehat{F}_i]\right)$ may be reduced by more simulation replications. Hence, with a sufficiently large simulation budget the size of the subset may be as small as one provided that $\mathrm{E}\left(\mathrm{E}[Y_i(\widehat{F}_i)|\widehat{F}_i]\right)$ is distinct for each i. However, $\mathrm{E}\left(\mathrm{E}[Y_i(\widehat{F}_i)|\widehat{F}_i]\right) \neq \mathrm{E}[Y_i(F_i^c)]$ in general, and therefore $\arg\max_i \mathrm{E}\left(\mathrm{E}[Y_i(\widehat{F}_i)|\widehat{F}_i]\right)$ may not be $\arg\max_i \mathrm{E}[Y_i(F_i^c)]$. The bias of $\mathrm{E}\left(\mathrm{E}[Y_i(\widehat{F}_i)|\widehat{F}_i]\right)$ is larger when the number of real-world observations is smaller, causing this fomulation to pose greater input model risk.

The last category of procedures directly attacks the problem of finding $\arg\max_i \mathrm{E}[Y_i(F_i^c)]$. Corlu and Biller (2013) present a subset selection procedure that includes the real-world best system in the subset assuming that $\max_i \mathrm{E}[Y_i(F_i^c)]$ is at least $\delta > 0$ better than the rest of the systems' true means. This procedure is distinguished from the subset selection procedure in Corlu and Biller (2015) in that it does not average $\mathrm{E}[Y_i(\widehat{F}_i)|\widehat{F}_i]$ over the distribution of $\widehat{F}_i$, but uses δ to

control the resolution to which the procedure can successfully separate the real-world best from the rest with a given statistical guarantee. Under the same indifference-zone (IZ) setting, Song et al. (2015) discuss a ranking-and-selection approach that guarantees the probability of correctly selecting $\arg\max_i \mathrm{E}[Y_i(F_i^c)]$ in the presence of input model risk. Both Corlu and Biller (2013) and Song et al. (2015) find that δ has an unknown nonzero lower bound, which is an increasing function of input uncertainty reflecting the fact that the procedures may not distinguish the real-world best system from the rest if the mean difference is too small relative to input uncertainty. To put it differently, for δ below an unknown threshold the probability of correctly selecting the optimal (or including the optimal in the subset) has an upper bound less than 1 so that even with infinite simulation effort we may not achieve the desired statistical guarantee. Further, assuming an IZ mean configuration makes both procedures conservative, because they are designed to provide the statistical guarantee for the case where all suboptimal systems' means are $\arg\max_i \mathrm{E}[Y_i(F_i^c)] - \delta$. When $F_1^c, F_2^c, \ldots, F_k^c$ are assumed known, this only makes us spend more simulation budget than necessary to correctly select the optimal solution with the target probability. In the presence of input model risk, however, the problem is much more severe and we may conclude that we cannot provide the target probability guarantee at all when in fact we could if we did not assume an IZ configuration.

Our input-output uncertainty comparisons (IOU-C) procedure belongs in the third category. *However, we focus on comparisons of systems, not selection, and we do not assume any configuration for the system means*, which differentiates our approach from Corlu and Biller (2013) and Song et al. (2015). By extending the multiple comparisons with the best (MCB) framework of Chang and Hsu (1992) to incorporate input model risk, IOU-C provides k joint confidence intervals (CIs) on the true mean differences between each system and the best of the rest that account for both stochastic and input uncertainties. With any given target probability guarantee, the CIs that contain 0 indicate systems that are statistically inseparable from the real-world optimal.

We restrict our attention to the case where all systems share the same input distributions, i.e., $F_i^c = F^c$ and $\widehat{F}_i = \widehat{F}$ for $i = 1, 2, \ldots, k$, which is a common setting for DOvS problems. For instance,

we may compare k scheduling rules for an emergency department given the same patient arrival process. In this case, the estimation error of $\widehat{F}$ from the common real-world data affects all k systems' simulation outputs. We call this the *common-input-data (CID) effect. One of the novel contributions of this paper is to model and estimate the joint distribution of CID effects to devise an efficient comparisons procedure that exploits it.* The IOU-C procedure has the following strengths:

1. Since our focus is not on selection of a single best system, we do not need to assume mean configurations *a priori* to provide the desired statistical guarantee. Naturally, if the real-world system means are well-separated then the comparisons become easy and the resulting subset may include only one system, which is the real-world optimal; we do not need to settle for a system that best hedges input model risk nor the system with the best performance averaged over both stochastic and input uncertainties.
2. MCB provides parsimonious comparisons: it is more efficient than all pairwise comparisons of k systems, and $k-1$ comparisons with a control is not sufficient as we do not know the identity of the real-world optimal system. Moreover, the MCB CIs provide a confidence bound on how far each system's performance could be from the best of the rest. This is useful when there is a secondary criterion to consider; if the best system's main performance measure is marginally better than the rest, but the secondary performance measure is much worse than the next best, then we may choose the next best. This applies to MCB in general, not only to IOU-C.
3. Narrow MCB CI widths make the size of the subset of systems that are indistinguishable from the best small. When there is no input uncertainty, the MCB CI widths can be reduced by simply increasing the simulation effort. However, input uncertainty makes the CI widths nonzero even with infinite simulation effort. Our biggest contribution in this paper is to make the comparisons as sharp as possible given the limited real-world input data, and thereby to provide a small subset even in the presence of input uncertainty, by exploiting the CID effects and common random numbers.

4. A large subset size may indicate either 1) the systems' performance measures are not too different so any system in the subset could be selected (narrow MCB CIs) or 2) input model risk is overwhelming so that it is difficult to separate the optimal system from the rest (wide MCB CIs). In the former case, we can apply a procedure that selects a defensive best with respect to input model risk among the remaining systems in the subset, which should be much less conservative than choosing a defensive best from all k systems. In the latter case, it may be appropriate to postpone the decision until additional real-world data are available (if possible) or approach the problem differently since the defensive choice is likely to be very conservative.

The remainder of the paper is organized as follows. In Section 2, we present a simple DOvS example to illustrate the difficulties that arise when there is input model risk and highlight the key factors for designing sharp comparisons. In Section 3, we introduce the general framework for IOU-C procedures. In Section 4, we show how to account for the joint effect of input model risk on all k systems' outputs. We revisit the IOU-C procedure in Section 5 to provide computation details for each step. In Section 6, we show under mild conditions that the IOU-C procedure provides the desired probability guarantee asymptotically as the real-world sample size and simulation effort increase. Performance of the procedure is tested in Section 7.

2. Illustration

We use a modified version of the (s, S) inventory problem from Koenig and Law (1985) to provide insights on DOvS under input model risk. Suppose we have $k = 4$ candidate (s, S) inventory policies, $(s, S) = (20, 50), (20, 55), (20, 60)$ and $(20, 65)$, where each solution is evaluated based on the expected cost of operation over a 30-day period. Unknown to us, the true daily demand is i.i.d. Poisson with mean $\lambda^c = 26$. Figure 1 shows the expected costs of all four policies. Under λ^c, $(20, 55)$ is the optimal policy that minimizes the real-world expected cost. Since in reality λ^c is unknown, suppose we estimate it from real-world observations of the demand. Figure 1 illustrates how the expected costs of the four policies are affected by the CID effects with four particular cases

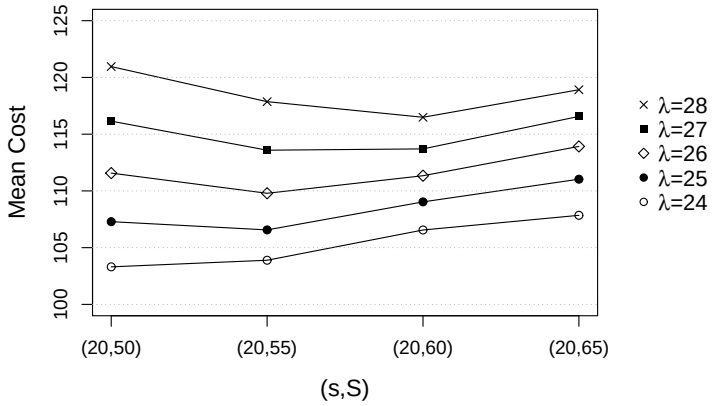

Figure 1 Mean cost of inventory policies for $\lambda^c = 26$ and $\widehat{\lambda} = 24, 25, 27$ and 28.

of estimated mean demand $\widehat{\lambda} = 24, 25, 27$ and 28. When $\widehat{\lambda} = 25$ or 27, the true optimal solution, $(20, 55)$, still minimizes the cost. However, $(20, 50)$ is optimal under $\widehat{\lambda} = 24$, whereas $(20, 60)$ is optimal under $\widehat{\lambda} = 28$. Note that $(20, 60)$ is the defensive best solution given $\mathcal{U} = \{24, 25, 26, 27, 28\}$. Meanwhile, notice that $(20, 65)$ does not minimize the expected cost at any value of $\widehat{\lambda}$, which means that $(20, 65)$ is ruled out as the true optimal solution even if λ^c is unknown.

In a realistic DOvS problem we do not obtain multiple values of $\widehat{\lambda}$; we have only one value of $\widehat{\lambda}$ estimated from the real-world data. If all solutions are affected exactly the same way by input uncertainty, then the cost plot for any $\widehat{\lambda}$ would be parallel to that for λ^c and for any value of $\widehat{\lambda}$ the solution $(20, 55)$ would minimize the expected cost. However, as depicted in Figure 1, the solutions can be affected differently when $\widehat{\lambda}$ varies, especially when $\widehat{\lambda}$ is far from λ^c, which makes the true optimal solution no longer minimize the expected cost given $\widehat{\lambda}$.

The insights obtained from this illustration are three-fold. First, if we use a procedure that assumes all input models are correct, then we may select a suboptimal solution as the best and falsely provide a much higher statistical guarantee than what is actually attained. In a realistic DOvS setting, each system's performance measure is estimated via simulation replications, which introduces stochastic error. This example shows that even if we spend infinite simulation effort to eliminate stochastic errors, input uncertainty may cause us to select a suboptimal solution as the best if we select a system conditional on the estimated $\widehat{\lambda}$.

Secondly, even in the presence of input uncertainty, we may be able to provide the same level of statistical guarantee as the DOvS procedure without input uncertainty (perhaps with increased simulation effort) if the CID effects are similar across systems. Therefore, it is important to estimate the *joint* distribution of the CID effects to make sharp comparisons.

Finally, some solutions are so inferior that we can rule them out even in the presence of input model risk. This corroborates the use of a CI procedure to identify a subset of near-optimal solutions even if we are interested in selecting a defensive system with respect to input model risk.

Our IOU-C procedure provides a set of solutions that are statistically inseparable from the real-world optimal solution where the size of the set depends on the MCB CI widths. Hence, if the systems are affected similarly by input uncertainty, the procedure should take advantage of it to provide CIs as narrow as possible. In the next section, we introduce the basic framework of IOU-C followed by a model to capture the joint effects of input uncertainty on systems' outputs in Section 4.

3. Framework for IOU Comparisons

In this section, we provide a high-level framework for IOU-C procedures by extending MCB to account for input model risk. Without loss of generality, we concentrate on a maximization problem in the remainder of the paper.

As mentioned in Section 1, when the simulation is run using estimated distribution $\widehat{F}$, the conditional mean of the output, $\mathrm{E}[Y_i(\widehat{F})|\widehat{F}]$, is a functional of $\widehat{F}$. To simplify the notation, we define $\eta_i(\widehat{F}) = \mathrm{E}[Y_i(\widehat{F})|\widehat{F}]$, so $\eta_i(F^c) = \mathrm{E}[Y_i(F^c)|F^c]$. Thus, $Y_i(\widehat{F})$ can be represented as

$$Y_i(\widehat{F}) = \mathrm{E}[Y_i(\widehat{F})|\widehat{F}] + \varepsilon_i(\widehat{F}) = \eta_i(F^c) + b_i(\widehat{F}, F^c) + \varepsilon_i(\widehat{F}), \tag{1}$$

where $b_i(\widehat{F}, F^c) \equiv \eta_i(\widehat{F}) - \eta_i(F^c)$ and $\varepsilon_i(\widehat{F})$ has mean 0 and finite variance $\sigma_i^2(\widehat{F})$. We do not require normality of the simulation output. Notice that $b_i(\widehat{F}, F^c)$ captures the *common-input-data (CID) effect* on system i. If we knew that $b_i(\widehat{F}, F^c) = b(\widehat{F}, F^c), \forall i$, then we could simply ignore input model risk as all systems are affected exactly the same.

Much as common random numbers (CRN) make stochastic errors, $\varepsilon_1(\widehat{F}), \varepsilon_2(\widehat{F}), \ldots, \varepsilon_k(\widehat{F})$, correlated, the CID effects cause $\eta_1(\widehat{F}), \eta_2(\widehat{F}), \ldots, \eta_k(\widehat{F})$, to be correlated. However, CRN and CID are different in nature. CRN are employed across different systems as a variance reduction technique; we typically assume $\text{Corr}(\varepsilon_i(\widehat{F}), \varepsilon_\ell(\widehat{F})) \equiv \rho_{i,\ell}(\widehat{F}) > 0$, which sharpens the comparison of systems i and ℓ by reducing the variance of the difference in simulation outputs, $Y_i(\widehat{F}) - Y_\ell(\widehat{F})$. Hence, if in fact CRN causes $\rho_{i,\ell}(\widehat{F}) < 0$, then we can choose not to employ CRN and run independent simulations. However, the CID effect is a property of the problem itself; we compare different system designs/policies under the same real-world stochastic processes. Therefore, even if the CID effect causes negative correlation between $b_i(\widehat{F}, F^c)$ and $b_\ell(\widehat{F}, F^c)$, we cannot eliminate such correlation. Our challenge is in accounting for the CID effect and exploiting it when it is favorable.

The following theorem by Chang and Hsu (1992) lets us obtain MCB CIs when F^c is known.

THEOREM 1. [Chang and Hsu 1992] *Let $\widehat{\eta}_i(F^c)$ be an unbiased estimator of $\eta_i(F^c)$ for $i = 1, 2, \ldots, k$, $x^+ = \max(0, x)$ and $x^- = \max(0, -x)$. If for each i individually*

$$\Pr\{\widehat{\eta}_i(F^c) - \widehat{\eta}_\ell(F^c) - (\eta_i(F^c) - \eta_\ell(F^c)) \geq -w_{i\ell}, \text{ for all } \ell \neq i\} \geq 1 - \alpha, \tag{2}$$

then we can make the joint probability statement

$$\Pr\left\{\eta_i(F^c) - \max_{\ell \neq i} \eta_\ell(F^c) \in [D_i^-, D_i^+], \text{ for all } i\right\} \geq 1 - \alpha,$$

where $D_i^+ = \left(\min_{\ell \neq i}[\widehat{\eta}_i(F^c) - \widehat{\eta}_\ell(F^c) + w_{i\ell}]\right)^+, \mathcal{I} = \{i : D_i^+ > 0\}$, and

$$D_i^- = \begin{cases} 0, & \text{if } \mathcal{I} = \{i\} \\ -\left(\min_{\ell \in \mathcal{I}, \ell \neq i}[\widehat{\eta}_i(F^c) - \widehat{\eta}_\ell(F^c) - w_{\ell i}]\right)^-, & \text{otherwise.} \end{cases}$$

Theorem 1 states that MCB CIs can be constructed from multiple comparisons with a fixed control system— that is, Equation (2)— by treating each system i as a control. Note that D_i^+ is positive only if for all $\ell \neq i$ the upper CI bound of $\eta_i(F^c) - \eta_\ell(F^c)$ from (2) is positive. We can conclude that the systems with $D_i^+ = 0$ are inferior to the best with probability $\geq 1 - \alpha$. If there is only one system with $D_i^+ > 0$, then we can conclude that system is the best with probability $\geq 1 - \alpha$.

Otherwise, all systems i with $D_i^+ > 0$ form a subset of the possible best and the value of D_i^+ is how much better than the best each one might be.

When F^c is known, $\widehat{\eta}_i(F^c)$ is simply $\bar{Y}_i(F^c)$, where the bar indicates a sample average from n replications. Then the interval widths, $w_{i\ell}$, depend only on the joint distribution of $\{\bar{\varepsilon}_i(F^c)\}_{i=1}^k$. Hence, as n increases, $w_{i\ell}$ decreases. In the presence of estimated input distributions, $w_{i\ell}$ should depend on both stochastic and input uncertainty; the interval widths are larger and the increase depends on how differently system i and ℓ are affected by input uncertainty. Clearly, more systems are likely to have $D_i^+ > 0$ when the $w_{i\ell}$ are larger making it more difficult to determine the inferior systems. Therefore, we desire to make $w_{i\ell}$ as small as possible given input uncertainty while still preserving the statistical guarantee. Estimating the distribution of CID effects helps greatly in this regard as opposed to using a conservative probability inequality such as the Bonferroni inequality. Even when the interval widths are large, if $\eta_i(F^c)$ is much smaller than some $\eta_\ell(F^c)$, then we may still have $D_i^+ = 0$. Thus, the difficulty of the comparisons depends on the true system means, $\eta_i(F^c), i = 1, 2, \ldots, k$, as well as input uncertainty. This echoes the inventory example in Section 1: when input uncertainty is small, i.e., $\widehat{\lambda}$ is closer to λ^c, the real-world best solution is still optimal at $\widehat{\lambda} \neq \lambda^c$; on the other hand, if a solution like $(20, 65)$ is inferior by enough, it remains suboptimal for any value of $\widehat{\lambda}$.

From (1), $\bar{Y}_i(\widehat{F}) - \bar{Y}_\ell(\widehat{F}) - (\eta_i(F^c) - \eta_\ell(F^c)) = b_i(\widehat{F}, F^c) - b_\ell(\widehat{F}, F^c) + \bar{\varepsilon}_i(\widehat{F}) - \bar{\varepsilon}_\ell(\widehat{F})$. Therefore, if $\bar{Y}_i(\widehat{F})$ is used as $\widehat{\eta}_i(F^c)$, then the left-hand side of (2) can be rewritten as

$$\begin{aligned}
&\Pr\left\{\bar{Y}_i(\widehat{F}) - \bar{Y}_\ell(\widehat{F}) - (\eta_i(F^c) - \eta_\ell(F^c)) \geq -w_{i\ell}, \forall \ell \neq i\right\} \\
&= \Pr\left\{b_i(\widehat{F}, F^c) - b_\ell(\widehat{F}, F^c) + \bar{\varepsilon}_i(\widehat{F}) - \bar{\varepsilon}_\ell(\widehat{F}) \geq -w_{i\ell}, \forall \ell \neq i\right\}.
\end{aligned}$$

If $w_{i\ell} = w_{i\ell}^{(1)} + w_{i\ell}^{(2)}$ for some $w_{i\ell}^{(1)}, w_{i\ell}^{(2)} > 0$, then

$$\begin{aligned}
&\Pr\left\{b_i(\widehat{F}, F^c) - b_\ell(\widehat{F}, F^c) + \bar{\varepsilon}_i(\widehat{F}) - \bar{\varepsilon}_\ell(\widehat{F}) \geq -(w_{i\ell}^{(1)} + w_{i\ell}^{(2)}), \forall \ell \neq i\right\} \\
&\geq \Pr\left\{b_i(\widehat{F}, F^c) - b_\ell(\widehat{F}, F^c) \geq -w_{i\ell}^{(1)}, \bar{\varepsilon}_i(\widehat{F}) - \bar{\varepsilon}_\ell(\widehat{F}) \geq -w_{i\ell}^{(2)}, \forall \ell \neq i\right\} \\
&= \mathrm{E}\left[\Pr\left\{b_i(\widehat{F}, F^c) - b_\ell(\widehat{F}, F^c) \geq -w_{i\ell}^{(1)}, \bar{\varepsilon}_i(\widehat{F}) - \bar{\varepsilon}_\ell(\widehat{F}) \geq -w_{i\ell}^{(2)}, \forall \ell \neq i \,\middle|\, \widehat{F}\right\}\right] \\
&= \mathrm{E}\left[\mathbf{1}\left\{b_i(\widehat{F}, F^c) - b_\ell(\widehat{F}, F^c) \geq -w_{i\ell}^{(1)}, \forall \ell \neq i\right\} \Pr\left\{\bar{\varepsilon}_i(\widehat{F}) - \bar{\varepsilon}_\ell(\widehat{F}) \geq -w_{i\ell}^{(2)}, \forall \ell \neq i \,\middle|\, \widehat{F}\right\}\right], \qquad (3)
\end{aligned}$$

where $\mathbf{1}\{\cdot\}$ is the indicator function. Note that (3) holds since conditional on $\widehat{F}$, $b_i(\widehat{F},F^c)-b_\ell(\widehat{F},F^c)$ is constant. Given the conditional distribution of $\bar{\varepsilon}_i(\widehat{F})-\bar{\varepsilon}_\ell(\widehat{F})$, suppose for any $\widehat{F}$ and $0<\alpha_2<1/2$ we can find $w_{i\ell}^{(2)},\forall\ell\neq i$, such that

$$\Pr\left\{\bar{\varepsilon}_i(\widehat{F})-\bar{\varepsilon}_\ell(\widehat{F})\geq -w_{i\ell}^{(2)},\forall\ell\neq i\middle|\,\widehat{F}\right\}=1-\alpha_2. \tag{4}$$

Then term (3) becomes $\Pr\left\{b_i(\widehat{F},F^c)-b_\ell(\widehat{F},F^c)\geq -w_{i\ell}^{(1)},\forall\ell\neq i\right\}\cdot(1-\alpha_2)$. Therefore, by finding $w_{i\ell}^{(1)},i\neq\ell$, that satisfy $\Pr\{b_i(\widehat{F},F^c)-b_\ell(\widehat{F},F^c)\geq -w_{i\ell}^{(1)},\forall\ell\neq i\}=1-\alpha_1$ for $0<\alpha_1<1/2$, we have $\Pr\{\bar{Y}_i(\widehat{F})-\bar{Y}_\ell(\widehat{F})-(\eta_i(F^c)-\eta_\ell(F^c))\geq -(w_{i\ell}^{(1)}+w_{i\ell}^{(2)}),\forall\ell\neq i\}\geq(1-\alpha_1)(1-\alpha_2)$. Hence, the overall statistical guarantee of the IOU-C procedure is $(1-\alpha_1)(1-\alpha_2)$ from Theorem 1. The following is a general framework for IOU-C procedures:

IOU-C Procedure

1. Select $0<\alpha_1,\alpha_2<1/2$ such that $1-\alpha=(1-\alpha_1)(1-\alpha_2)$ for given $0<\alpha<1/2$.
2. Collect real-world observations from F^c and compute its estimator $\widehat{F}$.
3. For each system i, use $\widehat{F}$ as an input model and run n replications, $Y_{i1}(\widehat{F}),Y_{i2}(\widehat{F}),\ldots,Y_{in}(\widehat{F})$. Compute $\bar{Y}_i(\widehat{F})=\Sigma_{j=1}^n Y_{ij}(\widehat{F})/n$.
4. **(Interval widths due to CID effects)** For each system i, find $w_{i\ell}^{(1)}>0,\forall\ell\neq i$, that satisfy $\Pr\{b_i(\widehat{F},F^c)-b_\ell(\widehat{F},F^c)\geq -w_{i\ell}^{(1)},\forall\ell\neq i\}=1-\alpha_1$.
5. **(Interval widths due to stochastic error)** For each system i, find $w_{i\ell}^{(2)}>0,\forall\ell\neq i$, that satisfy $\Pr\{\bar{\varepsilon}_i(\widehat{F})-\bar{\varepsilon}_\ell(\widehat{F})\geq -w_{i\ell}^{(2)},\forall\ell\neq i|\widehat{F}\}=1-\alpha_2$.
6. For each system i, set $w_{i\ell}=w_{i\ell}^{(1)}+w_{i\ell}^{(2)},\forall\ell\neq i$. Use Theorem 1 to derive $1-\alpha$ simultaneous comparisons CIs.

Accounting for stochastic error in an MCB procedure has been well-studied (Hsu 1996). Estimating the distribution of $b_i(\widehat{F},F^c)$ marginally has been addressed in the input uncertainty quantification literature. However, to find interval widths $w_{i\ell}^{(1)},i\neq\ell$, we need to estimate the joint distribution of $b_i(\widehat{F},F^c),i=1,2,\ldots,k$ focusing on how differently the systems are affected by $\widehat{F}$; this is new.

In the following section we introduce a model to represent $\eta_i(\widehat{F})$ that enables us to estimate the unknown distributions of the stochastic errors and the CID effects. In Section 5, we employ it to obtain interval widths that fully account for the estimation errors.

4. Model of CID Effects

From this section on, we focus on the case when F^c has a known parametric distribution family, F, with an unknown parameter vector, $\boldsymbol{\theta}^c$. Non-parametric F^c or unknown parametric distribution family is relevant future work. Thus, $\widehat{F} = F(\cdot|\widehat{\boldsymbol{\theta}})$ with estimated parameter $\widehat{\boldsymbol{\theta}}$ based on real-world observations. Similarly, $Y_i(\widehat{F}) = Y_i(\widehat{\boldsymbol{\theta}}), b_i(\widehat{F}, F^c) = b_i(\widehat{\boldsymbol{\theta}}, \boldsymbol{\theta}^c)$, and $\varepsilon_i(\widehat{F}) = \varepsilon_i(\widehat{\boldsymbol{\theta}})$. Note that F may be a collection of distributions of all input processes. For instance, q independent real-world input processes can be represented by a set of parametric distributions $F = \{\mathcal{F}_1, \mathcal{F}_2, \ldots, \mathcal{F}_q\}$ with parameter vector $\boldsymbol{\theta} = \{\boldsymbol{\vartheta}_1, \boldsymbol{\vartheta}_2, \ldots, \boldsymbol{\vartheta}_q\}$. If we use p to denote the dimension of $\boldsymbol{\theta}^c$, then $p \geq q$. Suppose we collect $m_1, m_2, \ldots, m_q$ real-world observations from each of q input processes and compute the maximum likelihood estimator (MLE) $\widehat{\boldsymbol{\theta}} = \{\widehat{\boldsymbol{\vartheta}}_1, \widehat{\boldsymbol{\vartheta}}_2, \ldots, \widehat{\boldsymbol{\vartheta}}_q\}$ of F. We use $m = (m_1 + m_2 + \cdots + m_q)/q$ to represent the average number of observations from the q input processes.

Further suppose η_i is a smooth function of $\widehat{\boldsymbol{\theta}}$ that is continuously differentiable at $\boldsymbol{\theta}^c$. Then, using the first-order Taylor series approximation

$$\eta_i(\widehat{\boldsymbol{\theta}}) \approx \eta_i(\boldsymbol{\theta}^c) + \beta_i^\top(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^c), \tag{5}$$

where $\beta_i \equiv \nabla\eta_i(\boldsymbol{\theta}^c)$. If we assume (5) to be exact, then $b_i(\widehat{\boldsymbol{\theta}}, \boldsymbol{\theta}^c) = \beta_i^\top(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^c)$ and β_i represents how much $\eta_i(\widehat{\boldsymbol{\theta}})$ is affected by each element of $\widehat{\boldsymbol{\theta}}$. This model is also used in Cheng and Holland (1997, 1998) and Lin et al. (2015) to quantify marginal input uncertainty in simulation output. Under (5), the distribution of $b_i(\widehat{\boldsymbol{\theta}}, \boldsymbol{\theta}^c)$ is characterized by the sampling distribution of $\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^c$ and β_i. Typically, knowing the exact sampling distribution of $\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^c$ is difficult, but under some regularity conditions we can approximate it with the asymptotic distribution of $\widehat{\boldsymbol{\theta}}$ (Amemiya 1985):

$$\sqrt{m}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^c) \xrightarrow{D} N(\mathbf{0}, \Sigma(\boldsymbol{\theta}^c)) \text{ as } m \to \infty, \tag{6}$$

where $\Sigma(\boldsymbol{\theta}^c)$ is the asymptotic variance-covariance matrix of $\widehat{\boldsymbol{\theta}}$. When F^c is a set of $q>1$ input distributions, we can define the asymptotic distribution in (6) by assuming the ratios of the numbers of observations, $m_1/m, m_2/m, \ldots, m_q/m$, converge to positive constants as $m \to \infty$ (Cheng and Holland 1997). In this case, Σ becomes a function of the limiting ratios as well as $\boldsymbol{\theta}^c$.

Given $\{\beta_1, \beta_2, \ldots, \beta_k\}$, we can approximate the joint distribution of $\{\beta_i^\top(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^c)\}_{i=1}^k$ by the following multivariate normal distribution:

$$\begin{pmatrix} \beta_1^\top(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^c) \\ \beta_2^\top(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^c) \\ \vdots \\ \beta_k^\top(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^c) \end{pmatrix} \sim \mathrm{N}\left(\mathbf{0}, \frac{1}{m} \begin{bmatrix} \beta_1^\top\Sigma(\boldsymbol{\theta}^c)\beta_1 & \beta_1^\top\Sigma(\boldsymbol{\theta}^c)\beta_2 & \cdots & \beta_1^\top\Sigma(\boldsymbol{\theta}^c)\beta_k \\ \beta_2^\top\Sigma(\boldsymbol{\theta}^c)\beta_1 & \beta_2^\top\Sigma(\boldsymbol{\theta}^c)\beta_2 & \cdots & \beta_2^\top\Sigma(\boldsymbol{\theta}^c)\beta_k \\ \vdots & \vdots & \ddots & \vdots \\ \beta_k^\top\Sigma(\boldsymbol{\theta}^c)\beta_1 & \beta_k^\top\Sigma(\boldsymbol{\theta}^c)\beta_2 & \cdots & \beta_k^\top\Sigma(\boldsymbol{\theta}^c)\beta_k \end{bmatrix} \right). \tag{7}$$

This joint distribution is key since it includes the CID effects across all k systems. Under Model (5), $b_i(\widehat{\boldsymbol{\theta}}, \boldsymbol{\theta}^c) - b_\ell(\widehat{\boldsymbol{\theta}}, \boldsymbol{\theta}^c) = (\beta_i - \beta_\ell)^\top(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^c)$, which captures how differently two systems i and ℓ are affected by the common input model $F(\cdot|\widehat{\boldsymbol{\theta}})$. The joint distribution of $(\beta_i - \beta_\ell)^\top(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^c), \forall \ell \neq i$ can be derived from (7). For instance, if $i=1$, the joint distribution of $\left\{(\beta_1 - \beta_2)^\top(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^c), (\beta_1 - \beta_3)^\top(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^c), \ldots, (\beta_1 - \beta_k)^\top(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^c)\right\}$ is

$$\mathrm{N}\left(\mathbf{0}, \frac{1}{m} \begin{bmatrix} (\beta_1 - \beta_2)^\top\Sigma(\boldsymbol{\theta}^c)(\beta_1 - \beta_2) & (\beta_1 - \beta_2)^\top\Sigma(\boldsymbol{\theta}^c)(\beta_1 - \beta_3) & \cdots & (\beta_1 - \beta_2)^\top\Sigma(\boldsymbol{\theta}^c)(\beta_1 - \beta_k) \\ (\beta_1 - \beta_3)^\top\Sigma(\boldsymbol{\theta}^c)(\beta_1 - \beta_2) & (\beta_1 - \beta_3)^\top\Sigma(\boldsymbol{\theta}^c)(\beta_1 - \beta_3) & \cdots & (\beta_1 - \beta_3)^\top\Sigma(\boldsymbol{\theta}^c)(\beta_1 - \beta_k) \\ \vdots & \vdots & \ddots & \vdots \\ (\beta_1 - \beta_k)^\top\Sigma(\boldsymbol{\theta}^c)(\beta_1 - \beta_2) & (\beta_1 - \beta_k)^\top\Sigma(\boldsymbol{\theta}^c)(\beta_1 - \beta_3) & \cdots & (\beta_1 - \beta_k)^\top\Sigma(\boldsymbol{\theta}^c)(\beta_1 - \beta_k) \end{bmatrix} \right). \tag{8}$$

Clearly, Model (5) is an approximation as it drops terms nonlinear in $(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^c)$, and therefore does not fully capture the the finite-sample bias in $\bar{Y}_i(\widehat{\boldsymbol{\theta}})$ in general. If we assume that $\eta_i(\widehat{\theta})$ is twice differentiable in $\boldsymbol{\theta}^c$, then the remainder not captured by Model (5) is $O_p(||\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^c||^2) = O_p(m^{-1})$, which converges faster than input uncertainty as m increases and therefore does not affect our asymptotic argument for the IOU-C procedure in Section 6.

We also need the joint distribution of the stochastic errors across all k systems to form the CIs. To make the overall interval widths as narrow as possible, we adopt CRN and run the same number

of replications, n, for all k systems using $F(\cdot|\widehat{\boldsymbol{\theta}})$. Since IOU-C is based on MCB and MCB CIs are constructed from Equation (2), the variance-covariance matrix of $\{\varepsilon_i(\widehat{\boldsymbol{\theta}}) - \varepsilon_\ell(\widehat{\boldsymbol{\theta}})\}_{\ell=1,\ell\neq i}^{k}$, which we denote by $V_i(\widehat{\boldsymbol{\theta}})$, is of interest.

For the traditional MCB problem without input uncertainty, Nelson (1993) shows that we can obtain the exact coverage probability for MCB CIs under a normality assumption when the variance-covariance matrix of the stochastic errors has the following property known as sphericity:

$$V(\widehat{\boldsymbol{\theta}}) = \begin{bmatrix} 2\psi_1(\widehat{\boldsymbol{\theta}}) + \tau^2(\widehat{\boldsymbol{\theta}}) & \psi_1(\widehat{\boldsymbol{\theta}}) + \psi_2(\widehat{\boldsymbol{\theta}}) & \cdots & \psi_1(\widehat{\boldsymbol{\theta}}) + \psi_k(\widehat{\boldsymbol{\theta}}) \\ \psi_1(\widehat{\boldsymbol{\theta}}) + \psi_2(\widehat{\boldsymbol{\theta}}) & 2\psi_2(\widehat{\boldsymbol{\theta}}) + \tau^2(\widehat{\boldsymbol{\theta}}) & \cdots & \psi_2(\widehat{\boldsymbol{\theta}}) + \psi_k(\widehat{\boldsymbol{\theta}}) \\ \vdots & \vdots & \ddots & \vdots \\ \psi_1(\widehat{\boldsymbol{\theta}}) + \psi_k(\widehat{\boldsymbol{\theta}}) & \psi_2(\widehat{\boldsymbol{\theta}}) + \psi_k(\widehat{\boldsymbol{\theta}}) & \cdots & 2\psi_k(\widehat{\boldsymbol{\theta}}) + \tau^2(\widehat{\boldsymbol{\theta}}) \end{bmatrix}, \tag{9}$$

where $\tau^2(\widehat{\boldsymbol{\theta}}) > \sqrt{k\sum_{i=1}^{k}\psi_i(\widehat{\boldsymbol{\theta}})} - \sum_{i=1}^{k}\psi_i(\widehat{\boldsymbol{\theta}})$ to ensure that $V(\widehat{\boldsymbol{\theta}})$ is positive definite. Sphercity causes the structure of $V_i(\widehat{\boldsymbol{\theta}})$ to simplify so that all $k-1$ variance terms become $2\tau^2(\widehat{\boldsymbol{\theta}})$ and all pairwise covariances are $\tau^2(\widehat{\boldsymbol{\theta}})$, which reduces estimating the full $V(\widehat{\boldsymbol{\theta}})$ to estimating $\tau^2(\widehat{\boldsymbol{\theta}})$. Nelson (1993) shows mathematically and empirically that the MCB procedure assuming sphericity provides robust coverage even when the true variance-covariance matrix of the simulation outputs departs significantly from the sphericity assumption. Moreover, Nelson and Matejcik (1995) propose an MCB procedure that provides exact finite-sample coverage when sphericity and normality are assumed.

The IOU-C procedure introduced in the following sections allows $V(\widehat{\boldsymbol{\theta}})$ to have a general structure. The procedure can be simplified under the sphericity assumption. We comment on these changes when relevant.

5. Computing Confidence Interval Widths

In this section, we provide the details of Steps 4–5 of the IOU-C procedure in Section 3. We start by estimating the joint distributions of the CID effects and stochastic errors using Model (5) in Section 4 and compute $w_{i\ell}^{(1)}$ and $w_{i\ell}^{(2)}$ from them. In Section 6, we show these CI widths provide the desired statistical guarantee asymptotically as m and n increase under mild conditions.

5.1. Interval Widths due to CID effects

As derived in Section 3, $w_{i\ell}^{(1)}$ is determined by the distribution of the difference between CID effects of two systems, $b_i(\widehat{\boldsymbol{\theta}}, \boldsymbol{\theta}^c) - b_\ell(\widehat{\boldsymbol{\theta}}, \boldsymbol{\theta}^c)$. Under Model (5), $b_i(\widehat{\boldsymbol{\theta}}, \boldsymbol{\theta}^c) - b_\ell(\widehat{\boldsymbol{\theta}}, \boldsymbol{\theta}^c) = (\beta_i - \beta_\ell)^\top(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^c)$ and therefore, $w_{i\ell}^{(1)}$ is the α_1-quantile of $(\beta_i - \beta_\ell)^\top(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^c)$. However, it turns out to be difficult to find a pivotal quantity that provides the distribution of $(\beta_i - \beta_\ell)^\top(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^c)$. On the other hand, there are several methods to estimate gradients $\beta_1, \beta_2, \ldots, \beta_k$ (Fu 2015) and the distribution of $\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^c$ can be approximated by its asymptotic distribution (6). We propose two variations of IOU-C procedures in the following.

One way to approximate the joint distribution of the CID effects is to plug estimates, $\widehat{\beta}_1, \widehat{\beta}_2, \ldots, \widehat{\beta}_k$, into (7) and compute $w_{i\ell}^{(1)}$ by finding multivariate normal quantiles. We call this the *plug-in IOU-C* procedure and show that it provides the desired statistical guarantee asymptotically in Section 6 given our choice of gradient estimator.

The plug-in procedure ignores the estimation errors of $\widehat{\beta}_1, \widehat{\beta}_2, \ldots, \widehat{\beta}_k$. While this does not affect the asymptotic guarantee, it may result in lower coverage probability than desired for small m and n. Alternatively, we propose the *all-in IOU-C* procedure that solves $\binom{k}{2}$ optimization problems to obtain interval widths that incorporate the estimation errors of the gradients.

Let $\mathcal{B}_i^\top = \{(\beta_i - \beta_1)^\top, (\beta_i - \beta_2)^\top, \ldots, (\beta_i - \beta_{i-1})^\top, (\beta_i - \beta_{i+1})^\top, \ldots, (\beta_i - \beta_k)^\top\}$. We denote the following optimization problem as $\mathcal{P}_{i\ell}$:

$$\begin{aligned} &\min && (\beta_i - \beta_\ell)^\top(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^c) && (10)\\ &\text{subject to} && \mathcal{B}_i \in \mathrm{CR}_{1,\alpha_{11}}, && (11)\\ & && (\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^c) \in \mathrm{CR}_{2,\alpha_{12}} && (12) \end{aligned}$$

where $\mathrm{CR}_{1,\alpha_{11}} \subset \mathbb{R}^{(k-1)p}$ contains $\mathcal{B}_i$ with probability $1-\alpha_{11}$ and $\mathrm{CR}_{2,\alpha_{12}} \subset \mathbb{R}^p$ includes $\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^c$ with probability $1-\alpha_{12}$. Note that $\widehat{\boldsymbol{\theta}}$ in (10) and (12) is a random variable rather than the particular realization of $\widehat{\boldsymbol{\theta}}$ estimated from the m real-world observations. Thus, we treat $\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^c$ as a p-dimensional vector of decision variables for $\mathcal{P}_{i\ell}$. In words, $\mathcal{P}_{i\ell}$ finds the smallest inner product,

$(\beta_i - \beta_\ell)^\top(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^c)$, given separate confidence regions for $\beta_i - \beta_\ell$ and $\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^c$. Clearly, $\Pr\{\mathcal{B}_i \in \mathrm{CR}_{1,\alpha_{11}}, (\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^c) \in \mathrm{CR}_{2,\alpha_{12}}\} \geq 1 - \alpha_{11} - \alpha_{12}$. Suppose we choose $\alpha_{11}, \alpha_{12} > 0$ such that $\alpha_1 = \alpha_{11} + \alpha_{12}$. Then, by equating the optimal soluiton of $\mathcal{P}_{i\ell}$ with $-w_{i\ell}^{(1)}$ for all $\ell \neq i$ for each i we achieve

$$\Pr\{(\beta_i - \beta_\ell)^\top(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^c) \geq -w_{i\ell}^{(1)}, \forall \ell \neq i\} \geq 1 - \alpha_1. \tag{13}$$

Note that the joint confidence region for $\mathcal{B}_i$ in (11) is shared by all $\mathcal{P}_{i\ell}$ for all $\ell \neq i$, which is essential to provide the joint probability guarantee in (13). If for each $\mathcal{P}_{i\ell}$ we use a confidence region just for $\beta_i - \beta_\ell$ instead of (11), then we cannot guarantee (13).

In the development in Sections 5.1.1–5.1.2, we provide asymptotically valid $\mathrm{CR}_{1,\alpha_{11}}$ and $\mathrm{CR}_{2,\alpha_{12}}$, and employ a tool for incorporating the effect of CRN to reformulate $\mathcal{P}_{i\ell}$ as

$$\begin{aligned}
\text{min} \quad & (\beta_i - \beta_\ell)^\top(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^c) \\
\text{subject to} \quad & (\mathcal{B}_i - \widehat{\mathcal{B}}_i)^\top \mathcal{V}_i^{-1}(\mathcal{B}_i - \widehat{\mathcal{B}}_i) \leq \chi^2_{(k-1)p,\alpha_{11}}, && (14) \\
& m(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^c)^\top \Sigma^{-1}(\widehat{\boldsymbol{\theta}})(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^c) \leq \chi^2_{p,\alpha_{12}}, && (15)
\end{aligned}$$

where $\mathcal{V}_i$ is a consistent estimator of the variance-covariance matrix of $\widehat{\mathcal{B}}_i$. Here, $\chi^2_{\nu,\beta}$ is the $(1-\beta)$-quantile of the χ^2 distribution with ν degrees of freedom.

Although $\mathcal{P}_{i\ell}$ is a non-convex optimization problem and difficult to solve to optimality in general, it has a bilinear objective function with separable convex constraints for $\mathcal{B}_i$ and $\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^c$. Once $\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^c$ is fixed, the resulting problem has a quadratic constraint with a linear objective function, which can be solved to optimality easily; given $\boldsymbol{\theta}^c - \widehat{\boldsymbol{\theta}}$, the optimal value of $\mathcal{B}_i$ obtained from the Karush-Kuhn-Tucker (KKT) conditions for the modified problem is $\widehat{\mathcal{B}}_i - \mathcal{V}_i\boldsymbol{\nu}_\ell\sqrt{\chi^2_{(k-1)p,\alpha_{11}}/\boldsymbol{\nu}_\ell^\top\mathcal{V}_i\boldsymbol{\nu}_\ell}$, where $\boldsymbol{\nu}_\ell = \mathbf{e}_\ell \otimes (\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^c)$ for $(k-1)\times 1$ ℓth unit directional vector $\mathbf{e}_\ell$ and $A \otimes B$ represents the Kronecker product of matrices A and B. Thus, the corresponding optimal objective function value is

$$\begin{aligned}
&(\widehat{\beta}_i - \widehat{\beta}_\ell)^\top(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^c) - \sqrt{\chi^2_{(k-1)p,\alpha_{11}}\boldsymbol{\nu}_\ell^\top\mathcal{V}_i\boldsymbol{\nu}_\ell} \\
&= (\widehat{\beta}_i - \widehat{\beta}_\ell)^\top(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^c) - \sqrt{\chi^2_{(k-1)p,\alpha_{11}}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^c)^\top\mathcal{V}_i(\ell,\ell)(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^c)}, \qquad (16)
\end{aligned}$$

where the equality is obtained by plugging in $\boldsymbol{\nu}_\ell = \mathbf{e}_\ell \otimes (\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^c)$ and given $\mathcal{V}_i(\ell,\ell)$, the $p \times p$ block diagonal matrix of $\mathcal{V}_i$ corresponding to $\widehat{\beta}_i - \widehat{\beta}_\ell$.

In the following section, we propose a gradient estimation method that we use for both plug-in and all-in IOU-C procedures as well as the confidence regions for $\mathcal{P}_{i\ell}$ derived from the gradient estimators. Readers may skip Sections 5.1.1–5.1.2 without loss of continuity.

5.1.1. Gradient Estimation and Confidence Regions of $\mathcal{P}_{i\ell}$ From (6), we have $\mathrm{CR}_{2,\alpha_{12}} = \left\{\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^c : m(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^c)^\top \Sigma^{-1}(\widehat{\boldsymbol{\theta}})(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^c) \le \chi^2_{p,\alpha_{12}}\right\}$, which provides $1-\alpha_{12}$ asymptotic coverage probability as m increases under mild conditions (Amemiya 1985). Notice that $\Sigma(\boldsymbol{\theta}^c)$ in (6) is replaced with its plug-in estimator, $\Sigma(\widehat{\boldsymbol{\theta}})$.

Defining $\mathrm{CR}_{1,\alpha_{11}}$ that provides the correct asymptotic probability coverage requires more effort. We fit Model (5) by linear regressions to obtain $\widehat{\beta}_1, \widehat{\beta}_2, \ldots, \widehat{\beta}_k$ as well as $\mathrm{CR}_{1,\alpha_{11}}$. Suppose we sample the regression design points, $\widehat{\boldsymbol{\theta}}^{(1)}, \widehat{\boldsymbol{\theta}}^{(2)}, \ldots, \widehat{\boldsymbol{\theta}}^{(B)}$, from $\mathrm{N}(\widehat{\boldsymbol{\theta}}, \Sigma(\widehat{\boldsymbol{\theta}})/m)$, which is the estimated asymptotic distribution of $\widehat{\boldsymbol{\theta}}$ in (6) where $\boldsymbol{\theta}^c$ is replaced by $\widehat{\boldsymbol{\theta}}$. For each $\widehat{\boldsymbol{\theta}}^{(b)}$, we run one replication from each system to obtain $Y_1(\widehat{\boldsymbol{\theta}}^{(b)}), Y_2(\widehat{\boldsymbol{\theta}}^{(b)}), \ldots, Y_k(\widehat{\boldsymbol{\theta}}^{(b)})$ using CRN. For each system i, we fit a linear regression model with response vector $\mathbf{Y}_i = \left(Y_i(\widehat{\boldsymbol{\theta}}^{(1)}), Y_i(\widehat{\boldsymbol{\theta}}^{(2)}), \ldots, Y_i(\widehat{\boldsymbol{\theta}}^{(B)})\right)^\top$ and design matrix $\mathbf{X} = [\mathbf{1}_B : \mathbf{C}]$, where $\mathbf{1}_B$ is a B-dimensional column vector of 1's and $\mathbf{C}$ is the $B \times p$ matrix whose rows are $(\widehat{\boldsymbol{\theta}}^{(1)} - \widehat{\boldsymbol{\theta}})^\top, (\widehat{\boldsymbol{\theta}}^{(2)} - \widehat{\boldsymbol{\theta}})^\top, \ldots, (\widehat{\boldsymbol{\theta}}^{(B)} - \widehat{\boldsymbol{\theta}})^\top$. From the least squares method, the estimator of β_i is $\widehat{\beta}_i = (\mathbf{C}^\top(\mathbf{I}_B - \mathbf{1}_B\mathbf{1}_B^\top)\mathbf{C})^{-1}\mathbf{C}^\top\mathbf{Y}_i$, where $\mathbf{I}_B$ is the $B \times B$ identity matrix. Note that $(\mathbf{C}^\top(\mathbf{I}_B - \mathbf{1}_B\mathbf{1}_B^\top)\mathbf{C})^{-1}$ is the lower $p \times p$ submatrix of $(\mathbf{X}^\top\mathbf{X})^{-1}$, which we denote by $\mathbf{S}_{pp}$.

For finite m and B, $\widehat{\beta}_i$ is not an unbiased estimator of $\beta_i \equiv \nabla\eta_i(\boldsymbol{\theta}^c)$ as the design points of the regressions are centered at $\widehat{\boldsymbol{\theta}}$ instead of $\boldsymbol{\theta}^c$. The following proposition states the asymptotic normality and consistency results of our gradient estimator when m and B increase at the right rate under Assumption 1 in Section 6; this result is of independent interest.

PROPOSITION 1. As $m \to \infty$, for all i 1) $\sqrt{B/m}(\widehat{\mathcal{B}}_i - \mathcal{B}_i)\Big|\,\widehat{\boldsymbol{\theta}}$ converges weakly uniformly to $\mathrm{N}(\mathbf{0}, V_i(\boldsymbol{\theta}^c) \otimes \Sigma^{-1}(\boldsymbol{\theta}^c))$ in probability, if $B = m^\gamma$ for $0 < \gamma < 2$, and 2) $\widehat{\beta}_i \xrightarrow{p} \beta_i$, if $B = m^\gamma$ for $\gamma > 1$.

The proof of Proposition 1 can be found in Section EC.5 of the electronic companion of this paper. In general, a sequence of distributions $\{F_m(\cdot)\}$ is said to converge weakly uniformly to $F^c(\cdot)$ if $\sup_{\mathbf{x}} |F_m(\mathbf{x}) - F^c(\mathbf{x})| \to 0$. Suppose we replace $F_m(\cdot)$ with conditional distribution $F(\cdot|\widehat{\boldsymbol{\theta}})$. If $\sup_{\mathbf{x}} |F(\mathbf{x}|\widehat{\boldsymbol{\theta}}) - F^c(\mathbf{x})| \xrightarrow{p} 0$, then $F(\cdot|\widehat{\boldsymbol{\theta}})$ is said to converge weakly uniformly to $F^c(\cdot)$ in probability. Proposition 1 applies this convergence scheme for $\widehat{\mathcal{B}}_i$ whose distribution is conditional on $\widehat{\boldsymbol{\theta}}$ guaranteeing that for any $\widehat{\boldsymbol{\theta}}$ estimated from the real-world sample of size m, the confidence region for β_i provides the same coverage.

Notice that different rates are required for B to obtain asymptotic normality of $\sqrt{B/m}(\widehat{\mathcal{B}}_i - \mathcal{B}_i)$ and consistency of $\widehat{\beta}_i$, respectively. Design points for the regression sampled from $\mathrm{N}(\widehat{\boldsymbol{\theta}}, \Sigma(\widehat{\boldsymbol{\theta}})/m)$ become more and more concentrated near $\widehat{\boldsymbol{\theta}}$ as m increases. Therefore, if we increase B too slowly, then $\mathbf{X}^\top\mathbf{X}$ becomes singular and we cannot achieve consistency of $\widehat{\beta}_i$. On the other hand, asymptotic normality of $\sqrt{B/m}(\widehat{\mathcal{B}}_i - \mathcal{B}_i)$ holds for smaller B than consistency because of the scaling factor, $\sqrt{B/m}$. At the same time, if we increase B too fast relative to m, then the bias in $\widehat{\beta}_i$ due to the nonlinearity of $\eta_i(\widehat{\boldsymbol{\theta}})$ does not fade away fast enough as m increases; therefore $\sqrt{B/m}(\widehat{\beta}_i - \beta_i)$ no longer has a finite mean. Although consistency of $\widehat{\beta}_i$ is of independent interest, the normality result in Proposition 1 is what provides the asymptotic probability guarantee for the plug-in and all-in IOU-C procedures. See Section 6 for further discussion.

The proposed gradient estimator is not a typical choice of stochastic gradient estimator, however, it is particularly advantageous for IOU-C as it is easy to form confidence region $\mathrm{CR}_{1,\alpha_{11}}$ for $\mathcal{B}_i$. Lemma 2 in Section EC.5 shows $\mathbf{S}_{pp}$ is a consistent estimator of $\Sigma^{-1}(\boldsymbol{\theta}^c)$ under Assumption 1. Let $e_{ib} = Y_i(\widehat{\boldsymbol{\theta}}^{(b)}) - \widehat{\beta}_i^\top(\widehat{\boldsymbol{\theta}}^{(b)} - \widehat{\boldsymbol{\theta}})$ for $i = 1, 2, \ldots, k$ and $b = 1, 2, \ldots, B$ and $\mathbf{e}_{-i,b} = (e_{1b}, e_{2b}, \ldots, e_{(i-1)b}, e_{(i+1)b}, \ldots, e_{kb})^\top - e_{ib}\mathbf{1}_{k-1}$, where $\mathbf{1}_{k-1}$ is $(k-1)\times 1$ vector of ones. From these, we can compute $\widehat{V}_i$, the sample variance-covariance matrix of $\mathbf{e}_{-i,1}, \mathbf{e}_{-i,2}, \ldots, \mathbf{e}_{-i,B}$. Lemma 5 in Section EC.5 shows that $\widehat{V}_i$ is a consistent estimator of $V_i(\boldsymbol{\theta}^c)$ under Assumption 1. Thus, $\mathrm{CR}_{1,\alpha_{11}} = \{\mathcal{B}_i : (\mathcal{B}_i - \widehat{\mathcal{B}}_i)^\top \mathcal{V}_i^{-1}(\mathcal{B}_i - \widehat{\mathcal{B}}_i) \le \chi^2_{(k-1)p,\alpha_{11}}\}$ is an asymptotic $1-\alpha_{11}$ confidence region for $\mathcal{B}_i$, where $\mathcal{V}_i \equiv \widehat{V}_i \otimes \mathbf{S}_{pp}$. Therefore, $\mathcal{V}_i(\ell,\ell)$ in (16) is $\widehat{V}_i(\ell,\ell)\mathbf{S}_{pp}$, where $\widehat{V}_i(\ell,\ell)$ is the ℓth diagonal element of $\widehat{V}_i$.

If we assume sphericity for $V(\widehat{\boldsymbol{\theta}})$, then

$$\mathcal{V}_i = 2\widehat{\tau}^2 \begin{bmatrix} 1 & 1/2 & \cdots & 1/2 \\ 1/2 & 1 & \cdots & 1/2 \\ \vdots & \vdots & \ddots & \vdots \\ 1/2 & 1/2 & \cdots & 1 \end{bmatrix} \otimes \mathbf{S}_{pp}, \tag{17}$$

where $\widehat{\tau}^2 = \sum_{i=1}^{k}\sum_{b=1}^{B}(e_{ib} - \bar{e}_{i\cdot} - \bar{e}_{\cdot b} + \bar{\bar{e}})^2/(k-1)(B-p-2)$ given $\bar{e}_{i\cdot} = B^{-1}\sum_{b=1}^{B} e_i(\widehat{\boldsymbol{\theta}}^{(b)})$ for $i = 1, 2, \ldots, k$, $\bar{e}_{\cdot b} = k^{-1}\sum_{i=1}^{k} e_i(\widehat{\boldsymbol{\theta}}^{(b)})$ for $b = 1, 2, \ldots, B$, and $\bar{\bar{e}} = k^{-1}\sum_{i=1}^{k}\bar{e}_{i\cdot}$. Therefore, $\mathcal{V}_i$ does not depend on i and $\mathcal{V}_i(\ell,\ell) = 2\widehat{\tau}^2\mathbf{S}_{pp}$ for any $\ell \neq i$, which further simplifies (16).

Note that using the same design points, $\widehat{\boldsymbol{\theta}}^{(1)}, \widehat{\boldsymbol{\theta}}^{(2)}, \ldots, \widehat{\boldsymbol{\theta}}^{(B)}$, or applying CRN for the linear regressions is not necessary; the framework of IOU-C procedure allows fitting k regressions completely independently to obtain the joint confidence region for $\mathcal{B}_i$. However, doing so makes $w_{i\ell}, i \neq \ell$, wider causing the procedure to be less efficient.

Remark: In fact, the joint asymptotic coverage probability of $\mathrm{CR}_{1,\alpha_{11}}$ and $\mathrm{CR}_{2,\alpha_{12}}$ is greater than $1 - \alpha_{11} - \alpha_{12}$. In Lemma 6 in Section EC.5, we show that the two events $\{\mathcal{B}_i \in \mathrm{CR}_{1,\alpha_{11}}\}$ and $\{\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^c \in \mathrm{CR}_{2,\alpha_{12}}\}$ are asymptotically independent, therefore, their joint asymptotic coverage probability is actually $(1-\alpha_{11})(1-\alpha_{12})$.

Remark: The sample size B required for consistency of our gradient estimator $\widehat{\beta}_i$ is smaller than that of the simultaneous perturbation stochastic approximation gradient estimator (SPSA, Spall 1992). The SPSA gradient estimator converges to the true gradient in $O_p(B^{-1/3})$ when $\boldsymbol{\theta}^c$ is known by choosing the optimal perturbation constant. Since $\boldsymbol{\theta}^c$ is unknown for our problem, we estimate β_i using an estimator of $\nabla\eta_i(\widehat{\boldsymbol{\theta}})$ whose bias cannot be improved from $O(m^{-1})$. By selecting the perturbation constant to match the bias, the variance of the SPSA estimator becomes $O(m/B)$. As a result, $B = O(m^3)$ to balance the variance and bias, which is larger than our choice for B.

5.1.2. Reformulation of $\mathcal{P}_{i\ell}$ Employing our choices for the confidence regions, constraints (11) and (12) of $\mathcal{P}_{i\ell}$ are replaced by (14) and (15). Since $\mathrm{CR}_{2,\alpha_{12}}$ is a symmetric hyperellipsoid with respect to the origin,

$$-\inf_{\mathcal{B}_i\in\mathrm{CR}_{1,\alpha_{11}},(\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}^c)\in\mathrm{CR}_{2,\alpha_{12}}}(\beta_i-\beta_\ell)^\top(\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}^c)=\sup_{\mathcal{B}_i\in\mathrm{CR}_{1,\alpha_{11}},(\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}^c)\in\mathrm{CR}_{2,\alpha_{12}}}(\beta_i-\beta_\ell)^\top(\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}^c).$$

In other words, $[-w_{i\ell}^{(1)},w_{i\ell}^{(1)}]$ provides a symmetric two-sided $(1-\alpha_{11})(1-\alpha_{12})$ CI for $(\beta_i-\beta_\ell)^\top(\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}^c)$. Because we only need one-sided CIs for our procedure, we can set $\Pr\left\{(\widehat{\beta}_i-\widehat{\beta}_\ell)^\top(\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}^c)\geq -w_{i\ell}^{(1)},\forall\ell\neq i\right\}=1-(\alpha_{11}+\alpha_{12}-\alpha_{11}\alpha_{12})/2$, i.e., $\alpha_1=(\alpha_{11}+\alpha_{12}-\alpha_{11}\alpha_{12})/2$, which makes $w_{i\ell}^{(1)}$ smaller.

The following proposition lets us focus on the boundary points of the feasible region of $\mathcal{P}_{i\ell}$ to obtain its optimal solution.

Proposition 2. For all $i\neq\ell$, the optimal objective function value of

$$\begin{aligned}
&\min && (\beta_i-\beta_\ell)^\top(\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}^c)\\
&\text{subject to} && (\mathcal{B}_i-\widehat{\mathcal{B}}_i)^\top\mathcal{V}_i^{-1}(\mathcal{B}_i-\widehat{\mathcal{B}}_i)=\chi^2_{(k-1)p,\alpha_{11}},\\
& && m(\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}^c)^\top\Sigma^{-1}(\widehat{\boldsymbol{\theta}})(\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}^c)=\chi^2_{p,\alpha_{12}},
\end{aligned}$$

is the same as the optimal objective function value of $\mathcal{P}_{i\ell}$.

The proof of Proposition 2 can be found in Section EC.1, which relies on showing that an optimal solution of $\mathcal{P}_{i\ell}$ is found at the boundary of two hyperellipsoidal constraints on $\mathcal{B}_i$ and $\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}^c$. From Proposition 2, we devise a random search algorithm for $\mathcal{P}_{i\ell}$ in Section EC.2.

5.2. Interval Widths due to Stochastic Error

In Section 3, we showed that the interval widths due to input uncertainty and the interval widths due to stochastic error can be obtained separately if we find $w_{i\ell}^{(2)},\forall\ell\neq i$ that satisfy (4) for any given $\widehat{F}=F(\cdot|\widehat{\boldsymbol{\theta}})$. Since we focus on the asymptotic coverage of the IOU-C intervals as $n,m\to\infty$, the joint distribution of $\{\bar{\varepsilon}_i(\widehat{\boldsymbol{\theta}})-\bar{\varepsilon}_i(\widehat{\boldsymbol{\theta}})\}_{\ell=1,\ell\neq i}^k$ converges to

$N(\mathbf{0}, V_i(\boldsymbol{\theta}^c))$ under Assumption 1. Similar to $\widehat{V}_i$ in Section 5.1, we can compute the estimator, $\widehat{\mathsf{V}}_i(\widehat{\boldsymbol{\theta}})$, of $V_i(\boldsymbol{\theta}^c)$ by computing the sample variance-covariance matrix of $\mathbf{Y}_{-i,j} = \left(Y_{ij}(\widehat{\boldsymbol{\theta}}) - Y_{1j}(\widehat{\boldsymbol{\theta}}), Y_{ij}(\widehat{\boldsymbol{\theta}}) - Y_{2j}(\widehat{\boldsymbol{\theta}}), \cdots, Y_{ij}(\widehat{\boldsymbol{\theta}}) - Y_{i-1,j}(\widehat{\boldsymbol{\theta}}), Y_{ij}(\widehat{\boldsymbol{\theta}}) - Y_{i+1,j}(\widehat{\boldsymbol{\theta}}), \cdots, Y_{ij}(\widehat{\boldsymbol{\theta}}) - Y_{kj}(\widehat{\boldsymbol{\theta}})\right)^\top$ for $j = 1, 2, \ldots, k$. Given $\widehat{\mathsf{V}}_i(\widehat{\boldsymbol{\theta}})$, $(w_{i1}^{(2)}, w_{i2}^{(2)}, \cdots, w_{i,i-1}^{(2)}, w_{i,i+1}^{(2)}, \cdots, w_{ik}^{(2)})$ is a $(k-1)$-dimensional $1-\alpha_2$ vector quantile of $\mathrm{N}(\mathbf{0}, \widehat{\mathsf{V}}_i(\widehat{\boldsymbol{\theta}})/n)$, which can be obtained via the Multidimensional Quantile Estimation algorithm in Section EC.3. Since $\sqrt{n}\bar{\varepsilon}_i(\widehat{\boldsymbol{\theta}}), i = 1, 2, \ldots, k$ converges weakly to a normal distribution conditional on $\widehat{\boldsymbol{\theta}}$ under Assumption 1, $w_{i\ell}^{(2)}, \forall \ell \neq i$, satisfy (4) for any $\widehat{\boldsymbol{\theta}}$ as n increases.

Both $\widehat{\mathsf{V}}_i(\widehat{\boldsymbol{\theta}})$ and $\widehat{V}_i$ are consistent estimators of $V_i(\boldsymbol{\theta}^c)$. However, the former is obtained from n replications of each system and the latter is from the the regression residuals. To provide a correct statistical guarantee with our procedure, the replications should be run independently from the regressions using different random numbers so that $w_{i\ell}^{(1)}$ and $w_{i\ell}^{(2)}$ are independent conditional on $\widehat{\boldsymbol{\theta}}$.

Under the sphericity assumption on $V(\boldsymbol{\theta}^c)$, estimating $V_i(\boldsymbol{\theta}^c)$ simplifies to estimating $\tau^2(\boldsymbol{\theta}^c)$ from n replications of the k systems. Nelson and Matejcik (1995) provide an MCB procedure with fixed interval lengths by assuming normality of simulation outputs and sphericity of $V(\boldsymbol{\theta}^c)$ when $\boldsymbol{\theta}^c$ is known. Applying their method, we can set $w_{i\ell}^{(2)} = \sqrt{\frac{2}{n}} s T_{k-1,(k-1)(n-1),1/2,\alpha_2}, \forall \ell \neq i$, where $T_{k-1,(k-1)(n-1),1/2,\alpha_2}$ is the $1-\alpha_2$ quantile of a multivariate-t distribution of dimension $k-1$ with $(k-1)(n-1)$ degrees of freedom and common correlation $1/2$ and

$$s^2 = \frac{1}{(k-1)(n-1)} \sum_{i=1}^{k} \sum_{j=1}^{n} \left(Y_{ij}(\widehat{\boldsymbol{\theta}}) - \bar{Y}_i(\widehat{\boldsymbol{\theta}}) - \bar{Y}_{\cdot j}(\widehat{\boldsymbol{\theta}}) + \bar{\bar{Y}}(\widehat{\boldsymbol{\theta}})\right)^2, \tag{18}$$

given $\bar{Y}_{\cdot j}(\widehat{\boldsymbol{\theta}}) = k^{-1} \sum_{i=1}^{k} Y_{ij}(\widehat{\boldsymbol{\theta}})$ for $j = 1, 2, \ldots, n$, and $\bar{\bar{Y}}(\widehat{\boldsymbol{\theta}}) = k^{-1} \sum_{i=1}^{k} \bar{Y}_i(\widehat{\boldsymbol{\theta}})$.

Another benefit of the sphericity assumption is that we can account for the estimation error in s^2. For any finite n, $\widehat{\mathsf{V}}_i(\widehat{\boldsymbol{\theta}})$ has estimation error. If we assume normality of simulation output, then $w_{i\ell}^{(2)} = \sqrt{\frac{2}{n}} s T_{k-1,(k-1)(n-1),1/2,\alpha_2}$ gives the exact $1-\alpha_2$ coverage for the interval widths due to stochastic error for any finite n.

5.3. Plug-in and All-in IOU Comparisons Procedures

We first present the all-in IOU-C procedure that incorporates the interval width computation discussed in Sections 5.1–5.2. See Assumption 1.(ii) for the definition of $\Sigma(\cdot)$ in Step 3.(b).

All-in IOU-C

1. Select $0<\alpha_1,\alpha_2<1/2$ such that $1-\alpha=(1-\alpha_1)(1-\alpha_2)$ for given $0<\alpha<1/2$.
2. Compute $\widehat{\boldsymbol{\theta}}$. Using $F(\cdot|\widehat{\boldsymbol{\theta}})$ as the common input model, run n replications of all k systems with CRN, $Y_{i1}(\widehat{\boldsymbol{\theta}}),Y_{i2}(\widehat{\boldsymbol{\theta}}),\ldots,Y_{in}(\widehat{\boldsymbol{\theta}})$, for $i=1,2,\ldots,k$. Compute $\bar{Y}_i(\widehat{\boldsymbol{\theta}})=\Sigma_{j=1}^{n}Y_{ij}(\widehat{\boldsymbol{\theta}})/n$.
3. **(Interval widths due to CID effects)**
   (a) Select $0<\alpha_{11},\alpha_{12}<1$ such that $\alpha_1=(\alpha_{11}+\alpha_{12}-\alpha_{11}\alpha_{12})/2$.
   (b) Generate $\widehat{\boldsymbol{\theta}}^{(1)},\widehat{\boldsymbol{\theta}}^{(2)},\ldots,\widehat{\boldsymbol{\theta}}^{(B)}$ from $\mathrm{N}(\widehat{\boldsymbol{\theta}},\Sigma(\widehat{\boldsymbol{\theta}})/m)$.
   (c) For each $b=1,2,\ldots,B$, use CRN to simulate all k systems to obtain $Y_1(\widehat{\boldsymbol{\theta}}^{(b)}),Y_2(\widehat{\boldsymbol{\theta}}^{(b)}),\ldots,Y_k(\widehat{\boldsymbol{\theta}}^{(b)})$.
   (d) For each i, fit a linear regression to compute $\widehat{\beta}_i$ using response $\mathbf{Y}_i=\left(Y_i(\widehat{\boldsymbol{\theta}}^{(1)}),Y_i(\widehat{\boldsymbol{\theta}}^{(2)}),\ldots,Y_i(\widehat{\boldsymbol{\theta}}^{(B)})\right)^\top$ and design matrix $[\mathbf{1}_B:\mathbf{C}]$, where the bth row of $\mathbf{C}$ corresponds to $\widehat{\boldsymbol{\theta}}^{(b)\top}$.
   (e) From the residuals of the regressions in (d), compute $\widehat{V}_i$ for $i=1,2,\ldots,k$.
   (f) For each combination $(i,\ell),i\neq\ell$, solve $\mathcal{P}_{i\ell}$ and set $-w_{i\ell}^{(1)}$ equal to its optimal solution.
4. **(Interval widths due to stochastic error)** From the replications in Step 2, compute $\widehat{\mathbb{V}}_i$ for $i=1,2,\ldots,k$ and apply the Multidimensional Quantile Estimation algorithm in Section EC.3 to find $w_{i\ell}^{(2)},\forall\ell\neq i$.
5. For each combination $(i,\ell),i\neq\ell$, set $w_{i\ell}=w_{i\ell}^{(1)}+w_{i\ell}^{(2)}$. Use Theorem 1 to derive $1-\alpha$ simultaneous comparisons CIs.

As mentioned in Section 5.1–5.2, we can simplify Steps 3(e) and 4 by assuming sphericity of $V(\boldsymbol{\theta}^c)$.

The plug-in procedure has the same steps as the all-in IOU-C except that Steps 3(e)–(f) are replaced by the following step:

Plug-in IOU-C

3. (e) For each i, plug in $\widehat{\beta}_1, \widehat{\beta}_2, \ldots, \widehat{\beta}_k$ to (8) (by replacing 1 with i). Compute the $(k-1)$-dimensional $(1-\alpha_1)$ quantile $\{w_{i\ell}, \forall \ell \neq i\}$ of the plug-in distribution.

We can once again apply the Multidimensional Quantile Estimation algorithm in this step.

As pointed out in Section 3, the CIs obtained from IOU-C are wider than MCB CIs as we have the interval widths due to input uncertainty added to those due to stochastic error. Not surprisingly, the CIs from all-in IOU-C are wider than those from plug-in IOU-C as the former account for the estimation error in the gradients. Splitting $1-\alpha$ into $1-\alpha_1$ and $1-\alpha_2$ to obtain the interval widths due to input and stochastic uncertainty independently makes the combined interval widths from the all-in procedure larger. Moreover, the optimal solution to $\mathcal{P}_{i\ell}$ tends to provide conservative $w_{i\ell}^{(1)}$. In fact, solving $\mathcal{P}_{i\ell}$ for all $\ell \neq i$ for fixed $\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^c$ has the same implication as obtaining $1-\alpha_{11}$ simultaneous CIs for $(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^c)^\top(\widehat{\beta}_i - \widehat{\beta}_\ell)$ for all $\ell \neq i$ given $\mathrm{CR}_{1,\alpha_{11}}$, which is known as Scheffé's method (Seber and Lee 2003). Scheffé's method tends to be conservative because it is designed to provide valid simultaneous CIs for any set of linear combinations of the regression coefficients. This conservatism also affects the CIs obtained from IOU-C. We empirically compare the performance of both all-in and plug-in IOU-C procedures in Section 7.

6. Asymptotic Validity

In this section, we provide the conditions under which the plug-in and all-in IOU-C procedures provide an asymptotic $1-\alpha$ probability guarantee as m, B and n increase.

As mentioned in Section 5.1.1, $\widehat{\beta}_i$ is not an unbiased estimator of β_i for finite m since the design points of the regression are centered at $\widehat{\boldsymbol{\theta}}$. Moreover, Model (5) approximates $\eta_i(\widehat{\boldsymbol{\theta}})$ with a linear function of $\widehat{\boldsymbol{\theta}}$, while $\eta_i(\widehat{\boldsymbol{\theta}})$ is typically nonlinear in $\widehat{\boldsymbol{\theta}}$. Therefore, for both plug-in and all-in procedures we need 1) asymptotic normality of MLE $\widehat{\boldsymbol{\theta}}$, 2) consistency of $\widehat{\beta}_i$, and 3) Model (5) to be an exact representation of $\eta_i(\widehat{\boldsymbol{\theta}})$ in the limit $(m \to \infty)$ to prove an asymptotic guarantee. For the all-in procedure, we need not only consistency of $\widehat{\beta}_i$, but also its asymptotic normality stated in Proposition 1 to ensure $w_{i\ell}^{(1)}, i \neq \ell$, provide CIs with the correct asymptotic coverage. For both procedures, the Central Limit Theorem for $\bar{Y}_i(\widehat{\boldsymbol{\theta}}), i = 1, 2, \ldots, k$, as $n \to \infty$ provides the correct

asymptotic coverage of $w_{i\ell}^{(2)}, i \neq \ell$. Below, we state a list of assumptions under which we show the asymptotic probability guarantee of plug-in and all-in IOU-C.

Assumption 1.

(i) For $F(\cdot|\boldsymbol{\theta})$, we have $\boldsymbol{\theta}^c, \widehat{\boldsymbol{\theta}} \in \Theta$, where Θ is a compact set in $\mathbb{R}^p$.

(ii) All of the necessary regularity conditions are satisfied for $\widehat{\boldsymbol{\theta}}$ to have, as $m \to \infty$,

(a) **(Consistency)** $\widehat{\boldsymbol{\theta}} \xrightarrow{a.s.} \boldsymbol{\theta}^c$.

(b) **(Asymptotic normality)** $\sqrt{m}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^c) \xrightarrow{D} \mathrm{N}(0, \Sigma(\boldsymbol{\theta}^c))$, where $\Sigma(\boldsymbol{\theta}^c) = I^{-1}(\boldsymbol{\theta}^c)$ and $I(\cdot)$ is the Fisher information matrix of $F(\cdot|\boldsymbol{\theta}^c)$.

(iii) For each i, $\eta_i(\cdot)$ is twice continuously differentiable in a neighborhood of $\boldsymbol{\theta}^c$.

(iv) For any $\boldsymbol{\theta} \in \Theta, \varepsilon_i(\boldsymbol{\theta})|\boldsymbol{\theta} \sim (0, \sigma_i^2(\boldsymbol{\theta}))$, where $0 < \sigma_i^2(\boldsymbol{\theta}) < \infty$, and $\sigma_i^2(\boldsymbol{\theta})$ is a continuous function of $\boldsymbol{\theta}$ at $\boldsymbol{\theta} = \boldsymbol{\theta}^c$.

(v) For each i and $\widehat{\boldsymbol{\theta}}^{(b)} \overset{\text{i.i.d.}}{\sim} \mathrm{N}\left(\widehat{\boldsymbol{\theta}}, \Sigma(\widehat{\boldsymbol{\theta}})/m\right)$ given $\widehat{\boldsymbol{\theta}} \in \Theta$

(a) All elements of $\mathrm{E}\left[\sqrt{m}\sigma_i^2(\widehat{\boldsymbol{\theta}}^{(b)})(\widehat{\boldsymbol{\theta}}^{(b)} - \widehat{\boldsymbol{\theta}}) \middle| \widehat{\boldsymbol{\theta}}\right]$ are bounded in Θ.

(b) All elements of $\mathrm{E}\left[m\sigma_i^2(\widehat{\boldsymbol{\theta}}^{(b)})(\widehat{\boldsymbol{\theta}}^{(b)} - \widehat{\boldsymbol{\theta}})(\widehat{\boldsymbol{\theta}}^{(b)} - \widehat{\boldsymbol{\theta}})^\top \middle| \widehat{\boldsymbol{\theta}}\right]$ are bounded in Θ.

(c) $\mathrm{E}\left[\varepsilon_i^4(\widehat{\boldsymbol{\theta}}^{(b)}) \middle| \widehat{\boldsymbol{\theta}}\right]$ is bounded in Θ.

(d) There exists $u^\star \in \mathbb{R}$ such that

$$\mathrm{E}\left[\varepsilon_i^3(\widehat{\boldsymbol{\theta}}^{(b)})\left\{\mathrm{Var}\left(\varepsilon_i(\widehat{\boldsymbol{\theta}}^{(b)})\right)^{-1} + (\widehat{\boldsymbol{\theta}}^{(b)} - \widehat{\boldsymbol{\theta}})^\top \mathrm{Cov}\left(\varepsilon_i(\widehat{\boldsymbol{\theta}}^{(b)})(\widehat{\boldsymbol{\theta}}^{(b)} - \widehat{\boldsymbol{\theta}})\right)^{-1}(\widehat{\boldsymbol{\theta}}^{(b)} - \widehat{\boldsymbol{\theta}})\right\}^{3/2} \middle| \widehat{\boldsymbol{\theta}}\right] < u^\star$$

for all $\widehat{\boldsymbol{\theta}} \in \Theta$.

(vi) For each $(i, \ell), i \neq \ell$, we have an oracle to solve $\mathcal{P}_{i\ell}$ to optimality.

(vii) Given probability $0 < \delta < 1$ and the distribution of a D-dimensional multivariate normal random vector $\mathbf{Z} = \{Z^1, Z^2, \ldots, Z^D\}$, we find an exact δ-quantile $\mathbf{q}(\delta) = \{q^1(\delta), q^2(\delta), \ldots, q^D(\delta)\}$ such that $\Pr\{Z^d \leq q^d(\delta), \forall d\} = \delta$.

Assumption 1(i) is one of the regularity conditions required for Assumption 1(ii), however, we state it separately as it is frequently referred to in other conditions and the proofs in Section EC.5. The regularity conditions in Assumption 1(ii) can be found in Amemiya (1985). Assumption 1(iii)

provides smoothness conditions on $\eta_i(\widehat{\boldsymbol{\theta}})$ to guarantee consistency of $\widehat{\beta}_i$. The continuity of $\sigma_i^2(\widehat{\boldsymbol{\theta}})$ in Assumption 1 (iv) causes the dependence of the distribution of $\varepsilon_i(\widehat{\boldsymbol{\theta}})$ on $\widehat{\boldsymbol{\theta}}$ to fade away as $\widehat{\boldsymbol{\theta}}$ converges to $\boldsymbol{\theta}^c$, making $\widehat{V}_i$ and $\widehat{\mathsf{V}}_i(\widehat{\boldsymbol{\theta}})$ consistent estimators of $V_i(\boldsymbol{\theta}^c)$.

One challenge to show asymptotic validity of both plug-in and all-in IOU-C is that the resulting interval widths are conditional on the particular $\widehat{\boldsymbol{\theta}}$ computed from the real-world sample in Step 2. Thus, we need to show that the procedures provide the desired probability guarantee uniformly over $\widehat{\boldsymbol{\theta}}$ as $m \to \infty$. Assumption 1(v) enables us to obtain such consistency over $\widehat{\boldsymbol{\theta}} \in \Theta$. In particular, Assumption 1(v)(a)–(c) ensure the conditional moments of $(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^c)$ and $\varepsilon_i(\widehat{\boldsymbol{\theta}})$ on $\widehat{\boldsymbol{\theta}}$ to be bounded for any $\widehat{\boldsymbol{\theta}} \in \Theta$ so that they converge to the right moments in the limit. Assumption 1(v)(d) is a sufficient condition to apply the Berry-Esseen theorem in the proof of Lemma 4, which is an intermediate step to prove Proposition 1. The conditions in Assumption 1(v) are fairly mild: for instance, these moment conditions are satisfied if $\sigma_i^2(\widehat{\boldsymbol{\theta}})$ is a polynomial in $\widehat{\boldsymbol{\theta}}$.

Assumption 1(vi) may appear strong since we solve $\mathcal{P}_{i\ell}$ by the random search algorithm in Section EC.3. We show that the Random Search algorithm's probability of finding a solution within an ϵ optimality gap converges to 1 exponentially as its sample size L increases in Section EC.3. Nevertheless, we are not too concerned with the optimality gap of the random search algorithm due to the inherent conservatism of the all-in procedure discussed in Section 5.3. Even if it finds a suboptimal solution of $\mathcal{P}_{i\ell}$ and therefore makes $w_{i\ell}^{(1)}$ smaller than at optimality, the all-in procedure shows good empirical performance. In fact, the results in Section 7 show that the all-in procedure is still conservative even if we use the random search to solve $\mathcal{P}_{i\ell}$.

Assumption 1(vii) states that given the plug-in distribution of CID effects and $\mathsf{V}_i(\widehat{\boldsymbol{\theta}})$, we can find the exact multidimensional quantile vectors for $-w_{i\ell}^{(1)}$ and $-w_{i\ell}^{(2)}$, respectively. The Multidimensional Quantile Estimation algorithm in Section EC.3 samples Q points from the plug-in distribution to find the Monte Carlo estimator of the quantile. As $Q \to \infty$, the Monte Carlo estimator converges to the true $1-\delta$ quantile of the given multivariate normal distribution. If desired, one could use an upper confidence bound for this value.

Theorem 2 states that both plug-in and all-in procedures provide an asymptotic $1-\alpha$ probability guarantee under Assumption 1. The proof of the theorem can be found in Section EC.3.

THEOREM 2. *Under Assumption 1, if* $B = m^\gamma, 0 < \gamma < 2$, *then as* $m \to \infty, n \to \infty$ $\Pr\left\{\bar{Y}_i(\widehat{\boldsymbol{\theta}}) - \bar{Y}_\ell(\widehat{\boldsymbol{\theta}}) - (\eta_i(\boldsymbol{\theta}^c) - \eta_\ell(\boldsymbol{\theta}^c)) \geq -w_{i\ell}, \forall \ell \neq i\right\} \to (1-\alpha_1)(1-\alpha_2)$ *given 1)* $w_{i\ell}, i \neq \ell$, *from the plug-in IOU-C procedure, or 2)* $w_{i\ell}, i \neq \ell$, *from the all-in IOU-C procedure.*

Theorem 2 requires $B = m^\gamma$ for $0 < \gamma < 2$, which is the condition for asymptotic normality of $\sqrt{B/m}(\widehat{\mathcal{B}}_i - \mathcal{B}_i)$ in Proposition 3 in Section EC.5. In fact, we do not need consistency of $\widehat{\mathcal{B}}_i$ itself for either the all-in or plug-in IOU-C procedure. For the former, asymptotic normality is sufficient to provide confidence region for $\mathcal{B}_i$ with the correct asymptotic probability coverage. For the plug-in procedure, we essentially approximate the asymptotic distribution of $\beta_i^\top(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta})$ by $N(0, \widehat{\beta}_i^\top \Sigma(\widehat{\boldsymbol{\theta}})\widehat{\beta}_i/m)$. Hence, even if $\widehat{\beta}_i$ has $o_p(m^{1/2})$ error, $\widehat{\beta}_i^\top \Sigma(\widehat{\boldsymbol{\theta}})\widehat{\beta}_i/m = \beta_i^\top \Sigma(\boldsymbol{\theta}^c)\beta_i/m + o_p(1)$ and $\beta_i^\top(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \xrightarrow{D} N(0, \widehat{\beta}_i^\top \Sigma(\widehat{\boldsymbol{\theta}})\widehat{\beta}_i/m)$ by the continuous mapping theorem as $m \to \infty$. Note that $\widehat{\beta}_i = \beta_i + o_p(m^{1/2})$, if $B = m^\gamma, 0 < \gamma < 2$.

Both procedures require $k(B+n)$ simulations in total. Since γ can be arbitrarily close to 0, B need not be too large to provide the asymptotic probability guarantee. For the plug-in procedure, a different gradient estimation method may be used as long as it has $o_p(m^{1/2})$ error. An alternative approach to estimate the gradients is to use the method of Wieland and Schmeiser (2006), which uses the MLE computed from the input random variates generated from $F(\cdot|\widehat{\boldsymbol{\theta}})$ within each replication of $Y_i(\widehat{\boldsymbol{\theta}})$ as the corresponding design point to fit the regression models, and therefore does not require additional simulation effort beyond the n replications spent for each of k systems.

7. Experiment Results

In this section, we demonstrate the performance of the plug-in and all-in IOU-C procedures using an (s, S) inventory problem modified from Koenig and Law (1985) and compare them to the results from the conditional MCB procedure, which ignores input uncertainty by assuming $\widehat{\boldsymbol{\theta}} = \boldsymbol{\theta}^c$ and only accounts for the stochastic error representing the current state of practice. We present the results from these experiments graphically, deferring the detailed numerical results to Section EC.4.

We also summarize application of IOU-C to a series of test cases with known $\eta_i(\theta^c)$ in which we controlled mean and CID effect configurations of the systems as well as the nonlinearity of $\eta_i(\theta^c)$.

The objective function for the (s,S) inventory problem is the expected average cost per period over 30 periods. The problem has three stochastic input processes, demand per period, lead time, and yield of the delivered order. The demand per period is a sequence of i.i.d. Poisson random variables with a common mean of 10. The order and unit shipping costs are \$50 and \$0.5, respectively, and the holding cost of inventory and the back-order cost are \$1 and \$3 per unit, per period, respectively. The lead time until the placed order arrives follows a geometric distribution with probability of success 0.5. The actual number of units that arrive has a binomial distribution where the probability that each unit in the order arrives is 0.95. Therefore, $\boldsymbol{\theta}^c = (10, 0.5, 0.95)$.

We consider $k = 23$ (s,S) inventory policies: $\{(s,S) : s \in \{10,20,30,40,50,60,70\}, S \in \{50,60,70,80\}, s < S\} \cup (25,35)$. Table 1 shows the expected cost of each solution. Note that the expected cost of the optimal solution, $i = 1$: $(25,35)$, is \$0.21 lower than that of the second best, $i = 2$: $(10,50)$, which makes this example more interesting as the two lowest objective function values are close. Each replication of the simulator is a batch mean of 100 iterations of the 30-period simulation. This enhances the performance of IOU-C since it makes the joint distribution of $Y_i(\widehat{\boldsymbol{\theta}}^{(b)})$ and $\widehat{\boldsymbol{\theta}}^{(b)}$ close to multivariate normal. We consider the negative of expected cost as the objective, which results in a maximization problem.

We assume that the distribution families are known, but $\widehat{\boldsymbol{\theta}}$ is estimated from data using MLEs. At each run of IOU-C, $m = 100$ or $m = 400$ "real-world" observations from the true demand, lead time, and yield distributions are collected to estimate $\widehat{\boldsymbol{\theta}}$. This is impossible in a real-world problem as we only have one set of real-world observations.

We set $\gamma = 1.1$, which gives $B = \lceil m^{1.1} \rceil = 159$ for $m = 100$ for the number of design points to fit the linear regressions. A total of $L = 1{,}000$ values of $(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^c)$ were sampled in the random search algorithm (see Section EC.2) to approximate the optimal solutions of $\mathcal{P}_{i\ell}, i \neq \ell$. Each system was simulated $n = 100$ times using $\widehat{\boldsymbol{\theta}}$ to obtain the interval widths due to stochastic uncertainty. The

Table 1 Expected cost per period of 23 (s, S) inventory policies estimated from Monte Carlo simulations. The standard errors are presented in the parentheses.

i	1	2	3	4	5	6	7	8	9	10	11	12
s	25	10	20	30	40	10	20	30	40	50	10	20
S	35	50	50	50	50	60	60	60	60	60	70	70
E[cost]	56.28	56.48	63.25	74.92	90.75	72.04	80.26	92.79	108.74	129.64	90.37	99.94
	(0.01)	(0.01)	(0.01)	(0.02)	(0.02)	(0.01)	(0.01)	(0.02)	(0.02)	(0.03)	(0.02)	(0.02)
i	13	14	15	16	17	18	19	20	21	22	23	
s	30	40	50	60	10	20	30	40	50	60	70	
S	70	70	70	70	80	80	80	80	80	80	80	
E[cost]	113.79	130.90	152.46	176.32	114.67	126.11	141.38	159.94	182.32	207.23	236.52	
	(0.02)	(0.03)	(0.03)	(0.04)	(0.02)	(0.02)	(0.03)	(0.03)	(0.04)	(0.04)	(0.05)	

target probability guarantee is set to $1-\alpha = 0.9$, where $1-\alpha_1 = 0.9^{2/3}$ and $1-\alpha_2 = 0.9^{1/3}$ are used to obtain the interval widths due to CID effects and stochastic uncertainty, respectively. We set $\alpha_{11} = \alpha_{12} = 0.0703$ for the confidence regions in $\mathcal{P}_{i\ell}$ by solving $(\alpha_{11} + \alpha_{12} - \alpha_{11}\alpha_{12})/2 = \alpha_1$ as discussed in Section 5.1.2. We also did not assume sphericity of the simulation error variance-covariance matrix for the experiments presented in this section.

Figure 2 shows the results averaged across 1,000 runs of all three procedures when the number of "real-world" observations is $m = 100$. When creating the "real-world" data, we resampled the yield observations for 6 runs when all 100 yield observations were equal to 1 corresponding to the case where there is no evidence of input uncertainty based on the observations. The x-axis of Figure 2 represents $\eta_i(\boldsymbol{\theta}^c) - \max_{\ell \neq i} \eta_\ell(\boldsymbol{\theta}^c)$ for $i = 1, 2, \ldots, 23$. Sorted in increasing order, the x-axis marks $\eta_i(\boldsymbol{\theta}^c) - \max_{\ell \neq i} \eta_\ell(\boldsymbol{\theta}^c)$ of $i = 1, 2, 3, 6, 4, 7, 11, 5, 8, 12, 9, 13, 17, 18, 10, 14, 19, 15, 20, 16, 21, 22,$ and 23 from left to right. The right-hand-side of the y-axis represents the joint coverage probability of the IOU-C (or MCB) CIs and the left-hand-side shows the probability that each system is in $\mathcal{S}_0$, which is the set of solutions containing 0 in their IOU-C (or MCB) CIs.

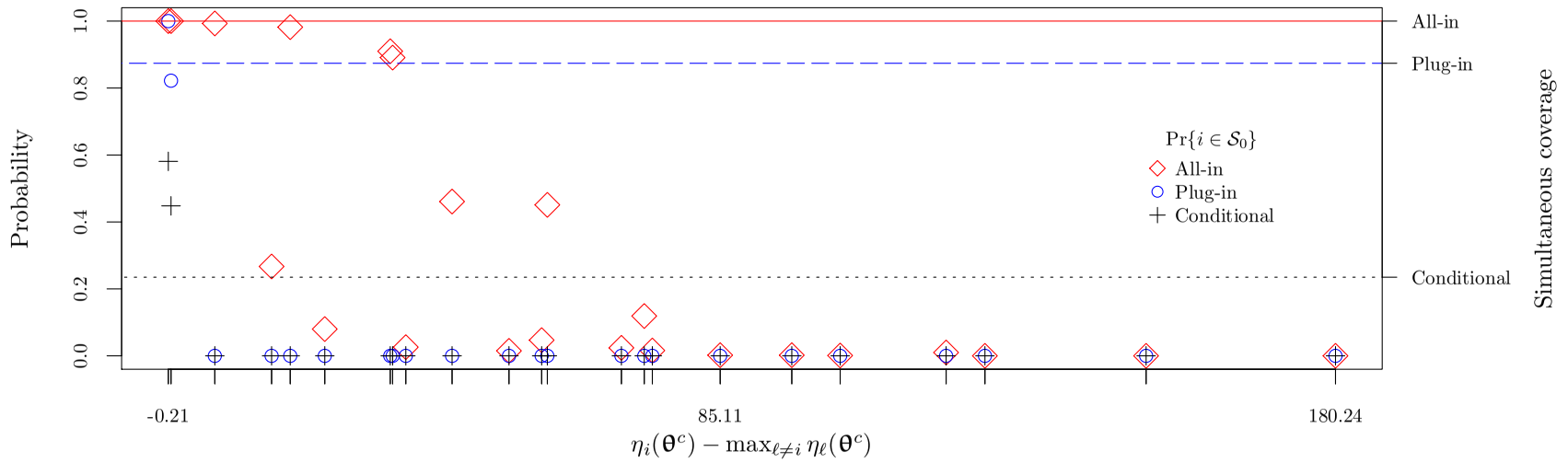

Figure 2: Simultaneous coverage probability and $\Pr\{i \in \mathcal{S}_0\}$ for each i when $m = 100$ from 1,000 runs of all-in IOU-C, plug-in IOU-C, and conditional MCB procedures.

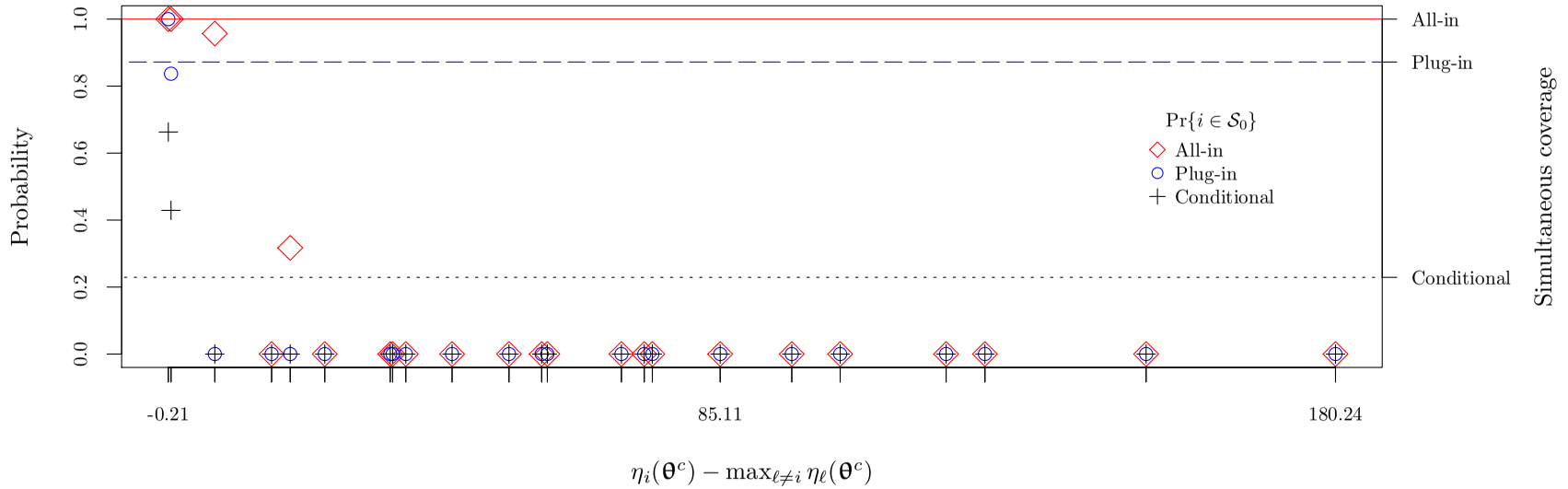

Figure 3: Simultaneous coverage probability and $\Pr\{i \in \mathcal{S}_0\}$ for each i when $m = 400$ from 1,000 runs of all-in IOU-C, plug-in IOU-C, and conditional MCB procedures.

Figure 2 shows that $i = 1$ is included in $\mathcal{S}_0$ in all runs of the all-in IOU-C procedure; however, several systems other than $i = 1$ are also frequently included in $\mathcal{S}_0$. On average, the all-in IOU-C procedure includes 7.30 systems in $\mathcal{S}_0$ and its estimated simultaneous coverage probability of CIs is 1.000 (solid line), which indicates that the all-in procedure is conservative as the desired coverage was set to 0.9. On the other hand, the plug-in IOU-C procedure only contains $i = 1$ and 2 in $\mathcal{S}_0$ while all other systems are correctly determined to be inferior in all 1,000 runs. The average subset size of the plug-in procedure is 1.82, which is much smaller than that of all-in IOU-C, yet the estimated simultaneous coverage probability of the plug-in procedure is 0.874 (dashed line). The true best system, $i = 1$, is included in $\mathcal{S}_0$ in all 1,000 runs, which shows that the plug-in IOU-C procedure has good performance despite ignoring the estimation error in the gradients. Figure 2 shows that the conditional MCB procedure contains only $i = 1$ or 2 in $\mathcal{S}_0$ in all 1,000 runs. However, it includes $i = 1$ in $\mathcal{S}_0$ only 58.1% of the time; 41.9% of the time $i = 1$ was ruled to be inferior to $i = 2$. This demonstrates that, depending on the real-world sample, the conditional MCB procedure may conclude the best system to be inferior. In a real-world experiment, it is impossible to know that such a false conclusion is made because all we have is the one set of real-world data. The average size of $\mathcal{S}_0$ is 1.03 for this procedure, which is the smallest among all three procedures since it ignores input uncertainty. Figure 2 also shows that the simultaneous MCB coverage probability of the conditional procedure is 0.235 (dotted line), which is far lower than 0.9.

Figure 3 shows the results from all three procedures when the "real-world" sample size is increased to $m = 400$ and B to $\lceil 400^{1.1} \rceil = 729$. The all-in IOU-C procedure still includes $i = 1$ and 2 in $\mathcal{S}_0$ in all 1,000 runs, while the average size of $\mathcal{S}_0$ drops to 3.27 as input uncertainty is smaller than when $m = 100$. The simultaneous coverage probability of the all-in IOU-C CIs remains at 1.000. The average size of $\mathcal{S}_0$ for the plug-in procedure is 1.84 and the simultaneous coverage probability of the CIs is 0.872. The average size of $\mathcal{S}_0$ for the conditional MCB procedure is 1.09, while $i = 1$ is still ruled as an inferior system 33.7% of the time. The simultaneous coverage probability of the CIs from conditional MCB is 0.229, which is still much lower than 0.9, although increased from

the $m = 100$ case. Upper and lower bounds from all three procedures for $m = 100$ and 400 are presented in Tables 2–3 in Section EC.4, respectively.

From the results of the all-in procedure when $m = 400$, $i = 6$ is never included in $\mathcal{S}_0$, whereas $i = 4$ is included in $\mathcal{S}_0$ 31.7% of the time, although $i = 6$ has a smaller mean than $i = 4$. The averages of $(\widehat{\beta}_4 - \widehat{\beta}_1)^\top \Sigma(\widehat{\boldsymbol{\theta}})(\widehat{\beta}_4 - \widehat{\beta}_1)/m$ and $(\widehat{\beta}_4 - \widehat{\beta}_2)^\top \Sigma(\widehat{\boldsymbol{\theta}})(\widehat{\beta}_4 - \widehat{\beta}_2)/m$ from 1,000 runs are 3.02 (standard error $= 0.02$) and 1.75(0.01), respectively. On the other hand, the averages of $(\widehat{\beta}_6 - \widehat{\beta}_1)^\top \Sigma(\widehat{\boldsymbol{\theta}})(\widehat{\beta}_6 - \widehat{\beta}_1)/m$ and $(\widehat{\beta}_6 - \widehat{\beta}_2)^\top \Sigma(\widehat{\boldsymbol{\theta}})(\widehat{\beta}_6 - \widehat{\beta}_2)/m$ are 0.393(0.005) and 0.795(0.012), respectively, which shows that the CID effects to $i = 6$ is closer to those to $i = 1$ or $i = 2$ making it easier to rule $i = 6$ to be inferior.

In this particular (s, S) inventory example, the plug-in IOU-C procedure shows excellent performance by including the best system in $\mathcal{S}_0$ with probability greater than $1 - \alpha$. However, for small m and B the plug-in procedure may fail to include the best system in $\mathcal{S}_0$ if its gradient estimate is poor. The all-in IOU-C procedure is protected against such an error by accounting for the estimation error in the gradients at the price of its conservatism. For small m and B, $w_{i\ell}^{(1)}$ from the all-in procedure is large reflecting that the gradient estimator has large uncertainty and includes more systems than the plug-in procedure for finite m and B. We also tested versions of the all-in and plug-in procedures under the sphericity assumption. While the overall trends remain the same, both all-in and plug-in procedures included more solutions in $\mathcal{S}_0$. Detailed results can be found in Section EC.4.

We close this section by summarizing what we learned from the controlled experiments in which we applied the all-in IOU-C procedure to compare $k = 10$ systems with 5-dimensional common multivariate normal input models ($p = 10$). For each i, $\eta_i(\widehat{\boldsymbol{\theta}}) = \eta_i(\boldsymbol{\theta}^c) + b_i \mathbf{1}_p^\top (\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^c) + c_i (\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^c)^\top (\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^c)$, where c_i was adjusted relative to b_i to control the nonlinearity of $\eta_i(\boldsymbol{\theta}^c)$. The case $c_i = 0, \forall i$ was also tested to see the performance when $\eta_i(\boldsymbol{\theta}^c)$ is truly linear. We tested different b_i to examine the impact of equal, increasing, and decreasing amounts of input uncertainty among k systems. The true means of the systems, $\eta_1(\boldsymbol{\theta}^c), \eta_2(\boldsymbol{\theta}^c), \ldots, \eta_{10}(\boldsymbol{\theta}^c)$, were set to be increasing and

equally spaced, where the difference in means of systems i and $i+1$ was controlled relative to b_i and c_i to test the impact of the relative size of input uncertainty to the mean differences.

The size of $\mathcal{S}_0$ is smaller when input uncertainty is relatively small compared to their mean differences. Also, the procedure was robust to the nonlinearity of $\eta_i(\boldsymbol{\theta}^c)$ compared to the linear case. When all k systems are affected exactly the same by input uncertainty, i.e., $b_i = b, \forall i$, all-in IOU-C effectively rules out the inferior systems. When better systems have higher input uncertainty, i.e., $b_i = b \times i, \forall i$, all-in IOU-C detects the inferior systems quite well as they also have smaller input uncertainty. The systems close to $i = 10$ are more frequently included in $\mathcal{S}_0$ since they have closer means to the best and higher input uncertainty, which makes it difficult to tell them apart from the best. When inferior systems have higher input uncertainty, that is, $b_i = b \times (k - i + 1), \forall i$, more systems are included in $\mathcal{S}_0$ since it is difficult to rule out the inferior systems with high input uncertainty. All of these findings coincide with the results from the (s, S) inventory problem.

Acknowledgments

Supported by the NSF Grant No. CMMI-1068473. We thank the Area Editor, Associate Editor and three anonymous reviewers for their careful reviews and suggestions.

References

Amemiya T (1985) *Advanced Econometrics* (Cambridge, MA, USA: Harvard University Press).

Barton RR (2012) Tutorial: Input uncertainty in output analysis. *Proceedings of the 2012 Winter Simulation Conference*, 1–12 (Piscataway, NJ, USA: IEEE Press).

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

Chang JY, Hsu JC (1992) Optimal designs for multiple comparisons with the best. *Journal of Statistical Planning and Inference* 30(1):45–62.

Cheng RCH, Holland W (1997) Sensitivity of computer simulation experiments to errors in input data. *Journal of Statistical Computation and Simulation* 57(1-4):219–241.

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

Corlu C, Biller B (2013) A subset selection procedure under input parameter uncertainty. *Proceedings of the 2013 Winter Simulation Conference*, 463–473 (Piscataway, NJ, USA: IEEE Press).

Corlu CG, Biller B (2015) Subset selection for simulations accounting for input uncertainty. *Proceedings of the 2015 Winter Simulation Conference*, 437–446, WSC '15 (Piscataway, NJ, USA: IEEE).

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

Fan W, Hong LJ, Zhang X (2013) Robust selection of the best. *Proceedings of the 2013 Winter Simulation Conference*, 868–876 (Piscataway, NJ, USA: IEEE Press).

Fu M (2015) *Handbook of Simulation Optimization*. International Series in Operations Research & Management Science (New York: Springer).

Gao S, Xiao H, Zhou E, Chen W (2017) Robust ranking and selection with optimal computing budget allocation. *Automatica* 81:30 – 36.

Hsu J (1996) *Multiple Comparisons: Theory and Methods* (Boca Raton, FL, USA: Chapman & Hall).

Koenig LW, Law AM (1985) A procedure for selecting a subset of size m containing the l best of k independent normal populations, with applications to simulation. *Communications in Statistics – Simulation and Computation* B14(3):719–734.

Lam H (2016) Input uncertainty and robust analysis in stochastic simulation. *Proceedings of the 2016 Winter Simulation Conference*, 178–192 (Piscataway, NJ, USA: IEEE Press).

Lin Y, Song E, Nelson BL (2015) Single-experiment input uncertainty. *Journal of Simulation* 9(3):249–259.

Nelson BL (1993) Robust multiple comparisons under common random numbers. *ACM Transactions on Modeling and Computer Simulation* 3(3):225–243.

Nelson BL, Matejcik FJ (1995) Using common random numbers for indifference-zone selection and multiple comparisons in simulation. *Management Science* 41(12):1935–1945.

Scarf H (1958) A min-max solution of an inventory problem. *Studies in the Mathematical Theory of Inventory and Production* 201–209.

Seber GAF, Lee AJ (2003) *Confidence Intervals and Regions* (Hoboken, NJ: John Wiley & Sons, Inc.).

Song E, Nelson BL, Hong LJ (2015) Input uncertainty and indifference-zone ranking & selection. *Proceedings of the 2015 Winter Simulation Conference*, 414–424 (Piscataway, NJ, USA: IEEE Press).

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

Spall JC (1992) Multivariate stochastic approximation using a simultaneous perturbation gradient approximation. *IEEE Transactions on Automatic Control* 37(3):332–341.

Wieland JR, Schmeiser BW (2006) Stochastic gradient estimation using a single design point. *Proceedings of the 2006 Winter Simulation Conference*, 390–397 (Piscataway, NJ, USA: IEEE Press).

Author Biographies

Eunhye Song is the Harold and Inge Marcus Early Career Assistant Professor in the Department of Industrial and Manufacturing Engineering at Penn State University. Her research area is simulation analysis methodology with emphasis on large-scale optimization via simulation, simulation uncertainty quantification and optimization via simulation under input model risk. Her website is `www.eunhyesong.info`.

Barry L. Nelson is the Walter P. Murphy Professor of the Department of Industrial Engineering and Management Sciences at Northwestern University and a Distinguished Visiting Scholar at Lancaster University in England. His research focus is on the design and analysis of computer simulation experiments on models of discrete-event and stochastic systems. Nelson is a Fellow of INFORMS and IIE. Further information can be found at `www.iems.northwestern.edu/~nelsonb/`.